Meta-Learning Preferences for Multilingual LLM Alignment

Jiaying Lin[1]*, **Seongho Son**[2,*], **Nam Phuong Tran**[3],
Long Tran-thanh[1], **Ilija Bogunovic**[4], **Debmalya Mandal**[1]
[1]Department of Computer Science, University of Warwick, Coventry, United Kingdom
[2]Department of Computer Science, University College London, London, United Kingdom
[3]Inria centre at the University of Lille, Lille, France
[4]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland
{Jiaying.Lin, Long.Tran-Thanh, Debmalya.Mandal}@warwick.ac.uk
seong.son.22@ucl.ac.uk, phuong-nam.tran@inria.fr, ilija.bogunovic@unibas.ch

Abstract

Unequal availability of human preference data across languages poses a significant challenge for aligning large language models in multilingual settings. To address the lack of sufficient data in low-resource language alignment, we propose a meta-learning framework for Reinforcement Learning from Human Feedback and Direct Preference Optimization. By leveraging preference data from other languages, our framework learns a transferable initialization that enables effective adaptation to a target language with minimal data. We provide theoretical guarantees for both the meta-reward modeling and meta-policy optimization settings, and empirically demonstrate the effectiveness of our approach on multilingual benchmarks. In an extremely low-resource setting with only 100 target-language preference samples, our approach achieves up to 33.4% win-rate improvements over baseline methods, and consistently outperforms baselines across multiple target languages and model scales. Our approaches retain these advantages across different combinations of meta-training languages and varying linguistic distances from the target languages.

1 Introduction

Alignment of Large Language Models (LLMs) have shown great progress, using methods such as Reinforcement Learning with Human Feedback (RLHF) [60, 37] and Direct Preference Optimization (DPO) [39]. However, these approaches heavily rely on a large amount of high-quality human preference data, whose availability is severely unequal *across languages*. As a result, LLMs often exhibit significantly weaker performance in low-resource languages, failing to provide global accessibility.

Existing works have investigated improving multilingual LLMs through approaches such as fine-tuning to a specific target language [12] or aggregating preference data across multiple languages and training models on them [28]. While these approaches are effective, they rely on strong assumptions, either requiring sufficient target-language data or a close linguistic relationship between languages. In practice, direct transfer of general capabilities across languages is suboptimal, particularly, when only a small amount of data is available for a given language.

In this work, we propose a meta-learning approach for preference-based multilingual alignment of LLMs. We formulate multilingual alignment as an *adaptation problem*, by treating each language in the dataset as a task. Our approach leverages data from multiple languages to meta-train the LLM policy, that can be rapidly adapted to a new target language using only a small number of preference

*Equal contribution. Correspondence to seong.son.22@ucl.ac.uk.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

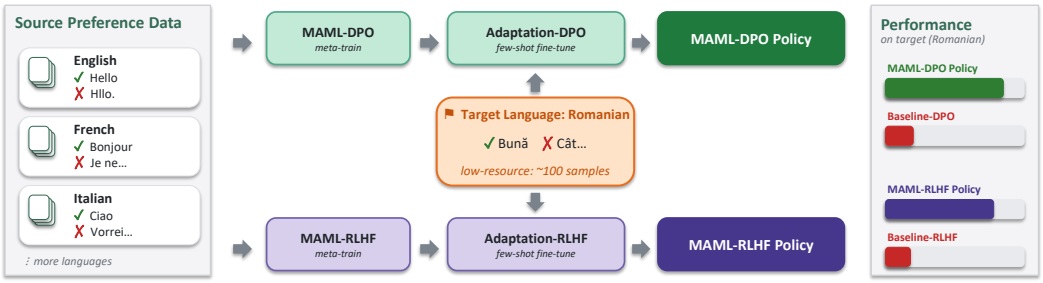


Figure 1: When the target language has low resource availability (e.g. Romanian) and has a very small number of data samples, such as 100, conventional approaches fail to properly fine-tune LLMs. On the other hand, our approach first meta-trains the model using other languages with sufficient data, allowing the model to boost sample efficiency in the target language. Our MAML-DPO/MAML-RLHF methods achieve significant performance improvement over the baseline methods, given the same amount of datapoints from the target language.

comparisons. We visually summarize our approach in Figure 1. For the first time, we instantiate this framework for both RLHF (via meta-reward learning) and DPO (via meta-policy learning) using gradient-based meta-learning.

Multilingual preference alignment naturally exhibits a task-level adaptation structure, where each language can be viewed as a distinct task with unevenly distributed preference data. Existing methods commonly extend RLHF or DPO through joint training over multilingual data, leaving task-level adaptation largely underexplored. Gradient-based meta-learning is well suited to this setting, as it learns an initialization that can efficiently adapt across related tasks [18, 61]. We close this gap by formulating multilingual preference alignment as a meta-learning problem and instantiating it as MAML-RLHF and MAML-DPO. The resulting algorithms meta-train across languages and adapt to a target language with as few as 100 preference samples. On the theory side, the Bradley-Terry losses that govern preference learning sit outside the regression/classification setting of existing MAML convergence theory, and we establish the first convergence guarantees for MAML in this setting, characterising the regime in which the meta-trained initialisation provably accelerates target-language adaptation over both target-only and multitask baselines.

We empirically evaluate our methods on various popular benchmarks across multiple target languages [9, 59, 20, 12]. Using open-source LLMs with 270M and 7B parameters, we demonstrate that meta-trained models substantially outperform standard RLHF and DPO baselines in low-resource settings. Notably, with only 100 target-language preference samples, our approach improves win rates by up to 33.4%. In addition, the benefits of meta-training persist as the amount of adaptation data increases to more than 1k samples, still more stably improving performances across benchmarks than the baseline method.

In this paper, we provide four main contributions:

- We propose a unified meta-learning framework for aligning LLMs using preference data. This framework treats the multilingual preference alignment as a task distribution problem, where each language corresponds to a task. The framework effectively addresses the issue of uneven preference data availability and significantly improves sample efficiency for low-resource languages.
- We provide convergence guarantees of our reward modeling and preference optimization algorithms, theoretically proving the efficiency of our approach.
- Using multilingual instruction-following datasets, we empirically demonstrate the greatly improved sample efficiency of our approach, outperforming baselines by up to 33.4% in win rates.
- We show that our method retains advantage across different combinations of meta-training and target languages, suggesting that our method learns a broadly transferable initialization for adaptation.

2 Preliminaries

Offline learning via human preference. In the offline setting, we assume access to a dataset $\mathcal{D} = (x, y_w, y_\ell)$, where x is a prompt and (y_w, y_ℓ) are two possible completions of the prompt x. A human labeler then indicates a preference between the two responses according to a reward function $r^\star(x, y)$ and the Bradley–Terry (BT) model:

$$P(y_1 \succ y_2|x) = \sigma(r^\star(x, y_1) - r^\star(x, y_2)) = \frac{\exp(r^\star(x, y_1))}{\exp(r^\star(x, y_1)) + \exp(r^\star(x, y_2))}, \tag{1}$$

where σ is the sigmoid function, $\sigma(z) = 1/(1 + \exp(-z))$. We aim to learn an aligned policy from the offline preference dataset $\mathcal{D}$. To this end, we study RLHF and DPO.

Reinforcement learning from human feedback. As proposed by [60], RLHF consists of two phases: reward learning and policy optimization. Let r_ϕ be a reward function parameterized by ϕ. The RLHF loss is

$$\mathcal{L}^{\text{RLHF}}(\phi) \triangleq -\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}} \left[\log \left(\sigma \left(r_\phi(x, y_w) - r_\phi(x, y_\ell)\right)\right)\right]. \tag{2}$$

After learning the reward function (denoted $r_{\widehat{\phi}}$), the last step is to optimize a policy using the following regularized objective with respect to $r_{\widehat{\phi}}$:

$$\widehat{\pi} = \arg\max_\pi \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi(\cdot|x)}} \left[r_{\widehat{\phi}}(x, y) - \beta \log \frac{\pi(y|x)}{\pi^{\text{SFT}}(y|x)}\right], \tag{3}$$

where $\beta > 0$ is a regularization coefficient that penalizes deviation from the reference policy π^{SFT}, a supervised fine-tuned policy.

Direct preference optimization. Direct Preference Optimization (DPO), proposed by [39], is an alternative to RLHF. DPO updates the policy directly from preference data, skipping the separate reward-modeling step. The main idea is that the reward can be written implicitly in terms of the optimal policy and π^{SFT}, which leads to a loss defined purely by preference comparisons. In particular, let π_θ be a policy parameterized by θ. Then, the DPO loss is

$$\mathcal{L}_{\text{DPO}}(\theta) \triangleq -\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}} \left[\log \left(\sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi^{\text{SFT}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_\ell|x)}{\pi^{\text{SFT}}(y_\ell|x)}\right)\right)\right]. \tag{4}$$

3 Multilingual preference alignment as a meta-learning problem

Multilingual preference alignment is fundamentally a *multi-task adaptation* problem. Each language τ has its own preference distribution, shaped by elements such as grammar and culture. On the other hand, large-scale comparison data is collected for only a handful of high-resource languages. Thus, the relevant question is not "how do we align a *single* language?" but "how do we align a *target* language efficiently when preference data is available only from *other* languages?". In this point of view, the problem maps to the meta-learning regime: each language is a task drawn from a distribution P_{TASK}, with taskwise preference loss $\mathcal{L}(\omega; \tau)$. The goal here is a meta-initialisation from which a few gradient steps on a small target-language dataset $\mathcal{D}^{\tau_{\text{new}}}$ recover a near-optimal aligned policy. The Model-Agnostic Meta-Learning (MAML) framework [18] formalizes this objective through the one-step meta-learning loss

$$\mathcal{L}_M(\omega) \;\triangleq\; \mathbb{E}_{\tau\sim P_{\text{TASK}}}\big[\mathcal{L}\big(\omega - \alpha\, \nabla_\omega \mathcal{L}(\omega; \tau);\, \tau\big)\big], \tag{5}$$

with $\alpha > 0$ the inner-loop learning rate. Adaptation to a new task τ_{new} then runs gradient descent on the empirical loss $\widehat{\mathcal{L}}(\,\cdot\,; \tau_{\text{new}})$ initialized at the meta-solution $\omega_M^\star = \arg\min_\omega \mathcal{L}_M(\omega)$:

$$\widehat{\omega}_{t+1}^{\tau_{\text{new}}} \;=\; \widehat{\omega}_t^{\tau_{\text{new}}} - \eta\, \nabla_\omega \widehat{\mathcal{L}}\big(\omega_t^{\tau_{\text{new}}}; \tau_{\text{new}}\big), \qquad \widehat{\omega}_0^{\tau_{\text{new}}} = \omega_M^\star, \tag{6}$$

with adaptation step size $\eta > 0$. Instantiating $\mathcal{L}(\omega; \tau)$ as the per-language RLHF reward-modelling loss or the per-language DPO policy loss yields the two algorithms developed in this paper: MAML-RLHF and MAML-DPO.

Algorithm 1 Meta-reward learning

Input: Dataset $\mathcal{D} = \{(x^i, y_w^i, y_\ell^i, \tau^i)\}_{i=1}^N$, #iterations T, initial reward model ϕ_1.

Initialize $\phi \leftarrow \phi_1$
for $t = 1, \ldots, T$ **do**
  Choose a batch $\mathcal{B}_t \subseteq \mathcal{I}$ of i.i.d. tasks from distribution P_{TASK} with size $|\mathcal{B}_t| = B$
  for each task $\tau \in \mathcal{B}_t$ **do**
    Sample $\mathcal{D}_{\text{in}}^\tau \subseteq \mathcal{D}^\tau$
    Compute empirical loss $\widehat{\mathcal{L}}^{\text{RLHF}}(\cdot; \tau)$ with $\mathcal{D}_{\text{in}}^\tau$
    $\phi_{t+1}^\tau = \phi_t - \alpha \nabla_\phi \widehat{\mathcal{L}}^{\text{RLHF}}(\phi_t; \tau)$
    Sample $\mathcal{D}_{\text{o}}^\tau \subseteq \mathcal{D}^\tau$
    Compute empirical loss $\widehat{\mathcal{L}}^{\text{RLHF}}(\phi_{t+1}^\tau; \tau)$ with $\mathcal{D}_{\text{o}}^\tau$
  end for
  $\phi_{t+1} = \phi_t - \frac{\eta}{B} \sum_{\tau \in \mathcal{B}_t} \nabla_\phi \widehat{\mathcal{L}}^{\text{RLHF}}(\phi_{t+1}^\tau; \tau)$
end for

Output: $\widehat{\phi}_M = \phi_{t^\star}$ where $t^\star \sim \text{Unif}\{1, \ldots, T\}$.

Both RLHF and DPO are based on a Bradley-Terry sigmoid log-likelihood structure, which is different from the regression and classification settings on which existing MAML convergence theory is built [17, 61]. The per-task strong-convexity constant of the DPO loss depends on both the feature-coverage constant and the reference policy log-ratio bound, and the meta-objective $\mathcal{L}_M$ becomes strongly convex only when the inner learning rate α is restricted in terms of these constants. For DPO, the reference policy term further couples the loss to an external SFT model. We give the first convergence analysis of MAML for both losses in Section 4 (Theorems 4.3 and 4.6), and characterize the regime in which the meta-trained initialization provably accelerates target language adaptation over both target-only and multitask baselines (Theorems E.3 and F.3).

In both algorithms, meta-learning is anchored at the preference learning stage. Formally, a task (language) τ is drawn from a distribution P_{TASK}, and we assume access to a multi-task dataset $\mathcal{D} = \{(x^i, y_w^i, y_\ell^i, \tau^i)\}_{i=1}^N$. We also denote $\mathcal{D}^\tau \subset \mathcal{D}$ as the subset of the dataset corresponding to task τ. Slightly abusing notation, we write $(x, y_w, y_\ell) \sim \mathcal{D}^\tau$ to denote uniform sampling from the dataset $\mathcal{D}^\tau$ for task τ.

3.1 MAML-RLHF

For RLHF, the per-task object that carries the cross-lingual preference signal is the reward function r_ϕ. We therefore meta-learn r_ϕ across languages and adapt it to the target language at deployment time, leaving SFT and policy optimisation as standard single-task stages. Concretely, the task-wise reward-modelling loss in (5) takes the form

Algorithm 2 Reward adaptation

Input: Dataset $\mathcal{D}^{\tau_{\text{new}}}$, meta reward model $\widehat{\phi}_M$, #gradient steps $\widetilde{m}$, step size η.

Initialize $\phi \leftarrow \widehat{\phi}_M$
for $i = 1, \ldots, \widetilde{m}$ **do**
  $\phi \leftarrow \phi - \eta \nabla_\phi \widehat{\mathcal{L}}^{\text{RLHF}}(\phi; \tau_{\text{new}})$.
end for

Output: ϕ.

$$\mathcal{L}^{\text{RLHF}}(\phi; \tau) \triangleq -\mathbb{E}_{(x, y_w, y_\ell) \sim \mathcal{D}^\tau} \left[\log \left(\sigma \left(r_\phi(x, y_w) - r_\phi(x, y_\ell) \right) \right) \right]. \tag{7}$$

Given $\mathcal{L}^{\text{RLHF}}(\phi, \tau)$, we solve the meta-learning problem $\phi_M^\star \in \arg\min_\phi \mathcal{L}_M^{\text{RLHF}}(\phi)$, where

$$\mathcal{L}_M^{\text{RLHF}}(\phi) \triangleq \mathbb{E}_{\tau \sim P_{\text{TASK}}(\cdot)} \left[\mathcal{L}^{\text{RLHF}}(\phi^\tau; \tau) \right], \quad \phi^\tau \triangleq \phi - \alpha \nabla_\phi \mathcal{L}^{\text{RLHF}}(\phi; \tau). \tag{8}$$

Algorithm 1 utilizes the MAML framework [18, 17] to approximate the solution of (8). It learns a meta-initialization $\widehat{\phi}_M$ that can be quickly adapted to a new task using few samples.

At each meta-iteration t, we sample a batch of i.i.d. tasks $\mathcal{B}_t \sim P_{\text{TASK}}$. For each task $\tau \in \mathcal{B}_t$, we first sample an inner-loop dataset $\mathcal{D}_{\text{in}}^\tau \subset \mathcal{D}^\tau$ and perform one gradient step on the empirical RLHF loss $\widehat{\mathcal{L}}^{\text{RLHF}}(\cdot; \tau)$ starting from ϕ_t, yielding an adapted parameter ϕ_{t+1}^τ. We then evaluate the adapted parameter on a fresh outer dataset $\mathcal{D}_{\text{o}}^\tau \subset \mathcal{D}^\tau$, which is sampled independently of $\mathcal{D}_{\text{in}}^\tau$, to compute

the outer empirical loss, and use its gradient to update the meta-parameter. Finally, we aggregate gradients across tasks and take a meta-update step $\phi_{t+1} = \phi_t - \frac{\eta}{B} \sum_{\tau \in \mathcal{B}_t} \nabla_\phi \widehat{\mathcal{L}}^{\mathrm{RLHF}}(\phi_{t+1}^{\tau}; \tau)$.

In the adaption phase, we are given a new task τ_{new}, and a preference dataset $\mathcal{D}^{\tau_{\text{new}}} = \{(x^i, y_w^i, y_\ell^i)\}_{i=1}^n$ where y_w^i (resp. y_ℓ^i) is the winning (resp. losing) completion among the pair of completions y_w^i, y_ℓ^i on the task τ_{new} given prompt x. The reward model r_{ϕ_M} is updated as in Algorithm 2, by performing gradient descent on the empirical loss function $\widehat{\mathcal{L}}^{\mathrm{RLHF}}(\phi; \tau_{\text{new}})$. This results in an adapted reward model $r_\phi^{\tau_{\text{new}}}$. Given the adapted reward model ϕ, we can train the KL-regularized policy specific to task τ_{new} using (3).

3.2 MAML-DPO

For DPO, the policy is the only learnable object in the preference-learning stage, so meta-learning naturally lands on the policy itself. The policy is meta-trained on preference datasets from non-target languages, then adapted to the target language by running DPO on $\mathcal{D}^{\tau_{\text{new}}}$.

