Learning from One Another: Toward Complementary Knowledge Transfer via Localized Preference Deltas

Chi-Ping Su [*1] **Jing-Hong Hu** [*1] **Ching-Hsun Tseng** [2] **Shin-Jye Lee** [1]

Abstract

Modern language models increasingly exhibit complementary local strengths, creating heterogeneous model ecosystems where useful capabilities are distributed across models rather than concentrated in a single dominant teacher. This challenges existing transfer pipelines, which typically assume that supervision should originate from a globally stronger model. We propose **Delta Preference Transfer (DPT)**, a localized synthetic preference construction framework for preference-based transfer across language models. DPT constructs chosen/rejected preference pairs within the response space of a single generator by introducing controlled local failures into the rejected response, encouraging the target model to learn localized behavioral refinements rather than directly imitating a globally stronger teacher. This enables targeted preference transfer without requiring external judges, reward models, or globally ordered teacher–student assumptions during either data construction or training. Across diverse settings, DPT consistently improves mathematical reasoning performance and achieves competitive or superior results compared with existing preference-based transfer and knowledge distillation baselines. Our analyses further suggest that DPT provides a practical step toward complementary knowledge transfer in heterogeneous language model ecosystems.

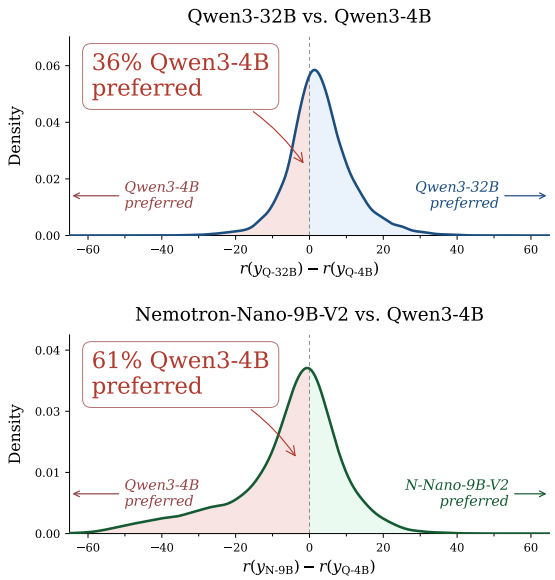

Figure 1. Reward-score difference distributions between Qwen3-4B responses and responses from two globally stronger models. Negative regions indicate prompts where Qwen3-4B is preferred, revealing substantial instance-level preference variation across models.

1. Introduction

Modern language models are increasingly developed through diverse architectures, data mixtures, and post-training recipes. As a result, models increasingly exhibit complementary local strengths that are not fully captured by aggregate benchmark evaluations (Ong et al., 2024; Jiang et al., 2023). Although larger frontier models still achieve the strongest average performance, model scale alone no longer fully determines response quality or specialization. Recent compact open models remain highly competitive with substantially larger systems on specific reasoning and instruction-following tasks (Team et al., 2025; Abouelenin et al., 2025; Yang et al., 2025). As useful capabilities become increasingly distributed across heterogeneous models, this raises an intriguing question: ***Can language models effectively learn from one another's complementary strengths?***

Knowledge distillation (KD) provides the classical framework for model transfer, where a student learns from a stronger teacher model (Hinton et al., 2015; Gou et al., 2021; Su et al., 2025). More recent preference-based approaches extend this paradigm through chosen/rejected preference

[*]Equal contribution [1]National Yang Ming Chiao Tung University [2]Delta Research Center. Correspondence to: Shin-Jye Lee <camhero@gmail.com>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

pairs rather than exact behavior imitation (Rafailov et al., 2023; Tunstall et al., 2023; Zhang et al., 2024a; Li et al., 2025). Despite these differences, both paradigms typically rely on a globally ordered teacher–student assumption in which larger or stronger models are treated as uniformly preferable supervision sources.

Recent work increasingly challenges this assumption, showing that weaker models can still provide useful supervision signals for stronger ones (Burns et al., 2023; Tao & Li, 2025; Roth et al., 2024), and that the most effective generator is not always the largest or strongest model (Kim et al., 2025). To better understand this phenomenon, we compare response-level reward differences on 20,000 prompts randomly sampled from NuminaMath-CoT between Qwen3-4B and two stronger models: the larger same-family Qwen3-32B and the heterogeneous Nemotron-Nano-9B-V2. Responses are evaluated using Skywork-Reward-V2-Llama-3.1-8B-40M (Liu et al., 2025). Both models substantially outperform Qwen3-4B on aggregate capability benchmarks such as MMLU (Yang et al., 2025; NVIDIA, 2025). However, as shown in Figure 1, Qwen3-4B responses remain preferred on 36% of prompts against Qwen3-32B and 61% against Nemotron-Nano-9B-V2. These results suggest that globally stronger models do not uniformly dominate at the response level and can still be locally outperformed on a substantial fraction of prompts.

This creates a fundamental challenge for existing transfer pipelines. *Size-based* preference construction heuristics (Zhang et al., 2024a; Li et al., 2025; Tunstall et al., 2023) implicitly assume that outputs from larger models should always be preferred, while outputs from smaller models are treated as rejected responses. Such heuristics are efficient, but can produce incorrect preference directions when models exhibit complementary local strengths. *Judge-based* pipelines (Cui et al., 2023; Lambert et al., 2024; Zhu et al., 2025) partially address this issue by evaluating candidate responses instance-by-instance, but require additional external scoring stages to determine preference direction at scale. Consequently, existing pipelines still struggle to construct scalable and robust transfer signals for heterogeneous model ecosystems.

We therefore revisit this problem from a different perspective. Rather than focusing on selecting which model should supervise another, we instead aim to construct learning signals that enable the target model to benefit from complementary strengths present across heterogeneous models. Inspired by recent findings from (Geng et al., 2025), which suggest that preference learning can remain effective even when preference pairs are generated from weaker models as long as meaningful quality differences exist, we hypothesize that constructing localized quality differences within a single generator model's response space may provide a more robust interface for heterogeneous capability transfer.

