DIVERSE PREFERENCE LEARNING FOR CAPABILITIES AND ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

As LLMs increasingly impact society, their ability to represent diverse perspectives is critical. However, recent studies reveal that alignment algorithms such as RLHF and DPO significantly reduce the diversity of LLM outputs. Not only do aligned LLMs generate text with repetitive structure and word choice, they also approach problems in more uniform ways, and their responses reflect a narrower range of societal perspectives. We attribute this problem to the KL divergence regularizer employed in preference learning algorithms. This causes the model to systematically overweight majority opinions and sacrifice diversity in its outputs. To address this, we propose Diverse Preference Learning, which decouples the entropy and cross-entropy terms in the KL penalty — allowing for fine-grained control over LLM generation diversity. From a capabilities perspective, LLMs trained using Diverse Preference Learning attain higher accuracy on difficult repeated sampling tasks and produce outputs with greater semantic and lexical diversity. From an alignment perspective, they are capable of representing a wider range of societal viewpoints and display improved logit calibration. Notably, Diverse Preference Learning resembles, but is a Pareto improvement over standard temperature scaling.

1 INTRODUCTION

Large language models (LLMs) are increasingly impacting society, now generating a significant portion of online content (Thompson et al., 2024). As LLMs become more integrated into how people consume information (Bick et al., 2024) and approach tasks (Deloitte, 2024), their ability to represent diverse perspectives is critical.

For example, consider an LLM answering the following multiple-choice question:

> The best way to reduce income inequality is:
> (A) Increase minimum wage
> (B) Expand access to education and job training
> (C) Implement universal basic income
> (D) Lower taxes on the wealthy to stimulate job creation

Imagine a survey showing people's preferences as: A (55%), B (20%), C (15%), and D (10%). How should an LLM respond to this question? Ideally, we may prefer it to reflect the range of views in the population. If an LLM assigns 99% probability to majority option A, it fails to represent the diversity of perspectives. With LLMs becoming important information sources, this may reinforce dominant narratives at the expense of minority views.

However, recent studies reveal that alignment algorithms such as RLHF and DPO significantly reduce the diversity of LLM outputs. This leads to mode collapse towards majority preferences, as described in the example above (Kirk et al., 2024; Padmakumar & He, 2024; Rafailov et al., 2024; Christiano et al., 2023). On multiple choice questions, this manifests as overconfidence and poor calibration (Tian et al., 2023; Kadavath et al., 2022). In a generative setting, this results in repetitive responses, as illustrated in Figure 1. Here, the DPO model frequently describes doctors with the same name and relationship to the patient. Lastly, in problem-solving settings, we show that

Prompt: Please write a two sentence story about a doctor and a close family member

DPO

Dr. Amelia watched with a heavy heart as her elderly father, tears streaming down his face, admitted to no longer being able to care for himself at home. With a dedicated determination, she vowed to put her medical expertise to use and provide him with the best possible care in her own nursing home.

Dr. Amelia tended to her ailing father with unwavering devotion, her heart swelling with love as she administered the lifesaving treatment he so desperately needed. Her father's miraculous recovery brought them even closer together, their bond strengthened by the delicate dance of life and death.

DPO $t = 1.4$

Amidst the sterile hospital bureaucracy, Dr. Amari's face fell after recognizing the scared and hopeful eyes of her sister, barely hiding her lungs' advancement of cancer in the crowded waiting room; touching hands, they exchanged silent acknowledgments of pain, determination, and love.

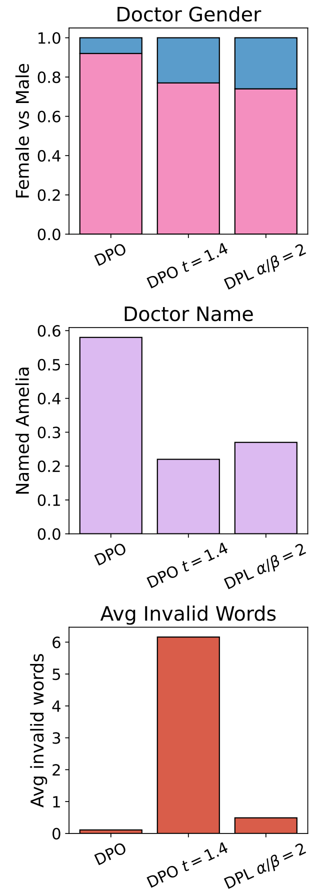
Dr. Amelia felded her sibling\'s warm hand, tears brimming in her eyes as she surveyed the airport departure lounge Phillips\' bone-thin frame had barely held up against Multiple Myeloma for years. Their embrace was an ignorant权dmermiment of infirm}[como handsome hero][COL shootColors(new Array manuscript Josh Goldman I L…

CDPO $\alpha/\beta = 2$

Dr. Alanna sat quietly in her hospital office, the sun casting golden rays through the window over her quiet, motionless mother. Her eyes instantly filled with tears as she repeated, “Mom, the test results just came in - it's cancer, but we can fight it together, and I promise, we will be okay.”

Dr. Amelia's heart ached as herened IT technician brother, Tom, was rushed into the hospital Pacific City Medical System, the very place where they'd grown up, with heart failure. "Please save my brother," Amelia begged imperatively, her voice trembling with emotion. "I'll do everything in my power to save him," Amelia reassured, activation the defibrillator.

(a) Example stories generated by DPO, DPO with temperature scaling, and our algorithm, DPL. We highlight Doctor name, gender, and textual aberration features shown in the plots on the right.



(b) Feature-level diversity and lexical integrity statistics on 100 generated stories.

Figure 1: **Diverse Preference Learning increases output diversity while preserving quality.** DPO responses are well-formed but lack diversity (e.g. same doctor name, gender, and family relationship to patient). With temperature scaling ($t = 1.4$), DPO generates responses with more diversity at the cost of fluency and token-level aberrations. In particular, temperature scaling results in many non-word tokens. Meanwhile, DPL at global temperature $\alpha/\beta = 2$ similarly increases diversity, but with significantly less degradation.

diversity loss harms models' ability to answer difficult questions across multiple samples. While standard token-level temperature scaling is effective at correcting for micro-scale diversity loss (e.g. for next-tokens on multiple-choice questions), it leads to rapid degradation of fluency and quality in multi-token generations(Kadavath et al., 2022).

We attribute this diversity loss to the KL-divergence term in preference learning, which strongly biases models towards majority preferences and sacrifices diversity in their outputs. In Section 3, we analyze RLHF and DPO from a social choice perspective. We prove that when the preferences of different groups conflict, the probability of generating majority-preferred outputs far exceeds the population preference for that output. This mode collapse has consequences for the social perspectives language models represent but also for lexical and logical diversity.

To address this issue, we propose splitting the KL penalty into distinct entropy and cross-entropy terms. This enables diversity to be controlled independently from the model's bias towards a reference policy. We call this method Diverse Preference Learning (DPL). DPL resembles but improves upon standard temperature scaling, increasing diversity at the sequence level rather than token-by-token. This increases macro-scale diversity while avoiding the rapid quality degradation caused by standard temperature scaling.

In summary, we provide the following contributions:

1. We identify KL-regularization as a cause of diversity loss in aligned language models. We connect this to a social choice analysis, where we prove that the KL divergence term heavily biases the model towards majority-preferred outputs.
2. We propose Diverse Preference Learning (DPL), which decouples the entropy and cross-entropy terms in the KL penalty, allowing for independent control of generation diversity. We prove this enables proportional representation of population preferences.
3. We demonstrate empirically that DPL improves output diversity in chat domains, best-of-N accuracy on difficult math problems, and logit calibration on common multiple-choice benchmarks.

Our work connects popular methods for LLM alignment with notions of diversity and representation. DPL advances LLMs that are attuned to the diversity of preferences in society while also displaying improved capability in a number of settings.

2 Related Work

Diversity loss in aligned LLMs. Prior work has studied diversity loss caused by LLM alignment algorithms (Kirk et al., 2024; Park et al., 2023; Xiao et al., 2024; Wang et al., 2023) as well as its impact on humans who use these models (Padmakumar & He, 2024; Ding et al., 2023; Doshi & Hauser, 2024). In Appendix B, we evaluate against Wang et al. (2023), who use other f-divergences as regularizers to avoid the mode-seeking property of KL divergence. Meanwhile, we analyze policies using social choice theory and propose entropy regularization to restore population representation. Xiao et al. (2024) also investigate entropy regularization to improve preference representation in aligned LLMs, however they do not perform experiments in a generative setting. Lastly, Sun & van der Schaar (2024) propose an inverse reinforcement learning-based approach to mitigate mode collapse during alignment, but focus on learning from demonstrations rather than pairwise preferences.

Social choice theory and alignment. Several recent works explore the intersection of social choice theory and AI alignment. Siththaranjan et al. (2024) analyze the reward learning portion of RLHF as a case of the Borda Count voting rule. In contrast, we characterize trained policies. Munos et al. (2024); Swamy et al. (2024), and Chakraborty et al. (2024) develop preference learning algorithms to better handle intransitive preferences. While these approaches also lead to proportional representation under some conditions, they require complex multi-agent reinforcement learning setups to train, and are not studied in a linguistic diversity or problem-solving setting.

Temperature scaling. Token-level temperature scaling is a common tool for controlling diversity of LLM outputs. Previous work applies temperature scaling to improve LLM calibration (Kadavath et al., 2022; Tian et al., 2023; Xie et al., 2024) and best-of-N coding ability Chen et al. (2021). In contrast, our method performs global temperature scaling over the entire sequence. Shih et al. (2023) develop a procedure for global temperature scaling for LLMs using reinforcement learning. Meanwhile, we develop an offline supervised learning algorithm for preference learning. Recent work has also combined high-temperature sampling with token-level heuristics to preserve quality while maintaining diversity ((Nguyen et al., 2024)).

Entropy bonuses. Entropy bonuses are common in reinforcement learning algorithms to encourage exploration (Haarnoja et al., 2018; Eysenbach & Levine, 2022; Lee et al., 2024). In contrast, we use entropy to restore diversity and proportional representation in an alignment setting. Additionally, entropy bonuses in RL tend to be orders of magnitude smaller than what we consider.