For a task τ, the task-wise loss function of policy learning is

$$\mathcal{L}_{\mathrm{DPO}}(\theta; \tau) \triangleq -\mathbb{E}_{(x, y_w, y_\ell) \sim \mathcal{D}^\tau} \left[\log \left(\sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi^{\mathrm{SFT}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_\ell|x)}{\pi^{\mathrm{SFT}}(y_\ell|x)} \right) \right) \right]. \tag{9}$$

Given $\mathcal{L}^{\mathrm{DPO}}(\theta, \tau)$ we solve the meta-learning problem $\phi_M^\star \in \arg\min_\theta \mathcal{L}_M^{\mathrm{DPO}}(\theta)$, where

$$\mathcal{L}_M^{\mathrm{DPO}}(\theta) \triangleq \mathbb{E}_{\tau \sim P_{\mathrm{TASK}}(\cdot)} \left[\mathcal{L}^{\mathrm{DPO}}(\theta^\tau; \tau) \right], \quad \theta^\tau \triangleq \theta - \alpha \nabla_\theta \mathcal{L}^{\mathrm{DPO}}(\theta; \tau). \tag{10}$$

Similar to MAML-RLHF, we employ a MAML framework (Algorithm 1) to approximate the solution for (10) by using DPO loss instead. In the adaptation phase, however, in contrast to meta-RLHF, we can directly update the meta-policy π_M for the new task τ_{new} using the empirical DPO loss [39] i.e. $\widehat{\mathcal{L}}_{\mathrm{DPO}}(\pi; \mathcal{D}_{\tau_{\text{new}}})$. In particular, starting from the meta-policy π_M, we perform gradient descent steps for the loss function $\widehat{\mathcal{L}}_{\mathrm{DPO}}$ and dataset $\mathcal{D}^{\tau_{\text{new}}}$, which is described as in Algorithm 3.

Both approaches described in Section 3.1, 3.2 aim to transfer preference structure across languages for few-shot adaptation, but they apply meta-learning at different points in the pipeline. In practice, RLHF and DPO exhibit different stability and compute tradeoffs even in the standard setting, and this also carries over to the meta setting. We therefore treat MAML-RLHF and MAML-DPO as complementary methods and evaluate how effectively each method adapts to the target language.

Algorithm 3 Policy adaptation

Input: Dataset $\mathcal{D}^{\tau_{\text{new}}}$, meta policy $\widehat{\theta}_M$, number of gradient steps $\widetilde{m}$, and step size η.

1: Initialize $\theta \leftarrow \widehat{\theta}_M$.
2: **for** $i = 1, \ldots, \widetilde{m}$ **do**
3: $\quad \theta \leftarrow \theta - \eta \nabla_\theta \widehat{\mathcal{L}}^{\mathrm{DPO}}(\theta; \tau_{\text{new}})$
4: **end for**

Output: $\pi_{\tau_{\text{new}}}^{\mathrm{DPO}} = \pi_\theta$.

4 Theoretical analysis

In this section, we initiate the study of how accurately Algorithm 1 (i.e., the MAML algorithm) can estimate the solution of the meta-learning problem in both the RLHF and DPO settings. When clear from the context, we drop the subscripts RLHF and DPO and write $\mathcal{L}$ for the corresponding loss. In particular, we use $\mathcal{L}(\phi) = \mathcal{L}^{\mathrm{RLHF}}(\phi)$ for reward learning and $\mathcal{L}(\theta) = \mathcal{L}^{\mathrm{DPO}}(\theta)$ for policy learning.

For theoretical tractability, we focus on a linear reward class and a log-linear policy class, a standard setup in theoretical studies (e.g., [36, 42, 56, 25]). In particular, we study linear reward and log-linear policy functions based on a d-dimensional feature map ψ satisfying $\max_{(x,y)} \|\psi(x, y)\|_2^2 \leq 1$. We also define $\Delta\psi(x, y_w, y_\ell) \triangleq (\psi(x, y_w) - \psi(x, y_\ell))$.

4.1 Convergence of MAML-RLHF

In this section, we analyze the convergence rate of Algorithm 1 in the RLHF setting. For some constant $B_\phi > 0$, we consider the linear reward function class

$$\mathcal{F} = \{r_\phi(x, y) = \phi^\top \psi(x, y) \mid \|\phi\|_2 \leq B_\phi\}, \tag{11}$$

with the corresponding task-wise loss

$$\mathcal{L}(\phi;\tau) = -\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}_\tau}\left[\log\sigma\left(\phi^\top\Delta\psi(x,y_w,y_l)\right)\right].$$

We make the following assumptions on the data distribution.

Assumption 4.1 (Uniform coverage of data distribution)**.** For any task τ, there exists a constant $\nu > 0$ s.t.

$$\mathbb{E}_{(x,y_w,y_\ell)\sim D^\tau}[(\Delta\psi(x,y_w,y_\ell)\Delta\psi(x,y_w,y_\ell)^\top] \succcurlyeq \nu I.$$

Assumption 4.1 is the standard uniform coverage condition in offline reinforcement learning. Here it is used to ensure convexity of the meta-loss function.

Assumption 4.2 (Variance of tasks)**.** Let $\nabla_\phi\mathcal{L}(\phi) = \mathbb{E}_{\tau\sim P_{\text{TASK}}}[\nabla_\phi\mathcal{L}(\phi,\tau)]$. Then there exists $\sigma > 0$ such that

$$\mathbb{E}_{\tau\sim P_{\text{TASK}}}\left[\|\nabla_\phi\mathcal{L}(\phi) - \nabla_\phi\mathcal{L}(\phi,\tau)\|_2^2\right] \le \sigma^2.$$

Assumption 4.2 bounds task diversity and is standard in the MAML literature [17].

Main results. We state a convergence guarantee for MAML applied to the meta objective for RLHF.

Theorem 4.3. *Suppose Assumptions 4.1 and 4.2 hold, and Algorithm 1 is run for $T = O\left(\frac{L}{\alpha^2\sigma^2}\right)$ iterations, and the batch size $B = 1/\alpha^2$, dataset sizes $D_o = \left|\mathcal{D}_o^i\right| = O(1), D_{in} = \left|\mathcal{D}_{in}^i\right| = 1/\sigma$. Then for the linear model:*

$$\mathbb{E}\left[\left\|\phi_M^\star - \widehat{\phi}_M\right\|_2\right] \le \frac{2\alpha\sigma}{m_M},$$

where $m_M = \frac{c\cdot\nu}{\nu+2(4+c)(1+\exp(2B_\phi))}$ for a constant $c > 0$.

Proof. Provided in Appendix C. □

Theorem 4.3 indicates that the proposed first-order method can approximate the solution of the MAML-RLHF objective up to a prescribed error. In particular, it provides a guarantee on how close the output of MAML-RLHF is to the meta-solution of (8) in Euclidean distance. Moreover, Zou et al. [61] shows that the meta-solution is, in expectation, close (in Euclidean distance) to the task-wise solution (see their Theorem 2). This implies that the output of MAML-RLHF also yields an initialization close to the task-wise solution.

4.2 Convergence of MAML-DPO

We now analyze the convergence rate of Algorithm 1 in the DPO setting. Given constant $B_\theta > 0$, we define the log-linear (softmax) policy with feature map ψ as

$$\Pi = \left\{\pi_\theta(y \mid x) = \frac{\exp(\theta^\top\psi(x,y))}{\sum_{y'}\exp(\theta^\top\psi(x,y'))} \;\middle|\; \|\theta\|_2 \le B_\theta\right\}. \tag{12}$$

Using this formulation, the population DPO loss for a task τ [36] can be written as

$$\mathcal{L}(\theta;\tau) = \mathbb{E}_{(x,y_w,y_\ell)\sim D^\tau}\left[\log\left(1+\exp(-z_\theta(x,y_w,y_\ell))\right)\right], \tag{13}$$

where $z_\theta(x,y_w,y_\ell) \triangleq \beta\theta^\top\Delta\psi(x,y_w,y_\ell) - J(x,y_w,y_\ell)$, and $J(x,y_w,y_\ell) \triangleq \beta\log\frac{\pi^{\text{SFT}}(y_w|x)}{\pi^{\text{SFT}}(y_\ell|x)}$. We further make the following assumptions.

Assumption 4.4 (Bounded reference log-ratio)**.** For all τ and $(x,y_w,y_\ell)\in\mathcal{D}^\tau$, there exists $J_0 < \infty$ such that $\left|\log\frac{\pi^{\text{SFT}}(y_w|x)}{\pi^{\text{SFT}}(y_\ell|x)}\right| \le J_0$. Equivalently, $|J(x,y_w,y_\ell)| \le \beta J_0$.

For all triplet (x,y_w,y_ℓ), define the uniform bound on Π:

$$|z_\theta(x,y_w,y_\ell)| \le 2\beta B_\theta + \beta J_0 \triangleq Z_{\max}. \tag{14}$$

Assumption 4.4 is mild because, in practice, both y_w and y_ℓ typically occur with non-negligible probability under the reference policy, so imposing a uniform bound on the log-ratio is a standard coverage condition.

Next, we assume the following bound on the variability of task-wise gradients.

Assumption 4.5. Let $\nabla_\theta\mathcal{L}(\theta) = \mathbb{E}_{\tau\sim P_{\text{TASK}}}[\nabla_\theta\mathcal{L}(\theta,\tau)]$. Then there exists $\sigma > 0$ such that

$$\mathbb{E}_{\tau\sim P_{\text{TASK}}}\left[\|\nabla_\theta\mathcal{L}(\theta) - \nabla_\theta\mathcal{L}(\theta,\tau)\|_2^2\right] \le \sigma^2.$$

Main result. The following theorem gives a finite-sample convergence guarantee for first-order MAML applied to the MAML-DPO objective.

Theorem 4.6. *Assume Assumptions 4.1, 4.4, 4.5 hold. Let* $\alpha \leq \frac{1}{\beta^2\left(2+\frac{8(1+\exp(Z_{\max}))}{\nu}\right)}$. *Suppose Algorithm 1 is run for* $T = O\left(\frac{L}{\alpha^2\sigma^2}\right)$ *iterations, and the batch size* $B = 1/\alpha^2$, *dataset sizes* $D_o = \left|\mathcal{D}_o^i\right| = O(1), D_{in} = \left|\mathcal{D}_{in}^i\right| = \beta/\sigma$. *Then for the log-linear model, we obtain*

$$\left\|\theta_M^\star - \widehat{\theta}_M\right\|_2 \leq \frac{2\alpha\sigma}{m_M},$$

where $m_M = \beta^2 \frac{c\nu^3}{(1+\exp(Z_{\max}))^3}$ *for some constant* $c > 0$.

Proof. Provided in Appendix D. □

Remark 4.7. Theorem 4.6 is the DPO counterpart of Theorem 4.3. It shows that Algorithm 1 can also approximate the minimizer of the DPO meta objective, i.e., $\widehat{\theta}_M$ converges to $\theta_M^\star$, with accuracy controlled by the inner-loop step size α, the task-gradient variability σ, and the conditioning of the meta objective through m_M (the strong-convexity constant of $\mathcal{L}_M$ prescribed in the theorems). Together, these results indicate that the same meta-learning pipeline can be applied to both RLHF (meta-reward learning) and DPO (meta-policy learning), yielding comparable convergence guarantees under analogous coverage and task-variance assumptions.

Remark 4.8. We analyze an unconditional improvement of MAML-DPO's adaptation speed over baseline DPO in Appendix F. As long as the inner learning rate is small enough, $\alpha \leq \min\left\{1/(8\beta^2),\, m_M/(2\sqrt{2}\,\mu_\tau)\right\}$ (where μ_τ is the strong-convexity constant of the per-task DPO loss; see Lemma D.3), Theorem F.3 shows the adaptation of MAML-DPO is faster than baseline DPO. This result further shows that MAML-DPO obtains greater acceleration when the gap between the SFT-initialized policy and preference-optimal policy is bigger, which corresponds to cases where the SFT policy is not yet good enough on the target task. The comparison of MAML-DPO to Multitask-DPO is analyzed in Theorem E.3: we show that the relative improvement of MAML-DPO over Multitask-DPO depends on a task-geometry parameter Φ that measures the alignment between the MAML correction direction and the direction from the multitask solution $\theta_{\text{base}}^\star$ to the unweighted task-optima centroid $\overline{\theta}^\star = \mathbb{E}_\tau[\theta_\tau^\star]$. When Φ is positive and α is small enough, MAML-DPO shows faster adaptation than Multitask-DPO.

5 Experiments

We evaluate the proposed meta-learning frameworks in a multilingual instruction-following setting, where each language is treated as a separate task. Across multiple model sizes and languages, we assess three key aspects of the proposed approach: (i) sample efficiency under limited target-language data, (ii) performance when sufficient adaptation data is available, and (iii) robustness to linguistic distance and different meta-training language groups.

5.1 Experiment setup

We provide further details of our experiment settings in Appendix A, and provide code of our experiments. [2]

Datasets. We use two multilingual preference datasets for the experiments. We use the Okapi dataset [12], derived from Alpaca-64K [45], which provides preference datasets of over 26 languages with over 40k preference pairs per language. We also use `multilingual/orca_dpo_pairs` publicly available on Huggingface, which provides at least 3.4k preference pairs across 7 languages.

Language models. We evaluate our methods under two different model configurations. In the first configuration, we use BLOOM [5] with 7.1B parameters for both the RLHF and DPO settings, which follows a GPT-3-like architecture and was originally pretrained on a multilingual corpus. To examine performance under smaller model scales, we additionally consider a lightweight configuration based on Gemma3 [50], which supports efficient fine-tuning for rapid adaptation. In this configuration, the

[2] Source code is available at https://github.com/geronest/maml-preferences

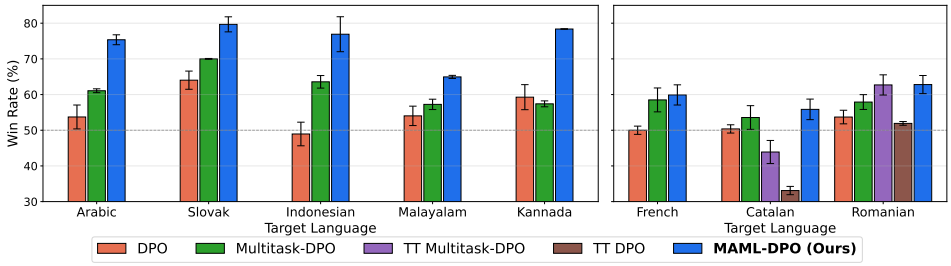


Figure 2: [Left] Win rate of Gemma 270M models adapted to 100 samples. MAML-DPO models significantly outperform baselines with higher win rates. [Right] Win rate of BLOOM 7.1B models adapted to 100 samples. MAML-DPO models show comparable or higher win rates than the baselines, including Translate-Test (TT) with `gpt-4o` as translator.

policy model is replaced by a Gemma3 model with 270 million parameters, and a lightweight BLOOM model with 1.7B parameters is used as the reward model in RLHF. To improve training efficiency, we apply LoRA adapters for both BLOOM models during fine-tuning to reduce memory consumption.

Baselines. In order to examine the efficiency of our approach, we compare against *standard* RLHF and DPO pipelines trained with the same amount of target-language preference data as the meta-trained models. We additionally include a *multitask* baseline that uses the same multilingual training data as the meta-learning methods, but removes the inner-loop adaptation step (i.e., setting the inner loop learning rate $\alpha = 0$ in Algorithm 1). This baseline allows us to isolate the effect of meta-learning from standard multitask pretraining. To provide a stronger cross-lingual transfer baseline, we also include translation-test approaches, which uses `gpt-4o` to translate target language prompts and responses to a high-resource language (e.g. French) and translate it back to the target language. Further details of the experiment configurations are provided in Appendix A.

Evaluation metrics. We evaluate how quickly and robustly the models adapt to the limited target language data using separately trained and fixed reward models, referred to as the *Judge reward models (RM)*. Each configuration uses its corresponding reward model architecture as the Judge RM, which is fully fine-tuned on the entire preference dataset and used exclusively for evaluation. The Judge RMs compute *win rates* by comparing model-generated responses against a fixed base policy, defined as the initial policy model without any training or adaptation. During evaluation, all policy models generate responses to the same set of 1,000 prompts sampled from the test split of policy optimization dataset. Appendix B also includes the *reward accuracies* of models in the validation split of the preference dataset to examine how the performances of reward function evolve during the training.

5.2 Main results

How do meta-trained models perform under limited adaptation data? We first evaluate adaptation performance in an extremely low-resource situation, using only 100 preference samples from the target language. In this setting, the model repeatedly adapts to a fixed and small set of samples, resulting in both limited data quantity and limited data diversity during adaptation. We evaluate model adaptation on eight target languages, reporting results for three Romance languages from the same family as the meta-training tasks and for five languages from different families in Figure 2. We compare the win rates of different models against the base model using the judge RM as described in Section 5.1. We report standard errors over three random seeds and show that MAML-DPO consistently outperforms baseline DPO and translation-test baseline. Figure 2-left shows that when adapting to target languages from different language families, MAML-DPO significantly outperforms the multi-task baseline. RLHF experiment results in Appendix B.2 show similar pattern, where MAML-RLHF models achieve consistently higher win rates across all target languages. Figure 2-right demonstrates the performance of MAML-DPO comparable or better than the baselines when adapted to European languages, despite the translate-test baseline's use of a strong model `gpt-4o`.

How do meta-trained languages affect the adaptation performance? We investigate the effect of languages used for meta-training on the adaptation performance in Figure 3. We test three different groups for meta-training: Romanian, Catalan, Italian, Spanish (Group 1, Romance); Vietnamese, Indonesian, Arabic, Hindi (Group 2, mixed non-Romance with, spanned high, medium and low

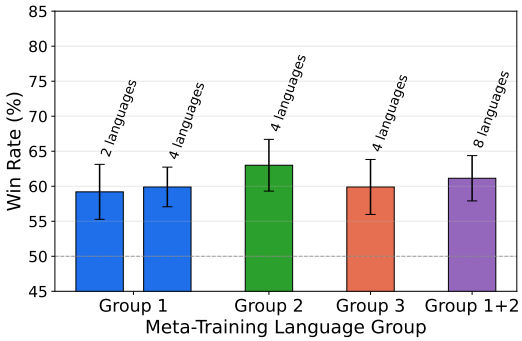


Figure 3: Win rates of BLOOM 7.1B models with different language groups for the MAML-DPO stage, adapted to 100 French samples. Neither the number of languages nor linguistic distance significantly affects adaptation performance.

resource); Chinese, Bengali, Malayalam, Kannada (Group 3, mixed non-Romance, paired a single high-resource language with three low-resource ones). The adaptation target is fixed to French and the goal is to test whether adaptation to French depends on linguistic similarity between meta-training and target languages, so both non-Romance groups were chosen to be distant from French, with the number of source languages fixed at four. We additionally vary the number of meta-training languages by sub-sampling Group 1 to its first two members (Romanian, Catalan) for the "Group 1 (2 languages)" bar in Figure 3 . Across three random seeds, the win rates of all three groups overlap within standard error, and Group 2 shows the highest point estimate despite being linguistically distant from French. This result is consistent with our theoretical bound (56), which depends on task-gradient variance σ^2 and the SFT preference-loss gap R_{SFT} rather than on linguistic distance to the target. Increasing the number of meta-training languages produces a small additional gain.

Table 1: Win rates of Gemma-270M models adapted to 4k and 40k samples from each target language. MAML-DPO models consistently outperform the baselines, followed by multitask method.