Building on this idea, we propose Delta Preference Transfer (DPT), a localized synthetic preference construction framework for preference-based transfer across language models. Rather than determining which model should supervise another, our method constructs *localized preference deltas*—small but reliable quality differences between otherwise highly similar responses—within the response space of a single model through controlled failure-mode generation. This enables preference transfer through localized behavioral improvements rather than direct imitation of a globally stronger teacher, providing a more robust transfer signal for heterogeneous language models. Through experiments across homogeneous and heterogeneous transfer settings, we show that DPT consistently improves downstream mathematical reasoning performance and outperforms existing preference-based transfer methods as well as conventional KD baselines across diverse model pairings. Our main contributions are summarized as follows:

- We introduce **Delta Preference Transfer (DPT)**, a preference construction framework that facilitates complementary capability transfer through localized preference deltas rather than globally ordered teacher–student supervision.
- We propose controlled failure-mode generation for constructing reliable chosen/rejected preference pairs with small but meaningful local quality differences, enabling self-contained preference transfer without external judges.
- We demonstrate that DPT consistently outperforms existing preference-based transfer methods and conventional knowledge distillation baselines across multiple mathematical reasoning benchmarks and transfer settings.

2. Related Work

2.1. Global Teacher–Student Transfer

Existing transfer methods for language models typically assume that supervision should originate from a globally stronger teacher. KD establishes the classical transfer paradigm in which a student learns by imitating the predictive behavior of a stronger teacher model (Hinton et al., 2015; Gou et al., 2021). At LLM scale, distillation has been extended to sequence-level generation and reasoning transfer (Gu et al., 2026; Ko et al., 2024). Preference-based distillation further relaxes exact imitation by transferring supervision through chosen/rejected response pairs rather than token-level matching. Representative approaches such as Zephyr, PLaD, and DPKD construct preference data directly from teacher and student responses (Tunstall et al.,

2023; Zhang et al., 2024a; Li et al., 2025). Despite these differences, existing transfer pipelines largely share the assumption that the teacher should be globally preferable to the student. Recent studies, however, suggest that transfer effectiveness does not always scale monotonically with teacher capability due to capacity gaps and heterogeneous specialization across models (Zhang et al., 2025). Our work extends this perspective by reconsidering transfer under heterogeneous capability distributions, where useful supervision may arise from localized improvements rather than globally dominant teachers.

2.2. Preference Optimization and Construction

Preference optimization provides a scalable alternative to RLHF by training directly on relative preference pairs. Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplifies preference learning into a single-stage objective, enabling broad adoption of preference-based post-training. Subsequent work has improved optimization objectives, data mixtures, and post-training recipes (Lambert et al., 2024; Yang et al., 2024). As preference optimization became increasingly standardized, the central challenge shifted toward scalable and reliable preference construction. UltraFeedback and Tulu 3 generate candidate responses from multiple language models and select preferred responses using LLM judges or reward models (Cui et al., 2023; Lambert et al., 2024; Zhu et al., 2025). These pipelines enable large-scale AI feedback, but often rely either on model-scale heuristics or expensive judge-based scoring pipelines to determine preference direction. Moreover, judge-based annotations can inherit calibration errors and stylistic biases from the evaluator model (Thakur et al., 2025; Ye et al., 2024). In contrast, our method constructs preference direction structurally through localized response perturbations, avoiding the need for external preference judges or globally ordered model rankings during DPT data construction and training.

2.3. Supervision Beyond Stronger Teachers

Recent work increasingly challenges the assumption that useful supervision must originate from a globally stronger teacher. Self-improvement and self-rewarding approaches demonstrate that language models can generate supervision signals for their own post-training through self-play, self-critique, and self-generated preferences (Chen et al., 2024; Yuan et al., 2025; Lyu et al., 2025). Further, weak-to-strong generalization shows that weak supervision can still elicit stronger capabilities in larger models (Burns et al., 2023; Tao & Li, 2025). Complementary transfer studies similarly suggest that different pretrained models may contain useful knowledge even when they are weaker in benchmark performance (Roth et al., 2024), while synthetic data studies show that the most effective data generator is not always the largest or strongest model (Kim et al., 2025). Together, these findings suggest that effective learning signals can arise from models regardless of their relative size or overall capability. More recently, delta learning showed that preference optimization can remain effective even when preference pairs are generated from weaker models, provided meaningful local quality differences exist between chosen and rejected responses (Geng et al., 2025). Our work extends this perspective by constructing localized preference deltas directly within a single model's response space. Rather than requiring one model to dominate another overall, our method only assumes that locally preferable behavioral directions can be reliably instantiated and learned.

3. Methodology

3.1. Background

Preference-based Transfer. Preference-based transfer trains a target model using chosen/rejected response pairs rather than exact imitation of a teacher distribution. Given an instruction x, the training data consists of a preferred response y^+ and a dispreferred response y^-. The transfer objective encourages the target model to increase the relative likelihood of y^+ over y^-:

$$\pi_\theta(y^+ \mid x) > \pi_\theta(y^- \mid x). \tag{1}$$

Unlike classical KD, which transfers knowledge through distribution matching, preference-based transfer encodes transferable behavior implicitly through relative quality differences between responses. Consequently, the effectiveness of transfer depends heavily on how preference direction is constructed.

Direct Preference Optimization. We train models using DPO (Rafailov et al., 2023), which optimizes a policy to prefer chosen responses over rejected ones relative to a frozen reference policy π_{ref}:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,\, y^+,\, y^-)} \left[\log \sigma \left(\beta\, \Delta r_\theta(y^+, y^- \mid x)\right)\right]. \tag{2}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, $\beta > 0$ controls regularization against the reference policy, and $\Delta r_\theta(y^+, y^- \mid x)$ is the DPO reward margin induced by the target and reference policies. The resulting optimization signal depends directly on the quality and structure of the underlying preference pairs, making preference construction a critical component of effective transfer.

