ORPO-Distill: Mixed-Policy Preference Optimization for Cross-Architecture LLM Distillation

Aasheesh Singh*
Phi Labs, Quantiphi
Toronto, ON, M5V 2L1, Canada
aasheesh.singh@quantiphi.com

Vishal Vaddina
Phi Labs, Quantiphi
Toronto, ON, M5V 2L1, Canada
vishal.vaddina@quantiphi.com

Dagnachew Birru
Phi Labs, Quantiphi
Marlborough, MA, 01752, USA
dagnachew.birru@quantiphi.com

Abstract

We introduce ORPO-Distill, a general-purpose method for cross-architecture LLM distillation that formulates the problem as a preference optimization task. Unlike standard CoT distillation, the approach transfers knowledge through diverse reasoning traces. It employs an Odds-Ratio Preference Optimization objective that contrasts teacher and student traces for more effective learning, and adopts a mixed-policy strategy for utilizing student-generated outputs, outperforming both off- and on-policy alternatives. Experiments on five datasets and multiple student models show consistent improvements over conventional black-box KD baselines.

1 Introduction

Knowledge distillation (KD) has emerged as a key approach for compressing LLMs into smaller, more efficient task-specific student models. While white-box KD techniques depend on shared token vocabularies and teacher logits, limiting them to same-architecture settings, black-box KD enables cross-architecture transfer by sampling teacher outputs for supervision. In this work, we introduce *ORPO-Distill*, a general-purpose technique for cross-architecture LLM distillation that reformulates the process as a preference optimization problem. ORPO-Distill integrates three key insights:

1. Distilling from *diverse* reasoning traces improves supervision over single CoT distillation.
2. Framing distillation as a *preference optimization* task using ORPO objective, where teacher-generated positive reasoning and student-generated negative reasoning strengthens contrastive learning.
3. *Mixed-policy* update of student negative traces outperforms both off- and on-policy strategies.

We evaluate ORPO-Distill across five QA benchmarks and multiple student architectures, emphasizing each component's contribution to the pipeline and overall improvements over black-box KD baselines.

2 Methodology

Related Work. Hsieh et al. [2023] proposed augmenting teacher CoT reasoning traces with labels to provide additional supervision in a multi-task learning fashion. Recent black-box KD methods

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

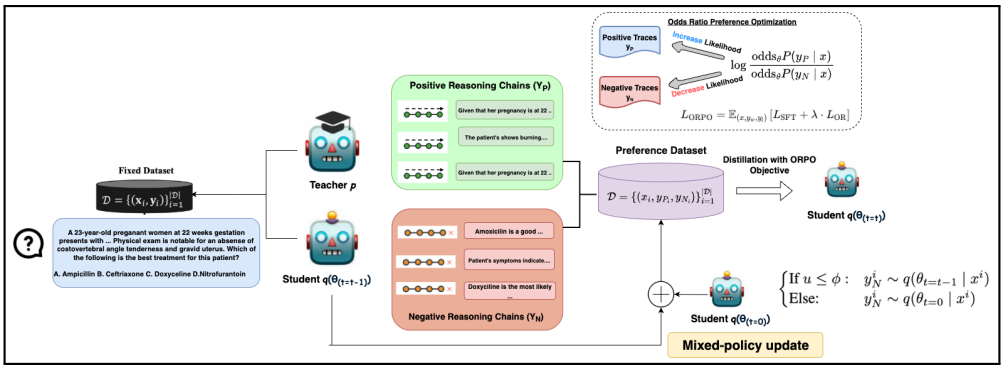

Figure 1: Overview of our ORPO-Distill method. Given an input prompt, teacher and student model generate diverse positive and negative reasoning traces forming the preference dataset for ORPO contrastive distillation, which is updated in a mixed-policy fashion over training epochs.