3 Theoretical Analysis and Method

In this section, we perform a theoretical analysis of the RLHF and DPO objectives through the lens of social choice theory. In settings where subpopulations disagree about the relative merit of different options, we prove that standard RLHF and DPO amplify majority preferences by several orders of magnitude. This causes mode collapse to the majority-preferred output, which decreases output diversity and worsens logit calibration.

In particular, we prove that this phenomenon is caused by these objectives' KL-regularization term performing two independent functions. First, it maximizes the log-likelihood of generations under the reference policy (cross-entropy term), and second, it maximizes the diversity of the learned policy (entropy term). We then propose Diverse Preference Learning (DPL), which decouples the cross-entropy and entropy terms from the KL-regularization objective. This allows for distinct control over generation diversity and bias towards the reference policy.

3.1 RLHF and DPO Lead to Mode Collapse

First, we outline RLHF as a three-step alignment procedure

1. **Preference Collection:** Query humans to gather a dataset of preferences over pairs of LLM generations
$$\mathcal{D} = \{y_1 \succ y'_1, y_2 \succ y'_2, y_3 \succ y'_3, ...\} \tag{1}$$
2. **Reward Modeling:** Learn a reward model with Bradley-Terry likelihood
$$\max_r \mathbb{E}_{y \succ y' \sim \mathcal{D}}[\log \sigma(r(y) - r(y'))] \tag{2}$$
3. **KL-regularized Reinforcement Learning:** Train a new policy against the learned reward with regularization against a reference policy
$$\max_\pi \mathbb{E}_{y \sim \pi}[r(y)] - \beta D_{KL}(\pi || \pi_{ref}) \tag{3}$$

In the following proposition, we perform a social choice analysis of RLHF. We show that when a population has conflicting preferences about LLM generations, RLHF vastly overrepresents the majority preference. Appendix A contains a generalized, multi-outcome version of this proposition.

Proposition 3.1 (Two-Outcome RLHF Policy)**.** *Suppose a population of raters prefers completion $y \succ y'$ with probability p. Then RLHF (or DPO) with KL-regularization penalty β has the optimal policy*
$$\pi(y) \propto \pi_{ref}(y) p^{1/\beta}.$$

Thus, in RLHF's optimal policy $\pi(y) \propto \pi_{ref}(y)p^{1/\beta}$, the KL regularization term β controls both the relative weighting of the reference policy and the sharpness of the resulting distribution.

For both RLHF and DPO, empirical choices of β typically lie in the range $[0.01, 0.1]$ (Ouyang et al., 2022; Ahmadian et al., 2024; Rafailov et al., 2024; Tunstall et al., 2023). This results in a very peaky distribution that exponentiates population preferences to the 10th or 100th power. For example, if 80% of the population prefer y and 20% prefer y', then with $\beta = 0.1$, π^* generates y with 99.9999% probability and y' with 0.0001% probability.

Relationship to output diversity. Since models trained on RLHF and DPO fail to represent diverse preferences, they will also struggle to produce diverse outputs. While the social choice result in Proposition 3.1 is most obviously relevant to a model's representation of social perspectives, LLM preferences encode many other aspects of diversity as well.

In many cases, preference variation is the result of random noise that we wish to preserve. For example, if a person has no strong preference over which of two synonyms is used in a sentence, the Bradley-Terry model (Equation 2) predicts their preference distribution will look relatively uniform. However, the optimal RLHF policy in Proposition 3.1 would remove this diversity and choose one of the words nearly all of the time. This perspective is supported at the large-scale by Kobak et al. (2024), who form trendlines of word usage in the academic literature. They find that common words used by Chat-GPT begin to occur orders of magnitude more frequently after the introduction of academic papers — an example of failing to match the true distribution of human writing. We demonstrate that DPL models attain higher generation diversity in Section 4.1.

Relationship to problem-solving. Diversity loss can harm capabilities as well as representation. The RLHF policy from Proposition 3.1 may underperform on challenging problem-solving tasks that require multiple samples to tackle. For instance, consider a difficult geometry problem: 60% of preferred samples in the preference dataset approach it using synthetic geometry, while the remaining 40% rely on coordinate geometry. Although many geometry problems appear similar, each often requires a specific method. In this case, when the optimal RLHF policy encounters a problem that demands coordinate geometry, it may almost always try to use synthetic geometry. Across multiple samples, this RLHF policy would consistently fail, whereas a more diverse algorithm, such as DPL, would succeed. We empirically verify this intuition in Section 4.2.

Recent developments have emphasized the importance of repeated inference-time sampling — Google DeepMind's models recently achieved silver medal performance on the International Mathematical Olympiad (DeepMind, 2023), and OpenAI's o1 model achieved state-of-the-art performance on a number of benchmarks (OpenAI, 2023). Similarly, inference-time methods like Tree of Thought (Yao et al., 2023) show promising initial results. DPL increases problem-solving diversity across samples, which is critical for the success of these methods.

Relationship to calibration. Finally, we should expect the optimal policy in Proposition 3.1 to exhibit poor logit calibration. The RLHF objective trains models to place very high probability mass on their preferred option, resulting in high confidence on almost every generation. As a result, RLHF and DPO models tend to have high output confidence regardless of their accuracy on the task (Tian et al., 2023). We study this and DPL's ability to improve logit calibration on factual multiple-choice benchmarks in Section 4.3.

3.2 Diverse Preference Learning

Now, we introduce Diverse Preference Learning (DPL), which decouples the KL-regularization term into separate cross-entropy and entropy terms, allowing for separate control of diversity and bias towards the reference policy

$$\max_{\pi} \mathbb{E}_{y\sim\pi(y|x)}[r(x,y)] + \alpha H(\pi(\cdot|x)) - \beta H(\pi(\cdot|x), \pi_{ref}(\cdot|x)) \tag{4}$$

While the entropy parameter α optimizes for diversity, the cross-entropy penalty β maximizes the average log-likelihood under the reference policy π_{ref} of generations from the learned policy. This controls the strength of the reference model as a prior.

We also propose a DPO-style objective that bypasses the reinforcement learning step

$$\max_{\pi} \mathbb{E}_{y\succ y'\sim\mathcal{D}}[\log\sigma(\alpha\log\frac{\pi(y|x)}{\pi(y'|x)} - \beta\log\frac{\pi_{ref}(y|x)}{\pi_{ref}(y'|x)})] \tag{5}$$

with a derivation in Appendix A.

Proposition 3.2 (Two-Outcome DPL Policy)**.** *Suppose a population of raters prefers completion $y \succ y'$ with probability p. Then DPL with entropy bonus α and cross-entropy penalty β has the optimal policy*

$$\pi(y) \propto \pi_{ref}(y)^{\beta/\alpha} p^{1/\alpha}.$$

Proof in Appendix A. Note that standard RLHF and DPO are special cases of DPL with $\alpha = \beta$.

By choosing an appropriate α parameter, DPL avoids raising probabilities to high powers, thereby preventing mode collapse. Through the choice of α, DPL allows for fine-grained control over the diversity of the learned policy. From a social choice perspective, separately controlling a policy's diversity allows for increased representation of minority preferences. As $\alpha \to 1$, π becomes less focused on majority preferences. When $\alpha = 1$, we recover proportional representation.

Corollary 3.1. *Suppose a population of raters prefers completion $y \succ y'$ with probability p. Then DPL with entropy bonus $\alpha = 1$ is a proper scoring rule, weighted by a reference policy prior.*

In other words, DPL can produce well-calibrated policies. This means that the distribution over outputs matches the population preference distribution they are trained on.

DPL performs global temperature scaling. The α entropy bonus in DPL can be thought of as a sequence-level (or global) version of standard token-level temperature. Standard token-level temperature scales and renormalizes the distribution at each token

$$\pi'(y|x) = \prod_{i=1}^{N} \frac{\pi(y_i|y_{1:i-1})^{1/t}}{Z(y_{1:i-1})} \tag{6}$$

Meanwhile, following Proposition A.2, the optimal DPL policy has the form

$$\pi'(y|x) = \exp(\frac{1}{\alpha}r(x,y))\pi_{ref}^{\beta/\alpha}/Z \quad (7)$$

$$= \exp(\frac{1}{\beta}r(x,y))^{\beta/\alpha}\pi_{ref}^{\beta/\alpha}/Z \quad (8)$$

$$= \pi_{DPO}(y|x)^{\beta/\alpha}/Z \quad (9)$$

$$= \pi_{DPO}(y|x)^{1/(\alpha/\beta)}/Z \quad (10)$$

Meaning α/β resembles a global "temperature" on top of the DPO policy. Global temperature scales the probability of entire sequences rather than individual tokens.

While standard temperature scaling is a convenient heuristic to control diversity at inference time, it quickly degrades quality at temperatures above 1 (e.g. Figure 1). One important advantage of global temperature scaling is that it preserves the relative probability ordering of each sequences, whereas regular temperature scaling does not. This means that the "best" or majority-preferred sequences always remain the most likely outputs.

Empirical choices in training DPL models. Corollary 3.2 shows that with the right hyperparameters, DPL policies achieve proportional representation of preferences. However, in practice, we must negotiate important tradeoffs between performance and diversity. We find DPL performs best at a middle ground between standard preference learning's $\alpha = \beta$ and proportional representation at $\alpha = 1$. At low global temperatures, the policy lacks diversity, while at high temperatures, quality begins to degrade. Nevertheless, we note that while standard temperature scaling often leads to unintelligible sequences at temperatures just above one, DPL remains relatively stable at much higher global temperatures (Figure 1).

One reason we see a performance dropoff at large α could be due to dataset noise. In these cases, some overweighting of majority preferences can be beneficial. If preference variation is due to random error, a lower temperature implicitly denoises the policy in a manner akin to a majority voting rule. For example, there is a significant 37% cross-rater disagreement rate within the HH-RLHF preference dataset we use (Bai et al., 2022). However, much of this variation is thought to be due to random noise (Cai et al., 2024).

4 Experiments

In this section, we consider four experimental settings for evaluating our algorithm. First, we show that in general-purpose chat domains, DPL allows for increased diversity with less performance degradation than DPO with token-level temperature scaling. Second, we consider an application of high-temperature generation in best-of-N problem-solving settings. Finally, we evaluate DPL's logit calibration, finding reduced overconfidence and improved calibration on standard multiple-choice benchmarks.