3.2. Delta Preference Transfer

DPT constructs preference pairs directly within the response space of a single generator through controlled failure-mode generation. We define a localized preference delta as a small but reliable quality difference between two otherwise

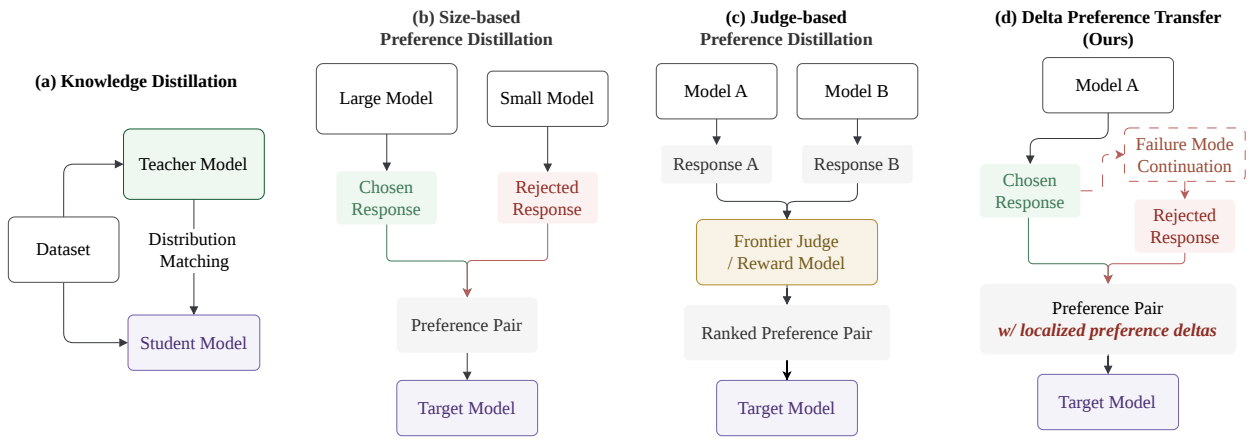

Figure 2. Comparison of knowledge transfer pipelines. DPT constructs localized preference deltas using a single-generator continuation process through controlled failure-mode generation, avoiding globally ordered teacher assumptions and external judge models during data construction.

Algorithm 1 DPT Data Generation

Require: Instruction set $\mathcal{X}$, failure mode set F, generator π_{gen}
Ensure: Preference dataset $\mathcal{D}$

```
D ← ∅
for each instruction x ∈ X do
    f ~ Uniform(F)
    y+ ~ π_gen(· | x)
    y− ~ π_gen(· | x, y+, f)
    D ← D ∪ {(x, y+, y−)}
end for
return D
```

highly similar responses. Rather than constructing unrelated chosen/rejected pairs, DPT preserves most of the original reasoning trajectory while introducing a targeted local degradation that induces a reliable preference ordering.

Formally, given an instruction x, the generator first produces a preferred response

$$y^+ \sim \pi_{\text{gen}}(\cdot \mid x), \tag{3}$$

then generates a corresponding rejected response conditioned on the chosen response and failure mode f:

$$y^- \sim \pi_{\text{gen}}(\cdot \mid x, y^+, f), \tag{4}$$

where f specifies the localized failure introduced into the rejected response. The resulting chosen and rejected responses therefore remain highly aligned semantically and stylistically while differing primarily in the localized reasoning behavior induced by the failure mode.

Figure 2 compares DPT with existing transfer pipelines. Classical knowledge distillation transfers knowledge through teacher–student distribution matching. Size-based preference distillation constructs preference pairs by directly assigning generator outputs as chosen responses (typically from larger models) and target-model outputs as rejected responses. Judge-based preference distillation instead ranks candidate responses using external reward models or frontier judges. In contrast, DPT constructs localized preference deltas within a single generator rollout through controlled failure-mode generation. This construction shifts preference transfer from selecting globally preferable supervision sources to learning localized complementary behaviors.

To construct localized preference deltas, the generator is prompted to produce a strong response and then continue with a corresponding plausible-but-flawed variant conditioned on a predefined failure mode $f \in F$. Specifically, we adopt three failure modes in this paper:

- **Subtle Hallucination**: introduces a plausible but unsupported concept, theorem, or proof step into the reasoning process.
- **Premature Conclusion**: terminates the reasoning process before completing a necessary intermediate inference or verification step.
- **Logical Inconsistency**: introduces a subtle contradiction between claims, assumptions, or intermediate conclusions.

The failure mode f is sampled uniformly for each instruction, encouraging diversity across the generated preference dataset. The full generation prompt and failure-mode templates are provided in Appendix A. Algorithm 1 summarizes the data generation procedure. The constructed preference pairs are used directly for DPO training via Equation 2. The

Table 1. Results on Qwen3-4B using generators of different scales and model families.

Generator	Method	MATH500	GSM8K	AIME24	AMC23	Gaokao23-EN	Avg	Δ
-	Base	91.2±0.4	94.4±0.3	50.0±6.9	85.0±2.0	82.9±0.8	80.7	–
Qwen3-32B	KD (SKL)	92.6±0.3	94.7±0.1	50.0±6.9	90.0±1.2	81.3±0.4	81.7	(+1.0)
	Size-based	91.6±0.3	94.5±0.1	43.3±4.7	82.5±2.4	81.8±0.5	78.7	(-2.0)
	Judge-based	91.6±0.5	94.5±0.2	53.3±6.3	95.0±3.1	83.1±0.4	83.5	(+2.8)
	Ours	**93.0**±0.1	**95.1**±0.2	**63.3**±2.7	**95.0**±2.0	**84.2**±0.7	**86.1**	**(+5.4)**
Qwen3-1.7B	KD (SKL)	90.4±1.1	89.2±2.3	40.0±6.9	92.5±1.2	81.0±1.0	78.6	(-2.1)
	Size-based	77.0±0.6	93.3±0.2	23.3±4.2	47.5±2.0	69.1±0.6	62.0	(-18.7)
	Judge-based	91.2±1.0	94.8±0.2	60.0±7.2	95.0±0.0	83.4±0.8	**84.9**	**(+4.2)**
	Ours	**92.2**±0.2	**95.1**±0.2	**60.0**±3.2	87.5±0.0	**83.6**±0.6	83.7	(+3.0)
Nemotron-Nano-9B-v2	Size-based	88.6±0.4	94.2±0.2	56.7±1.8	82.5±2.1	79.5±0.3	80.3	(-0.4)
	Judge-based	**92.4**±0.3	95.0±0.3	60.0±1.6	**95.0**±4.7	**84.2**±0.3	**85.3**	**(+4.6)**
	Ours	92.0±0.2	**95.1**±0.2	**66.7**±1.6	90.0±2.0	82.9±0.8	**85.3**	**(+4.6)**

reference policy π_{ref} is initialized from the baseline checkpoint of π_θ, allowing DPO to preserve the model's original capabilities while encouraging locally preferable behavioral updates. No external reward model, verifier, or cross-model comparison is required during training.