such as MAGDI Chen et al. [2024] and Payoungkhamdee et al. [2024] proposed utilizing contrastive learning with diverse positive and negative rationales generated from single or multiple teacher LLMs, but do not use student-generated outputs (SGOs) in their formulation. Prior literature Agarwal et al. [2024], Ko et al. [2024], Ko et al. [2025] on white-box KD methods provide growing evidence that utilizing SGOs during distillation provides significant gains. This is because it addresses the distribution mismatch between fixed output sequences seen during training and the auto-regressive generation by student at inference. These works further argue that updating SGOs during training through on- or off-policy updates improves performance. Wang et al. [2024] on the other hand, proposes a black-box KD technique that does utilize contrastive negative SGOs, but doesn't use policy updates from previous iteration student models.

Novelty. In this work therefore, we seek to bridge the gap between these white-box and black-box KD methods through a novel formulation that allows task-specific cross architecture distillation. In our methodology, we utilize diverse reasoning traces for supervision, combine positive teacher traces and negative student traces through a contrastive ORPO loss objective, and address distribution mismatch between training and auto-regressive inference through a mixed-policy sampling technique for generating negative student reasoning traces. More details about each of these components follows below. A pipeline diagram of our methodology is shown in Figure-1.

2.1 Odds Ratio Preference Optimization (ORPO)

ORPO, proposed by Hong et al. [2024], seeks to combine the traditional two-stage LLM training pipeline comprising of SFT followed by preference alignment stage into a single objective function. This is done by incorporating an odds ratio-based penalty to the conventional negative log-likelihood (NLL) term. This additional penalty term weighted by hyperparameter λ imposes a preference for one response over the other. Assuming y_P and y_N denote favoured positive trace and disfavoured negative trace respectively and q_θ denotes parametrized student model distribution, the objective term for an input sequence x would be defined as:

$$L_{SFT} = -\log q_\theta(y_P \mid x), \quad L_{OR} = -\log \sigma\left(\log \frac{\text{odds}\, q_\theta(y_P|x)}{\text{odds}\, q_\theta(y_N|x)}\right) \tag{1}$$

where, the odds term and the final ORPO objective is defined as:

$$\text{odds}\, q_\theta(y|x) = \frac{q_\theta(y|x)}{1 - q_\theta(y|x)}, \quad L_{ORPO} = L_{SFT} + \lambda L_{OR} \tag{2}$$

The authors of ORPO conclude that a small value $\lambda = 0.1$ is suited for slightly nudging model towards human preferences(say, helpfulness) whereas a large value such as $\lambda = 1$ is used for strong adaptation of one reasoning over other. Since this aligns with our goal of clipping incorrect generation paths in the student model reasoning, we set $\lambda = 1$ in our pipeline.

Algorithm 1 ORPO-Distill

Given: Teacher model p, Student model q_θ, Fixed dataset (X, L) denoting (prompt,label) pair
Hyperparameters: Policy fraction $\phi \in [0, 1]$, Diversity Param K, Odds-Ratio λ, LR η
Initialize: Sample K diverse positive traces $\{y_P^{(k)}\}_{k=1}^K \sim p(\cdot \mid x)\ \forall$ x $\in$ X to get Y_P
for epoch $e = 1, 2.\ldots., E$ **do**
  for iteration $t = 1, 2.\ldots., T$ **do**
    Generate a random number $u \sim \mathcal{U}(0, 1)$
    if $u \leq \phi$ **then**
      Sample K negative traces $\{y_n^{(k)}\}_{k=1}^K \sim q(\theta_{t=t-1} \mid x) \rightarrow \mathcal{B} = \{(x^i, y_P^i, y_N^i)\}_{i=1}^{|\mathcal{B}|}$
    else
      Sample K negative traces $\{y_n^{(k)}\}_{k=1}^K \sim q(\theta_{t=0} \mid x) \rightarrow \mathcal{B} = \{(x^i, y_P^i, y_N^i)\}_{i=1}^{|\mathcal{B}|}$
    end if
    Update θ to minimize L_{ORPO}: $L_{ORPO} = L_{SFT} + \lambda L_{OR}$
  end for
end for