4.1 Improving Diversity-Quality Tradeoffs

Currently, inference-time strategies such as token-level temperature scaling are the standard approach to sampling diverse outputs from aligned LLMs. However, at temperatures above 1, output quality degrades rapidly. In contrast, performing global temperature scaling using DPL leads to increased diversity without such a steep degradation in quality.

For these experiments, we LoRA finetune Mistral-7B-Instruct-v0.2 with DPO and DPL (Rafailov et al., 2024; Hu et al., 2021). We train on the HH-RLHF preference dataset for 5,000 steps (details in Appendix C.3) (Bai et al., 2022).

Quality metrics. We evaluate against three quality metrics. First, we use Arena-Hard, a popular LLM chatbot benchmark with high agreement with human annotators (Tianle Li*, 2024). Arena-Hard evaluates models on 500 queries, which mostly deal with programming, business, or software help. We use gpt-4o-mini-2024-07-18 as a judge. Models are assessed by win-rate against gpt-4-0314 outputs on the same responses. This metric measures general-purpose capabilities acquired by the model during training. For our second quality metric, we train a separate reward model on the

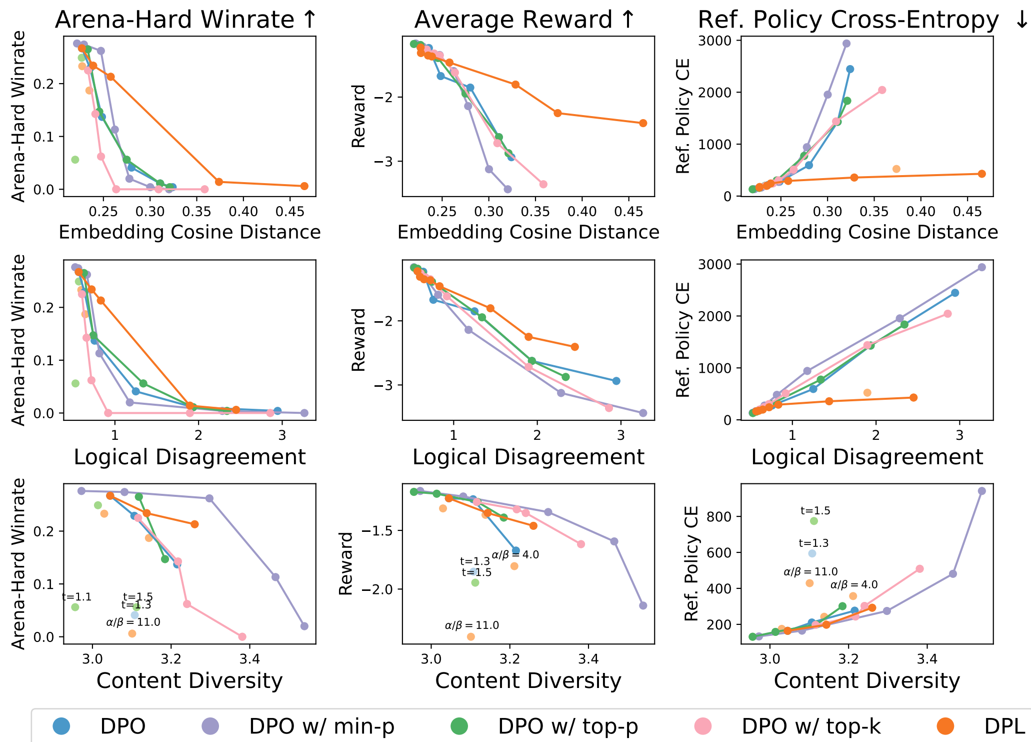


Figure 2: **Improved diversity-quality tradeoffs with DPL.** We construct diversity-quality Pareto curves contrasting DPO with token-level temperature scaling against DPL (by modulating the entropy term). We also plot the performance of DPO with min-p, top-p, and top-k sampling, which can improve diversity-quality tradeoffs when sampling at high temperatures. We plot points that lie below the Pareto curve in lighter shades. DPL Pareto-dominates DPO with standard temperature scaling across all nine metrics, and it outperforms all sampling methods on six.

HH-RLHF dataset (details in Appendix C.3). We then evaluate models by the average reward of their generations on a held-out test set of 500 inputs, for which we generate 16 responses each and compute the average reward. This metric measures how well DPO and DPL optimize generation quality on the dataset against the training metric. Finally, we include the cross-entropy with respect to the reference policy, which is the second part of the alignment optimization objective. The cross-entropy is again computed over the HH-RLHF test set. Ideally, models achieve high diversity while maintaining high average reward on the training dataset, low reference cross-entropy, and strong general-purpose chat capabilities.

Diversity metrics. We include three diversity metrics in Figure 2, with results against additional metrics in Appendix B. Drawing on Kirk et al. (2024); Tevet & Berant (2021), who perform diversity studies on LLM outputs, our diversity metrics roughly measure 1) general semantic diversity (embedding distance metric), 2) logical diversity or diversity of viewpoints (logical disagreement metric), and 3) diversity in response content and ideas (content diversity metric). For all diversity metrics, we compute per-input diversity between 16 generated responses for each of 500 inputs on a held-out test split of HH-RLHF. Our first metric evaluates expected cosine similarity between embeddings of model outputs. The second and third metrics follow the design of human diversity questionnaires in Tevet & Berant (2021). Here pools of 4 responses are presented to gpt-4o-mini and evaluated according to their logical agreement and content diversity on a scale of 1-5. See Appendix B for further details, example generations, and results against additional diversity metrics.

Results. In Figure 2, we construct diversity-quality Pareto curves contrasting DPO with token-level temperature scaling against DPL with different values of α, the entropy bonus. We perform a sweep across token-level temperature range $[1, 1.5]$. For min-p sampling, we choose $p_{base} = 0.1$ and token-level temperature range $[1.3, 4]$. For top-p sampling, we choose $p = 0.9$ and temperature range $[1.1, 1.7]$. For top-k sampling, we choose $k = 180$ and temperature range $[1.1, 2.5]$. We choose these ranges since beyond this, responses are nearly always nonsensical. We choose min-p p_{base}, top-p cutoff, and top-k cutoffs according to the recommendations in Nguyen et al. (2024). For DPL, we sweep across global temperature range $[1, 11]$. We run all results with $\beta = 0.1$. See Appendix B for additional baselines and hyperparameters. Across diversity and quality metrics, DPL with global temperature scaling is always competitive or Pareto-dominant.

We note that increasing diversity with DPL eventually results in quality loss. But importantly, unlike token-level temperature scaling, DPL never results in predominantly nonsensical generations, even at very high global temperatures. This means that in applications where diversity is particularly important, it can be achieved without rendering responses useless.

In most cases, increasing both global and token-level temperature leads to significant improvements in diversity. However, we note that the content diversity metric shows less variation. This is because at the highest temperatures, the content diversity metric rates completions as less diverse. We label points displaying this non-monotonic behavior in Figure 2. In contrast, the embedding distance and logical disagreement metrics exhibit purely monotonic relationships between temperature and diversity. We provide results with additional diversity metrics in Appendix B.

4.2 Best-of-N Problem-Solving

Previous work has found that increasing token-level temperature can improve LLM performance in a best-of-N problem-solving setting (Chen et al., 2021). Intuitively, this is because in a repeated sampling setting, it is desirable to test a diversity of strategies and potential solutions. This aligns with recent observations that effective use of inference-time compute can lead to greater gains than scaling model size (Snell et al., 2024). Similar approaches have allowed LLMs to solve previously unprecedented reasoning problems (OpenAI, 2024; DeepMind, 2023). While other alignment algorithms like RLHF and DPO cause mode collapse, DPL models exhibit an increased diversity of problem-solving strategies.

Setup and evaluation. We apply DPO and DPL to a Mistral-7B base model (HuggingFace, 2023) trained with supervised fine-tuning on the UltraChat dataset (Ding et al., 2023). We finetune with LoRA for one epoch, replicating the Zephyr training recipe, but with $\beta = 0.1$, which improved performance (Tunstall et al., 2023). This approach trains on the Ultrafeedback-200k dataset, a large preference dataset covering a broad suite of chat and reasoning tasks (Cui et al., 2024).

We evaluate against two mathematical reasoning datasets: the GSM8K grade-school math dataset (Cobbe et al., 2021) and the more challenging MATH dataset (Hendrycks et al., 2021b). We include few-shot chain-of-thought examples to prompt the model to perform reasoning step-by-step. We sample 128 completions on a random split of 200 problems from each dataset. We also divide problems into Easy, Medium, and Hard categories. For MATH, this corresponds to Level 1, Level 3, and Level 5 problems. For GSM8K, we run our evaluation on Mistral-Instruct-7B and group problems as easy if they take 4 or fewer samples to solve, medium if they take 5-64 samples to solve, and hard if they take more than 64 samples to solve.

Results. Figure 3 shows our results. On easy and medium questions, standard DPO performs the best. On GSM8K and MATH hard splits, both token-level temperature scaling and DPL improve performance over standard DPO at higher samples, with DPL performing best.

In the figure, higher temperature runs have a lower y-intercept – meaning they perform worse in a one-shot setting. This makes sense as DPO concentrates probability mass on the option likely to be best. However, on the hard splits, high-temperature runs cross with and surpass the DPO curve due to better exploration of the solution space.

However, quality quickly degrades at higher token-level temperatures. For example, DPO at $t = 1.2$ performs poorly in nearly all evaluations. In contrast, DPL with global temperature $\alpha/\beta = 1.2$ performs relatively well in all settings. The performance gap between DPL and temperature-scaled DPO is most pronounced in challenging, high-sample scenarios. In these cases, the level of diversity required would cause token-level temperature scaling to significantly degrade output quality, while DPL's sequence-level approach preserves coherence. We provide an extensive analysis of the relationship between problem difficulty, temperature, and best-of-N sampling in Appendix C.