4. Experiments

4.1. Experimental Setup

Data We construct preference pairs from 20k randomly sampled prompts drawn from NuminaMath-CoT (LI et al., 2024), following the UNI-DPO setup (Peng et al., 2026). The same prompt set is used across all methods for both preference construction and training to ensure fair comparison. For each prompt, the generator model produces both a chosen and rejected response under identical decoding conditions (temperature 0.6, top-p 0.95, top-k 20). The chosen response follows standard chain-of-thought reasoning, while the rejected response is generated by continuing from the chosen response with a localized failure-mode corruption sampled from three categories: subtle hallucination, premature conclusion, and logical inconsistency. All experiments use approximately 20k filtered preference pairs after removing malformed generations and duplicate responses.

Models. For generators, we evaluate Qwen3-family (Yang et al., 2025) models at scales ranging from 1.7B to 32B, together with the cross-family Nemotron-Nano-9B-v2 (NVIDIA, 2025) generator. Qwen3-4B serves as the primary target model for generator comparison experiments, while Qwen3-8B, Llama-3.1-8B, and Llama-3.2-3B (Grattafiori et al., 2024) are used for scaling and cross-family transfer evaluation.

Baselines. We compare against four baselines: (1) **Base**, the unmodified target model. (2) **KD (SKL)** (Ko et al., 2024), which matches the generator distribution using the symmetric KL distillation objective from DistiLLM. (3) **Size-based**, following PLaD (Zhang et al., 2024a), where generator responses are *always* treated as chosen and target-model responses as rejected. To ensure consistency across all generator settings, we apply this protocol even when the generator is weaker than the target model. This baseline evaluates whether directly preferring generator outputs is sufficient for effective transfer. (4) **Judge-based**, a reward-model-based preference construction strategy in which generator and target-model responses are ranked using Skywork-Reward-V2-Llama-3.1-8B-40M, with the higher-scoring response selected as chosen.

Training and Evaluation. For preference-based distillation methods, all models are fine-tuned using DPO with $\beta = 0.01$ for one epoch. We use AdamW with a cosine learning rate schedule, a peak learning rate of 1×10^{-5}, and 10% linear warmup. Training is conducted with a global batch size of 128 and LoRA adapters (rank 64, $\alpha = 128$, dropout 0.05) applied to the query, key, value, and output projections. For KD, we adopt the training setup from (Zhang et al., 2024b) while using the same LoRA configuration for fair comparison. We report pass@1 accuracy on MATH500, GSM8K, AIME24, AMC23, and Gaokao2023-EN using the same decoding configuration across all methods (temperature 0.6, top-p 0.95, maximum length 16,384 tokens). Results are averaged over five random seeds (0–4), with thinking mode enabled for the Qwen family models.

4.2. Results

Transfer with Different Generators. Table 1 presents results on Qwen3-4B using generators of varying scales and

Table 2. Results on diverse target model families and parameter scales.

Target Model	Generator	MATH500	GSM8K	AIME24	AMC23	Gaokao23-EN	Avg	Δ
Qwen3-8B	Base	92.0±0.6	95.3±0.3	60.0±1.6	87.5±1.2	85.7±1.2	84.1	–
	Qwen3-4B	**94.2**±0.7	**95.8**±0.1	63.3±1.6	**92.5**±2.0	**86.2**±0.6	86.4	(+2.3)
	Qwen3-32B	93.6±0.2	**95.8**±0.2	**73.3**±4.7	**92.5**±4.1	85.2±0.3	**88.1**	**(+4.0)**
Llama-3.1-8B	Base	38.6±0.8	39.5±0.8	3.3±0.0	22.5±2.4	23.9±2.3	25.9	–
	Qwen3-4B	**44.6**±0.5	**65.5**±1.0	**6.7**±2.7	**30.0**±2.4	**38.2**±1.2	**37.0**	**(+11.1)**
Llama-3.2-3B	Base	24.8±0.6	37.5±0.5	10.0±1.6	12.5±2.4	20.8±0.7	21.2	–
	Qwen3-4B	**50.4**±0.8	**82.6**±0.3	6.7±0.0	**17.5**±4.7	**41.6**±0.5	**39.8**	**(+18.6)**

Table 3. Ablation study on failure modes using Nemotron-Nano-9B-v2 as the generator and Qwen3-4B as the target model. Each row corresponds to a single failure mode used during preference-pair generation.

Failure Mode	MATH500	GSM8K	AIME24	AMC23	Gaokao23-EN	Avg	Δ
Baseline	91.2±0.4	94.4±0.3	50.0±6.9	85.0±2.0	82.9±0.8	80.7	–
Logical Inconsistency	**92.0**±0.2	95.0±0.0	53.3±3.2	**90.0**±1.2	83.4±0.5	82.7	(+2.0)
Premature Conclusion	91.6±0.3	95.1±0.4	56.7±4.2	87.5±1.2	83.6±0.6	82.9	(+2.2)
Subtle Hallucination	**92.0**±0.3	**95.3**±0.4	**60.0**±1.6	87.5±3.1	**83.9**±0.1	**83.7**	**(+3.0)**

model families. Our method remains consistently effective across all generator settings, achieving the strongest and most stable overall performance, with gains of up to +5.4 average points in the strong-generator setting. In the weaker-generator setting with Qwen3-1.7B, Judge-based transfer achieves the best performance (+4.2%), suggesting that reward-model ranking can preserve useful preference signals from weaker generators while filtering harmful supervision. Still, by encouraging the target model to learn from localized preference signals, our method achieves competitive performance without requiring additional judge-model inference or response-ranking steps during data construction. In contrast, both KD and Size-based transfer degrade substantially in the weaker-generator setting. Since the Size-based baseline directly treats generator responses as chosen responses, these results suggest that simply learning from weaker-generator outputs is insufficient for effective transfer. The strong performance of DPT with a substantially weaker generator therefore indicates that transfer effectiveness arises not only from response quality, but also from the structure of the constructed preference pairs.