2.2 Preference Dataset Creation

Dataset Structure. ORPO relies on creating a preference dataset consisting of ⟨Prompt, Chosen, Rejected⟩ triplets. We adopt the same structure, where *Chosen* is a teacher CoT trace leading to the positive or correct answer and *Rejected* is a student CoT trace leading to a negative or incorrect answer. Compared to a previous work Payoungkhamdee et al. [2024], which utilizes teacher CoT traces for creating both positive and negative traces for contrastive training, we first empirically conclude in (Table 1) on a subset of datasets that utilizing negative traces from student-generated outputs, infact yields better results for contrastive training using ORPO and thus forms the basis of our further methodology. For this work, we use multi-choice QA datasets with ground truth labels, enabling direct classification of traces as positive or negative. However, the approach generalizes to open-ended tasks given a task-specific definition of desired responses. Such definitions may rely on training small verifier or reward models, unit tests, or other heuristics, and represent an avenue for future work.

Table 1: Effect of sampling contrastive negative trace from Teacher vs Student model (Accuracy %)

Experiment	**Datasets**	
Off-Policy ORPO	MedQA	ARC-C
(p-$CoT_{Teacher}$, n-$CoT_{Teacher}$)	41.72	45.87
(p-$CoT_{Teacher}$, n-$CoT_{Student}$)	49.33	56.48

Diversity Sampling. We sample diverse reasoning chains for both positive and negative traces using temperature sampling with parameter $\tau = 0.8$ for K generations. Both teacher and student are prompted with the same `Reason-then-Answer` format, where the final answer is marked in a parseable boxed {} output. Note, that we do not induce any bias in the prompt while sampling negative traces such as by injecting the incorrect answer. Experiments with $K \in \{4, 8, 12\}$ showed diminishing gains beyond 8, so we fix $K = 8$ in the following experiments. To remove redundancy, we apply rejection sampling and discard traces with high ROUGE-L overlap over 0.80.

2.3 Mixed-Policy Update

We define updates to student-generated negative traces under three settings: off-policy - which corresponds to a fixed negative set generated using initial student model, on-policy - which corresponds to a new negative trace set sampled after every epoch using the latest checkpoint, and mixed-policy - which corresponds to using a combination of last checkpoint and base model traces. The pseudocode of our method is detailed in Algorithm-1 covering all these cases using a policy fraction parameter ϕ.

3 Experimentation

Datasets We conduct extensive experiments on five widely-used benchmark QA datasets across different domains such as: MedQA-USMLE for medical diagnostic reasoning, ARC-Challenge, Strat-

Table 2: Experimental results across different student models and datasets.

Experiments	Datasets (Accuracy %)					Avg Acc%
	MedQA	ARC-C	StrategyQA	OBQA	GSM8K	
TinyLlama 1.1B-Instruct						
Zero-shot CoT Eval	29.78	29.95	43.52	26.60	11.97	28.36
Single CoT Fine Tuning	32.10	32.63	46.25	29.05	31.56	34.32
Diverse CoT Fine Tuning	34.85	35.40	47.84	33.60	36.22	37.58
Off Policy ORPO	38.95	41.20	49.77	37.45	39.45	41.36
On Policy ORPO	35.11	38.01	49.24	35.60	36.88	38.97
Mixed Policy ORPO	**40.25**	**43.55**	**51.25**	**40.10**	**40.72**	**43.17**
InternLM 2.5 1.8B-Chat						
Zero-shot CoT Eval	35.82	37.12	54.15	27.40	41.02	39.10
Single CoT Fine Tuning	37.94	40.45	57.50	41.35	44.38	44.32
Diverse CoT Fine Tuning	40.56	42.15	58.66	54.50	47.50	48.67
Off Policy ORPO	49.33	56.48	59.39	53.20	51.25	53.93
On Policy ORPO	43.25	49.80	58.50	52.79	47.94	50.46
Mixed Policy ORPO	**50.43**	**59.32**	**61.75**	**55.22**	**52.47**	**55.84**
InternLM 2.5 7B-Chat Teacher						
Zero-shot CoT Eval Teacher	50.98	56.40	62.57	61.80	66.14	59.58

egyQA and OpenBookQA for general-purpose reasoning, and GSM8K for mathematical problem-solving reasoning.