Target	Algorithm	Samples	WR(%)
French	MAML-DPO	4000	**66.60**
	Multitask-DPO	4000	65.70
	DPO	4000	56.20
	MAML-DPO	40000	**75.40**
	Multitask-DPO	40000	65.58
	DPO	40000	54.00
Catalan	MAML-DPO	4000	**71.95**
	Multitask-DPO	4000	70.20
	DPO	4000	61.35
	MAML-DPO	40000	**74.50**
	Multitask-DPO	40000	70.30
	DPO	40000	62.90
Romanian	MAML-DPO	4000	**86.95**
	Multitask-DPO	4000	84.50
	DPO	4000	71.35
	MAML-DPO	40000	**88.20**
	Multitask-DPO	40000	80.15
	DPO	40000	75.30

Table 2: Win rate of BLOOM 7.1B models adapted to 8k/40k samples from each target language. MAML training stage also improves performance with a larger amount of data.

Target	Algorithm	Samples	WR(%)
French	MAML-DPO	8000	**65.17**
	Multitask-DPO	8000	56.90
	DPO	8000	64.92
	MAML-DPO	40000	71.4
	Multitask-DPO	40000	**73.74**
	DPO	40000	67.4
Catalan	MAML-DPO	8000	**66.75**
	Multitask-DPO	8000	63.93
	DPO	8000	59.46
	MAML-DPO	40000	68.16
	Multitask-DPO	40000	**71.08**
	DPO	40000	68.16
Romanian	MAML-DPO	8000	**64.47**
	Multitask-DPO	8000	63.00
	DPO	8000	59.12
	MAML-DPO	40000	67.29
	Multitask-DPO	40000	**69.56**
	DPO	40000	65.27

Do meta-trained models still have advantage under sufficient adaptation data? In this setting, we adapt models using 4K/8K and 40K samples for MAML-DPO, while for MAML-RLHF we use 8K/12K and 40K samples. Table 1 and Table 2 show the performances of Gemma3 270M and BLOOM 7.1B DPO models when sufficient amount of adaptation samples are provided, from 4K/8K to 40K. While MAML-DPO outperforms both baselines across all three target languages with 4K/8K adaptation samples, we note that Multitask-DPO achieves higher win rates than MAML-DPO at 40K samples in BLOOM 7.1B and the performance gap between the MAML-DPO models and standard DPO narrows in many cases. This is expected because the number of meta-training samples is less than 20K, whose advantageous effect on the adaptation stage is supposed to diminish as the number of adaptation datapoints exceeds it. MAML-RLHF in Table 12 and Table 13 outperforms standard RLHF in every setting and is competitive with multitask-RLHF, which is consistent with multitask-RLHF being a strong baseline when the target language is closely aligned with the Romance meta-training tasks. These results show that meta-training accelerates early adaptation while preserving competitive performance with sufficient target data.

6 Related work

While most LLMs possess multilingual capabilities, their performance across languages is unequal [4, 51, 13, 54]. Previous works have investigated improving multilingual capability of LLMs by curating data [8, 57, 26], adopting a new objective during pretraining [2], in-context learning [32, 6, 29], and fine-tuning on multilingual datasets [12, 28]. Improving few-shot capabilities of deep neural networks have been investigated from meta-learning perspectives [53, 18, 23]. Application of meta-learning to LLMs have been studied, but most of them discuss in-context learning scenarios [7, 10, 40, 30, 22, 33]. Hu et al. [24] meta-train a separate autoregressive model to train importance weighting of tokens in a document, which is then used to fine-tuning an LLM. A complementary line of work addresses multilingual preference alignment via gradient-level interventions, filtering conflicting per-language gradients during joint preference training to mitigate negative interference across languages [31]. Our approach does not modify the gradient of joint training but replace it with a meta-objective that explicitly optimises for fast *adaptation* to a held-out target language. To the best of our knowledge, our work is the first approach to addressing under-representation of low-resource languages by applying meta-learning to the preference learning pipeline, providing theoretical guarantees and demonstrating improved multilingual capabilities. See Appendix G for further discussion on related work.

7 Conclusion

We presented a meta-learning framework for aligning multilingual LLMs using human preference data. By treating each language as a task, we develop MAML-RLHF and MAML-DPO as two meta-learning algorithms that leverage preference data across languages to enable efficient adaptation to low-resource languages. We provided theoretical guarantees and empirically demonstrated substantial improvements in sample efficiency across multiple languages and model sizes. In particular, meta-trained models achieve higher win rates than standard RLHF and DPO baselines in low-resource languages, while maintaining or improving performance as more adaptation data become available. Importantly, these gains remain consistent across varying language distances between meta-training and target languages, and are not sensitive to the choice of meta-language combinations. Overall, our work demonstrates meta-learning as an effective way to address data imbalance in multilingual alignment. While we focused on preference learning, extending our approach to reinforcement learning stages and reducing end-to-end pipeline compute are promising directions for future work.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. Jiaying Lin was supported by the China Scholarship Council under Grant No. 202408310075.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, et al. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Rachel Bawden and François Yvon. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*, 2023.

[5] BigScience Workshop. BLOOM (revision 4ab0472), 2022. URL https://huggingface.co/bigscience/bloom.

[6] Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*, 2024.

[7] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.53. URL https://aclanthology.org/2022.acl-long.53/.

[8] Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. Tigerbot: An open multilingual multitask llm, 2023. URL https://arxiv.org/abs/2312.08688.

[9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[10] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X. Wang, and Eric Schulz. Meta-in-context learning in large language models, 2023. URL https://arxiv.org/abs/2305.12907.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[12] Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307, 2023.

[13] Guoliang Dong, Haoyu Wang, Jun Sun, and Xinyu Wang. Evaluating and mitigating linguistic discrimination in large language models: Perspectives on safety equity and knowledge equity. 2025.

[14] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf, 2024. URL https://arxiv.org/abs/2405.07863.

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

[16] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[17] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR, 2020.

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[19] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

[20] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[22] Youssef Attia El Hili, Albert Thomas, Malik Tiomoko, Abdelhakim Benechehab, Corentin Léger, Corinne Ancourt, and Balázs Kégl. Llms as in-context meta-learners for model and hyperparameter selection, 2025. URL https://arxiv.org/abs/2510.26510.

[23] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44 (9):5149–5169, 2021.

[24] Nathan Hu, Eric Mitchell, Christopher D Manning, and Chelsea Finn. Meta-learning online adaptation of language models. *arXiv preprint arXiv:2305.15076*, 2023.

[25] Kaixuan Ji, Jiafan He, and Quanquan Gu. Reinforcement learning from human feedback with active queries. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EScatQaRxz. Featured Certification.

[26] Aditya Kallappa, Palash Kamble, Abhinav Ravi, Akshat Patidar, Vinayak Dhruv, Deepak Kumar, Raghav Awasthi, Arveti Manjunath, Himanshu Gupta, Shubham Agarwal, Kumar Ashish, Gautam Bhargava, and Chandra Khatri. Krutrim llm: Multilingual foundational model for over a billion people, 2025. URL https://arxiv.org/abs/2502.09642.

[27] Siddharth Karamcheti, Laurel Orr, Jason Bolton, Tianyi Zhang, Karan Goel, Avanika Narayan, Rishi Bommasani, Deepak Narayanan, Tatsunori Hashimoto, Dan Jurafsky, et al. Mistral–a journey towards reproducible language model training, 2021.

[28] Wen Lai, Mohsen Mesgar, and Alexander Fraser. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*, 2024.

[29] Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, 2024.

[30] Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. Meta in-context learning makes large language models better zero and few-shot relation extractors, 2024. URL https://arxiv.org/abs/2404.17807.

[31] Jiangnan Li, Thuy Vu, Christian Herold, Amirhossein Tebbifakhr, Shahram Khadivi, and Gholamreza Haffari. Congrad: Conflicting gradient filtering for multilingual preference alignment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6371–6387, 2026.

[32] Junlei Li, Xiao Wei, Xiaobao Wang, Ning Zhuang, Longbiao Wang, and Jianwu Dang. Language-emphasized cross-lingual in-context learning for multilingual llm. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 327–339. Springer, 2024.

[33] Yichuan Li, Xiyao Ma, Sixing Lu, Kyumin Lee, Xiaohu Liu, and Chenlei Guo. Mend: Meta demonstration distillation for efficient and effective in-context learning, 2024. URL https://arxiv.org/abs/2403.06914.

[34] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[35] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[36] Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanović, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[38] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024.

[39] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Sanchit Sinha, Yuguang Yue, Victor Soto, Mayank Kulkarni, Jianhua Lu, and Aidong Zhang. Maml-en-llm: Model agnostic meta-training of llms for improved in-context learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2711–2720, 2024.

[41] Seongho Son, William Bankes, Sayak Ray Chowdhury, Brooks Paige, and Ilija Bogunovic. Right now, wrong then: Non-stationary direct preference optimization under preference drift. *arXiv preprint arXiv:2407.18676*, 2024.

[42] Seongho Son, William Bankes, Sayak Ray Chowdhury, Brooks Paige, and Ilija Bogunovic. Right now, wrong then: Non-stationary direct preference optimization under preference drift. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=c2xhvZA8Xq.

[43] Yuhao Sun, Yifan Zhang, Quandong Wang, QinZhuo Wu, Wei Liu, and Jian Luan. Spo: Self preference optimization with self regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5601–5614, 2025.

[44] Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. Game-theoretic regularized self-play alignment of large language models. *arXiv preprint arXiv:2503.00030*, 2025.

[45] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[46] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, et al. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

[48] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[49] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[50] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[51] Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*, 2024.

[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[53] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[54] Dandan Wang, Stephanie Jean Tsang, and Yadong Zhou. Performance unfairness of large language models in cross-language fact-checking. *Information Processing & Management*, 63 (4):104616, 2026. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2026.104616. URL https://www.sciencedirect.com/science/article/pii/S0306457326000087.

[55] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more, 2024. URL https://arxiv.org/abs/2407.16216.

[56] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=c1AKcA6ry1.

[57] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*, 2024.

[58] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

[59] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[60] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

[61] Yingtian Zou, Fusheng Liu, and Qianxiao Li. Unraveling model-agnostic meta-learning via the adaptation learning rate. In *International Conference on Learning Representations*, 2022.

Appendix

Table of Contents

A Further Details of Experimental Setting **17**

B Additional Experimental Results **19**
B.1 Additional Results of MAML-DPO with LoRA adapters 19
B.2 Additional Results of MAML-RLHF with LoRA adapters 22
B.3 Resource Consumption of MAML-DPO 25

C Convergence of Meta-RLHF **26**
C.1 Proof of Theorem 4.3 26
C.2 Technical Lemmas 27

D Convergence of Meta-DPO **29**
D.1 Proof of Theorem 4.6 29
D.2 Technical Lemmas 29

E Convergence Analysis of the Adaptation Phase for MAML-DPO **31**
E.1 Preliminaries and Task Geometry 31
E.2 Deriving the Initialization Advantage 31
E.3 Comparative Convergence Rate 33

F Comparative Convergence Analysis: MAML-DPO vs. Baseline-DPO **34**
F.1 Setup and Baseline-DPO Initialization 34
F.2 Initialization Distance Bounds 34
F.3 Comparative Convergence Rate 35
F.4 Discussion 36

G Additional Related Work **37**

H Impact Statement and Limitations **37**

A Further Details of Experimental Setting

In this section, we explain the remaining details of the experimental setting.

Datasets. For each language task, the Okapi dataset provides 52K instruction–response pairs for supervised fine-tuning, 42K preference comparisons for reward modeling or DPO, and 64K instructions for policy optimization. For meta-training with Okapi, we use 12K samples from each language for DPO and 16K samples from each language for reward modeling. For the target languages, we use 100 or more samples for each of SFT, reward modeling, and policy optimization stage. Unless specified, MAML-based or Multitask- approaches use one of French, Italian, Spanish, Catalan, Romanian as the target language, while the others are used for MAML/Multitask training stage. For `multilingual/orca_dpo_pairs`, we use 3 sets of 1K samples for each of SFT, preference optimization for policies, preference learning for judge reward functions. 2 sets of 200 samples are used for each of preference learning validation and win rate evaluation using judge reward models. For MAML/Multitask training, we use Arabic, German, Russian and Chinese, while using the remaining French, Spanish, Turkish for adaptation.

MAML-DPO. The base model is directly trained with MAML-DPO using 4 out of 5 languages being considered in Section 5.1. For BLOOM with 7.1B parameters, MAML-DPO is trained for 400 steps, while each step accumulates gradients over 40 iterations of choosing a task and sampling 6 datapoints. We use 3×10^{-5} for learning rate, while 3 is multiplied to the learning rate for inner loop updates. For each outer loop update, we use 2 consecutive inner loop steps. Once meta-training is complete, we apply adaptation DPO. We use batch size 20 and 2 gradient accumulation steps with 1×10^{-5} as the learning rate. For experiments with 100 and 8K samples, we train the models over 200 steps and use the model showing the lowest validation loss. For experiments with 40K samples, we train the model for 2K steps.

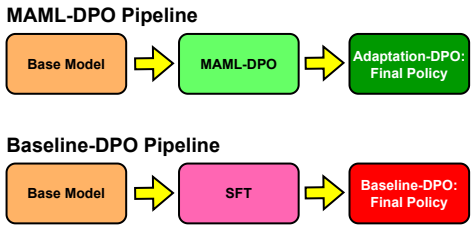


Figure 4: Visualization of DPO training pipelines. MAML-DPO pipeline does not use SFT, instead initializing the Adaptation-DPO policy with MAML-DPO algorithm on several non-target languages with sufficient data.

For Gemma-3 with 270M parameters, MAML-DPO is meta-trained for 300 steps. At each outer step, the model accumulates gradients over 40 iterations by selecting 2 tasks and sampling 2 data points per task. We use a single inner loop step for each outer loop update. The learning rate configuration is chosen from the pairs $(5 \times 10^{-5}, 5)$ and $(1 \times 10^{-5}, 3)$ for the outer learning rate and inner-loop learning coefficient. After meta-training, the model is adapted using DPO. For experiments with 100 samples, adaptation is performed for 100 steps using a batch size of 2 with 20 gradient accumulation steps. For experiments with 4k and 40k samples, adaptation is trained over 1000 steps with a batch size of 2, using 2 and 20 gradient accumulation steps, respectively. The adaptation learning rate is selected between 1×10^{-4} and 5×10^{-6}.

We visually describe the pipeline of our DPO training pipeline in Figure 4.

MAML-RLHF. Before reward modeling, MAML-SFT is first applied to the base model, followed by MAML-RM and then adaptation reward modeling. For BLOOM with 7.1B parameters, we use 100 outer steps for MAML-SFT with 3×10^{-5} as the learning rate and multiplying 3 to it for a single inner loop update. Each outer step includes 40 gradient accumulation steps of sampling 2 datapoints from each of 2 tasks. MAML-RM uses the same learning rate and datapoint sampling hyperparameters as MAML-SFT, except that MAML-RM is run for 400 steps. Adaptation reward modeling fine-tunes the MAML-RM trained model with 1×10^{-5} as the learning rate and 40 as the batch size without gradient accumulation. For 100 samples, we use 100 steps of reward modeling while we use 200 steps for 8K samples. The policy is first trained with SFT, using 1×10^{-4} as the learning rate and 40 as the batch size without gradient accumulation. For experiments with 8K samples, SFT is run for 200 steps. The final fine-tuning of MAML-RLHF pipeline is applying GRPO to the SFT model using the adapted reward model. We use 400 prompts for training, while each of

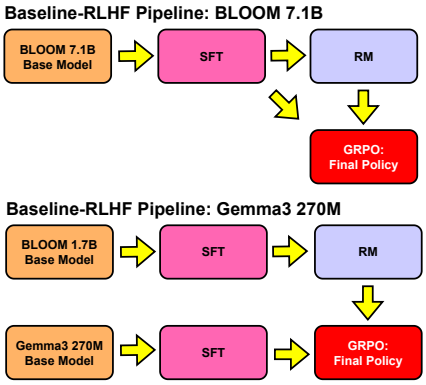


Figure 5: Visualization of baseline RLHF pipeline for both model architectures. For Gemma3 270M policies, BLOOM 1.7B is used for initializing the reward function.

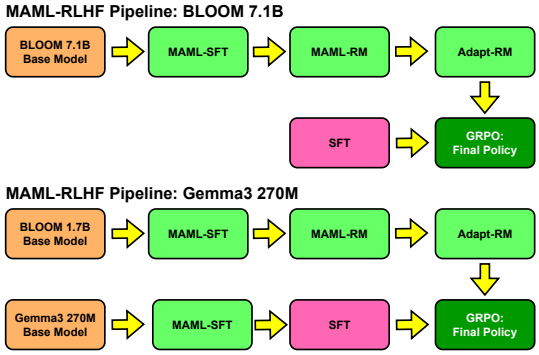


Figure 6: Visualization of MAML-RLHF pipeline for BLOOM 7.1B model and Gemma3 270M model. The main difference between the models is that for Gemma3 270M model, we use a separate BLOOM 1.7B model as the base model of reward functions.

100 training steps consists of generating 4 responses from each prompt. We use batch size 4 and 4 gradient accumulation steps with learning rate of 1×10^{-4}.

For Gemma 3 with 270M parameters, MAML-SFT is trained for 3000 outer steps with an outer learning rate of 5×10^{-4}, using an inner-loop learning rate scaled by a factor of 3. Each outer step samples 2 data points from each of 2 tasks per step, without extra accumulation steps. Initialized from the MAML-SFT model, the SFT policy is fine-tuned with a learning rate of 1×10^{-4} and a batch size of 4 in 3000 steps, without extra accumulation steps. MAML-RM is trained for 300 steps with 40 gradient accumulation steps, sampling 4 data points from each of 2 tasks per accumulation step. The learning rate configuration is selected from $(5 \times 10^{-4}, 3)$ and $(3 \times 10^{-4}, 1)$ for the outer learning rate and inner-loop coefficient. We use 8 data points with 5 gradient accumulation steps for adapting reward models. It is trained for 100, 300, and 1000 steps in the 100-, 12k-, and 40k-sample experiments, respectively, with adaptation learning rates selected between 5×10^{-4} and 5×10^{-5}. Using the adapted reward model for the target language, the SFT policy is further optimized via GRPO for 1000 steps with batch size 4 and no gradient accumulation, where the GRPO learning rate is selected between 5×10^{-5} and 5×10^{-6}.

Visualization of the baseline RLHF pipeline for both models is presented in Figure 5. See Figure 6 for the MAML-RLHF pipelines for each model architecture.

Compute Resources. For experiments with BLOOM 7.1B models, we use one NVIDIA H200 GPU. For experiments with Gemma3 270M models, we use NVIDIA Quadro RTX 6000 GPU.