The Nemotron-Nano-9B-v2 results further highlight the advantage of localized preference transfer. Despite not being uniformly stronger than the target model (Figure 1), our method successfully transfers useful behaviors from the Nemotron-Nano-9B-v2 generator and consistently outperforms Size-based baseline while matching the performance of the Judge-based method. Additionally, our method matches the average performance achieved using the much larger Qwen3-32B generator (within 1 average point) while attaining the strongest AIME24 result overall. These observations are consistent with the hypothesis that localized preference deltas can facilitate the transfer of complementary strengths between heterogeneous models that are not captured by direct imitation alone.

Scaling to Stronger Target Models and Cross-Family Transfer. Table 2 shows that localized preference transfer remains effective across both stronger and cross-family target models. Notably, Qwen3-8B continues to benefit substantially despite already achieving strong baseline performance, suggesting that localized preference deltas remain informative even when the target model is already highly capable. Using Qwen3-32B as the generator improves average performance by +4.0 points while achieving the strongest AIME24 result across all evaluated settings at 73.3.

Interestingly, the substantially smaller Qwen3-4B generator also yields strong improvements, including the highest MATH500 score among all evaluated generator settings for Qwen3-8B. This suggests that transfer effectiveness does not depend solely on global model scale or aggregate capability, but also on the localized complementary strengths captured by the generated preference deltas. As shown in Figure 1, different generators exhibit distinct response-level preference patterns even within the same model family, which may induce complementary learning signals for the target model.

Cross-family transfer to the Llama family further demonstrates the robustness of localized preference transfer. Despite substantial architectural and post-training differences between Qwen and Llama models, localized preference

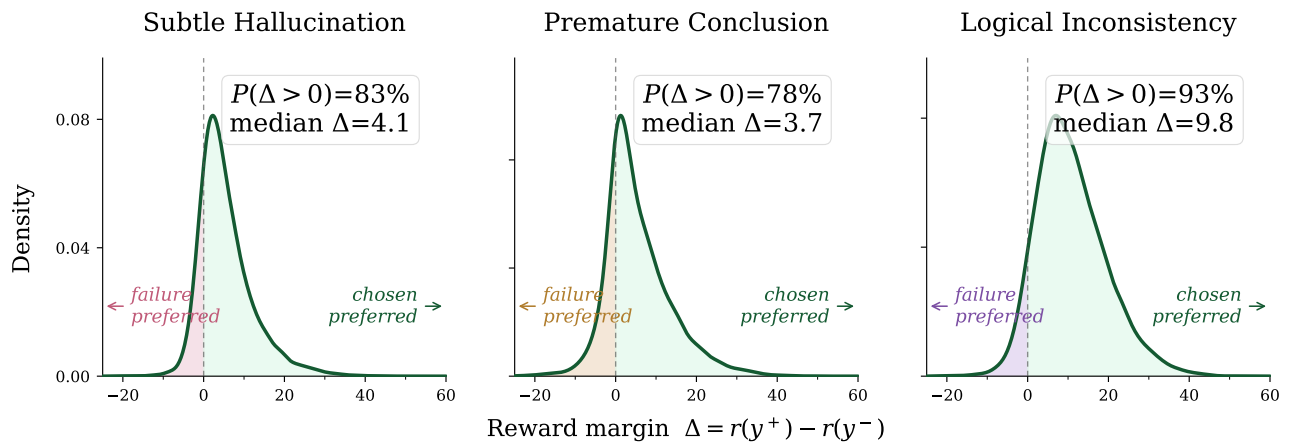

Figure 3. Reward-margin distributions induced by different failure modes for responses generated by Nemotron-Nano-9B-v2, evaluated using the Skywork-Reward-V2-Llama-3.1-8B-40M reward model. Subtle Hallucination induces smaller but consistently positive preference margins, whereas Logical Inconsistency produces larger and easier-to-separate reward gaps.

deltas remain highly transferable across model families, yielding gains of +11.1 for Llama-3.1-8B and +18.6 for Llama-3.2-3B. The particularly large improvements on smaller Llama models suggest that localized preference transfer may provide a lightweight mechanism for importing complementary reasoning behaviors across heterogeneous model ecosystems without requiring globally aligned teacher distributions.

5. Analysis

5.1. Ablation Study

Failure Modes. Table 3 analyzes the contribution of individual failure modes using Nemotron-Nano-9B-v2 as the generator and Qwen3-4B as the target model. All three failure modes consistently improve over the baseline, confirming that localized structured corruptions provide effective preference signals for transfer. The gains remain relatively stable across failure types, suggesting that DPT is not overly dependent on a single corruption strategy and can benefit from diverse forms of localized preference deltas.

Among the three failure modes, Subtle Hallucination achieves the strongest overall performance and the highest average score. Unlike conventional reasoning errors, these failures preserve most of the original reasoning trajectory while introducing fabricated justification or unsupported claims, producing more subtle preference distinctions. Premature Conclusion and Logical Inconsistency also yield consistent improvements, indicating that incomplete reasoning trajectories and locally contradictory reasoning both provide useful supervision signals for preference transfer. Overall, these results suggest that multiple forms of localized corruption can contribute complementary learning signals, while subtler failures may produce particularly informative preference supervision.

5.2. Quality of Generated Pairs

To estimate the quality of the generated preference pairs, we analyze the reward-margin distributions induced by each failure mode using the Skywork-Reward-V2-Llama-3.1-8B-40M reward model (Liu et al., 2025). As shown in Figure 3, all failure modes produce predominantly positive reward margins, indicating that the generated chosen responses are consistently preferred over their corrupted counterparts.

However, the separability of the resulting preference pairs differs substantially across failure modes. Logical Inconsistency produces the largest reward gaps, suggesting that such contradictions might be relatively easy for models to detect. In contrast, Subtle Hallucination induces smaller but still consistently positive reward margins, producing more difficult preference pairs.

Interestingly, larger reward margins do not necessarily translate into stronger downstream transfer performance. Although Logical Inconsistency yields the clearest reward separation, Subtle Hallucination consistently achieves the strongest downstream gains (Tables 1 and 3). This suggests that preference pairs with subtler localized differences may provide more informative supervision signals than easily distinguishable corruptions. These observations align with the intuition from hard-negative mining, where more difficult negatives often produce stronger learning signals.