Implementation Details We use `InternLM 2.5 7B-Chat` as the teacher model, selected for its strong performance and instruction-following ability with reliable adherence to the `Reason-then-Answer` format in zero-shot prompting. However, as discussed, our method is invariant to any teacher and student architecture combinations. We use two student models from different architecture families and sizes: `InternLM 2.5 1.8B-Chat`, `TinyLlama 1.1B-Instruct` to validate our method. We conduct full-parameter tuning for 5 epochs, with additional implementation and training details available in our codebase.

Experiments We design experiments to validate each component of our methodology as below, and report results across all datasets in Table 2.

1. **Diverse Reasoning Traces.** We validate this component by creating a single-trace CoT fine-tuning baseline against diverse-traces CoT fine-tuning using the same NLL objective. Both settings use traces generated from the teacher model. The latter equates to SeqKD, proposed by Kim and Rush [2016], in the SFT setting and has been used as a baseline in previous literature Payoungkhamdee et al. [2024], and Ko et al. [2024]. We also show baseline results of zero-shot CoT evaluation to measure the models out-of-box performance.
2. **Contrastive Training with ORPO.** To validate this component, we treat the previous best experiment, i.e., diverse CoT fine-tuning from teacher traces, as the baseline, and seek to improve it further by utilizing the student model generated negative traces using a contrastive ORPO objective. Note that we utilize the base student model (before distillation) at Epoch 0 to generate these negative traces without further updates, and define this as the *Off-Policy ORPO distillation* experiment in results.
3. **Mixed-Policy Updates.** The policy parameter ϕ defined in our pseudocode Algorithm-1 controls whether the method is off-policy($\phi = 0$), on-policy($\phi = 1$), or mixed-policy($\phi = 0.5$). This parameter is responsible for controlling the level of mixing between the base student model and the latest epoch checkpoint. For our mixed-policy experiments, we set $\phi = 0.5$ inspired from a similar setting in Agarwal et al. [2024], which amounts to an equally proportioned mixing between the two distributions.

4 Conclusion

Experiments across multiple student models and datasets show that ORPO-Distill, which leverages student-generated negative traces for contrastive distillation with mixed-policy updates, achieves

the best performance. In contrast, purely on-policy updates after every epoch degrade performance compared to the fixed-trace off-policy setting which aligns with findings from Ko et al. [2024]. We attribute this to the fact, that although recently sampled negative traces are of higher quality and closely resemble correct rationales, the overall distribution narrows, reducing diversity for contrastive learning. Mixed-policy updates mitigate this issue by anchoring the negative trace distribution to the initial student model (Epoch 0) through random sampling between the initial and most recent checkpoint. We foresee that sophisticated strategies such as curriculum based updates to the mixed-policy buffer and extension to open-ended tasks beyond QA remains an avenue for future work.

References

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*, 2023.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models. *arXiv preprint arXiv:2402.01620*, 2024.

Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. An empirical study of multilingual reasoning distillation for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7739–7751, 2024.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

Wei Wang, Zhaowei Li, Qi Xu, Yiqing Cai, Hang Song, Qi Qi, Ran Zhou, Zhida Huang, Tao Wang, and Li Xiao. Qcrd: Quality-guided contrastive rationale distillation for large language models. *arXiv preprint arXiv:2405.13014*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327, 2016.