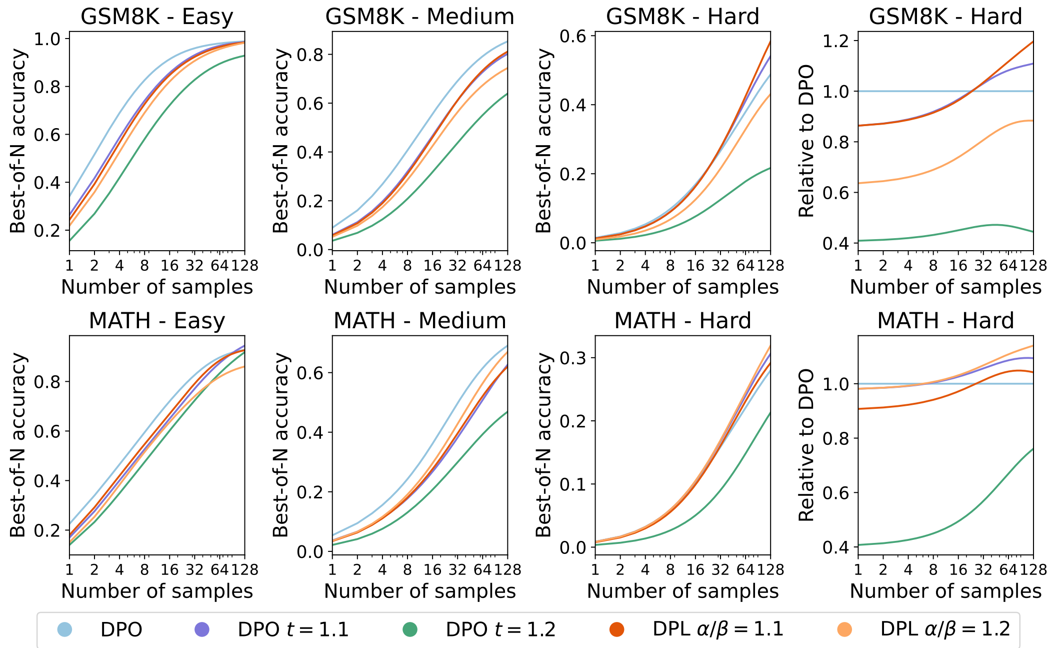


Figure 3: **DPL improves best-of-N mathematical problem-solving on difficult instances.** Left three columns show best-of-N accuracy across difficulty levels. Right column shows performance on hard problems relative to DPO at a given sample count. For easier problems, standard DPO ($t = 1$) performs well. However, hard repeated sampling tasks benefit from diverse solution strategies. On hard problems, both token-level temperature sampling and DPL improve best-of-N accuracy. However, DPL achieves a better quality-diversity tradeoff, especially at high temperatures where token-level scaling rapidly degrades quality. This makes DPL particularly effective for generating diverse yet high-quality solutions.

4.3 Logit Calibration

As predicted by the theory in Section 3, DPL models trained with higher global temperatures exhibit increased logit calibration and decreased overconfidence. They do this while preserving multiple-choice accuracy.

Setup and evaluation. We evaluate logit calibration using standard calibration metrics on multiple-choice datasets. Our baseline is a Mistral-7B base model (HuggingFace, 2023), finetuned with supervised fine-tuning (SFT) on the UltraChat dataset (Ding et al., 2023). We also evaluate a suite of DPL models trained on top of this base model with the training setup from Section 4.2. For each model, we track its Expected Calibration Error (ECE), Brier Score, and accuracy across questions. We prompt the model to begin its response with the token corresponding to its answer choice and use the normalized probabilities assigned to A, B, C, and D to evaluate its calibration (Appendix D).

Datasets. We evaluate against two standard multiple-choice datasets: TruthfulQA and MMLU. TruthfulQA is a benchmark designed to assess a model's ability to provide truthful answers in contexts where misconceptions are prevalent (Lin et al., 2022). MMLU tests a model's knowledge and reasoning across 57 diverse subjects, from philosophy to abstract algebra (Hendrycks et al., 2021a).

Results. As shown in Figure 4, models trained with higher global temperatures consistently display lower calibration error. While models trained using an global temperature of 1 (equivalent to DPO) are less calibrated than the base model, increasing the global temperature during training quickly enables the DPL models to surpass even the base model's calibration. DPL models also maintain similar accuracy to models trained using DPO. In fact, DPL models with global temperatures slightly greater than 1 consistently attain both higher accuracy and lower calibration error than their DPO counterparts.

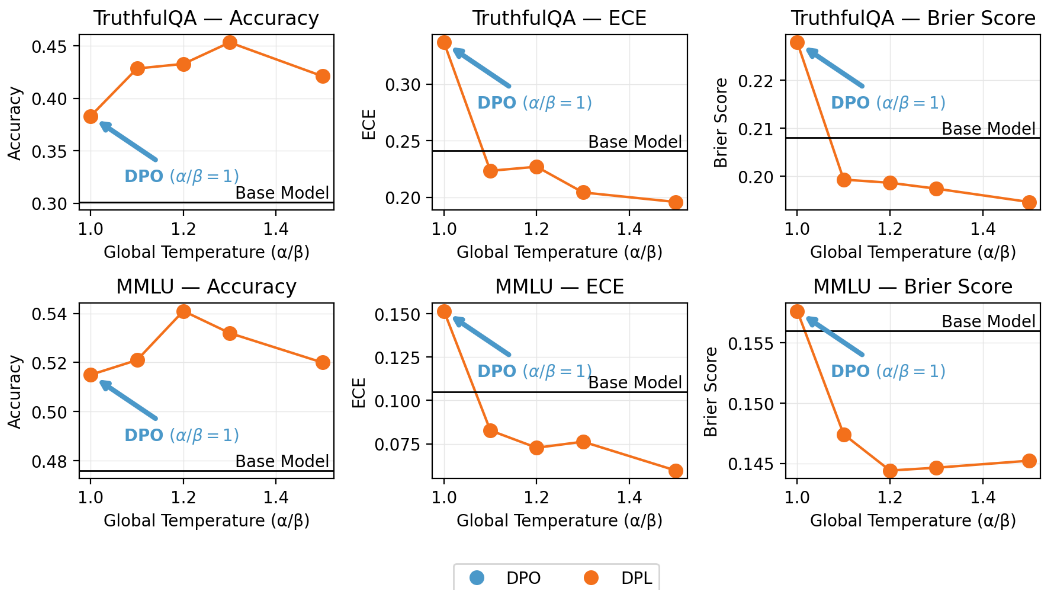


Figure 4: **DPL improves both calibration and accuracy on multiple-choice question (MCQ) datasets.** We plot model accuracy, Expected Calibration Error (ECE), and Brier Score for all models on both TruthfulQA and MMLU. The DPO model (equivalent to DPL with global temperature 1) displays significantly worse calibration than the base model. In contrast, DPL models consistently exhibit improved calibration without sacrificing accuracy.

5 Conclusion

In this paper, we presented Diverse Preference Learning (DPL), a novel approach to mitigating the loss of output diversity in aligned LLMs. By analyzing the role of the KL divergence regularizer in RLHF and DPO, we identified that the coupling of entropy and cross-entropy terms leads to overweighting of majority preferences and reduced output diversity. DPL addresses this issue by decoupling these terms, allowing for independent control over generation diversity and bias towards the reference policy.

Our experimental results demonstrate that LLMs trained with DPL outperform those trained with standard methods in several key areas. From a capabilities standpoint, DPL models produce outputs with greater semantic and lexical variety and achieve higher accuracy on challenging tasks that benefit from diverse sampling strategies. From an alignment perspective, these models can proportionately represent a wider range of societal viewpoints and exhibit improved logit calibration. Importantly, DPL achieves enhanced diversity with less quality degradation, offering a Pareto improvement over standard temperature scaling.

The introduction of DPL highlights the importance of diversity in the development and alignment of LLMs. DPL provides the most significant benefits in scenarios that require inference-time scaling or representation of diverse perspectives. By providing a mechanism to align LLMs without causing mode collapse, DPL contributes to the creation of more capable, reliable, and representative language models. In future work, it would be interesting to explore alternative, semantically grounded diversity metrics that could be integrated into preference learning algorithms — such as metrics constructed from embeddings or LLM judges.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms, 2024. URL https://arxiv.org/abs/2402.14740.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Alexander Bick, Adam Blandin, and David J Deming. The rapid adoption of generative ai. Technical report, National Bureau of Economic Research, 2024.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie Gu, and Guannan Zhang. Ulma: Unified language model alignment with human demonstration and point-wise preference, 2024. URL https://arxiv.org/abs/2312.02554.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences, 2024. URL https://arxiv.org/abs/2402.08925.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

DeepMind. Ai solves IMO problems at silver medal level, 2023. URL https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/. Accessed: 2024-10-02.

Deloitte. State of generative ai in the enterprise: Q3 report. Technical report, Deloitte, 2024. URL https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-generative-ai-in-enterprise.html. Accessed: 2024-09-28.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023. URL https://arxiv.org/abs/2305.14233.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.

Benjamin Eysenbach and Sergey Levine. Maximum entropy rl (provably) solves some robust rl problems, 2022. URL https://arxiv.org/abs/2103.06257.

John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pp. 1317–1339, 1989.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, 2018. URL https://arxiv.org/abs/1801.01290.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL https://arxiv.org/abs/2103.03874.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

HuggingFace. Mistral-7b-sft-beta model card. https://huggingface.co/HuggingFaceH4/mistral-7b-sft-beta, 2023. Accessed: 2024-10-01.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Dmitry Kobak, Rita González-Márquez, Emőke Ágnes Horvát, and Jan Lause. Delving into chatgpt usage in academic writing through excess vocabulary, 2024. URL https://arxiv.org/abs/2406.07016.

Seanie Lee, Minsu Kim, Lynn Cherif, David Dobre, Juho Lee, Sung Ju Hwang, Kenji Kawaguchi, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Moksh Jain. Learning diverse attacks on large language models for robust red-teaming and safety tuning, 2024. URL https://arxiv.org/abs/2405.18540.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback, 2024. URL https://arxiv.org/abs/2312.00886.

Minh Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2024. URL https://arxiv.org/abs/2407.01082.

OpenAI. Learning to reason with LLMs, 2023. URL https://openai.com/index/learning-to-reason-with-llms/. Accessed: 2024-10-02.

OpenAI. O1: Scalable Inference for Foundation Models, 2024. URL https://openai.com/o1/. Accessed: 2024-10-02.

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.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity?, 2024. URL https://arxiv.org/abs/2309.05196.