B Additional Experimental Results

B.1 Additional Results of MAML-DPO with LoRA adapters

Limited adaptation data. Based on Gemma3 270M, Table 3 provides the numerical results of adaptation performance for MAML-DPO using only 100 samples from non-Romance target languages, where MAML-DPO substantially outperforms both the multitask and standard DPO baselines. Since our meta-DPO policies are trained with 4 Romance languages, these results suggest that the proposed meta-learning approach can retain its adaptation advantage even when the target languages are linguistically distant from the meta-training tasks. This is particularly important because larger linguistic distances can reduce the transferability of shared multilingual representations, thereby weakening the effectiveness of standard multitask training.

Table 3: Win rate of Gemma3 270M adapted to 100 samples from each non-Romance target language in the DPO pipeline.

Target	Algorithm	Samples	WR (%)
Arabic	MAML-DPO	100	$\mathbf{75.36} \pm 1.39$
	Multitask	100	61.07 ± 0.54
	DPO	100	53.72 ± 3.35
Slovak	MAML-DPO	100	$\mathbf{79.68} \pm 2.11$
	Multitask	100	70.00 ± 0.12
	DPO	100	64.03 ± 2.55
Indonesian	MAML-DPO	100	$\mathbf{76.91} \pm 4.90$
	Multitask	100	63.57 ± 1.76
	DPO	100	48.95 ± 3.31
Malayalam	MAML-DPO	100	$\mathbf{64.95} \pm 0.43$
	Multitask-DPO	100	57.25 ± 1.45
	DPO	100	54.03 ± 2.72
Kannada	MAML-DPO	100	$\mathbf{78.37} \pm 0.07$
	Multitask-DPO	100	57.40 ± 0.83
	DPO	100	59.27 ± 3.51

Table 4 and Table 5 present the adaptation performance of MAML-DPO on Romance target languages based on Gemma with both Gemma 270M and Bloom 7.1B. Across all three target languages, MAML-DPO consistently demonstrates strong adaptation ability and substantially outperforms conventional DPO. Compared with the multitask baseline, MAML-DPO achieves either comparable or better performance. The ranking among these three algorithms is preserved after controlling for response length in the Romance-language evaluations. This indicates that our method remains competitive even when the target language is linguistically close to the meta-training languages, where multitask training is expected to be a relatively strong baseline.

Table 4: Raw and length-controlled win rates of Gemma3 270M adapted to 100 samples from each target language in the DPO pipeline. Δ denotes (LC-WR - raw WR), and Δlength denotes response-length difference between the adapted model and its paired comparison model in tokens.

Target	Algorithm	Samples	Raw WR (%)	LC-WR (%)	Δ	Δlength
French	MAML-DPO	100	$\mathbf{66.88} \pm 1.47$	**67.03**	$+0.15$	$+9.8$
	Multitask-DPO	100	66.28 ± 0.84	65.38	-0.90	$+11.6$
	DPO baseline	100	53.81 ± 5.76	53.78	-0.03	$+3.8$
Catalan	MAML-DPO	100	$\mathbf{70.55} \pm 0.72$	**74.23**	$+3.68$	-4.1
	Multitask-DPO	100	69.75 ± 0.29	72.18	$+2.43$	$+0.0$
	DPO baseline	100	63.45 ± 1.76	58.98	-4.47	$+29.3$
Romanian	MAML-DPO	100	$\mathbf{79.83} \pm 0.26$	**78.93**	-0.90	-43.8
	Multitask-DPO	100	59.51 ± 0.94	58.63	-0.88	-11.2
	DPO baseline	100	51.56 ± 2.97	51.85	$+0.28$	-3.6

Table 5: Raw and length-controlled win rates of BLOOM-7.1B adapted to 100 samples from each target language in the DPO pipeline. Δ denotes (LC-WR - raw WR), and Δlength denotes response-length difference between the adapted model and its paired comparison model in tokens.

Target	Algorithm	Samples	Raw WR (%)	LC-WR (%)	Δ	Δlength
French	MAML-DPO	100	**59.90**	**61.02**	+1.13	−19.2
	Multitask-DPO	100	58.48	59.63	+1.15	−19.6
	DPO baseline	100	52.05	52.19	+0.14	−5.8
Catalan	MAML-DPO	100	**55.85**	**57.34**	+1.49	−22.4
	Multitask-DPO	100	53.57	52.45	−1.12	+5.3
	DPO baseline	100	50.37	49.93	−0.44	+2.6
Romanian	MAML-DPO	100	**62.75**	**63.26**	+0.51	−8.0
	Multitask-DPO	100	58.15	59.18	+1.03	−19.0
	DPO baseline	100	57.33	57.64	+0.32	−6.4

Table 6: Refusal rates (%) of BLOOM-7.1B and Gemma3-270M in the 100-sample setting, evaluated over 1,000 prompts per target language. Refusals are classified by GPT-5.4-mini. All results are from a single seed.

Model	Target	MAML-DPO	Multitask-DPO	DPO baseline	Base model
BLOOM-7.1B	French	0.81	0.81	1.71	0.00
	Catalan	1.61	0.71	3.73	2.72
	Romanian	0.10	0.30	0.20	0.10
Gemma3-270M	French	0.50	0.40	0.20	0.20
	Catalan	0.00	0.00	0.00	0.30
	Romanian	0.00	0.40	0.20	0.20

In Table 6, refusal rates are low across both model families, remaining below 4% for BLOOM-7.1B and below 1% for Gemma3-270M. For BLOOM-7.1B, MAML-DPO refuses less often than the DPO baseline in all three target languages, while in Catalan both MAML-DPO and Multitask-DPO also show lower refusal rates than the base model. For Gemma3-270M, refusal rates remain similarly small, with no systematic increase under MAML-DPO relative to the other methods or the base model. These results suggest that the higher win rates achieved by MAML-DPO are not driven by increased refusal behaviour.

Out-of-distribution evaluation. To examine robustness beyond the training distribution, we follow the Okapi evaluation protocol in Dac Lai et al. [12] and report zero-shot performance on ARC [9], HellaSwag [58], and MMLU [21], which are multiple-choice benchmarks and assess robustness to out-of-distribution shifts in format and content.

In Table 7, we evaluate BLOOM 7.1B models trained with DPO algorithms in OOD benchmarks. While both MAML-DPO and baseline DPO models have used 10K samples for adaptation to the target language, baseline DPO models show higher accuracy than the base model in fewer columns than MAML-DPO. This shows the robustness of MAML-DPO training pipeline to OOD evaluations.

We further provide OOD benchmark evaluation of DPO models based on Gemma3 270M architecture in Table 8. Similarly to Table 7, MAML-DPO models show improvements in more benchmarks and metrics than baseline DPO models. This indicates that MAML-DPO provides stable improvement across different model architectures and data domains.

Table 7: Out-of-distribution benchmark evaluation results of BLOOM 7.1B models. Bold numbers indicate where the trained model outperformed the base model. With 10k samples given for adaptation, MAML-DPO maintains its robustness to OOD datapoints and gains improvement in more diverse benchmarks and metrics than the baseline method.

Language	Algorithm	Samples	ARC		Hellaswag		MMLU	
			Acc.	Norm. Acc.	Acc.	Norm. Acc.	Acc.	Norm. Acc.
French	Base model	-	0.3293	0.3644	0.4259	0.5452	0.2884	0.297
	MAML-DPO	10000	**0.3362**	**0.3704**	**0.434**	**0.5464**	**0.2885**	**0.2977**
	DPO	10000	0.3251	**0.3721**	0.4339	**0.5477**	0.2820	0.2887
Catalan	Base model	-	0.3199	0.3482	0.3956	0.4992	0.2796	0.2887
	MAML-DPO	10000	**0.3242**	**0.3611**	**0.401**	0.4981	**0.2811**	**0.2898**
	DPO	10000	**0.3344**	**0.3670**	**0.3981**	0.4958	0.2730	0.2829
Romanian	Base model	-	0.2091	0.2708	0.2814	0.3186	0.2557	0.2736
	MAML-DPO	10000	**0.2134**	0.2639	**0.2864**	**0.3254**	**0.2599**	**0.2739**
	DPO	10000	0.**2125**	0.2699	**0.2864**	**0.3266**	**0.2596**	0.2719

Table 8: Out-of-distribution (OOD) benchmark evaluation results of Gemma 270M models under baseline and MAML-DPO. Bold numbers indicate where the trained model outperformed the base model.

Language	Algorithm	Samples	ARC		Hellaswag		MMLU	
			Acc.	Norm. Acc.	Acc.	Norm. Acc.	Acc.	Norm. Acc.
French	Base model	-	0.210	0.249	0.271	0.261	0.251	0.266
	MAML-DPO	40000	**0.222**	**0.253**	**0.274**	**0.263**	0.250	0.261
	DPO	40000	**0.212**	0.243	0.271	0.259	0.249	0.262
Catalan	Base model	-	0.212	0.239	0.265	0.281	0.242	0.252
	MAML-DPO	40000	**0.214**	0.225	**0.282**	**0.29**	**0.257**	**0.257**
	DPO	40000	0.200	0.239	**0.274**	0.281	0.24	0.244
Romanian	Base model	-	0.191	0.240	0.283	0.276	0.260	0.268
	MAML-DPO	40000	**0.194**	**0.246**	**0.294**	0.259	**0.262**	0.264
	DPO	40000	0.186	0.233	**0.284**	0.263	0.256	0.259

Table 9: Accuracy on the French subsets of MGSM and MMMLU in the 100-sample setting. BLOOM-7.1B results are averaged over three random seeds.

Model	Method	MGSM (Acc. %)	MMMLU (Acc. %)
Gemma-270M	MAML-DPO	**3.20**	**28.90**
	Multitask-DPO	0.40	26.50
	DPO baseline	1.20	25.90
	Base model	1.60	26.60
BLOOM-7.1B	MAML-DPO	4.27	28.68
	Multitask-DPO	4.00	**28.70**
	DPO baseline	4.00	28.48
	Base model	**4.40**	28.15

We additionally evaluate our methods on MGSM and MMMLU and results are shown in Table 9. For Gemma-270M, MAML-DPO obtains the highest accuracy on both benchmarks, indicating that few-shot adaptation does not sacrifice OOD capability. For BLOOM-7.1B the three methods are statistically indistinguishable: the spread between them is at most 0.3% against a standard error near 1.0%, and every method sits within 0.6% of the untrained model on both benchmarks — marginally above it on MMMLU, marginally below on MGSM.

B.1.1 Text generation examples

We further provide a representative example to illustrate the qualitative differences between adaptation methods at the Gemma-3-270M scale. The prompt is in Romanian:

Prompt (Romanian): *"Sugerați 3 strategii de marketing pe care companiile ar trebui să implementeze pentru a crește conversiile clienților pe site-ul lor."*

English translation: "Suggest three marketing strategies companies should implement to increase customer conversions on their website."

We preserve the original wording of the model outputs, remove HTML tags, and use [...] to indicate omitted text.

Base model (Gemma-3-270M)

"1. Provide three strategies on how your company should achieve this 2. How will you implement it? [...] The first step is to research your product [...]"

The base model ignores the Romanian instruction, responds in English, and largely returns questions or generic guidance rather than three concrete marketing strategies.

DPO baseline

"Strategii de marketing includ strategiile de marketing corectă. [...] personalizarea proiectelor și meniuul online [...]"

The model attempts to answer in Romanian, but the output is grammatically degraded and does not provide three clearly distinguishable, conversion-oriented strategies.

Multitask-DPO

"Strategii de marketing includ trei tipuri de marketing: Marketing înlocuitor / Marketing reprezentant / Marketing personalizare [...] precum festivaluri, concerte [...]"

Multitask-DPO better follows the requested three-item structure, but several proposed strategies are semantically inappropriate for website conversion and the response drifts toward unrelated offline activities.

MAML-DPO (ours; 100 Romanian examples)

"1. Promovă o socializare inteligente. [...] creșterea conversiției de clienți [...] Promovând newsletter clienților."

MAML-DPO remains primarily in Romanian and makes a more explicit connection between online promotional actions, newsletters, and customer conversion, although malformed phrasing and repetition remain.

Overall, this example illustrates the relative behaviour of the four models: the base model fails target-language following, DPO produces vague and degraded Romanian, Multitask-DPO better follows the requested format but proposes semantically unsuitable strategies, while MAML-DPO more directly connects its recommendations to conversion-oriented actions after adaptation on only 100 Romanian examples. This qualitative example is intended as an illustration and our aggregate evaluation remains the primary evidence.

B.2 Additional Results of MAML-RLHF with LoRA adapters

Limited adaptation data. Table 10 reports the win rates of the corresponding policies optimized with these reward models in the RLHF pipeline. MAML-RLHF achieves the highest mean win rate across all three Romance target languages, suggesting that the benefits of our approach are not limited to reward modeling and can also be reflected to policy. At the same time, the margins over the multitask-RLHF are relatively small in Catalan and French. In both cases, the multitask-RLHF substantially improves over conventional RLHF, suggesting that pretraining on Romance preference data provides a strong basis for adapting to these targets. Against this competitive baseline, MAML-RLHF still performs best on average.

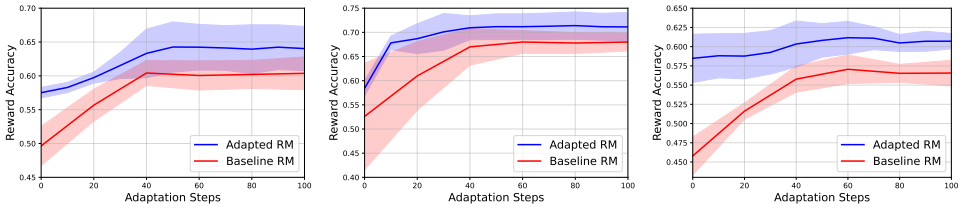


Figure 7: Average reward accuracies with standard deviation bands of BLOOM 1.7B reward functions over 3 random seeds during adaptation on the target language with 100 samples. [Left] Arabic. [Center] Slovak. [Right] Indonesian. In all settings, MAML reward functions show consistently higher reward accuracy than the baseline reward functions in adaptation.

Table 10: Win rate of Gemma3 270M adapted to 100 samples from each target language in the RLHF pipeline.

Target	Algorithm	Samples	WR (%)
French	MAML-RLHF	100	$\mathbf{66.23} \pm 1.32$
	Multitask-RLHF	100	64.55 ± 1.40
	RLHF	100	49.12 ± 2.21
Catalan	MAML-RLHF	100	$\mathbf{63.28} \pm 1.35$
	Multitask-RLHF	100	62.97 ± 1.07
	RLHF	100	50.03 ± 3.30
Romanian	MAML-RLHF	100	$\mathbf{85.23} \pm 0.32$
	Multitask-RLHF	100	63.41 ± 2.34
	RLHF	100	60.28 ± 6.17

In Figure 7, we evaluate Gemma 270M settings on Arabic, Slovak, and Indonesian, which come from language families different from the Romance meta-training languages. The averaged curves with standard deviation bands show that MAML-RM continues to outperform the baseline across all three targets. This suggests that the benefit of meta-training is not limited to closely related Romance languages, but also transfers to linguistically distant target languages while preserving a similar adaptation trend.

Table 11 reports the corresponding policy evaluation results for these non-Romance languages. Overall, MAML-RLHF improves over both baselines on Arabic and Indonesian, with larger margins over multitask RLHF than those observed for the Romance targets. Similar to MAML-DPO on non-Romance languages, this comparison suggests that our approach yields more robust transfer across language families, whereas multitask appears more sensitive to the language composition during preference pretraining. The improvement is more modest in Slovak, where conventional RLHF already achieves a strong win rate, resulting limited room for further improvement through GRPO even when MAML-RLHF provides stronger reward-model adaptation.

Table 11: Win rate of Gemma3 270M adapted to 100 samples from each non-Romance target language in the RLHF pipeline.

Target	Algorithm	Samples	WR (%)
Arabic	MAML-RLHF	100	$\mathbf{76.90} \pm 6.29$
	Multitask-RLHF	100	59.30 ± 6.80
	RLHF	100	50.71 ± 4.35
Slovak	MAML-RLHF	100	$\mathbf{77.30} \pm 3.81$
	Multitask-RLHF	100	75.60 ± 0.92
	RLHF	100	76.22 ± 0.65
Indonesian	MAML-RLHF	100	$\mathbf{82.07} \pm 1.66$
	Multitask-RLHF	100	44.17 ± 0.42
	RLHF	100	48.67 ± 1.47

Sufficient adaptation data. Table 12 and Table 13 report the win-rate results for models fine-tuned using the RLHF pipeline using 12k/8k and 40k target language samples. With both model configurations, MAML-RLHF consistently outperforms conventional RLHF in all cases, showing the effectiveness of the meta-learned initialization for RLHF-based adaptation. With Gemma3 270M, MAML-RLHF achieves higher win rates than multitask RLHF on Catalan and Romanian under both data settings, while multitask RLHF performs better on French. This suggests that MAML-RLHF provides a stronger adaptation advantage in most target-language settings, although multitask training can remain competitive when the target language is closely aligned with the meta-training languages.

Table 12: Win rate evaluation of Gemma 270M models adapted with 12k and 40k samples from each target language under baseline, multi-task, and MAML RLHF settings.

Target	Algorithm	Samples	WR(%)
French	MAML-RLHF	12000	72.05
	Multitask-RLHF	12000	**73.45**
	RLHF	12000	69.25
	MAML-RLHF	40000	72.30
	Multitask-RLHF	40000	**74.90**
	RLHF	40000	64.90
Catalan	MAML-RLHF	12000	**63.00**
	Multitask-RLHF	12000	57.70
	RLHF	12000	50.30
	MAML-RLHF	40000	**69.65**
	Multitask-RLHF	40000	65.80
	RLHF	40000	58.20
Romanian	MAML-RLHF	12000	**87.25**
	Multitask-RLHF	12000	85.10
	RLHF	12000	76.15
	MAML-RLHF	40000	**88.95**
	Multitask-RLHF	40000	87.00
	RLHF	40000	82.60

Table 13: Win rate evaluation of BLOOM 7.1B models trained with RLHF pipelines. MAML-RLHF models achieve at least 9% in performance improvement over baseline models.

Target	Algorithm	Samples	WR(%)
French	MAML-RLHF	8000	**69.6**
	RLHF	8000	60.1
Catalan	MAML-RLHF	8000	**73.2**
	RLHF	8000	56.8
Romanian	MAML-RLHF	8000	**65.0**
	RLHF	8000	52.7

Out-of-distribution evaluation. We evaluate the OOD benchmark performances of Gemma 3 270M models trained with the RLHF pipeline in Table 14. While MAML-RLHF models show improvements in more benchmarks and metrics than the baseline RLHF models in French and Catalan, the same pattern is not present in Romanian. We attribute this to the diminishing advantage of MAML-RLHF as the amount of adaptation dataset becomes larger. We also note that compared to DPO models, RLHF models are more unstable, due to training with reward functions increasing the possibility of overfitting and losing generalization.