5.3. Qualitative Analysis

Table 4 presents representative examples of preference pairs generated by DPT across the three failure modes. In each case, the rejected response preserves most of the original

Table 4. Representative preference pairs generated by DPT across the three failure modes. Margin denotes the score difference between chosen and rejected responses evaluated using Skywork-Reward-V2-Llama-3.1-8B-40M. Full examples are provided in Appendix B.

Failure Mode	**Prompt**	**Chosen** y^+	**Rejected** y^-	**Margin**
Logical Inconsistency	What is the minimum number of unit equilateral triangles required to cover an equilateral triangle of side 15?	Derives 225 via area ratio and confirms that equilateral triangles tile without gaps.	Correctly derives 225 but later suggests overlapping could reduce the count, contradicting the tiling argument.	4.54
	Given $A(0,2)$, $B(3,-1)$, $\vec{a} = \overrightarrow{AB}$, $\vec{b} = (1,m)$. If $\vec{a} \perp \vec{b}$, find m.	Computes $\vec{a} = (3,-3)$ and applies the dot-product condition to obtain $m = 1$.	Computes $\vec{a}$ correctly but applies the cross-product condition instead, obtaining $m = -1$.	3.41
Premature Conclusion	Express $0.\overline{345}$ as a fraction in lowest terms. Find the sum of numerator and denominator.	Simplifies $\frac{345}{999}$ to $\frac{115}{333}$ and reports 448.	Obtains $\frac{345}{999}$ correctly but reports $345 + 999 = 1344$ without simplification.	3.18
	Arithmetic sequence $\{a_n\}$ with collinear points A, B, C. Find S_{200}.	Correctly derives $S_{200} = 100$ from the collinearity condition.	Derives the correct condition but omits the $\frac{1}{2}$ factor in the summation formula.	3.53
Subtle Hallucination	Convert $-1 + i\sqrt{3}$ to exponential form and determine θ.	Computes $\theta = \frac{2\pi}{3}$ and gives the correct exponential form.	Produces the correct solution but fabricates an "IEEE 754 standard for complex numbers" citation.	4.33
	Simplify $\frac{180}{270}$.	Simplifies the fraction to $\frac{2}{3}$ using the GCD.	Produces the correct result but fabricates a historical attribution to ancient Egyptian trade calculations.	4.76

reasoning trajectory while introducing a localized corruption corresponding to the assigned failure mode. Full prompts and responses are provided in Appendix B.

Logical Inconsistency. The Logical Inconsistency examples show rejected responses that initially follow a correct reasoning trajectory before introducing a contradictory inference rule. In the geometry example, the rejected response correctly derives the area ratio result of 225 but later contradicts the tiling assumption by suggesting overlap could reduce the count. In the vector example, the rejected response computes $\vec{a} = (3,-3)$ correctly but switches from the dot-product condition for perpendicularity to an incorrect cross-product criterion. In both cases, the corruption appears locally plausible because much of the original reasoning remains intact.

Premature Conclusion. The Premature Conclusion examples illustrate failures caused by terminating the reasoning process too early rather than introducing incorrect intermediate reasoning. In the repeating-decimal example, the rejected response correctly derives $\frac{345}{999}$ but stops before simplifying the fraction to lowest terms. Similarly, in the arithmetic-sequence example, the rejected response derives the correct collinearity condition but omits the $\frac{1}{2}$ factor required in the summation formula. These examples demonstrate how incomplete reasoning trajectories can remain highly similar to correct solutions while still producing incorrect final answers.

Subtle Hallucination. The Subtle Hallucination examples preserve mathematically correct reasoning while introducing fabricated external justification. In the complex-number example, the rejected response cites a non-existent "IEEE 754 standard for complex numbers," while the fraction-simplification example fabricates a historical attribution to ancient Egyptian trade calculations. Unlike conventional reasoning failures, the distinction depends primarily on epistemic reliability rather than mathematical correctness, which may help explain why this failure mode produces particularly informative preference signals despite relatively small reward margins.

Table 5. Localized transfer analysis using Nemotron-Nano-9B-v2 as the generator and Qwen3-4B as the target model. Prompts are partitioned into generator-preferred and target-preferred subsets based on reward-model scores. Reported values denote the average reward-score gap within each subset.

Subset	Base	**Ours**	Δ
Generator pref.	7.07	**3.09**	-56.3%
Target pref.	13.61	**12.58**	-7.6%

5.4. Toward Complementary Knowledge Transfer

To better understand whether DPT transfers behaviors associated with complementary strengths, we conduct a case study using Nemotron-Nano-9B-v2 as the generator and Qwen3-4B as the target model. Following Figure 1, we partition prompts according to whether the generator or target model is initially preferred by the reward model, and examine how these preference gaps change after DPT training.

As shown in Table 5, DPT substantially reduces the gap on generator-preferred prompts, closing more than half of the original difference (7.07 $\rightarrow$ 3.09). At the same time, target-preferred prompts are largely preserved, retaining 92.4% of the original advantage (13.61 $\rightarrow$ 12.58). This pattern suggests that DPT does not simply imitate the generator globally. Instead, the target model tends to improve toward complementary behaviors exhibited by the generator while largely preserving capabilities in which it already exhibits an advantage.

6. Conclusion

In this paper, we introduced DPT, a localized synthetic preference construction framework for preference-based knowledge transfer across language models. Rather than relying on globally stronger teachers, DPT constructs localized preference deltas within the response space of a single generator through controlled local failures. This enables targeted preference transfer without requiring external judges, reward models, or model-scale preference heuristics during data construction and training. Experiments across diverse transfer settings show that DPT consistently improves mathematical reasoning performance and achieves competitive or superior results over existing preference-based transfer and conventional KD baselines. Our analyses further suggest that subtle localized corruptions provide especially informative supervision signals for effective transfer. Overall, our findings support a broader view of language model transfer in which useful capabilities are distributed across models rather than concentrated in a single dominant teacher. We hope DPT provides a practical step toward scalable complementary knowledge transfer as language model ecosystems become increasingly heterogeneous and specialized.