Peter S. Park, Philipp Schoenegger, and Chongyang Zhu. Diminished diversity-of-thought in a standard large language model, 2023. URL https://arxiv.org/abs/2302.07267.

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.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

CP Robert. Monte carlo statistical methods, 1999.

Andy Shih, Dorsa Sadigh, and Stefano Ermon. Long horizon temperature scaling, 2023. URL https://arxiv.org/abs/2302.03686.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf, 2024. URL https://arxiv.org/abs/2312.08358.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Hao Sun and Mihaela van der Schaar. Inverse-rlignment: Inverse reinforcement learning from demonstrations for llm alignment, 2024. URL https://arxiv.org/abs/2405.15624.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback, 2024. URL https://arxiv.org/abs/2401.04056.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1763–1775, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.103. URL https://aclanthology.org/2024.findings-acl.103.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023. URL https://arxiv.org/abs/2305.14975.

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/.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023. URL https://arxiv.org/abs/2310.16944.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints, 2023. URL https://arxiv.org/abs/2309.16240.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization, 2024. URL https://arxiv.org/abs/2405.16455.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling, 2024. URL https://arxiv.org/abs/2409.19817.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

APPENDIX

A THEORETICAL RESULTS

Proposition 3.1 (Two-Outcome RLHF Policy). *Suppose a population of raters prefers completion $y \succ y'$ with probability p. Then RLHF (or DPO) with KL-regularization penalty β has the optimal policy*

$$\pi(y) \propto \pi_{ref}(y)p^{1/\beta}. \tag{11}$$

Proof. From Rafailov et al. (2024), RLHF and DPO share the same optimal policy. So, it suffices to analyze the RLHF objective. First, consider the Bradley-Terry reward modeling objective

$$\max_r p \log \sigma(r(y) - r(y')) + (1-p) \log \sigma(r(y') - r(y)) \tag{12}$$

Because log-loss is a proper scoring rule, we know that the optimal r^* will result in distribution-matching

$$p = \sigma(r^*(y) - r^*(y')) = \frac{\exp(r^*(y))}{\exp(r^*(y)) + \exp(r^*(y'))}.$$

Following Rafailov et al. (2024), the solution to the KL-regularized RL problem gives us optimal policy

$$\pi(y) = \pi_{ref}(y) \exp(r^*(y))^{1/\beta}/Z \tag{13}$$

for a normalizing constant Z. We may rewrite this as

$$\pi(y) = \pi_{ref}(y)(\frac{\exp(r^*(y))}{\exp(r^*(y)) + \exp(r^*(y'))})^{1/\beta}/Z' \tag{14}$$

$$= \pi_{ref}(y)p^{1/\beta}/Z'. \tag{15}$$

for a new normalizing constant Z'. □

Proposition 3.2 (Two-Outcome DPL Policy). *Suppose a population of raters prefers completion $y \succ y'$ with probability $p(y \succ y')$. Then DPL with entropy bonus α and cross-entropy penalty β has the optimal policy*

$$\pi(y) \propto \pi_{ref}(y)^{\beta/\alpha}p^{1/\alpha}. \tag{16}$$

Proof. From Rafailov et al. (2024), RLHF and DPO share the same optimal policy. So it suffices to analyze the RLHF objective. First, consider the Bradley-Terry reward modeling objective

$$\max_r p \log \sigma(r(y) - r(y')) + (1-p) \log \sigma(r(y') - r(y)) \tag{17}$$

Because log-loss is a proper scoring rule, we know that the optimal r^* will result in distribution-matching $p = \sigma(r(y) - r(y')) = \exp(r(y))/(\exp(r(y)) + \exp(r(y')))$.

Following Proposition A.2, we obtain optimal policy

$$\pi(y) = \pi_{ref}(y)^{\beta/\alpha} \exp(r(y))^{1/\alpha}/Z \tag{18}$$

for a normalizing constant Z. We may rewrite this as

$$\pi(y) = \pi_{ref}(y)^{\beta/\alpha}(\frac{\exp(r(y))}{\exp(r(y)) + \exp(r(y'))})^{1/\alpha}/Z' \tag{19}$$

$$= \pi_{ref}(y)^{\beta/\alpha}p^{1/\alpha}/Z'. \tag{20}$$

for a new normalizing constant Z'. □

Proposition A.1 (Multi-Outcome DPL Policy). *Suppose a population of raters have preferences over completions that can be modeled by Plackett-Luce (i.e. there exists an r such that $\forall y_1, ..., y_N$ we have $p(y_1 \succ y_2, ..., y_N) = \frac{\exp(r(y))}{\sum_{i=1}^{N} \exp(r(y))}$). Then DPL with entropy bonus α and cross-entropy penalty β has the optimal policy*

$$\pi(y) \propto \pi_{ref}(y)^{\beta/\alpha} p_{best}(y)^{1/\alpha} \tag{21}$$

where $p_{best}(y) = p(y \succ_{y' \neq y} y')$ is the proportion of people for whom y is the most preferred completion.

Proof. With more than two completions, it is possible to form preference distributions that cannot be perfectly fit by a Bradley-Terry model (e.g. intransitive preferences). So, we first assume that the preference distribution can be represented by a Bradley-Terry model, which ensures that the reward learning step results in distribution-matching $p(y \succ y) = \frac{\exp(r(y))}{\exp(r(y)) + \exp(r(y'))}$.

Following Proposition A.2, we have optimal policy $\pi(y) = \pi_{ref}(y)^{\beta/\alpha} \exp(r(y))^{1/\alpha}/Z$. We may rewrite this as

$$\pi(y) = \pi_{ref}(y)^{\beta/\alpha} (\frac{\exp(r(y))}{\sum_{y'} \exp(r(y'))})^{1/\alpha}/Z' \tag{22}$$

$$= \pi_{ref}(y)^{\beta/\alpha} p_{best}(y)^{1/\alpha}/Z' \tag{23}$$

for a new normalizing constant Z. □

Proposition A.2 (DPL DPO Derivation). *The following objective*

$$\max_{\pi} \mathbb{E}_{y \succ y' \sim \mathcal{D}} [\log \sigma(\alpha \log \frac{\pi(y|x)}{\pi(y'|x)} - \beta \log \frac{\pi_{ref}(y|x)}{\pi_{ref}(y'|x)})] \tag{24}$$

shares the same optimal policy as the DPL RLHF objective

$$\max_{\pi} \mathbb{E}_{y \sim \pi(y|x)} [r(x, y)] + \alpha H(\pi(\cdot|x)) - \beta H(\pi_{ref}(\cdot|x), \pi(\cdot|x)). \tag{25}$$

Proof. For the RLHF objective, we begin by training a reward model using Bradley-Terry to find

$$r^*(x, y) = \arg\max_{r} \mathbb{E}_{y \succ y' | x \sim \mathcal{D}} [\log \sigma(r(x, y) - r(x, y'))] \tag{26}$$

Next, we find the optimal policy for the DPL RLHF objective. Let us represent policies $\pi(\cdot|x)$ and $\pi_{ref}(\cdot|x)$ and reward function $r(x, \cdot)$ as vectors $\boldsymbol{\pi}, \boldsymbol{\pi_{ref}}, \boldsymbol{r}$. We rewrite the DPL RLHF objective as

$$\max_{\boldsymbol{\pi}} \boldsymbol{\pi}^\top \boldsymbol{r} - \alpha \boldsymbol{\pi}^\top \log \boldsymbol{\pi} - \beta \boldsymbol{\pi}^\top \log \boldsymbol{\pi_{ref}} \tag{27}$$

$$= \boldsymbol{\pi}^\top (\boldsymbol{r} - \beta \log \boldsymbol{\pi_{ref}} - \alpha \log \boldsymbol{\pi_{ref}}) \tag{28}$$

We solve this constrained optimization problem using the Lagrangian:

$$\mathcal{L}(\boldsymbol{\pi}, \lambda) = \boldsymbol{\pi}^\top \boldsymbol{r} - \alpha \boldsymbol{\pi}^\top \log \boldsymbol{\pi} - \beta \boldsymbol{\pi}^\top \log \boldsymbol{\pi_{ref}} + \lambda \left(\sum_i \pi_i - 1 \right). \tag{29}$$

Taking the derivative with respect to $\boldsymbol{\pi}$ and setting it to zero, we obtain:

$$0 = \nabla_{\boldsymbol{\pi}} L(\boldsymbol{\pi}, \lambda) = \boldsymbol{r} - \alpha(\mathbf{1} + \log \boldsymbol{\pi}) - \beta \log \boldsymbol{\pi_{ref}} + \lambda \mathbf{1}. \tag{30}$$

Now, solving for $\boldsymbol{\pi}$, we have:

$$\alpha(\mathbf{1} + \log \boldsymbol{\pi}) = \boldsymbol{r} - \beta \log \boldsymbol{\pi_{ref}} + \lambda \mathbf{1} \tag{31}$$

$$\mathbf{1} + \log \boldsymbol{\pi} = \frac{\boldsymbol{r} - \beta \log \boldsymbol{\pi_{ref}} + \lambda \mathbf{1}}{\alpha} \tag{32}$$

$$\log \boldsymbol{\pi} = \frac{\boldsymbol{r} - \beta \log \boldsymbol{\pi_{ref}} + (\lambda - \alpha) \mathbf{1}}{\alpha} \tag{33}$$

$$\boldsymbol{\pi} = \exp \left(\frac{\boldsymbol{r} - \beta \log \boldsymbol{\pi_{ref}} + (\lambda - \alpha) \mathbf{1}}{\alpha} \right) \tag{34}$$

which implies the optimal DPL RLHF policy for all x

$$\pi^*(y|x) = \exp(\frac{1}{\alpha} r^*(x,y))\pi_{ref}(y|x)^{\beta/\alpha}/Z(x) \tag{35}$$

where $Z(x) = \exp(\alpha - \lambda)$ is a normalizing constant such that the probabilities over outputs sum to one.