Table 14: Out-of-distribution benchmark evaluation results of Gemma 270M models under baseline and MAML-RLHF. Bold numbers indicate where the trained model outperformed the base model.

Language	Algorithm	Samples	ARC		Hellaswag		MMLU	
			Acc.	Norm. Acc.	Acc.	Norm. Acc.	Acc.	Norm. Acc.
French	Base model	-	0.210	0.249	0.271	0.261	0.251	0.266
	MAML-RLHF	40000	**0.213**	**0.250**	**0.273**	**0.291**	0.239	0.251
	RLHF	40000	0.210	0.248	0.270	**0.264**	**0.260**	0.258
Catalan	Base model	-	0.212	0.239	0.265	0.281	0.242	0.252
	MAML-RLHF	40000	**0.214**	**0.257**	**0.279**	**0.287**	**0.247**	**0.253**
	RLHF	40000	**0.230**	**0.271**	0.263	0.265	0.238	**0.268**
Romanian	Base model	-	0.191	0.240	0.283	0.276	0.260	0.268
	MAML-RLHF	40000	0.188	**0.244**	0.255	0.251	0.230	0.246
	RLHF	40000	**0.207**	**0.241**	0.271	0.263	0.241	0.247

Table 15: Resource consumption of different training stages based on Bloom 7.1B model. Note that the *Adaptation-DPO/RM* is also used for conventional DPO/RM. All the processes run with a single H200 GPU.

Method	Training steps	Samples	Wall-clock time	GPU memory consumption
MAML-DPO	400	12,000	23 hours	max. 137.2GB
Multitask-DPO	400	12,000	23 hours	max. 128.8GB
Adaptation-DPO	100	100	0.5 hour	max. 121.8GB
MAML-RM	400	16,000	10 hours	max. 135.8GB
Multitask-RM	400	16,000	8 hours	max. 131.6GB
Adaptation-RM	100	100	0.25 hour	max. 114.8GB

B.3 Resource Consumption of MAML-DPO

We further report the resource consumption of different approaches in Table 15. With the first-order approximation to the MAML objective, the additional cost of our MAML-based methods remains moderate compared with the corresponding multitask models trained on the same amount of multilingual preference data. Specifically, MAML-DPO mainly increases peak GPU memory usage, while MAML-RM incurs a small increase in both wall-clock time and memory consumption. These results suggest that the cost of our approaches remains manageable with respect to both training time and GPU memory requirements.

C Convergence of Meta-RLHF

C.1 Proof of Theorem 4.3

We first show that the task-specific loss function has Lipschitz-continuous gradient. Recall the definition of $\mathcal{L}(\phi;\tau)$ and its gradient $\nabla_\phi \mathcal{L}(\phi;\tau)$.

$$\begin{aligned}\mathcal{L}(\phi;\tau) &= -\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\left[\log\sigma\left(\phi^\top\Delta\psi(x,y_w,y_\ell)\right)\right]\\ \nabla_\phi\mathcal{L}(\phi;\tau) &= -\mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\left[\sigma(-\phi^\top\Delta\psi(x,y_w,y_\ell))\cdot\Delta\psi(x,y_w,y_\ell)\right]\end{aligned}$$

Now,

$$\begin{aligned}|\nabla_\phi\mathcal{L}(\phi;\tau)-\nabla_\phi\mathcal{L}(\phi';\tau)| &\le \mathop{\mathbb{E}}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\Big[\Big\|\sigma(-\phi^\top\Delta\psi(x,y_w,y_\ell))\cdot\Delta\psi(x,y_w,y_\ell)\\ &\qquad -\ \sigma(-\phi'^\top\Delta\psi(x,y_w,y_\ell))\cdot\Delta\psi(x,y_w,y_\ell)\Big\|\Big]\\ &\le \mathop{\mathbb{E}}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\left[\left|\sigma(-\phi^\top\Delta\psi(x,y_w,y_\ell))-\sigma(-\phi'^\top\Delta\psi(x,y_w,y_\ell))\right|\cdot\|\Delta\psi(x,y_w,y_\ell)\|\right]\\ &\le 2\mathop{\mathbb{E}}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\left[\left\|(\phi-\phi')^\top\Delta\psi(x,y_w,y_\ell))\right\|\right]\\ &\le 2\mathop{\mathbb{E}}_{(x,y_w,y_\ell)\sim\mathcal{D}^\tau}\left[\|\phi-\phi'\|\,\|\Delta\psi(x,y_w,y_\ell)\|\right]\\ &\le 4\|\phi-\phi'\|.\end{aligned}$$

Additionally, let $\widetilde{\nabla}_\phi\mathcal{L}(\phi;\tau)=\frac{1}{|\mathcal{D}^\tau_{\text{in}}|}\sum_{j\in\mathcal{D}^\tau_{\text{in}}}\nabla_\phi\mathcal{L}(\phi;\tau,j)$ be a stochastic gradient estimator for task τ. Since $\|\nabla_\phi\mathcal{L}(\phi;\tau)\|_2\le 2$, standard concentration inequality gives us,

$$\mathbb{E}_{\mathcal{D}^\tau_{\text{in}}}\left[\left\|\widetilde{\nabla}_\phi\mathcal{L}(\phi;\tau)-\nabla_\phi\mathcal{L}(\phi;\tau)\right\|_2^2\right]\le O\left(\frac{1}{|\mathcal{D}^\tau_{\text{in}}|}\right)\triangleq\widetilde{\sigma}^2.$$

We can now apply Theorem 4.15 from [17] to conclude that as long as the number of iterations $T\ge\Delta\min\left\{\frac{4}{\varepsilon^2},\frac{4}{\sigma^2(16\alpha^2+B^{-1})}+\frac{4(BD_o+D_{\text{in}})}{\widetilde{\sigma}^2}\right\}$ with $\Delta=\mathcal{L}_M(\phi_1)-\min_\phi\mathcal{L}_M(\phi)$, we have

$$\mathbb{E}\left[\left\|\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M)\right\|_2\right]\le O\left(\sqrt{\sigma^2\left(16\alpha^2+\frac{1}{B}\right)+\frac{\widetilde{\sigma}^2}{BD_o}+\frac{\widetilde{\sigma}^2}{D_{\text{in}}}}\right)+\varepsilon$$

Now we substitute the value of $\widetilde{\sigma}$ and choose $D_{\text{in}}=1/\sigma, D_o=O(1), B=1/\alpha^2$, and $\varepsilon=\alpha\sigma$. Then we obtain

$$\mathbb{E}\left[\left\|\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M)\right\|_2\right]\le O(\alpha\sigma).$$

Moreover, $\Delta=\mathcal{L}_M(\phi_1)-\min_\phi\mathcal{L}_M(\phi)\le O(L)$ as feature norms are bounded by L. Then the number of iterations T is bounded by $O\left(\frac{L}{\alpha^2\sigma^2}\right)$.

By lemma C.1, $\mathcal{L}_M(\phi)$ is a m_M-strongly convex function for $m_M=\frac{c\cdot m_0}{4+c+m_0}$.

$$\mathcal{L}_M(\phi^\star_M)\ge\mathcal{L}_M(\widehat{\phi}_M)+\left\langle\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M),\phi^\star_M-\widehat{\phi}_M\right\rangle+\frac{m_M}{2}\left\|\phi^\star_M-\widehat{\phi}_M\right\|_2^2$$

After rearranging and using Cauchy-Schwartz inequality, we obtain the following bound.

$$\begin{aligned}\frac{m_M}{2}\left\|\phi^\star_M-\widehat{\phi}_M\right\|_2^2 &\le \mathcal{L}_M(\phi^\star_M)-\mathcal{L}_M(\widehat{\phi}_M)+\left\|\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M)\right\|_2\left\|\phi^\star_M-\widehat{\phi}_M\right\|_2\\ &\le \left\|\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M)\right\|_2\left\|\phi^\star_M-\widehat{\phi}_M\right\|_2\end{aligned}$$

This implies $\mathbb{E}\left[\left\|\phi^\star_M-\widehat{\phi}_M\right\|_2\right]\le\mathbb{E}\left[\frac{2}{m_M}\left\|\nabla_\phi\mathcal{L}_M(\widehat{\phi}_M)\right\|_2\right]\le\frac{2\alpha\sigma}{m_M}$. □

C.2 Technical Lemmas

We dedicate this section to proving that the meta-objective $\mathcal{L}_M(\phi)$ is strongly convex under the conditions of Theorem 4.3. This property is formalized in the following lemma.

Lemma C.1. *Suppose Assumption 4.1 hold and* $m_0 = \frac{\nu}{2(1+\exp(2B_\phi))}$. *If* $\alpha \leq \alpha_0 = \frac{m_0}{2(4+m_0)}$, *then* $\nabla^2_\phi \mathcal{L}_M(\phi)$ *is positive semidefinite. Moreover, if* $\alpha \leq \frac{m_0}{2(4+c+m_0)}$ *for some constant* $c > 0$ *then* $\nabla^2_\phi \mathcal{L}_M(\phi) \succcurlyeq \frac{c \cdot m_0}{4+c+m_0} I$.

Proof. We first derive the third derivative of the task specific loss function. Moreover, we will also write $z = \phi^T \Delta\psi$.

$$\begin{aligned}\nabla^3_\phi \mathcal{L}(\phi;\tau) &= \nabla_\phi \left(\mathbb{E}_{(x,y_w,y_\ell)}[\sigma(z)(1-\sigma(z))\Delta\psi\Delta\psi^T]\right) \\ &= \mathbb{E}_{(x,y_w,y_\ell)}\left[\frac{d}{dz}(\sigma(z)(1-\sigma(z)))\nabla_\phi z \cdot \Delta\psi\Delta\psi^T\right]\end{aligned}$$

Since $\nabla_\phi z = \Delta\psi$ and $\frac{d}{dz}(\sigma(z)(1-\sigma(z))) = \sigma(z)(1-\sigma(z))(1-2\sigma(z))$, we have:

$$\nabla^3_\phi \mathcal{L}(\phi;\tau) = \mathbb{E}_{(x,y_w,y_\ell)}\left[\sigma(z)(1-\sigma(z))(1-2\sigma(z))\Delta\psi \otimes \Delta\psi\Delta\psi^T\right] \tag{15}$$

Where $\otimes$ represents the tensor product.

Let $\phi_\tau = \phi - \alpha\nabla_\phi \mathcal{L}(\phi;\tau)$ denote the adapted parameter for task τ. The gradient of the meta-loss is given as follows.

$$\nabla_\phi \mathcal{L}_M(\phi) = \mathbb{E}_{\tau\sim P_{\text{TASK}}}\left[(I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau))\nabla_\phi \mathcal{L}(\phi_\tau;\tau)\right] \tag{16}$$

Differentiating the gradient with respect to ϕ we get the Hessian of the meta-loss.

$$\begin{aligned}\nabla^2_\phi \mathcal{L}_M(\phi) &= \mathbb{E}_\tau\left[\nabla_\phi\left((I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau))\nabla_\phi \mathcal{L}(\phi_\tau;\tau)\right)\right] \\ &= \mathbb{E}_\tau\left[\nabla_\phi(I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla_\phi \mathcal{L}(\phi_\tau;\tau) + (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla_\phi(\nabla_\phi \mathcal{L}(\phi_\tau;\tau))\right] \\ &= \mathbb{E}_\tau\left[-\alpha\nabla_\phi(\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla_\phi \mathcal{L}(\phi_\tau;\tau) + (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla_\phi(\nabla_\phi \mathcal{L}(\phi_\tau;\tau))\right]\end{aligned}$$

The first term involves the third derivative tensor $\nabla^3_\phi \mathcal{L}(\phi;\tau)$, and for the second term, we calculate $\nabla_\phi(\nabla_\phi \mathcal{L}(\phi_\tau;\tau))$.

$$\begin{aligned}\nabla_\phi(\nabla_\phi \mathcal{L}(\phi_\tau;\tau)) &= \nabla^2_\phi \mathcal{L}(\phi_\tau;\tau) \cdot \nabla_\phi \phi_\tau && (17)\\ &= \nabla^2_\phi \mathcal{L}(\phi_\tau;\tau) \cdot \nabla_\phi(\phi - \alpha\nabla_\phi \mathcal{L}(\phi;\tau)) && (18)\\ &= \nabla^2_\phi \mathcal{L}(\phi_\tau;\tau) \cdot (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) && (19)\end{aligned}$$

Substituting back we obtain.

$$\begin{aligned}\nabla^2_\phi \mathcal{L}_M(\phi) = \mathbb{E}_\tau\big[&-\alpha\nabla^3_\phi \mathcal{L}(\phi;\tau) \cdot \nabla_\phi \mathcal{L}(\phi_\tau;\tau) \\ &+(I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla^2_\phi \mathcal{L}(\phi_\tau;\tau) \cdot (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau))\big]\end{aligned}$$

The following is the complete expression for the Hessian, which consists of two terms.

$$\nabla^2_\phi \mathcal{L}_M(\phi) = \mathbb{E}_\tau[T_1(\tau) + T_2(\tau)] \tag{20}$$

where

$$\begin{aligned}T_1(\tau) &= -\alpha\nabla^3_\phi \mathcal{L}(\phi;\tau) \cdot \nabla_\phi \mathcal{L}(\phi_\tau;\tau) && (21)\\ T_2(\tau) &= (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) \cdot \nabla^2_\phi \mathcal{L}(\phi_\tau;\tau) \cdot (I - \alpha\nabla^2_\phi \mathcal{L}(\phi;\tau)) && (22)\end{aligned}$$

Since $\nabla^2_\phi \mathcal{L}_M(\phi) = \mathbb{E}_\tau[T_1(\tau) + T_2(\tau)]$, for the Hessian to be positive semidefinite, we need.

$$\mathbb{E}[w^\top T_2(\tau) w] - |\mathbb{E}[w^\top T_1(\tau) w]| \geq 0$$

For any unit vector w, Lemma C.2 shows $|w^T T_1(\tau) w| \leq 8\alpha$, and $w^\top T_2(\tau) w \geq (1-\alpha)^2 m_0$.

Then the sufficient condition is the following.

$$(1-\alpha)^2 m_0 > 8\alpha. \tag{23}$$

Therefore, a sufficient condition is $\alpha < \frac{m_0}{2(4+m_0)}$.

In order to show positive definiteness, the required condition is

$$(1-\alpha)^2 m_0 > 8\alpha + 2c\alpha, \tag{24}$$

for some $c > 0$. The rest of the proof is almost identical to the positive-semidefinite setting. □

Lemma C.2. *Assume Assumption 4.1. Fix any task τ and stepsize $\alpha > 0$. Let $T_1(\tau)$ and $T_2(\tau)$ be defined as in the proof of Lemma C.1. Then,*

$$\|T_1(\tau)\|_2 \leq 8\alpha, \quad \textit{and}, \quad \lambda_{\min}(T_2(\tau)) \geq (1-\alpha)^2 m_0,$$

where $m_0 = \frac{\nu}{2(1+\exp(2B_\phi))}$.

Proof. Bound for $T_1(\tau)$. We first expand $T_1(\tau)$ as follows:

$$T_1(\tau) = -\alpha f(z)\,(\Delta\psi^\top v)\,\Delta\psi\Delta\psi^\top, \quad v \triangleq \nabla_\phi \mathcal{L}(\phi_\tau;\tau), \quad f(z) = \sigma(z)(1-\sigma(z))(1-2\sigma(z)), \quad z = \phi^\top \Delta\psi.$$

Using $\|\Delta\psi\Delta\psi^\top\|_F = \|\Delta\psi\|_2^2$, and Cauchy–Schwarz,

$$\|T_1(\tau)\|_F \leq \alpha |f(z)| |\Delta\psi^\top v| \|\Delta\psi\|_2^2 \leq \alpha |f(z)| \|\Delta\psi\|_2^3 \|v\|_2.$$

Moreover, from the gradient expression $\nabla_\phi \mathcal{L}(\cdot;\tau) = -\mathbb{E}[\sigma(-z)\Delta\psi]$, we have

$$\|v\|_2 = \|\nabla_\phi L(\phi_\tau;\tau)\|_2 \leq \mathbb{E}\|\Delta\psi\|_2 \leq 2.$$

Finally, $|f(z)| \leq 1/2$. Therefore,

$$\|T_1(\tau)\|_2 \leq \|T_1(\tau)\|_F \leq \alpha \cdot \frac{1}{2} \cdot 2^3 \cdot 2 = 8\alpha,$$

which completes the proof.

Bound for $T_2(\tau)$. By Lemma C.3, we have that $\nabla_\phi^2 \mathcal{L}(\phi;\tau) \succcurlyeq m_0$ and $\nabla_\phi^2 \mathcal{L}(\phi;\tau) \preccurlyeq I$. For any unit w let $u = (I - \alpha\nabla_\phi^2 \mathcal{L}(\phi;\tau))w$, we have that $\|u\|_2^2 \leq (1-\alpha)^2$. Then

$$w^\top T_2(\tau) w = u^\top \nabla_\phi^2 \mathcal{L}(\phi_\tau;\tau) u \geq m_0 \|u\|_2^2 \geq m_0(1-\alpha)^2.$$

In other words, $\|T_2(\tau)\|_2 \geq m_0(1-\alpha)^2$ as claimed. □

Lemma C.3. *For any task τ, the loss function $\mathcal{L}(\phi;\tau)$ is positive semidefinite i.e. $\nabla_\phi^2 \mathcal{L}(\phi;\tau) \succcurlyeq 0$. Additionally, if Assumption 4.1 hold, then it is positive definite i.e.*

$$\nabla_\phi^2 \mathcal{L}(\phi)\|_2 \preccurlyeq I \quad \textit{and} \quad \nabla_\phi^2 \mathcal{L}(\phi;\tau) \succcurlyeq \frac{\nu}{2(1+\exp(2B_\phi))} I,$$

for any ϕ with $\|\phi\|_2 \leq B_\phi$.

Proof. First, as $\|\Delta\psi(\cdot)\|_2^2 \leq 2$, one can show that $\|\nabla_\phi^2 \mathcal{L}(\phi)\|_2 \leq 1$.

Since $\sigma(u) \geq 0\ \forall u$, the multiplication of two sigmoid functions satisfy $\sigma(u)\sigma(-u) \geq 0$. The matrix $\psi\psi^\top$ is positive semidefinite. Hence, $\nabla_\phi^2 \mathcal{L}(\phi;\tau) \succcurlyeq 0$.