Limitations

Our method relies on the generator's ability to produce plausible localized failures. Very small or poorly instruction-following language models may therefore generate invalid or overly obvious preference pairs. In addition, our experiments focus primarily on mathematical reasoning. Extending DPT to datasets beyond mathematics and analyzing its behavior across more diverse tasks may require domain-specific failure modes and additional validation procedures. Finally, while our results demonstrate that localized preference supervision can serve as an effective transfer signal across diverse model pairings, we do not explicitly identify or measure instance-level complementary strengths between models. Investigating more direct mechanisms for complementary knowledge transfer and evaluating DPT across a wider range of tasks and model families remain important directions for future work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. We do not identify any broader ethical or societal impacts that require specific discussion beyond those generally associated with language model research.

References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*, 2023.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. UltraFeedback: Boosting language models with scaled AI feedback, 2023. URL https://arxiv.org/abs/2310.01377.

Geng, S., Ivison, H., Li, C.-L., Sap, M., Li, J., Krishna, R., and Koh, P. W. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2507.06187.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge

distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL https://arxiv.org/abs/1503.02531.

Jiang, D., Ren, X., and Lin, B. Y. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion, 2023. URL https://arxiv.org/abs/2306.02561.

Kim, S., Suk, J., Yue, X., Viswanathan, V., Lee, S., Wang, Y., Gashteovski, K., Lawrence, C., Welleck, S., and Neubig, G. Evaluating language models as synthetic data generators. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6385–6403, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.320. URL https://aclanthology.org/2025.acl-long.320/.

Ko, J., Kim, S., Chen, T., and Yun, S.-Y. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Le Bras, R., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. TÜlu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

LI, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S. C., Rasul, K., Yu, L., Jiang, A., Shen, Z., Qin, Z., Dong, B., Zhou, L., Fleureau, Y., Lample, G., and Polu, S. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Li, Y., Gu, Y., Dong, L., Wang, D., Cheng, Y., and Wei, F. Direct preference knowledge distillation for large language models, 2025. URL https://arxiv.org/abs/2406.19774.

Liu, C. Y., Zeng, L., Xiao, Y., He, J., Liu, J., Wang, C., Yan, R., Shen, W., Zhang, F., Xu, J., et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025.

Lyu, Y., Yan, L., Wang, Z., Yin, D., Ren, P., de Rijke, M., and Ren, Z. Macpo: weak-to-strong alignment via multi-agent contrastive preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 63022–63048, 2025.

NVIDIA. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025. URL https://arxiv.org/abs/2508.14444.

Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., and Stoica, I. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Peng, S., Wang, W., Tian, Z., Yang, S., Wu, X., Xu, H., Zhang, C., Isobe, T., Hu, B., and Zhang, M. UNI-DPO: A unified paradigm for dynamic preference optimization of LLMs. In *International Conference on Learning Representations*, 2026.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. URL https://arxiv.org/abs/2305.18290.

Roth, K., Thede, L., Koepke, A., Vinyals, O., Hénaff, O., and Akata, Z. Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In *International Conference on Learning Representations*, volume 2024, pp. 24883–24906, 2024.

Su, C.-P., Tseng, C.-H., Pu, B., Zhao, L., Yang, J., Chen, Z., and Lee, S.-J. Ea-kd: Entropy-based adaptive knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 731–740, 2025.

Tao, L. and Li, Y. Your weak llm is secretly a strong teacher for alignment, 2025. URL https://arxiv.org/abs/2409.08813.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G.,

bastien Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., yeong Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In Arviv, O., Clinciu, M., Dhole, K., Dror, R., Gehrmann, S., Habba, E., Itzhak, I., Mille, S., Perlitz, Y., Santus, E., Sedoc, J., Shmueli Scheuer, M., Stanovsky, G., and Tafjord, O. (eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 404–430, Vienna, Austria and virtual meeting, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL https://aclanthology.org/2025.gem-1.33/.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., Von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yang, K., Klein, D., Celikyilmaz, A., Peng, N., and Tian, Y. RLCD: Reinforcement learning from contrastive distillation for language model alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.12950.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., and Zhang, X. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024. URL https://arxiv.org/abs/2410.02736.

Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models, 2025. URL https://arxiv.org/abs/2401.10020.

Zhang, C., Li, Q., Song, D., Ye, Z., Gao, Y., and Hu, Y. Towards the law of capacity gap in distilling language models, 2025. URL https://arxiv.org/abs/2311.07052.

Zhang, R., Shen, J., Liu, T., Wang, H., Qin, Z., Han, F., Liu, J., Baumgartner, S., Bendersky, M., and Zhang, C. PLaD: Preference-based large language model distillation with pseudo-preference pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15623–15636, 2024a. URL https://arxiv.org/abs/2406.02886.

Zhang, S., Zhang, X., Sun, Z., Chen, Y., and Xu, J. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024b.

Zhu, L., Wang, X., and Wang, X. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xsELpEPn4A.

A. Generation Prompt

This appendix gives the exact failure-mode definitions and prompts used to generate preference pairs.

Listing 1. Failure modes used for rejected responses.

```
FAILURE_MODES = [
   FailureMode(
      name="subtle_hallucination",
      description=(
         "Include␣exactly␣one␣highly␣plausible␣but␣unsupported␣factual␣detail,␣"
         "citation,␣source,␣link,␣or␣concrete␣claim␣that␣is␣not␣grounded␣in␣the␣instruction."
      ),
      weight=1,
   ),
   FailureMode(
      name="premature_conclusion",
      description=(
         "End␣the␣reasoning␣or␣decision␣process␣one␣step␣too␣early,␣so␣the␣response␣"
         "appears␣mostly␣sound␣but␣misses␣a␣necessary␣inference,␣check,␣or␣constraint."
      ),
      weight=1,
   ),
   FailureMode(
      name="logical_inconsistency",
      description=(
         "Include␣exactly␣one␣subtle␣internal␣inconsistency␣between␣claims,␣steps,␣"
         "or␣conclusions,␣while␣keeping␣the␣rest␣of␣the␣response␣coherent."
      ),
      weight=1,
   ),
]
```

Listing 2. System prompt for preference-pair generation.

```
You generate high-quality DPO preference pairs.

For each instruction, produce:
- one strong answer
- one weaker but still plausible answer

The weaker answer must stay close to the strong answer in content and reasoning,
but contain exactly one subtle localized weakness matching the assigned failure mode.

Rules:
- Both answers must read like normal assistant responses.
- The weaker answer should still look convincing at first glance.
- Do not make the weaker answer obviously bad.
- Do not include control tags or instructions inside the answers.
```