Finally, we analyze the optimal policy for the DPO-style objective. Applying the substitution trick from Rafailov et al. (2024), we can reparameterize reward functions in terms of their optimal policies according to Equation A.2.

$$\pi(y|x) = \exp(\frac{1}{\alpha} r(x,y))\pi_{ref}(y|x)^{\beta/\alpha}/Z(x) \tag{36}$$

$$\log \pi(y|x) = \frac{1}{\alpha} r(x,y) + \frac{\beta}{\alpha}\log \pi_{ref}(y|x) - \log Z(x) \tag{37}$$

$$\alpha \log \pi(y|x) = r(x,y) + \beta \log \pi_{ref}(y|x) - \alpha \log Z(x) \tag{38}$$

$$r(x,y) = \alpha \log \pi(y|x) - \beta \log \pi_{ref}(y|x) + \alpha \log Z(x) \tag{39}$$

Observe that

$$r(x,y) - r(x,y') = \alpha \log \pi(y|x) - \beta \log \pi_{ref}(y|x) + \alpha \log Z(x) \tag{40}$$

$$- [\alpha \log \pi(y'|x) - \beta \log \pi_{ref}(y'|x) + \alpha \log Z(x)] \tag{41}$$

$$= \alpha \log \frac{\pi(y|x)}{\pi(y'|x)} - \beta \log \frac{\pi(y|x)}{\pi(y'|x)} \tag{42}$$

since the normalizing constants cancel.

We now substitute our reward function reparameterization into the Bradley-Terry objective to get the DPL DPO objective

$$\pi^*(y|x) = \max_{\pi} \mathbb{E}_{y \succ y'|x \sim \mathcal{D}}[\log \sigma(\alpha \log \frac{\pi(y|x)}{\pi(y'|x)} - \beta \log \frac{\pi_{ref}(y|x)}{\pi_{ref}(y'|x)})] \tag{43}$$

which satisfies the relationship $\pi^*(y|x) = \exp(\frac{1}{\alpha} r^*(x,y))\pi_{ref}(y|x)^{\beta/\alpha}/Z(x)$. $\square$

B Diversity-Quality Tradeoffs Experiment

B.1 Additional Diversity-Quality Results and Details on Diversity Metrics

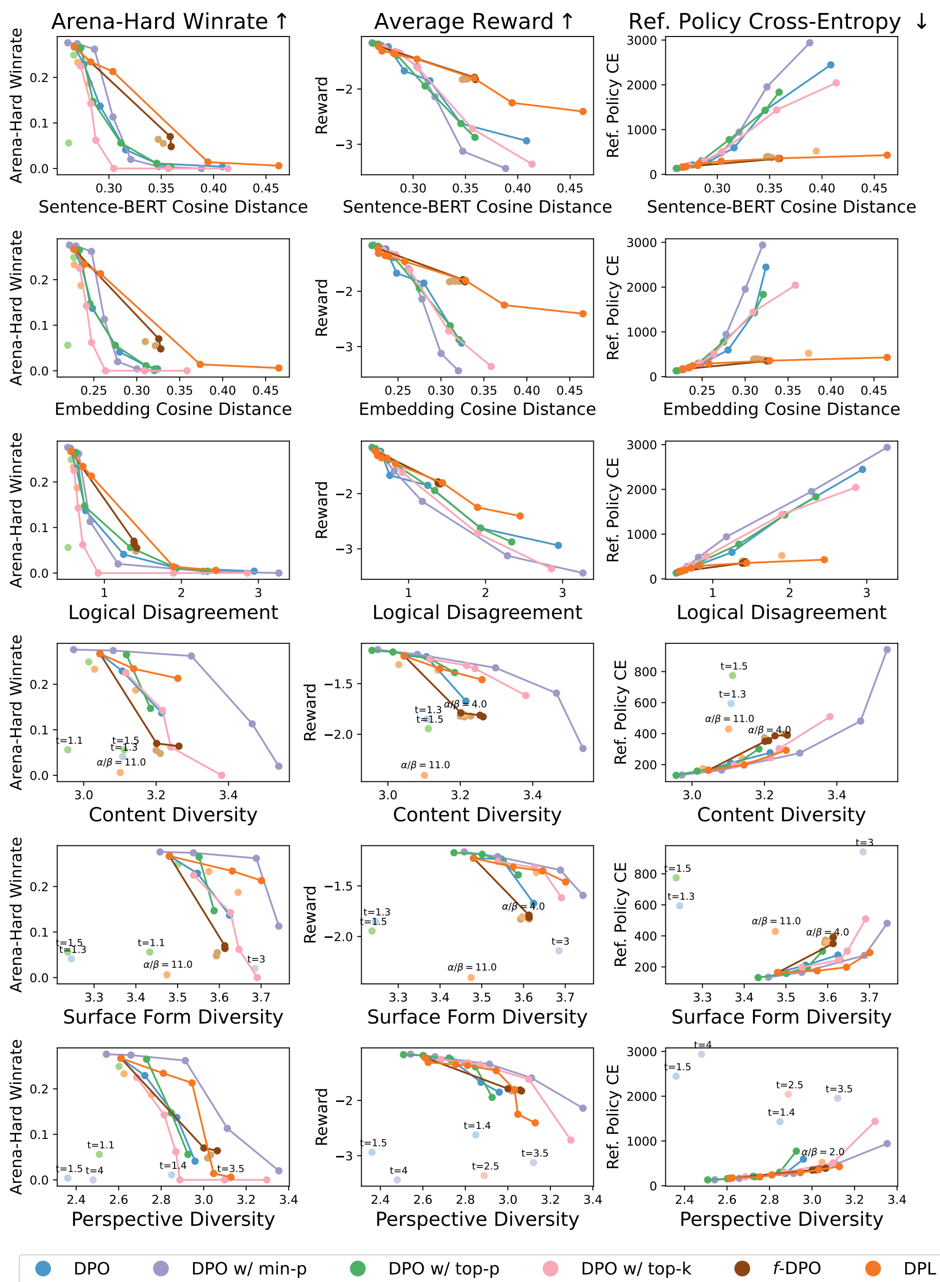


Figure 5: **Diversity-quality tradeoffs with additional diversity metrics.** Here, Sentence-BERT is an additional expected cosine distance metric in addition to "Embedding Cosine Distance" which uses the OpenAI Embeddings API. The remaining four diversity metrics are evaluated with an LLM judge on pools of responses. For DPL, we perform a global temperature sweep in range $t \in [1, 11]$. For all other methods, we sweep until outputs degenerate into unintelligible text. For some diversity metrics, such as content, surface form, and perspective diversity, low-quality generations are often rated as less diverse (see examples below). We label points off the Pareto frontier to help identify these cases.

Cosine distance diversity metrics. In Figure 2 and Figure 5, the Sentence-BERT Cosine Distance and Embedding Cosine Distance measure the expected cosine distance between response embedding pairs according to two different embedding models: 1) Sentence-BERT-Large (Reimers & Gurevych, 2019) and OpenAI's "text-embedding-3-small" model. For each of 500 held-out prompts from HH-RLHF, we sample $N = 16$ inputs and calculate the mean cosine distance between all pairs. Concretely, this is expressed in the formula

$$\mathbb{E}_{y_1,\ldots,y_N \sim \pi(y|x), x \sim D} \left[\frac{1}{N^2} \sum_{i,j} 1 - \frac{\phi(y_i)^\top \phi(y_j)}{||\phi(y_i)|| \cdot ||\phi(y_j)||} \right] \tag{44}$$

where $\phi(\cdot)$ is the embedding map acting on response y.

LLM-based diversity metrics. For the remaining diversity metrics, which we split the 16 responses into smaller pools and use "gpt-4o-mini-mini-2024-07-18" as a judge for response diversity. Here we provide the list of diversity evaluation prompts used for metrics in Figures 2 and 5.

Logical Agreement Evaluation Prompt

You will be presented with 2 responses to the same prompt. Your task is to analyze their logical agreement on a scale of 1-5, where 1 means completely divergent approaches or ideas and 5 means nearly identical logical frameworks or conclusions.

Prompt: {input}

Responses: {response_list}

Please provide a numerical rating (1-5) of their logical agreement. Output your response as Rating: X, where X is your rating. Afterwards, on the next line, please provide a brief explanation of your rating.

For the logical agreement prompt, we then subtract the resulting score from 5 to turn it into a logical disagreement metric.

Content Diversity Evaluation Prompt

You will be presented with 4 responses to the same prompt. Your task is to analyze their content diversity on a scale of 1-5, where 1 means nearly identical content or conclusions and 5 means completely divergent content or ideas.

Prompt: {input}

Responses: {response_list}

How diverse are the contents of the proposed responses? Please provide a numerical rating (1-5). Output your response as Rating: X, where X is your rating. Afterwards, on the next line, please provide a brief explanation of your rating.

Surface Form Diversity Prompt

You will be presented with 4 responses to the same prompt. Your task is to analyze their surface form diversity on a scale of 1-5, where 1 means nearly identical textual organization or structure, and 5 means completely varied textual arrangements or styles.

Prompt: {input}

Responses: {response_list}

How diverse are the surface forms of the proposed responses? Please provide a numerical rating (1-5). Output your response as Rating: X, where X is your rating. Afterwards, on the next line, please provide a brief explanation of your rating.

Perspective Diversity Prompt

You will be presented with 4 responses to the same prompt. Your task is to analyze their perspective diversity on a scale of 1-5, where 1 means nearly identical viewpoints or approaches and 5 means completely varied perspectives or approaches to the topic.

Prompt: {input}

Responses: {response_list}

How diverse are the perspectives of the proposed responses? Please provide a numerical rating (1-5). Output your response as Rating: X, where X is your rating. Afterwards, on the next line, please provide a brief explanation of your rating.

B.2 Example Generations

B.3 Training Setup

DPO and DPL Finetuning and Inference. For both DPO and DPL, we LoRA-finetune Mistral-7B-Instruct-v0.2 on HH-RLHF for 5,000 steps with batch size 8. We use LoRA rank $r_{LoRA} = 16$, regularization $\alpha_{LoRA} = 16$, and dropout $p_{LoRA} = 0.05$. We use learning rate $1e - 5$, 150 warmup steps, and max conversation length of 512 tokens.