Additionally, for any u, $\sigma(u)\sigma(-u) = \frac{1}{1+\exp(u)}\frac{1}{1+\exp(-u)} = \frac{1}{2+\exp(u)+\exp(-u)} \geq \frac{1}{2+2\exp(|u|)}$. Moreover, $|z_\phi(x, y_w, y_\ell)| = \left|\phi^\top(\psi(x,y_w) - \psi(x,y_\ell)\right| \leq 2B_\phi$. Moreover, under assumption 4.1 the matrix $\psi\psi^\top$ is positive-definite with constant ν. Therefore, $\nabla_\phi^2 \mathcal{L}(\phi;\tau) \succcurlyeq \frac{\nu}{2(1+\exp(2B_\phi))} I$. □

D Convergence of Meta-DPO

D.1 Proof of Theorem 4.6

The key step of the proof of Theorem 4.6 is similar to that of Theorem 4.3.

By Lemma D.3, for all τ, we have that the smoothness constant of task-wise gradient as

$$\|\nabla_\theta \mathcal{L}(\theta, \tau) - \nabla_\theta \mathcal{L}(\theta', \tau)\|_2 \leq \beta^2 \|\theta - \theta'\|_2. \tag{25}$$

Moreover, by Proposition D.4, we have that $\|\nabla_\theta \mathcal{L}(\theta, \tau)\|_2 \leq 2\beta$.

Additionally, let $\widetilde{\nabla}_\phi \mathcal{L}(\theta; \tau) = \frac{1}{|\mathcal{D}_{\text{in}}^\tau|} \sum_{j \in \mathcal{D}_{\text{in}}^\tau} \nabla_\theta \mathcal{L}(\theta; \tau, j)$ be a stochastic gradient estimator for task τ. Since by Proposition D.4, we have that $\|\nabla_\theta \mathcal{L}(\theta, \tau)\|_2 \leq 2\beta$, standard concentration inequality gives us,

$$\mathbb{E}_{\mathcal{D}_{\text{in}}^\tau}\left[\left\|\widetilde{\nabla}_\theta \mathcal{L}(\theta; \tau) - \nabla_\theta \mathcal{L}(\theta; \tau)\right\|_2^2\right] \leq O\left(\frac{\beta^2}{|\mathcal{D}_{\text{in}}^\tau|}\right) \triangleq \widetilde{\sigma}^2.$$

We can now apply Theorem 4.15 from [17] to conclude that as long as the number of iterations $T \geq \Delta \min\left\{\frac{4}{\varepsilon^2}, \frac{4}{\sigma^2(16\alpha^2 + B^{-1})} + \frac{4(BD_o + D_{\text{in}})}{\widetilde{\sigma}^2}\right\}$ with $\Delta = \mathcal{L}_M(\theta_1) - \min_\theta \mathcal{L}_M(\theta)$, we have

$$\mathbb{E}\left[\left\|\nabla_\theta \mathcal{L}_M(\widehat{\theta}_M)\right\|_2\right] \leq O\left(\sqrt{\sigma^2\left(16\alpha^2 + \frac{1}{B}\right) + \frac{\widetilde{\sigma}^2}{BD_o} + \frac{\widetilde{\sigma}^2}{D_{\text{in}}}}\right) + \varepsilon.$$

Now we substitute the value of $\widetilde{\sigma}$ and choose $D_{\text{in}} = \beta/\sigma, D_o = O(1), B = 1/\alpha^2$, and $\varepsilon = \alpha\sigma$. Then we obtain

$$\mathbb{E}\left[\left\|\nabla_\theta \mathcal{L}_M(\widehat{\theta}_M)\right\|_2\right] \leq O(\alpha\sigma).$$

Moreover, $\Delta = \mathcal{L}_M(\theta_1) - \min_\theta \mathcal{L}_M(\theta) \leq O(L)$ as feature norms are bounded by L. Then the number of iterations T is bounded by $O\left(\frac{L}{\alpha^2\sigma^2}\right)$.

By Theorem D.1, $\mathcal{L}_M(\theta)$ is m_M-strongly convex with $m_M = (1 - \alpha\beta^2)^2 \mu_\tau - 4\alpha\beta^4$, where μ_τ is the per-task strong-convexity constant of Lemma D.3. Standard calculation gives us that, with $\alpha = \frac{1}{\beta^2\left(2 + \frac{8(1+\exp(Z_{\max}))}{\nu}\right)}$, we have $m_M \geq \beta^2 \frac{c\nu^3}{(1+\exp(Z_{\max}))^3}$ for some constant $c > 0$. We have the following

$$\mathcal{L}_M(\theta_M^\star) \geq \mathcal{L}_M(\widehat{\theta}_M) + \left\langle \nabla_\theta \mathcal{L}_M(\widehat{\theta}_M), \theta_M^\star - \widehat{\theta}_M \right\rangle + \frac{m_M}{2}\left\|\theta_M^\star - \widehat{\theta}_M\right\|_2^2.$$

After rearranging and using Cauchy-Schwartz inequality, we obtain the following bound.

$$\begin{aligned}\frac{m_M}{2}\left\|\theta_M^\star - \widehat{\theta}_M\right\|_2^2 &\leq \mathcal{L}_M(\theta_M^\star) - \mathcal{L}_M(\widehat{\theta}_M) + \left\|\nabla_\theta \mathcal{L}_M(\widehat{\theta}_M)\right\|_2 \left\|\theta_M^\star - \widehat{\theta}_M\right\|_2 \\ &\leq \left\|\nabla_\theta \mathcal{L}_M(\widehat{\theta}_M)\right\|_2 \left\|\theta_M^\star - \widehat{\theta}_M\right\|_2.\end{aligned}$$

This implies $\mathbb{E}\left[\left\|\theta_M^\star - \widehat{\theta}_M\right\|_2\right] \leq \frac{2}{m_M}\mathbb{E}\left[\left\|\nabla_\theta \mathcal{L}_M(\widehat{\theta}_M)\right\|_2\right] \leq \frac{2\alpha\sigma}{m_M}$. □

D.2 Technical Lemmas

We first show the the meta loss function $\mathcal{L}_M(\theta)$ is strongly convex under condition of Theorem 4.6.

Theorem D.1 (Strong convexity of DPO loss)**.** *Let μ_τ be the per-task strong-convexity constant of Lemma D.3. Define*

$$m_M \triangleq (1 - \alpha\beta^2)^2 \mu_\tau - 4\alpha\beta^4. \tag{26}$$

If $m_M > 0$, then

$$\nabla_\theta^2 \mathcal{L}_M(\theta) \succcurlyeq m_M I.$$

Proof. Let $\theta_\tau = \theta - \alpha\nabla_\theta\mathcal{L}(\theta,\tau)$. Then for any unit w,

$$\begin{aligned}\nabla_\theta\mathcal{L}_M(\theta) &= \mathbb{E}_\tau\big[(I-\alpha\nabla^2_\theta\mathcal{L}(\theta,\tau))\cdot\nabla_\theta\mathcal{L}(\theta_\tau,\tau)\big],\\ \nabla^2_\theta\mathcal{L}_M(\theta) &= \mathbb{E}_\tau[T_1(\tau)+T_2(\tau)],\end{aligned} \tag{27}$$

where

$$T_1(\tau) = -\alpha\nabla^3_\theta\mathcal{L}(\theta;\tau)\cdot\nabla_\theta\mathcal{L}(\theta_\tau;\tau), \tag{28}$$

$$T_2(\tau) = (I-\alpha\nabla^2_\theta\mathcal{L}(\theta;\tau))\cdot\nabla^2_\theta\mathcal{L}(\theta_\tau;\tau)\cdot(I-\alpha\nabla^2_\theta\mathcal{L}(\theta;\tau)). \tag{29}$$

Now, we have that

$$w^\top\nabla^2\mathcal{L}_M(\theta)w = \mathbb{E}_\tau\big[w^\top(T_1(\tau)+T_2(\tau))w\big] \geq \mathbb{E}_\tau[w^\top T_2(\tau)w] - \mathbb{E}_\tau|w^\top T_1(\tau)w|.$$

Apply Lemma D.2 to obtain

$$w^\top\nabla^2\mathcal{L}_M(\theta)w \geq (1-\alpha\beta^2)^2\mu_\tau - 4\alpha\beta^4 = m_M.$$

Thus $\nabla^2\mathcal{L}_M(\theta) \succcurlyeq m_M I$, i.e. $\mathcal{L}_M(\theta)$ is m_M-strongly convex. □

Lemma D.2 (Bounds on the meta-Hessian terms)**.** *Under Assumption 4.1, we have that*

$$|w^\top T_1(\tau)w| \leq 4\alpha\beta^4, \qquad w^\top T_2(\tau)w \geq (1-\alpha\beta^2)^2\,\mu_\tau.$$

Proof. For T_2, by Lemma D.3 we have that $\nabla^2_\theta\mathcal{L}(\theta,\tau) \preccurlyeq L_\tau I = \beta^2 I$ and $\nabla^2_\theta\mathcal{L}(\theta_\tau,\tau) \succcurlyeq \mu_\tau I$; thus,

$$w^\top T_2(\tau)w = \big((I-\alpha\nabla^2_\theta\mathcal{L}(\theta,\tau))w\big)^\top\nabla^2_\theta\mathcal{L}(\theta_\tau,\tau)\big((I-\alpha\nabla^2_\theta\mathcal{L}(\theta,\tau))w\big) \geq \mu_\tau\|(I-\alpha\nabla^2_\theta\mathcal{L}(\theta,\tau))w\|^2 \geq (1-\alpha\beta^2)^2\mu_\tau.$$

For T_1, by Cauchy-Schwarz for tensor contraction,

$$|w^\top T_1(\tau)w| \leq \alpha\,\|\nabla^3_\theta L(\theta;\tau)\|_{\mathrm{op}}\,\|\nabla_\theta\mathcal{L}(\theta_\tau;\tau)\|\,\|w\|^2.$$

Apply Lemma D.4: $\|\nabla^3\mathcal{L}(\cdot)\|_{\mathrm{op}} \leq 2\beta^3$ and $\|\nabla\mathcal{L}(\theta_\tau;\tau)\| \leq 2\beta$, hence $|w^\top T_1(\tau)w| \leq 4\alpha\beta^4$. □

Lemma D.3 (Task-wise smoothness and strong convexity)**.** *Define the task-wise smoothness and strong-convexity constants as*

$$L_\tau \triangleq \beta^2, \qquad \mu_\tau \triangleq \frac{\beta^2\,\nu}{2\big(1+e^{Z_{\max}}\big)}.$$

Then $\nabla^2_\theta\mathcal{L}(\theta;\tau) \preccurlyeq L_\tau I$, and under Assumption 4.4 (so $|z_\theta| \leq Z_{\max}$) and Assumption 4.1 we also have $\nabla^2_\theta\mathcal{L}(\theta;\tau) \succcurlyeq \mu_\tau I$.

Proof. Upper bound. Note that $\sigma(u)\sigma(-u) \leq 1/4$,

$$\|\nabla^2\mathcal{L}(\theta;\tau)\|_2 = \beta^2\,\mathbb{E}\big[\sigma(z_\theta)\sigma(-z_\theta)\Delta\psi\Delta\psi^\top\big] \leq \beta^2\,\mathbb{E}\big[\sigma(z_\theta)\sigma(-z_\theta)\,\|\Delta\psi\Delta\psi^\top\|\big] \leq \beta^2\cdot\frac{1}{4}\cdot 4 = \beta^2.$$

Lower bound. For any scalar u,

$$\sigma(u)\sigma(-u) = \frac{1}{2+e^u+e^{-u}} \geq \frac{1}{2+2e^{|u|}} = \frac{1}{2(1+e^{|u|})}.$$

Using (14), $|z_\theta| \leq Z_{\max}$, hence $\sigma(z_\theta)\sigma(-z_\theta) \geq \frac{1}{2(1+e^{Z_{\max}})}$. Therefore

$$\nabla^2\mathcal{L}(\theta;\tau) \succcurlyeq \beta^2\cdot\frac{1}{2(1+e^{Z_{\max}})}\,\mathbb{E}[\Delta\psi\Delta\psi^\top] \succcurlyeq \beta^2\cdot\frac{1}{2(1+e^{Z_{\max}})}\,\nu I,$$

where the last step uses Assumption 4.1. □

Lemma D.4 (Uniform bound on the third derivative of the task loss)**.** *For all τ,*

$$\|\nabla^3_\theta\mathcal{L}(\theta;\tau)\|_{\mathrm{op}} \leq 2\beta^3,$$
$$\|\nabla_\theta\mathcal{L}(\theta;\tau)\|_2 \leq 2\beta.$$

Proof. For the scalar loss $\ell(z) = \log(1+e^{-z})$, we have $\ell'''(z) = -\sigma(z)\sigma(-z)(1-2\sigma(z))$ so $|\ell'''(z)| \leq 1/4$. For a single sample, $z_\theta = \beta\,\theta^\top\Delta\psi - J$ is affine in θ with $\nabla_\theta z_\theta = \beta\Delta\psi$. Hence, for unit vectors a, b, c,

$$|\nabla^3_\theta\ell(z_\theta)[a,b,c]| = |\ell'''(z_\theta)|\cdot|\langle\beta\Delta\psi,a\rangle|\cdot|\langle\beta\Delta\psi,b\rangle|\cdot|\langle\beta\Delta\psi,c\rangle| \leq \frac{1}{4}\,\beta^3\,\|\Delta\psi\|^3 \leq \frac{1}{4}\,\beta^3\cdot 8 = 2\beta^3,$$

using $\|\Delta\psi\| \leq 2$. Taking expectation over $\mathcal{D}^\tau$ preserves the bound.

For the gradient, $\nabla_\theta\mathcal{L}(\theta;\tau) = -\mathbb{E}[\beta\sigma(-z_\theta)\Delta\psi]$, so $\|\nabla_\theta\mathcal{L}(\theta;\tau)\| \leq \beta\,\mathbb{E}\|\Delta\psi\| \leq 2\beta$. □

E Convergence Analysis of the Adaptation Phase for MAML-DPO

In this section, we provide a comparative convergence analysis of the adaptation phase. We prove that the meta-learned initialization $\hat{\theta}_M$ yields a faster expected convergence rate to a target task's optimal policy θ^*_τ compared to a baseline multitask initialization θ_{base}.

E.1 Preliminaries and Task Geometry

Let the target task $\tau \sim P_{\text{TASK}}$ be drawn from the distribution of languages. During adaptation, we optimize the empirical DPO loss using gradient descent with step size η:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}^{\text{DPO}}(\theta_t; \tau) \tag{30}$$

From Lemma D.3, the task-wise DPO loss $\mathcal{L}^{\text{DPO}}(\theta; \tau)$ is L_τ-smooth and μ_τ-strongly convex.

We define the baseline initialization as the minimizer of the joint expected risk across all tasks:

$$\theta^*_{\text{base}} = \arg\min_\theta \mathbb{E}_{\tau \sim P_{\text{TASK}}}[\mathcal{L}^{\text{DPO}}(\theta; \tau)] \tag{31}$$

We define the optimal MAML initialization as the minimizer of the meta-objective:

$$\theta^*_M = \arg\min_\theta \mathbb{E}_{\tau \sim P_{\text{TASK}}}[\mathcal{L}^{\text{DPO}}(\theta - \alpha \nabla_\theta \mathcal{L}^{\text{DPO}}(\theta; \tau); \tau)] \tag{32}$$

By Theorem 4.6, our gradient-based algorithm produces $\hat{\theta}_M$ such that $||\theta^*_M - \hat{\theta}_M||_2 \leq \epsilon_{meta}$.

E.2 Deriving the Initialization Advantage

To prove faster convergence, we must first bound the expected squared distance of both initializations to the task-specific optimum $\theta^*_\tau = \arg\min_\theta \mathcal{L}^{\text{DPO}}(\theta; \tau)$.

Lemma E.1 (Baseline Initialization Distance)**.** *Let Assumption 4.5 hold, such that the variance of task gradients is bounded by σ^2. The expected distance from the baseline initialization θ^*_{base} to a sampled task optimum θ^*_τ is bounded by:*

$$\mathbb{E}_\tau[||\theta^*_{\text{base}} - \theta^*_\tau||_2^2] \leq \frac{\sigma^2}{\mu_\tau^2} \triangleq D^2_{\text{base}} \tag{33}$$

Proof. By the strong convexity of the task-specific loss $\mathcal{L}^{\text{DPO}}(\theta; \tau)$, the gradient is co-coercive. For any θ, the following inequality holds:

$$\langle \nabla \mathcal{L}^{\text{DPO}}(\theta; \tau) - \nabla \mathcal{L}^{\text{DPO}}(\theta^*_\tau; \tau), \theta - \theta^*_\tau \rangle \geq \mu_\tau ||\theta - \theta^*_\tau||_2^2$$

Because θ^*_τ is the exact minimizer for task τ, its gradient is zero ($\nabla \mathcal{L}^{\text{DPO}}(\theta^*_\tau; \tau) = 0$). Evaluating this at the baseline initialization θ^*_{base} yields:

$$\langle \nabla \mathcal{L}^{\text{DPO}}(\theta^*_{\text{base}}; \tau), \theta^*_{\text{base}} - \theta^*_\tau \rangle \geq \mu_\tau ||\theta^*_{\text{base}} - \theta^*_\tau||_2^2$$

Applying the Cauchy-Schwarz inequality to the left side:

$$||\nabla \mathcal{L}^{\text{DPO}}(\theta^*_{\text{base}}; \tau)||_2 ||\theta^*_{\text{base}} - \theta^*_\tau||_2 \geq \mu_\tau ||\theta^*_{\text{base}} - \theta^*_\tau||_2^2$$

Assuming $\theta^*_{\text{base}} \neq \theta^*_\tau$, we divide by $\mu_\tau ||\theta^*_{\text{base}} - \theta^*_\tau||_2$ to isolate the distance:

$$||\theta^*_{\text{base}} - \theta^*_\tau||_2 \leq \frac{1}{\mu_\tau} ||\nabla \mathcal{L}^{\text{DPO}}(\theta^*_{\text{base}}; \tau)||_2 \tag{34}$$

Squaring both sides and taking the expectation over the task distribution P_{TASK}:

$$\mathbb{E}_\tau[||\theta^*_{\text{base}} - \theta^*_\tau||_2^2] \leq \frac{1}{\mu_\tau^2} \mathbb{E}_\tau[||\nabla \mathcal{L}^{\text{DPO}}(\theta^*_{\text{base}}; \tau)||_2^2] \tag{35}$$

By definition, θ^*_{base} minimizes the expected joint loss, meaning the expected gradient across tasks is zero ($\mathbb{E}_\tau[\nabla \mathcal{L}^{\text{DPO}}(\theta^*_{\text{base}}; \tau)] = 0$). Therefore, the uncentered second moment is exactly equal to the variance of the task gradients, which is bounded by σ^2 (Assumption 4.5):

$$\mathbb{E}_\tau[||\theta^*_{\text{base}} - \theta^*_\tau||_2^2] \leq \frac{\sigma^2}{\mu_\tau^2} \triangleq D^2_{\text{base}}$$

□

Lemma E.2 (MAML Initialization Advantage)**.** *Let us define* $\overline{\theta}^* = \mathbb{E}_\tau[\theta_\tau^*]$, $G(\theta) = \mathbb{E}_\tau\left[\|\nabla\mathcal{L}(\theta;\tau)\|_2^2\right]$ *and* $H_{\text{base}} = \mathbb{E}_\tau[\nabla^2\mathcal{L}(\theta_{\text{base}}^*;\tau)]$. *There exists a deterministic vector* $\delta \in \mathbb{R}^d$ *and a scalar* $\Phi \in \mathbb{R}$*, both independent of* α*, such that for all sufficiently small* $\alpha > 0$*,*

$$\mathbb{E}_\tau\left\|\theta_M^* - \theta_\tau^*\right\|_2^2 = \mathbb{E}_\tau\left\|\theta_{\text{base}}^* - \theta_\tau^*\right\|_2^2 - 2\alpha\,\Phi + O(\alpha^2), \tag{36}$$

where

$$\delta = H_{\text{base}}^{-1}\,\nabla G(\theta_{\text{base}}^*), \qquad \Phi = -\left\langle \theta_{\text{base}}^* - \overline{\theta}^*,\, H_{\text{base}}^{-1}\,\nabla G(\theta_{\text{base}}^*)\right\rangle. \tag{37}$$

In particular, combining (36) with the variance bound $\mathbb{E}_\tau\|\theta_{\text{base}}^* - \theta_\tau^*\|_2^2 \le D_{\text{base}}^2 = \sigma^2/\mu_\tau^2$ *from Assumption 4.5 yields the inequality* $\mathbb{E}_\tau\|\theta_M^* - \theta_\tau^*\|_2^2 \le D_{\text{base}}^2 - 2\alpha\Phi + O(\alpha^2)$.