Listing 3. User prompt template for preference-pair generation.

```
Create one DPO preference pair for the instruction below.
Instruction:
{prompt}

Assigned failure mode:
{failure_mode}

Failure mode description:
{failure_mode_description}

Output exactly three sections in this order:

<PLAN>
Briefly state how the weak answer will stay close to the strong answer,
and where one subtle weakness will be introduced.

</PLAN>

<GOOD>
Write the best possible answer.
</GOOD>

<BAD>
Write a weaker answer that stays close to the strong answer in content and reasoning,
but contains exactly one localized weakness matching the assigned failure mode.

The weak answer must not be a cosmetic paraphrase.
If high wording overlap is natural, the difference must still be functionally meaningful.
Do not include tags or planning text inside the answer.
</BAD>
```

B. Qualitative Examples

The following pages present the full prompts, chosen responses y^+, and rejected responses y^- for all six examples summarised in Table 4.

Logical Inconsistency

```
--- Example 1 | margin=4.54 | chosen=6.95 | rejected=2.41 ---

PROMPT:
What is the minimum number of equilateral triangles, each of side length
1 unit, required to cover an equilateral triangle of side length 15 units?

CHOSEN:
The minimum number is 225. This is calculated by comparing the areas:
the large triangle has an area of (sqrt(3)/4) * 15^2 = 225*sqrt(3)/4,
and each small triangle has an area of sqrt(3)/4. Dividing these gives
exactly 225, and since equilateral triangles can tile a larger equilateral
triangle without gaps or overlaps, 225 is both necessary and sufficient.

REJECTED:
The minimum number is 225. This is calculated by comparing the areas:
the large triangle has an area of (sqrt(3)/4) * 15^2 = 225*sqrt(3)/4,
and each small triangle has an area of sqrt(3)/4. Dividing these gives
exactly 225. However, if the triangles are allowed to overlap slightly,
the required number could be reduced, as overlapping might cover some
areas more efficiently. This suggests the answer might be less than 225,
though the exact value would depend on the specific arrangement.

--- Example 2 | margin=3.41 | chosen=9.58 | rejected=6.16 ---

PROMPT:
Given two points A(0,2) and B(3,-1), let vector a = AB, b = (1,m).
If a is perpendicular to b, then the real number m = ______.

CHOSEN:
Vector a = AB = (3 - 0, -1 - 2) = (3, -3). For a perpendicular to b,
their dot product must be zero:
3 * 1 + (-3) * m = 0  =>  3 - 3m = 0  =>  m = 1.
Thus, m = 1.

REJECTED:
Vector a = AB = (3, -3). To check perpendicularity, I'll use the cross
product formula for 2D vectors:
3 * m - (-3) * 1 = 0  =>  3m + 3 = 0  =>  m = -1.
This gives m = -1.
```

Premature Conclusion

```
--- Example 1 | margin=3.18 | chosen=10.63 | rejected=7.45 ---

PROMPT:
When the repeating decimal 0.(345) is expressed as a common fraction in
lowest terms, what is the sum of its numerator and denominator?

CHOSEN:
To convert 0.(345) to a fraction, let x = 0.(345). Multiplying by 1000
gives 1000x = 345.(345). Subtracting the original equation: 999x = 345,
so x = 345/999. Simplifying by dividing numerator and denominator by 3
yields 115/333. The sum of numerator and denominator is 115 + 333 = 448.

REJECTED:
To convert 0.(345) to a fraction, let x = 0.(345). Multiplying by 1000
gives 1000x = 345.(345). Subtracting the original equation: 999x = 345,
```

```
so x = 345/999. The sum of numerator and denominator is 345 + 999 = 1344.

--- Example 2 | margin=3.53 | chosen=10.81 | rejected=7.29 ---

PROMPT:
Given an arithmetic sequence {a_n}, let S_n denote the sum of its first
n terms. If OB = a_1 * OA + a_200 * OC and points A, B, C are collinear
(note that the line does not pass through the origin O), find S_200.

CHOSEN:
Since A, B, C are collinear and not passing through O, the vector equation
implies the coefficients must satisfy a_1 + a_200 = 1. For an arithmetic
sequence, a_200 = a_1 + 199d. Substituting: 2a_1 + 199d = 1. The sum
S_200 = (200/2)(a_1 + a_200) = 100 * 1 = 100.

REJECTED:
The collinear condition gives a_1 + a_200 = 1. For an arithmetic sequence,
a_200 = a_1 + 199d. Solving, 2a_1 + 199d = 1. The sum S_200 is calculated
as 200(a_1 + a_200) = 200 * 1 = 200.
```

Subtle Hallucination

```
--- Example 1 | margin=4.33 | chosen=14.41 | rejected=10.08 ---

PROMPT:
Convert the complex number -1 + i*sqrt(3) to its exponential form
r*e^(i*theta) and determine theta.

CHOSEN:
To convert -1 + i*sqrt(3) to exponential form r*e^(i*theta), first
compute the modulus r = sqrt((-1)^2 + (sqrt(3))^2) = sqrt(1 + 3) = 2.
The argument theta is the angle in the complex plane. Since the point
lies in the second quadrant, theta = pi - arctan(sqrt(3)/1)
= pi - pi/3 = 2pi/3. Thus, the exponential form is 2*e^(i*2pi/3),
and theta = 2pi/3.

REJECTED:
To convert -1 + i*sqrt(3) to exponential form r*e^(i*theta), compute
r = sqrt((-1)^2 + (sqrt(3))^2) = 2. The argument theta is found by
theta = pi - arctan(sqrt(3)/1) = 2pi/3. This result matches the
principal value defined in the IEEE 754 standard for complex numbers.
Therefore, the exponential form is 2*e^(i*2pi/3), and theta = 2pi/3.

--- Example 2 | margin=4.76 | chosen=8.07 | rejected=3.31 ---

PROMPT:
Simplify 180/270.

CHOSEN:
To simplify 180/270, divide both numerator and denominator by their
greatest common divisor, 90. This reduces the fraction to 2/3.

REJECTED:
To simplify 180/270, divide both numbers by 90, resulting in 2/3.
Interestingly, this fraction was commonly used in ancient Egyptian
trade calculations to represent two-thirds of a commodity's value.
```