For all runs, we use regularization parameter $\beta = 0.1$. For DPO with token-level temperature scaling, we sample with temperatures $t = 1$, 1.1, 1.2, 1.3, 1.4, and 1.5. When combined with min-p sampling, we choose $p_{base} = 0.1$ and temperatures $t = 1.3$, 1.5, 2, 2.5, 3, 3.5, and 4. When combined with top-p sampling, we choose $p = 0.9$ and temperatures $t = 1.1$, 1.2, 1.3, 1.4, 1.5, 1.6, and 1.7. When combined with top-k sampling, we choose $k = 180$ and temperatures $t = 1.1$, 1.2, 1.3, 1.5, 2, and 2.5. We stop at these temperatures for token-level methods because beyond this, responses are nearly always incoherent strings of tokens. For DPL, we train with global temperatures $\alpha/\beta = 1$, 1.1, 1.2, 1.3, 1.5, 2, 4, and 11.

Reward Model Training and Inference. As one of our quality metrics, we measure the average reward obtained by model completions against a separately trained reward function. To obtain this reward model, we LoRA-finetune Mistral-7B-Instruct-v0.2 with a reward head using the Bradley-Terry loss on the HH-RLHF dataset.

At inference time, we sample 500 prompts from a held-out validation split of HH-RLHF that neither the language models nor the reward model were trained on. We use this to calculate the average reward achieved by each model's completions.

C Best-of-N Problem-Solving Experiment

C.1 Relationship Between Problem Difficulty and Optimal Temperature

In this section, we study the following questions

1. Why do harder problems generally benefit from higher temperatures? And why does this benefit disappear once the temperature becomes too large?
2. Why are high temperatures especially helpful in high best-of-N sample settings?

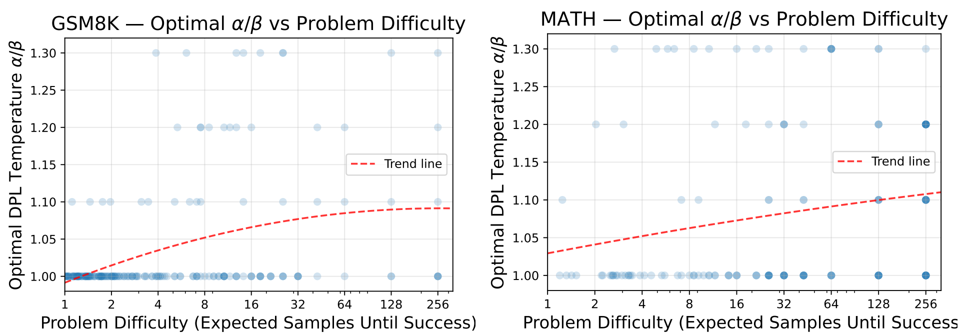


Figure 6: **Optimal DPL temperature as a function of problem difficulty.** On the x-axis, we measure problem difficulty as the expected number of samples needed for Mistral-Instruct-7B to achieve a correct answer. On the y-axis, we indicate which DPL temperature led to the best success rate on that problem. We then fit a polynomial trendline to the data, which finds a positive relationship between problem difficulty and higher optimal temperatures. Dots are plotted with transparency, meaning darker dots correspond to multiple problems.

Why harder problems benefit from higher temperatures, up to a point. We first consider the role of temperature in the single-sample (not best-of-N setting). As displayed in Figure 6, every problem has an optimal temperature that maximizes the probability of a successful solution. Increasing temperature flattens the distribution while decreasing it sharpens the distribution. Difficult problems are those where the trained model has a low probability of generating a correct solution.

During inference-time scaling procedures, our goal is to construct a sampling distribution that minimizes the number of samples needed to obtain a correct solution. For best-of-N, we receive feedback only upon generation completion, where samples must pass a final scoring function. This is akin to rejection sampling in statistics, where proposal distributions are chosen to minimize the number of samples required to approximate a target distribution.

In the rejection sampling literature, when little is known about the target distribution, it is well-known that flatter or high-entropy proposal distributions are best (Robert, 1999). Intuitively, highly concentrated proposal distributions that narrowly miss the target will achieve extremely low success rates. In contrast, flatter proposal distributions whose modes narrowly miss the target will still achieve reasonable success rates. The optimal proposal distribution scales in flatness inversely with one's confidence in its overlap with the target distribution (Geweke, 1989). However, an overly flat distribution is also inefficient, underestimating the true overlap with the target distribution.

This intuition is confirmed in Figure 6, where low-success problems are benefitted by DPL temperatures higher than 1, which produce flatter sampling distributions. However, when temperatures become too large, performance again degrades. Additional evidence can be found in the y-intercepts of Figure 3 and best-of-1 performance of Tables 1- 3. On easy and medium problems, DPO has a substantially higher single-sample success rate than temperature-scaled methods. However, on hard problems, this gap shrinks. Table 1 shows that on GSM8K and MATH hard splits, DPL is within 0.2% and 0.08% of DPL's best-of-1 success rate. This supports the finding that harder problems have higher optimal temperatures.

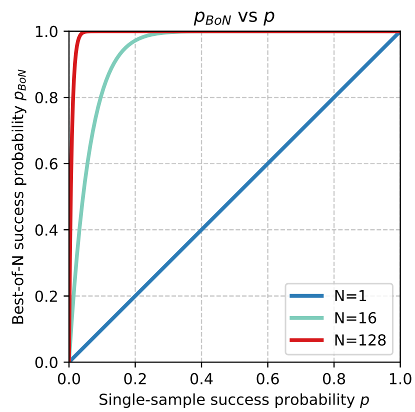


Figure 7: **Best-of-N success rate vs single-sample success rate as a function of N.** As N increases, the best-of-N success curve approaches a step function. Hard problems with low single-sample success rates benefit substantially from small increases in p. Meanwhile, the best-of-N success rate of problems past the "elbow" is already saturated, meaning that small decreases in p have little effect on p_{BoN}. The rapidly decreasing gradient of this curve favors high-variance strategies such as high-temperature sampling.

Why high temperatures are especially helpful in high sample best-of-N settings. For a given problem x, let use define the acceptable set A as the set of completions y that are correct. We can then define the probability of success for an LLM as $p = \pi(A|x) = \sum_{y \in A} \pi(y|x)$.

The best-of-N success rate can then be modeled by the CDF of a geometric distribution

$$p_{BoN} = 1 - (1 - p)^N. \tag{45}$$

Note that p_{BoN} is monotonic in single-sample success rate p. This means that for a given problem, the optimal temperature which maximizes single-sample success rate also maximizes best-of-N success rate. Thus the number of samples N does not influence the optimal sampling temperature for a problem.

Why, then, in Figure 3 do we observe high temperatures being differentially helpful at high N? While optimal temperatures do not change, the relative gains and losses from increasing temperature become asymmetrical.

Rather than single problem performance, we are interested in the aggregate success rate across a set of problems with different difficulty levels

$$\mathbb{E}_{x,A\sim D}[p_{BoN}] = \mathbb{E}_{x,A\sim D}[1 - (1 - \pi(A|x))^N]. \tag{46}$$

As shown in Figure 7, hard problems benefit more from small increases in p than easier problems suffer from an equivalent decrease in p. In aggregate, this causes the optimal temperature for the set of problems to trend upwards as N increases. Intuitively, at high N, we can afford to sacrifice some performance on easy problems (which will still succeed) to boost performance on hard problems (which benefit more from more diversity).

This behavior is a consequence of the concavity of the best-of-N success rate function $f(p) = 1 - (1-p)^N$. By Jensen's inequality, increasing the variance of p (e.g., via higher temperatures) can raise the aggregate success rate across problems, even though the single-problem optimal temperature remains unchanged (Boyd & Vandenberghe, 2004). This results in the "crossing" behavior seen in Figures 8 and 9. As N increases, $f(p)$ becomes more aggressively concave, and higher temperatures begin to outperform standard DPO.

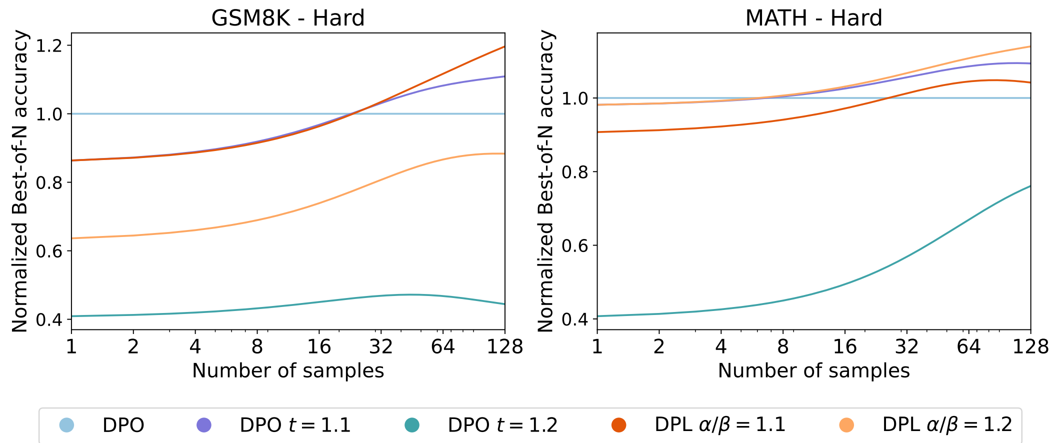


Figure 8: **Normalized best-of-N accuracy on difficult problem-solving instances.** We show best-of-N accuracy of each method as a fraction of the accuracy achieved by standard DPO. On these problems, higher temperature runs have a lower single-sample accuracy than standard DPO. However, some of them cross and eventually surpass DPO for large N. However, with too high of a token-level temperature, quality degradation is so significant that the curve never approaches standard DPO performance.