Proof. **Step 1: First-order expansion of the meta objective.**

By the integral form of Taylor's theorem applied to $s \mapsto \mathcal{L}\big(\theta - s\nabla\mathcal{L}(\theta;\tau);\tau\big)$ on $[0,\alpha]$,

$$\mathcal{L}\big(\theta - \alpha\nabla\mathcal{L}(\theta;\tau);\tau\big) = \mathcal{L}(\theta;\tau) - \alpha\,\|\nabla\mathcal{L}(\theta;\tau)\|_2^2 + \alpha^2\int_0^1 (1-s)\,\nabla\mathcal{L}(\theta;\tau)^\top\,\nabla^2\mathcal{L}\big(\theta - s\alpha\nabla\mathcal{L}(\theta;\tau);\tau\big)\,\nabla\mathcal{L}(\theta;\tau)\,ds. \tag{38}$$

By Lemma D.3, $\left\|\nabla^2\mathcal{L}(\,\cdot\,;\tau)\right\|_{\text{op}} \le \beta^2$, and by Lemma D.4 the paper, $\|\nabla\mathcal{L}(\theta;\tau)\|_2 \le 2\beta$ uniformly in τ on the relevant region. Hence the remainder in (38) is $O(\alpha^2)$ uniformly in τ. Taking expectation with respect to τ,

$$\mathcal{L}_{\text{MAML}}(\theta) = \mathbb{E}_\tau\left[\mathcal{L}(\theta;\tau)\right] - \alpha\,G(\theta) + R(\theta;\alpha), \qquad \sup_{\theta\in\mathcal{B}}|R(\theta;\alpha)| = O(\alpha^2). \tag{39}$$

Differentiating once more under the expectation (the integrand and its θ-derivative are bounded uniformly in τ),

$$\nabla\mathcal{L}_{\text{MAML}}(\theta) = \mathbb{E}_\tau\left[\nabla\mathcal{L}(\theta;\tau)\right] - \alpha\,\nabla G(\theta) + r(\theta;\alpha), \qquad \sup_{\theta\in\mathcal{B}}\|r(\theta;\alpha)\|_2 = O(\alpha^2). \tag{40}$$

Note that when $\alpha = 0$, $\nabla\mathcal{L}_{\text{MAML}}(\theta_{\text{base}}^*) = 0$ as θ_{base}^* minimizes $\mathbb{E}_\tau[\nabla\mathcal{L}(\theta;\tau)]$.

Step 2: Implicit-function expansion of $\theta_M^*(\alpha)$**.**

By Theorem D.1, $\mathcal{L}_{\text{MAML}}$ is m_M-strongly convex for α small, so $\theta_M^*(\alpha)$ is well-defined and unique. The map $F\colon(\theta,\alpha)\mapsto\nabla\mathcal{L}_{\text{MAML}}(\theta)$ is C^1 on a neighbourhood of θ_{base}^* for small α, and $\partial_\theta F(\theta_{\text{base}}^*,0) = H_{\text{base}} \succeq \mu I$ is invertible. The implicit function theorem yields a C^1 branch $\alpha\mapsto\theta_M^*(\alpha)$ with $\theta_M^*(0) = \theta_{\text{base}}^*$.

Write $\theta_M^*(\alpha) = \theta_{\text{base}}^* + \alpha\,\delta + O(\alpha^2)$. Inserting into the stationarity condition $\nabla\mathcal{L}_{\text{MAML}}(\theta_M^*(\alpha)) = 0$, expanding $\nabla F_{\text{base}}(\theta_M^*(\alpha)) = H_{\text{base}}\,(\alpha\delta) + O(\alpha^2)$, and using (40):

$$0 = \alpha H_{\text{base}}\,\delta - \alpha\,\nabla G(\theta_{\text{base}}^*) + O(\alpha^2).$$

Dividing by α and sending $\alpha\to 0$,

$$\delta = H_{\text{base}}^{-1}\,\nabla G(\theta_{\text{base}}^*). \tag{41}$$

Crucially, δ is a single deterministic vector, independent of τ.

Step 3: Expected squared distance to task optima.

Expanding the squared norm,

$$\|\theta_M^*(\alpha) - \theta_\tau^*\|_2^2 = \|\theta_{\text{base}}^* - \theta_\tau^*\|_2^2 + 2\alpha\,\langle\theta_{\text{base}}^* - \theta_\tau^*,\,\delta\rangle + O(\alpha^2),$$

where the $O(\alpha^2)$ remainder is uniform in τ since δ is deterministic. Taking $\mathbb{E}_\tau$ and pulling δ out of the expectation,

$$\begin{aligned}\mathbb{E}_\tau\left\|\theta_M^*(\alpha) - \theta_\tau^*\right\|_2^2 &= D_{\text{base}}^2 + 2\alpha\left\langle\theta_{\text{base}}^* - \overline{\theta}^*,\,\delta\right\rangle + O(\alpha^2)\\ &= D_{\text{base}}^2 - 2\alpha\,\Phi + O(\alpha^2),\end{aligned} \tag{42}$$

with $\Phi = -\left\langle\theta_{\text{base}}^* - \overline{\theta}^*,\,\delta\right\rangle = -\left\langle\theta_{\text{base}}^* - \overline{\theta}^*,\,H_{\text{base}}^{-1}\nabla G(\theta_{\text{base}}^*)\right\rangle$. This establishes (36)–(37). □

E.3 Comparative Convergence Rate

Theorem E.3 (Accelerated Adaptation via MAML-DPO)**.** *Let $\eta \leq 1/L_\tau$ and let $\epsilon > 0$ be the target accuracy threshold. Under Assumptions 4.1, 4.4, 4.5, the expected number of gradient steps t required to reach $\mathbb{E}[||\theta_t - \theta^*_\tau||_2^2] \leq \epsilon^2$ for the baseline (t_{base}) and MAML-DPO (t_{MAML}) pipelines satisfies:*

$$\mathbb{E}[t_{\text{base}}] - \mathbb{E}[t_{\text{MAML}}] \geq \frac{1}{\eta\mu_\tau} \ln\left(\frac{D^2_{\text{base}}}{D^2_{\text{base}} - 2\alpha\Phi + c_1 \cdot \alpha^2}\right). \tag{43}$$

for a universal constant $c_1 > 0$.

Proof. For a μ_τ-strongly convex and L_τ-smooth function optimized via gradient descent, the distance to the optimum contracts geometrically:

$$||\theta_t - \theta^*_\tau||_2^2 \leq (1 - \eta\mu_\tau)^t ||\theta_{init} - \theta^*_\tau||_2^2 \tag{44}$$

Taking expectations over the task distribution and solving for the minimum steps t to achieve ϵ^2 error:

$$t \geq \frac{1}{\eta\mu_\tau} \ln\left(\frac{\mathbb{E}_\tau[||\theta_{init} - \theta^*_\tau||_2^2]}{\epsilon^2}\right) \tag{45}$$

For the baseline pipeline, substituting $\theta_{init} = \theta^*_{\text{base}}$ yields:

$$t_{base} \geq \frac{1}{\eta\mu_\tau} \ln\left(\frac{D^2_{\text{base}}}{\epsilon^2}\right) \tag{46}$$

For the MAML-DPO pipeline, initializing at $\hat{\theta}_M$, we bound the distance using the triangle inequality and Theorem 4.6.

$$\mathbb{E}_\tau[||\hat{\theta}_M - \theta^*_\tau||_2^2] \leq D^2_{\text{MAML}} + \epsilon^2_{\text{meta}} \tag{47}$$

$$t_{\text{MAML}} \geq \frac{1}{\eta\mu_\tau} \ln\left(\frac{D^2_{\text{MAML}} + \epsilon^2_{\text{meta}}}{\epsilon^2}\right) \tag{48}$$

Taking the difference yields:

$$t_{\text{base}} - t_{\text{MAML}} \geq \frac{1}{\eta\mu_\tau} \ln\left(\frac{D^2_{\text{base}}}{D^2_{\text{MAML}} + \epsilon^2_{\text{meta}}}\right) \tag{49}$$

By Lemma E.2 we have $D^2_{\text{MAML}} = D^2_{\text{base}} - 2\alpha\Phi + O(\alpha^2)$ and by Theorem 4.6 we have, $\epsilon_{\text{meta}} = \frac{2\alpha\sigma}{m_M}$. Substituting these values we obtain the desired bound. □

Remark: If $\Phi > 0$ and α is sufficiently small, then $2\alpha\Phi - c_1 \cdot \alpha^2 < 0$ and we have $\mathbb{E}[t_{\text{base}}] - \mathbb{E}[t_{\text{MAML}}] > 0$. Recall that $\Phi = -\left\langle \theta^*_{\text{base}} - \overline{\theta}^*, \delta \right\rangle$ where δ is the MAML update direction, θ^*_{base} is the curvature-weighted centroid of the task optima, and $\overline{\theta}^* = \mathbb{E}[\theta^*_\tau]$ is the unweighted centroid. MAML update usually corrects for curvature-weighting bias and tends to push θ^*_M away from θ^*_{base} and towards solutions with large task-specific gradient. This means we should expect the MAML update δ to align with the direction $\overline{\theta}^* - \theta^*_{\text{base}}$, which makes $\Phi > 0$.

F Comparative Convergence Analysis: MAML-DPO vs. Baseline-DPO

The analysis in Appendix E compares MAML-DPO with multitask-DPO, which is recovered from Algorithm 1 by setting the inner-loop learning rate $\alpha = 0$. Multitask-DPO is a strong reference because it consumes the *same* multilingual preference data as MAML-DPO; consequently, the rigorous gap is at most $O(\alpha)$, in agreement with prior MAML theory in the mixed-linear-regression setting [61]. In our experiments, however, the most natural and most informative comparison is against *baseline-DPO*: a standard DPO pipeline that does not see any preference data outside the target language and is initialized solely from the target-language SFT model. We now provide a dedicated convergence analysis for this comparison. The resulting bound captures the benefit of *any* cross-lingual preprocessing—it applies equally to MAML-DPO and to multitask-DPO over baseline-DPO—and is controlled by the SFT preference-loss gap R_{SFT}, a quantity directly measurable from data and reflecting how far the SFT-initialized policy is from preference-optimal on the target task. The MAML-specific refinement over multitask-DPO is then the additional $O(\alpha)$ term established in Appendix E; the two analyses are complementary.

F.1 Setup and Baseline-DPO Initialization

Baseline-DPO. Let $\pi^{\mathrm{SFT}}_{\tau_{\mathrm{new}}}$ be the policy obtained by running supervised fine-tuning on target-language data only, and let $\theta_0^{\mathrm{BL}} \in \mathbb{R}^d$ be its log-linear parameter, i.e. $\pi_{\theta_0^{\mathrm{BL}}} = \pi^{\mathrm{SFT}}_{\tau_{\mathrm{new}}}$. Baseline-DPO performs gradient descent on the empirical DPO loss for τ_{new} starting from θ_0^{BL}:

$$\theta_{t+1}^{\mathrm{BL}} = \theta_t^{\mathrm{BL}} - \eta\, \nabla_\theta \widehat{\mathcal{L}}^{\mathrm{DPO}}(\theta_t^{\mathrm{BL}}; \tau_{\mathrm{new}}), \qquad \theta_0^{\mathrm{BL}} = \pi^{\mathrm{SFT}}_{\tau_{\mathrm{new}}}\text{-parameter}. \tag{50}$$

Crucially, θ_0^{BL} is a function only of target-language SFT data and is independent of the preference data $\bigcup_\tau \mathcal{D}^\tau$ used for meta-training. By the constraint $\|\theta\|_2 \le B_\theta$ defining the log-linear policy class Π in (12), both θ_0^{BL} and every task optimum $\theta_\tau^\star = \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}^{\mathrm{DPO}}(\theta;\tau)$ satisfy $\|\theta\|_2 \le B_\theta$ on the relevant feasible set.

Why this comparison is harder for MAML. Multitask-DPO, by definition, lies on the trajectory of MAML at $\alpha = 0$, which forces the gap to scale with α. Baseline-DPO instead lies *off* this trajectory: it is constructed without using any cross-lingual preference signal. The resulting gap can therefore be controlled by the difference between an SFT-only initialization and the manifold of task optima, a quantity that does not vanish as $\alpha \to 0$.

F.2 Initialization Distance Bounds

We separately bound the expected squared distance to the task optimum from each initialization.

Define the *SFT preference-loss gap* on the target task as

$$R_{\mathrm{SFT}} \;\triangleq\; \mathbb{E}_\tau\big[\mathcal{L}^{\mathrm{DPO}}(\theta_0^{\mathrm{BL}};\tau) - \mathcal{L}^{\mathrm{DPO}}(\theta_\tau^\star;\tau)\big]\,. \tag{51}$$

This is a quantity directly measurable from data: it is the average DPO loss of the SFT-initialized policy minus the optimal DPO loss for the target task, and it captures *how much room is left for preference learning after SFT*.

Lemma F.1 (Initialization distance for Baseline-DPO)**.** *Suppose $\theta_0^{\mathrm{BL}}, \theta_\tau^\star \in \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B_\theta\}$ for every $\tau \sim P_{\mathit{TASK}}$ (the standing constraint of the log-linear policy class Π in (12)). Then*

$$\mathbb{E}_\tau\big[\|\theta_0^{\mathrm{BL}} - \theta_\tau^\star\|_2^2\big] \;\le\; D_{\mathrm{BL}}^2 \;\triangleq\; \min\left(4B_\theta^2,\; \frac{2R_{\mathrm{SFT}}}{\mu_\tau}\right). \tag{52}$$

The first bound is the policy-class diameter; the second is loss-aware and tight whenever the SFT-initialized policy already incurs small DPO loss on the target task.

Proof. By the triangle inequality, $\|\theta_0^{\mathrm{BL}} - \theta_\tau^\star\|_2 \le \|\theta_0^{\mathrm{BL}}\|_2 + \|\theta_\tau^\star\|_2 \le 2B_\theta$, so $\mathbb{E}_\tau\|\theta_0^{\mathrm{BL}} - \theta_\tau^\star\|_2^2 \le 4B_\theta^2$. For the second bound, μ_τ-strong convexity (Lemma D.3) yields

$$\mathcal{L}^{\mathrm{DPO}}(\theta;\tau) \;\ge\; \mathcal{L}^{\mathrm{DPO}}(\theta_\tau^\star;\tau) + \tfrac{\mu_\tau}{2}\|\theta - \theta_\tau^\star\|_2^2,$$

since $\nabla\mathcal{L}^{\mathrm{DPO}}(\theta_\tau^\star;\tau) = 0$. Rearranging at $\theta = \theta_0^{\mathrm{BL}}$ and taking expectation gives $\mathbb{E}_\tau\|\theta_0^{\mathrm{BL}} - \theta_\tau^\star\|_2^2 \le 2R_{\mathrm{SFT}}/\mu_\tau$. The bound is the smaller of the two. □

The role of R_{SFT} is the central conceptual difference between this analysis and the multitask comparison in Appendix E: rather than a worst-case parameter-space diameter, we now control the baseline distance by the actual quality of the SFT initialization, which is independent of α and reflects *how much DPO learning the SFT model has not yet done.*

The bound in Lemma F.1 is the rigorous, α-independent counterpart of the variance bound $D_{\mathrm{base}}^2 = \sigma^2/\mu_\tau^2$ that holds for the multitask initialization $\theta_{\mathrm{base}}^\star$. The two are different: D_{base}^2 only quantifies how far the task optima are from the multitask minimizer, whereas D_{BL}^2 quantifies how far they are from a target-only SFT model that has never seen any non-target preference data.

Lemma F.2 (Distance for MAML-DPO)**.** *Let $\hat{\theta}_M$ be the output of Algorithm 1 under the conditions of Theorem 4.6. Then*

$$\mathbb{E}_\tau\left[\|\hat{\theta}_M - \theta_\tau^\star\|_2^2\right] \;\le\; \left(D_{\mathrm{MAML}} + \epsilon_{\mathrm{meta}}\right)^2 \;\le\; 2D_{\mathrm{MAML}}^2 + 2\epsilon_{\mathrm{meta}}^2, \tag{53}$$

where $D_{\mathrm{MAML}}^2 \triangleq D_{\mathrm{base}}^2 - 2\alpha\Phi + O(\alpha^2)$, $D_{\mathrm{base}}^2 = \sigma^2/\mu_\tau^2$, Φ is as in Lemma E.2, and $\epsilon_{\mathrm{meta}}^2 = (2\alpha\sigma/m_M)^2$.

Proof. By the triangle inequality, $\|\hat{\theta}_M - \theta_\tau^\star\|_2 \le \|\hat{\theta}_M - \theta_M^\star\|_2 + \|\theta_M^\star - \theta_\tau^\star\|_2$. Squaring and taking expectation over τ,

$$\mathbb{E}_\tau\big[\|\hat{\theta}_M - \theta_\tau^\star\|_2^2\big] \;\le\; \|\hat{\theta}_M - \theta_M^\star\|_2^2 + 2\|\hat{\theta}_M - \theta_M^\star\|_2 \cdot \mathbb{E}_\tau\big[\|\theta_M^\star - \theta_\tau^\star\|_2\big] + \mathbb{E}_\tau\big[\|\theta_M^\star - \theta_\tau^\star\|_2^2\big].$$

Applying Theorem 4.6 (which gives $\|\hat{\theta}_M - \theta_M^\star\|_2 \le \epsilon_{\mathrm{meta}}$), Jensen's inequality on the cross term ($\mathbb{E}_\tau\|\theta_M^\star - \theta_\tau^\star\|_2 \le \sqrt{\mathbb{E}_\tau\|\theta_M^\star - \theta_\tau^\star\|_2^2} = D_{\mathrm{MAML}}$), and Lemma E.2 (which gives $\mathbb{E}_\tau\|\theta_M^\star - \theta_\tau^\star\|_2^2 = D_{\mathrm{MAML}}^2$) yields the first inequality $(D_{\mathrm{MAML}} + \epsilon_{\mathrm{meta}})^2$. The second follows from $(a+b)^2 \le 2a^2 + 2b^2$. □

In the regime that motivates meta-learning, namely small task variance σ, $\theta_M^\star$ approaches the manifold of task optima while θ_0^{BL} does not. Concretely, $D_{\mathrm{MAML}}^2 \le \sigma^2/\mu_\tau^2$, which can be much smaller than $D_{\mathrm{BL}}^2 \le 4B_\theta^2$ whenever

$$\frac{\sigma}{\mu_\tau} \;\ll\; B_\theta, \tag{54}$$

i.e. task heterogeneity, measured in parameter space, is much smaller than the diameter of the policy class.

F.3 Comparative Convergence Rate

We now plug these distance bounds into the standard contraction inequality for strongly-convex smooth optimization and obtain the main comparative result of this section.

Theorem F.3 (Accelerated adaptation: MAML-DPO vs. Baseline-DPO)**.** *Let $\eta \le 1/L_\tau$ and $\epsilon > 0$. Under Assumption 4.1, 4.4, 4.5, and the standing assumption $\theta_0^{\mathrm{BL}}, \theta_\tau^\star \in \{\|\theta\|_2 \le B_\theta\}$, the expected number of gradient steps needed to reach $\mathbb{E}[\|\theta_t - \theta_\tau^\star\|_2^2] \le \epsilon^2$ for baseline-DPO (t_{BL}) and MAML-DPO (t_{MAML}) satisfies*

$$\mathbb{E}[t_{\mathrm{BL}}] - \mathbb{E}[t_{\mathrm{MAML}}] \;\ge\; \frac{1}{2\eta\mu_\tau} \ln\!\left(\frac{D_{\mathrm{BL}}^2}{(D_{\mathrm{MAML}} + \epsilon_{\mathrm{meta}})^2}\right), \tag{55}$$

where $D_{\mathrm{BL}}^2 = \min(4B_\theta^2,\, 2R_{\mathrm{SFT}}/\mu_\tau)$, $D_{\mathrm{MAML}}^2 = \sigma^2/\mu_\tau^2 - 2\alpha\Phi + c_1\alpha^2$, and $\epsilon_{\mathrm{meta}} = 2\alpha\sigma/m_M$. In particular, as long as $\alpha \le \min\left\{\frac{1}{8\beta^2},\, \frac{m_M}{2\sqrt{2}\,\mu_\tau}\right\}$ the right-hand side of (55) is at least

$$\mathbb{E}[t_{\mathrm{BL}}] - \mathbb{E}[t_{\mathrm{MAML}}] \;\ge\; \frac{1}{2\eta\mu_\tau} \ln\!\left(\frac{R_{\mathrm{SFT}}\,\mu_\tau}{2\sigma^2}\right), \tag{56}$$

which is also lower bounded by the looser worst-case form $\frac{1}{2\eta\mu_\tau}\ln(B_\theta^2\mu_\tau^2/\sigma^2)$ when $R_{\mathrm{SFT}} \ge 2B_\theta^2\mu_\tau$.

Proof. Tight contraction. $\mathcal{L}^{\mathrm{DPO}}(\,\cdot\,;\tau)$ is μ_τ-strongly convex and L_τ-smooth (Lemma D.3). Writing $\theta_{t+1} - \theta_\tau^\star = (I - \eta H_t)(\theta_t - \theta_\tau^\star)$ with $H_t = \int_0^1 \nabla^2\mathcal{L}^{\mathrm{DPO}}(\theta_\tau^\star + s(\theta_t - \theta_\tau^\star);\tau)\,ds$ satisfying $\mu_\tau I \preccurlyeq H_t \preccurlyeq L_\tau I$, the operator-norm bound $\|I - \eta H_t\|_2 \le 1 - \eta\mu_\tau$ holds for $\eta \le 1/L_\tau$, so

$$\|\theta_t - \theta_\tau^\star\|_2 \;\le\; (1 - \eta\mu_\tau)^t\|\theta_{\mathrm{init}} - \theta_\tau^\star\|_2, \qquad \|\theta_t - \theta_\tau^\star\|_2^2 \;\le\; (1 - \eta\mu_\tau)^{2t}\|\theta_{\mathrm{init}} - \theta_\tau^\star\|_2^2. \tag{57}$$

Taking expectation over τ, the smallest t that ensures $\mathbb{E}_\tau\|\theta_t - \theta_\tau^\star\|_2^2 \le \epsilon^2$ is

$$t \;\ge\; \frac{1}{2\eta\mu_\tau}\ln\!\left(\frac{\mathbb{E}_\tau\|\theta_{\text{init}} - \theta_\tau^\star\|_2^2}{\epsilon^2}\right). \tag{58}$$

Substituting $\theta_{\text{init}} = \theta_0^{\text{BL}}$ gives $t_{\text{BL}} \ge \frac{1}{2\eta\mu_\tau}\ln(D_{\text{BL}}^2/\epsilon^2)$ via Lemma F.1, and substituting $\theta_{\text{init}} = \hat{\theta}_M$ gives $t_{\text{MAML}} \ge \frac{1}{2\eta\mu_\tau}\ln\big((D_{\text{MAML}} + \epsilon_{\text{meta}})^2/\epsilon^2\big)$ via Lemma F.2. Subtracting yields (55).

Reduction to (56). We show that $(D_{\text{MAML}} + \epsilon_{\text{meta}})^2 \le 4D_{\text{base}}^2$ in the valid range of α.

(a) Bound on $|\Phi|$. By Cauchy-Schwarz, $|\Phi| \le \|\theta_{\text{base}}^\star - \overline{\theta}^\star\|_2 \cdot \|H_{\text{base}}^{-1}\nabla G(\theta_{\text{base}}^\star)\|_2$. By Jensen, $\|\theta_{\text{base}}^\star - \overline{\theta}^\star\|_2^2 = \|\mathbb{E}_\tau[\theta_\tau^\star] - \theta_{\text{base}}^\star\|_2^2 \le \mathbb{E}_\tau\|\theta_\tau^\star - \theta_{\text{base}}^\star\|_2^2 = D_{\text{base}}^2$. For the second factor, $\|H_{\text{base}}^{-1}\|_2 \le 1/\mu_\tau$ by strong convexity, and

$$\|\nabla G(\theta_{\text{base}}^\star)\|_2 = \big\|\mathbb{E}_\tau\big[2\nabla^2\mathcal{L}(\theta_{\text{base}}^\star;\tau)\,\nabla\mathcal{L}(\theta_{\text{base}}^\star;\tau)\big]\big\|_2 \le 2L_\tau\mathbb{E}_\tau\big[\|\nabla\mathcal{L}(\theta_{\text{base}}^\star;\tau)\|_2\big] \le 2L_\tau\sigma,$$

using $\nabla^2\mathcal{L} \preceq L_\tau I$ and the variance bound (Assumption 4.5, with $\mathbb{E}_\tau[\nabla\mathcal{L}(\theta_{\text{base}}^\star;\tau)] = 0$ by definition of $\theta_{\text{base}}^\star$). Thus $|\Phi| \le D_{\text{base}} \cdot (2L_\tau\sigma/\mu_\tau) = 2L_\tau D_{\text{base}}^2$. Substituting $L_\tau = \beta^2$, $|2\alpha\Phi| \le 4\alpha\beta^2 D_{\text{base}}^2$, which is at most $\frac{1}{2}D_{\text{base}}^2$ whenever $\alpha \le 1/(8\beta^2)$.

(b) Bound on ϵ_{meta}^2. From Theorem D.1, $m_M = \beta^2 c\nu^3/(1+\exp(Z_{\max}))^3 = \Theta(\mu_\tau \cdot \nu^2/(1+\exp(Z_{\max}))^2)$, hence $\sigma/m_M = \kappa_M \cdot \sigma/\mu_\tau = \kappa_M D_{\text{base}}$ for a constant $\kappa_M = O(1)$ that depends only on $(\nu, Z_{\max})$. Therefore $\epsilon_{\text{meta}}^2 = 4\alpha^2\kappa_M^2 D_{\text{base}}^2$, which is at most $\frac{1}{2}D_{\text{base}}^2$ for $\alpha \le 1/(2\sqrt{2}\kappa_M)$.

(c) Combining. For α smaller than $\min\{1/(8\beta^2), 1/(2\sqrt{2}\kappa_M)\}$ (and the smaller of these is implied by the validity ranges of Lemma E.2 and theorem 4.6), the higher-order $c_1\alpha^2$ term is also negligible, so

$$D_{\text{MAML}}^2 \;\le\; D_{\text{base}}^2 + |2\alpha\Phi| + c_1\alpha^2 \;\le\; \tfrac{3}{2}D_{\text{base}}^2,$$

and $(D_{\text{MAML}} + \epsilon_{\text{meta}})^2 \le 2D_{\text{MAML}}^2 + 2\epsilon_{\text{meta}}^2 \le 3D_{\text{base}}^2 + D_{\text{base}}^2 = 4D_{\text{base}}^2 = 4\sigma^2/\mu_\tau^2$. Plugging the SFT-aware $D_{\text{BL}}^2 \le 2R_{\text{SFT}}/\mu_\tau$ from Lemma F.1 into (55) yields $\frac{1}{2\eta\mu_\tau}\ln\big((2R_{\text{SFT}}/\mu_\tau)/(4\sigma^2/\mu_\tau^2)\big) = \frac{1}{2\eta\mu_\tau}\ln(R_{\text{SFT}}\mu_\tau/(2\sigma^2))$, establishing (56). The diameter form follows analogously by plugging $D_{\text{BL}}^2 \le 4B_\theta^2$. □

F.4 Discussion

We now position Theorem F.3 relative to the multitask comparison in Appendix E, link the bound to the empirical patterns observed in our experiments, and discuss its limitations.

(i) Cross-lingual preprocessing helps; the bound is shared by MAML and multitask. Because θ_0^{BL} never sees any non-target preference data, Theorem F.3 attributes the speed-up to *any* algorithm whose initialization θ_{init} achieves $\mathbb{E}_\tau\|\theta_{\text{init}} - \theta_\tau^\star\|_2^2 = O(\sigma^2/\mu_\tau^2)$. Both MAML-DPO and multitask-DPO satisfy this property under mild conditions: multitask-DPO directly minimizes $\mathbb{E}_\tau\mathcal{L}(\theta;\tau)$ (whose minimizer attains $D_{\text{base}}^2 = \sigma^2/\mu_\tau^2$), and MAML-DPO additionally trims this distance by an $O(\alpha)$ amount via Lemma E.2. The bound (56) is therefore best read as "*any* cross-lingual preprocessing dominates target-only adaptation by a $\log(R_{\text{SFT}}\mu_\tau/\sigma^2)$ factor in adaptation step count." The additional MAML-specific gain over multitask is captured separately by the $O(\alpha)$ analysis in Appendix E.

(ii) The gap does not vanish as $\alpha \to 0$. (56) depends only on R_{SFT}, μ_τ, σ, and η; it is bounded below by a positive constant uniformly in α as long as $R_{\text{SFT}}\mu_\tau \gg \sigma^2$. In contrast, the multitask gap from Appendix E is $O(\alpha)$ and collapses to zero as $\alpha \to 0$. This complementarity gives a two-tier story: the cross-lingual preprocessing gain is order-one and order-log in the parameters, while the meta-vs-multitask gain is order-α.

(iii) Harder target languages should benefit more from MAML-DPO. The asymptotic gap (56) scales with $\log R_{\text{SFT}}$, where R_{SFT} is the SFT-init DPO loss gap on the target task. A target language for which the base/SFT model performs poorly has a large R_{SFT}: the policy is far from optimal under preference comparisons, leaving more room for cross-lingual preprocessing to help.

The bound therefore predicts a *larger* MAML-vs-baseline win-rate gap on lower-resource or harder target languages. This matches Tables 1 and 3: the largest improvements appear on Romanian and Indonesian (lowest base-model competence among the targets we evaluate), while comparatively closer-to-MAML targets such as French show smaller gaps. The same monotone trend appears across model scales (270M $\rightarrow$ 7.1B) and adaptation budgets (100, 4k, 40k), all consistent with R_{SFT}-scaling.

(iv) Empirical magnitudes. Tables 1, 2, 3 and 4 show the MAML-DPO advantage over baseline-DPO is consistently large and substantially exceeds the MAML-DPO advantage over multitask-DPO in most of cases. Both regimes are predicted by our analysis: the large gap is the $\Theta(\ln(R_{\mathrm{SFT}}\mu_\tau/\sigma^2))$ separation of (56), while the smaller gap is the $O(\alpha)$ refinement of Appendix E.

(v) Limitations. The bound is in the population (infinite-sample) regime. A finite-sample extension is straightforward via standard concentration of the per-task gradient: with n adaptation samples, the empirical iterate satisfies $\mathbb{E}_\tau\|\theta_t - \theta_\tau^\star\|_2^2 \le (1-\eta\mu_\tau)^{2t} D_{\mathrm{init}}^2 + O(\sigma_{\mathrm{grad}}^2 \log(d/\delta)/(\mu_\tau^2 n))$ with high probability, where the second term is an irreducible sampling floor that is the same for both MAML-DPO and baseline-DPO; the step-count gap of Theorem F.3 is therefore preserved on top of this shared floor. Beyond sampling, the strong-convexity-and-smoothness framework forces the per-step contraction $(1-\eta\mu_\tau)$ to be the same for any vanilla-GD adaptation, regardless of initialization, so the gain manifests only through the initialization distance and not through a faster rate. Obtaining a faster rate would require either a different inner loop (e.g., Newton-type) or stronger structural assumptions such as a shared low-rank multilingual representation across tasks; under such an assumption, MAML-style methods can in principle achieve exponential rate improvements in the conditioning, which we leave for future work.

G Additional Related Work

Due to their wide range of capabilities, LLMs have rapidly gained popularity in both business and research, being developed as closed-source models [1, 47, 46, 11] as well as open-source ones [52, 48–50, 27, 3, 34]. Regardless of their public availability or parameter size, RLHF and preference optimization have been two most popular approaches to training LLMs with human feedback data. RLHF uses preference data consisting of human feedback to train reward functions and use them to train the policies [60, 37, 14, 55]. Direct Preference Optimization has been proposed as an alternative to RLHF, which eliminates the need to train a separate reward function [39, 38]. Due to its practical advantage, many works have proposed improvements of DPO. Most follow-up works to DPO have investigated alternative objective functions [16, 19, 35], trying to address well-known issues of DPOs, including ones coming from the use of Bradley-Terry model. Son et al. [41] proposes addressing temporal drift in the dataset by applying exponential discount to the samples in the past. Tang et al. [44] proposes regularizing the self-play optimization method [43], further stabilizing the training and improving performances on an independent benchmark such as AlpacaEval [15].

H Impact Statement and Limitations

This work proposes a meta-learning approach for LLM preference learning, designed to mitigate the unequal performance across different languages. By significantly improving sample efficiency in low-resource settings, demonstrating effective alignment with only 100 samples, our method addresses the data scarcity issue that often limits the access of linguistic communities with less population to AI services. Furthermore, our theoretical guarantees provide a rigorous foundation in cross-lingual transfer, supporting the development of more equitable, globally accessible, and representative AI systems without the excessive cost of massive human-labeled datasets for every target language.

However, like existing cross-lingual approaches, our method may inadvertently propagate biases, normative assumptions, and safety standards learned from high-resource source languages to low-resource target languages. Therefore, while the method improves alignment performance, it does not guarantee culturally appropriate or community-specific alignment. Beyond that, although the first-order approximation used in our MAML-based methods reduces the computational overhead, MAML

still introduces an inner–outer optimization structure, which requires additional hyperparameters, such as the inner-loop learning rate coefficient.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our contributions in Section 1. Unequal data availability across languages poses a substantial challenge for preference alignment in LLMs. To address this issue, we apply meta-learning to both DPO and RLHF in a multilingual language-learning setting. We evaluate our approach using two model scales and conduct empirical experiments across a broad range of language groups, covering both strictly limited-data and more sufficient-data settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix H, we discuss our limitations. Our approach applies MAML to DPO and RLHF, and therefore inherits an inner–outer optimization structure that requires additional hyperparameters. In addition, like existing cross-lingual approaches, our method may transfer biases from high-resource data to low-resource data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important

role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the theoretical analysis in Section 4. The full proofs are provided in Appendix C, Appendix D, Appendix E and Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the experimental setup in Section 5.1, including the datasets, language models, baselines, and evaluation metrics. Our code is available through the https://github.com/geronest/maml-preferences. Additional experimental details are provided in Appendix A, including pipeline visualizations, detailed hyperparameters, and compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce

the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets [12] we used are open-resource. Our code is available through the https://github.com/geronest/maml-preferences.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The data-splitting strategy is described in Section 5.1 and Appendix A. Detailed hyperparameters are provided in Appendix A. Other training and evaluation details can be found in our source code, available through the https://github.com/geronest/maml-preferences.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: We report error bars for all 100-sample settings in Section 5.2 and Appendix B, where performance is more sensitive to sampling variability. In sufficient-data settings, the variance is expected to be substantially reduced, and our conclusions are supported by consistent trends across languages. Given this reduced sensitivity and the high computational cost of repeated large-scale runs, we report error bars for the 100-sample experiments only.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide a resource consumption table in Appendix B.3, including the number of training steps,the number of used samples, wall-clock time and GPU memory consumption.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We confirm that this study strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both potential positive and negative societal impacts are discussed in Appendix H.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release new datasets or trained model weights with a high risk of misuse. We use publicly available datasets (Okapi and orca-dpo) and language models (Gemma3 and Bloom) under their existing licenses, and our released code is intended only to support reproducibility of the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: The paper credits creators of all datasets and language models used in Section 5.1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide our code and documentation in https://github.com/geronest/maml-preferences.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not conduct crowdsourcing experiments. We use publicly available datasets and follow the documentation provided by the original dataset creators.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve new research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.