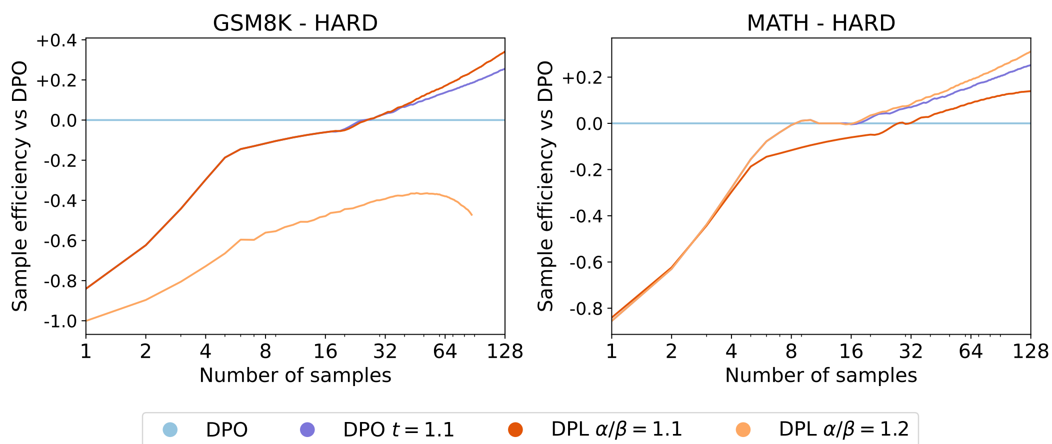


Figure 9: **Sample efficiency versus DPO.** For each number of standard DPO samples, we show how many samples are required for other methods to achieve the same best-of-N accuracy. At low samples, DPO is the most sample efficient method. However, at higher sampling budgets, DPL achieves the same accuracy as DPO with 25-35% computational savings. DPO $t = 1.2$ was not included in this graph as it lies far below all other methods.

One intuitive mechanism that drives this diversity-favoring non-convexity is the way that best-of-N treats duplicate solutions. As shown in Figure 10, redundant samples are a major source of sample inefficiency for DPO at large N. In the single sample case, putting high mass on likely completions increases the success probability. However, in the best-of-N setting, it is optimal to never sample the same solution twice, as it does not increase the probability of success. One advantage of high-temperature sampling in the best-of-N setting is simply that it produces fewer redundant samples.

Conclusion. This analysis provides theoretical grounding for our empirical finding that higher DPL temperatures benefit difficult problems in high-sample settings. It also suggests guidelines for practitioners: when compute allows for many samples, appropriately chosen high temperature sampling can be used to tackle the hardest problems in a dataset without significantly impacting overall performance.

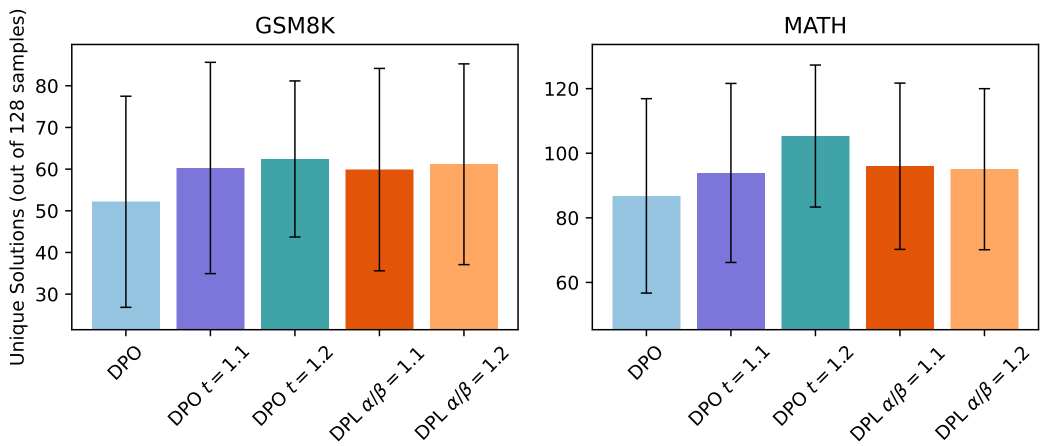


Figure 10: **Solution diversity and sampling efficiency.** For best-of-N sampling, repetition in final solutions is a major cause of sample inefficiency. At 128 samples, 30-40% of solutions are redundant and have already been sampled before. While convergent chain-of-thought reasoning can be helpful in settings such as majority voting, it is undesirable when a verifier is present. Both token-level temperature scaling and DPL increase the number of unique solutions sampled. For example, DPL samples 17% and 10% more unique solutions than DPO on GSM8K and MATH, respectively. This is equivalent to a 12% and 22% reduction in redundant solutions.

C.2 BEST-OF-N ACCURACY TABLES

We also provide results from Figure 3 in tabular form for an easier quantitative performance comparison.

Best-of-N	DPO t=1	DPO t=1.1	DPO t=1.2	DPL α/β=1.1	DPL α/β=1.2
GSM8K					
1	**1.43%**	1.24%	0.59%	1.24%	0.91%
4	**5.35%**	4.75%	2.25%	4.74%	3.53%
16	**16.92%**	16.37%	7.62%	16.30%	12.50%
64	37.98%	41.11%	17.77%	**42.43%**	32.91%
128	48.75%	54.07%	21.66%	**58.34%**	43.07%
MATH					
1	**0.86%**	0.85%	0.35%	0.78%	0.85%
4	**3.22%**	3.19%	1.37%	2.97%	3.20%
16	10.18%	10.44%	5.03%	9.89%	**10.50%**
64	22.12%	24.02%	14.81%	23.11%	**24.51%**
128	27.99%	30.62%	21.31%	29.16%	**31.91%**

Table 1: **Best-of-N accuracy on GSM8K and MATH dataset hard splits.** We provide Figure 3 results in tabular form for easier quantitative analysis. 64 samples and above on GSM8K and 16 samples and above on MATH, DPL outperforms standard and temperature-scaled DPO. At 128 samples on GSM8K, DPL achieves a 10% higher accuracy than standard DPO and a 4% higher accuracy than temperature-scaled DPO. At 128 samples on MATH, DPL achieves a 4% higher accuracy than DPO and 1% higher accuracy than temperature-scaled DPO. We expect these gaps to widen with more compute-efficient inference-time scaling methods than naive best-of-N.

Best-of-N	DPO t=1	DPO t=1.1	DPO t=1.2	DPL α/β=1.1	DPL α/β=1.2
GSM8K					
1	**8.95%**	6.06%	3.62%	5.79%	5.22%
4	**26.81%**	19.70%	12.39%	19.01%	17.39%
16	**55.63%**	46.67%	31.77%	46.05%	42.45%
64	**79.21%**	72.74%	55.10%	73.40%	67.33%
128	**85.30%**	80.17%	63.93%	81.14%	74.40%
MATH					
1	**5.40%**	3.67%	2.16%	3.43%	3.56%
4	**15.57%**	11.23%	7.59%	11.09%	11.61%
16	**36.15%**	26.74%	20.71%	27.62%	29.57%
64	**61.49%**	50.92%	39.31%	52.11%	56.41%
128	**69.07%**	62.66%	46.88%	62.01%	66.96%

Table 2: **Best-of-N accuracy on GSM8K and MATH dataset medium splits.** Standard DPO achieves the best accuracy in all settings here, although DPL begins to approach DPO at high sample counts, especially on the harder MATH dataset.

Best-of-N	DPO t=1	DPO t=1.1	DPO t=1.2	DPL α/β=1.1	DPL α/β=1.2
GSM8K					
1	**34.18%**	26.40%	15.58%	24.33%	21.86%
4	**68.94%**	58.70%	41.93%	56.60%	52.70%
16	**91.39%**	85.67%	72.19%	84.75%	82.04%
64	**97.97%**	97.00%	89.48%	96.51%	95.62%
128	**98.80%**	98.51%	92.86%	98.38%	98.22%
MATH					
1	**22.51%**	16.99%	14.01%	18.02%	15.09%
4	**46.52%**	40.07%	34.89%	42.09%	38.32%
16	**72.03%**	65.00%	60.18%	66.99%	63.60%
64	**89.72%**	87.51%	83.32%	88.49%	81.86%
128	92.77%	**94.49%**	91.84%	92.63%	86.05%

Table 3: **Best-of-N accuracy on GSM8K and MATH dataset easy splits.** Standard DPO is by far the best performer prior to saturation (accuracies near 100). At 128 samples on MATH however, temperature scaled DPO outperforms standard DPO by 2% and DPL matches standard DPO.

C.3 Training and Inference Setup

DPO and DPL Finetuning and Inference. For both DPO and DPL, our base model is a Mistral-7B base model that has been full-parameter supervised fine-tuned on the UltraChat dataset. We then LoRA-finetune this model on Ultrafeedback-200k for one epoch. We use batch size 4, LoRA rank $r_{LoRA} = 64$, regularization $\alpha_{LoRA} = 64$, and dropout $p_{LoRA} = 0.05$. We use learning rate $1e-5$, 150 warmup steps, and max conversation length of 1024 tokens. This replicates the Zephyr training recipe, except that we use $\beta = 0.1$ instead of $\beta = 0.01$, which substantially improved problem-solving performance. While more computationally expensive than the HH-RLHF training runs, we found it necessary to use a stronger fine-tuning recipe like Zephyr in order to get meaningful results in difficult mathematical problem-solving settings.

At inference-time, we use standard few-shot prompts to encourage chain-of-thought reasoning.

GSM8K few-shot prompt

```
As an expert problem solver solve step by step the following mathematical questions. Place your answer after the “####” symbol.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?
A: We start with 15 trees. Later we have 21 trees. The difference must be the number of trees they planted. So, they must have planted 21 - 15 = 6 trees. The answer is 6. #### 6

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5. #### 5

Q: {question}
A:
```

MATH few-shot prompt

```
Given a mathematics problem, determine the answer. Simplify your answer as much as possible. Output your answer in the format Answer: $\boxed{answer}$

Problem: Let $f(x)=x^3+3$ and $g(x) = 2x^2 + 2x +1$. What is $g(f(-2))$?
Answer: We note that $f(-2)=(-2)^3+3=-5$, so $g(f(-2))=g(-5)=2\cdot(-5)^2+2\cdot(-5)+1=41$ So, $\boxed{41}$
###
Problem: What is the greatest possible value of $x+y$ such that $x^2 + y^2 =90$ and $xy=27$?
Answer: We have $(x+y)^2=x^2+y^2+2xy=90+2\cdot 27=144$, so $x+y=12$ or $x+y=-12$. We want the larger value $x+y=\boxed{12}$
###
Problem: {problem}
```

D LOGIT CALIBRATION EXPERIMENT

Multiple Choice Prompt

Please answer the following multiple choice question. Start your answer with the capital letter corresponding to your choice:

{question}

Answer choices:
A. {option_a}
B. {option_b}
C. {option_c}
D. {option_d}

Your answer: