CroCo: Cross-Lingual Contrastive Preference Tuning on Self-Generations

Anonymous ACL submission

Abstract

Prior work establishes that controlled contrastiveness between self-generated responses from large language models, set via reward scores, improves downstream preference tuning in English. We extend this method to multiple languages and evaluate two models across a total of 14 high and low-resource languages on a diverse set of tasks. Our central finding is that **cro**ss-lingual **co**ntrastive preference tuning on self-generations (CroCo) transfers without language-specific preference annotation. A reward model trained on English preferences (atop a multilingual base) produces useful within-language rankings across most languages, and pairing in either a monolingual or multilingual setting improves over each model on the majority of setups while preventing the catastrophic forgetting of supervised fine-tuning. We observe that the gains require on-policy data. Off-policy responses reduce the benefit and online preference optimization fails to improve over the offline variant. Specifically, on structured tasks, our method matches or exceeds the base in 6/7 languages for `EuroLLM-9B` and 4/7 settings for `aya-3B`. On open-ended generation, both tuned models win against their respective base across 11 evaluated languages. Overall, we show promising directions for multilingual preference tuning.[1]

1 Introduction

Aligning large language models (LLMs) with human preferences is the standard final stage of post-training, and Direct Preference Optimization (DPO; Rafailov et al., 2023) is one of the dominant approaches. Recently, DPO has been applied to self-generated samples rather than human preferences (Guo et al., 2024; Xiao et al., 2025): a policy model is paired with a reward model (RM) that scores its on-policy responses to build preference pairs of *chosen* and *rejected* completions. Similarly, recent work has shifted attention from the optimizer to the data: Pan et al. (2025) show that chosen-response quality dominates downstream performance, Geng et al. (2025) establish that the *relative* quality gap drives improvement, and Xiao et al. (2025) identify a "sweet spot" in which the rejected response is sampled near a specific quartile of the reward distribution rather than at the minimum. These findings are exclusively in English.

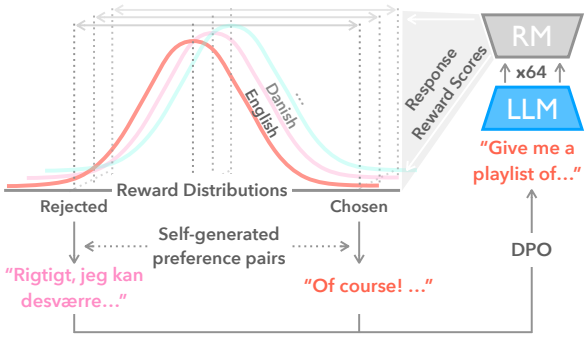


Figure 1: **Setup.** An LLM generates 64 responses per prompt per language; an external off-the-shelf RM scores these and we sample specific quartiles to construct *contrastive* preference pairs.

Extending preference tuning beyond English raises open questions. Prior multilingual work relies on translation-based preference signals (She et al., 2024), exploits the English/non-English capability gap as an implicit reward (Yang et al., 2025c,b), or reweights the DPO loss for noisy multilingual pairs (Pokharel et al., 2025). None of these establishes whether reward-distribution-based pair construction itself transfers across languages. We therefore ask: *Does contrastive preference tuning on self-generations transfer to a multilingual setting without language-specific preference annotation?* We examine this across monolingual and multilingual training regimes and two post-tranied models at different scales (3B and 9B parameters).

[1]The anonymized repository can be found here: https://anonymous.4open.science/r/CroCo-BFD5/.

Hypothesis. We posit that contrastive preference tuni transfers cross-lingually, because the DPO objective depends on the relative reward gap rather than absolute calibration. Consistent *within*-language ranking suffices despite cross-lingual miscalibration. This predicts that (i) an English-only RM — built atop a multilingual base, as is standard for open RMs (e.g., Liu et al., 2025) — suffices for multilingual tuning when scored on within-language samples, removing the need for per-language annotation, and (ii) on-policy data matters more than generator quality, since the contrastive signal is informative only when paired responses come from the policy's own distribution.

Contributions. ① Contrastive preference tuning transfers cross-lingually and across models: DPO on self-generations outperforms SFT baselines and existing multilingual preference-tuning methods (She et al., 2024; Yang et al., 2025b), while standard SFT causes catastrophic forgetting in both models. ② Multilingual preference tuning does not require multilingual preference annotation: an English-only RM (atop a multilingual base) drives consistent gains across most languages, and joint multilingual training matches or exceeds monolingual training for both models. ③ The method improves both structured and open-ended evaluation: multilingual `Paired` DPO matches or exceeds the base in 6/7 languages for `EuroLLM-9B` and 4/7 settings for `aya-3B` on EuroEval, and both DPO-tuned models beat their base in all 11 evaluated languages on m-ArenaHard 2.1. ④ Ablations on translation, prompt language, and on-policy vs. off-policy data confirm hypothesis (ii) and isolate which design choices are crucial, in line with Tajwar et al. (2024) and Shenfeld et al. (2026).

2 Problem Formulation

Preference Tuning. Let π_θ be a policy language model parameterized by θ, and π_{ref} a frozen reference model. Given a prompt x and a preference pair (y_c, y_r), where y_c is *chosen* over the *rejected* y_r, DPO (Rafailov et al., 2023) minimizes

$$\mathcal{L}_{\text{DPO}}(\theta) = \underset{(x,y_c,y_r)\sim\mathcal{D}}{-\mathbb{E}}\Big[\log\sigma\big(\Delta r_\theta\big)\Big], \quad (1)$$

where $\Delta r_\theta \coloneqq r_\theta(x, y_c) - r_\theta(x, y_r)$ is the reward margin, $r_\theta(x, y) \coloneqq \beta \log\big(\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)\big)$ is the implicit reward, and $\sigma(\cdot)$ is the sigmoid. The quality of the dataset $\mathcal{D} = \{(x^{(i)}, y_c^{(i)}, y_r^{(i)})\}_{i=1}^N$ is central to downstream performance.

Contrastive Preference Pairs. Following Xiao et al. (2025), we build $\mathcal{D}$ via on-policy self-generation. For each prompt x, the policy generates K candidates $\mathcal{Y}_x = \{y^{(k)}\}_{k=1}^K$, each scored by an external reward model $R\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. With μ_x, σ_x the mean and standard deviation of $\{R(x, y^{(k)})\}_{k=1}^K$, a preference pair is formed as

$$\begin{aligned} y_c &= \underset{y\in\mathcal{Y}_x}{\arg\max}\ R(x, y), \\ y_r &= \underset{y\in\mathcal{Y}_x}{\arg\min}\ \big|R(x, y) - (\mu_x - 2\sigma_x)\big|. \end{aligned} \quad (2)$$

In other words, rather than targeting the lowest-scoring candidate, y_r is selected as the sample in $\mathcal{Y}_x$ whose reward is nearest to $\mu_x - 2\sigma_x$, inducing a controlled level of contrastiveness between y_c and y_r. We show samples from each region of the reward distribution in Appendix A.

Multilingual Extension. Prior work establishes this construction only for English; we extend it to target languages $\mathcal{L} = \{\ell_1, \ldots, \ell_L\}$. Given an English prompt set $\mathcal{X}_{\text{eng}}$, we obtain parallel prompts $\mathcal{X}_\ell$ for each ℓ via machine translation. For every (x, ℓ), the policy generates K responses conditioned on the ℓ-language prompt, yielding a language-specific dataset $\mathcal{D}_\ell$. We study two settings: (1) **Monolingual**, tuning on each $\mathcal{D}_\ell$ independently, and (2) **Multilingual**, tuning jointly on $\mathcal{D} = \bigcup_{\ell\in\mathcal{L}} \mathcal{D}_\ell$. We use two models of different scales (3B/9B) to test robustness to model size.

3 Experimental Setup

3.1 Data

We stratify 20K instances from Dolci-Instruct-SFT - the instruction tuning corpus used to train OLMo3 (Olmo et al., 2025); Figure 2 shows the sampled domain distribution. We translate the English data into six European languages: Danish (`dan`), Dutch (`nld`), French (`fra`), German (`deu`), Italian (`ita`), and Spanish (`spa`), using `TranslateGemma-27B` (Finkelstein et al., 2026). Figure 3 shows token-length statistics per language.

Using `EuroLLM-9B`[2] (Ramos et al., 2026) or `aya-3B`[3] (Salamanca et al., 2026) as the on-policy model, we generate 64 responses per instance ($>$ 60 samples plateaus performance per Xiao et al.,

[2] https://huggingface.co/utter-project/EuroLLM-9B-Instruct-2512.

[3] https://huggingface.co/CohereLabs/tiny-aya-global.

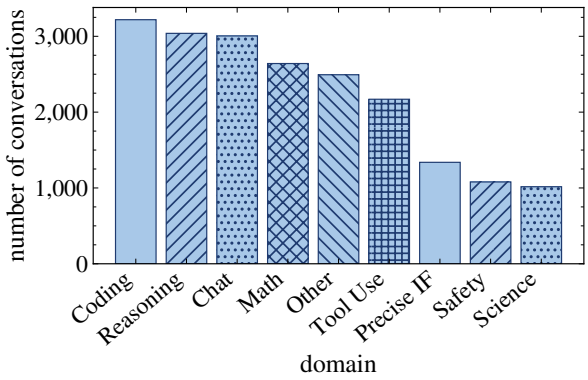


Figure 2: **Domain distribution of Dolci-Instruct-SFT.** Our 20K stratified sample covers nine task domains, with coding, reasoning, chat, and math accounting for the bulk of instances.

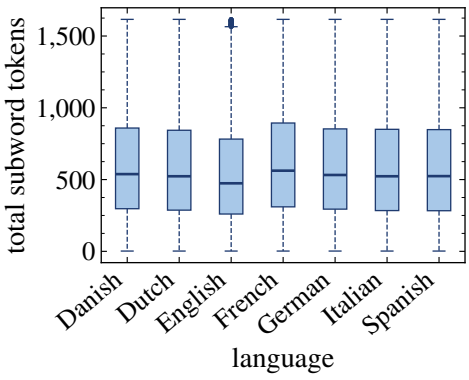


Figure 3: **Subword-token length distribution across languages.** We cap the 90th percentile at 1,616 tokens. Romance languages (French, Italian, Spanish) produce systematically longer translations than Germanic ones.

2025) at temperature $T = 0.7$ for `EuroLLM-9B` and $T = 0.1$ for `aya-3B`, producing 1.28M samples per language. Each is scored with `Skywork-Reward-V2-Qwen3-8B` (Liu et al., 2024, 2025), an RM whose preference training is English-only but whose model (Qwen3-8B) is multilingual (Yang et al., 2025a). We select this RM because English-preference-trained RMs of this kind transfer robustly across languages (Wu et al., 2024; Hong et al., 2025) and because it ranks sixth on RewardBench 2.0 (Malik et al., 2026).[4] Crucially, our hypothesis requires the RM to *score* responses consistently within and across each target language. We show this happens qualitatively in Appendix B.

Training Data Construction. We compare four construction strategies, in both monolingual and multilingual regimes, applied to both models: ① **In-Lang** / ② **All Lang (SFT)**: the translated in-language set, or the union across all languages, fine-tuned with standard SFT, without any preference signal. ③ **Max-R (SFT)**: For each prompt, only the highest-scoring response is kept and SFT applied: a best-of-K baseline that uses the reward signal but discards contrastiveness. ④ **Paired (DPO)**: Following Xiao et al. (2025), we form preference pairs following Equation (2), and apply DPO.

We verify in Appendix C that the multilingual `Paired` construction does not degenerate into selecting English as chosen and a non-English language as rejected, but selects across all languages.

[4] `https://huggingface.co/spaces/allenai/reward-bench`

3.2 Training

We fine-tune with LoRA (Hu et al., 2022) for all setups in TRL (von Werra et al., 2020).[5] For SFT, we train for 1 epoch with sequence length 4,096, global batch size 64, and learning rate 2×10^{-4} (cosine schedule, 5% warmup, weight decay 1×10^{-2}), optimizing the standard autoregressive cross-entropy loss over completions only.

For preference tuning, the policy π_θ also serves as the frozen reference π_{ref}. We train for 1 epoch with learning rate 5×10^{-6} (cosine schedule, 5% warmup, weight decay 1×10^{-2}), $\beta = 0.1$, and the same batch size and sequence length as SFT. Full training details are in Appendix D.

3.3 Evaluation

We evaluate with EuroEval (Smart, 2023; Saattrup Nielsen et al., 2025), a multilingual framework supporting all European languages. The suite comprises 32 datasets across the seven target languages (`dan`, `nld`, `eng`, `fra`, `deu`, `ita`, `spa`), covering reading comprehension, knowledge, commonsense reasoning, linguistic acceptability, and word-in-context tasks; full details are in Appendix E. For cross-lingual generalization analyses we additionally evaluate on Norwegian (`nor`), Portuguese (`por`), and Swedish (`swe`). For open-ended generation we use m-ArenaHard 2.1 (Section 4.2), where we evaluate on `dan`, `nld`, `eng`, `fra`, `deu`, `ita`, `spa`, Galician (`glg`), Irish (`gle`), Maltese (`mlt`), and Welsh (`cym`).

[5] We are aware of the gradient accumulation and CPU offloading bug found by Limozin et al. (2026) in SFT training using TRL; we detail in Appendix D how we are not affected.

		Baselines			Monolingual Post-training			Multilingual Post-training		
Language	model	Base	ICR 8B	MAPO 13B	In-lang (SFT)	Max-R (SFT)	Paired (DPO)	All Lang (SFT)	Max-R (SFT)	Paired (DPO)
dan (7)	aya-3B	48.4	-	-	↓4.0	↓4.2	↑0.1	↓4.6	↓7.4	↑0.1
	EuroLLM-9B	65.0	-	-	↓5.3	↑0.1	↑1.0	↓8.5	↑1.3	↑1.1
deu (4)	aya-3B	42.4	↓1.7	↓0.7	↓4.2	↓5.7	↓0.1	↓6.9	↓7.6	↓0.1
	EuroLLM-9B	47.4	↓6.7	↓5.7	↑0.4	↓2.0	↑1.0	↓3.8	↓0.9	↑1.2
eng (5)	aya-3B	54.1	↓1.0	0.0	↓2.8	↓6.3	↑0.2	↓7.9	↓7.0	↑0.4
	EuroLLM-9B	56.8	↓3.6	↓2.6	↓0.1	↓0.5	↓0.4	↓3.9	↑0.2	↑0.3
spa (4)	aya-3B	42.8	↓1.0	↑2.3	↓1.4	↓3.0	↓0.1	↓3.8	↓4.5	↓0.5
	EuroLLM-9B	51.8	↓10.0	↓6.7	↓2.9	↓1.7	↓1.3	↓4.8	↓2.1	↑0.9
fra (5)	aya-3B	45.9	↓3.6	↓1.1	↓5.2	↓6.8	↓0.1	↓7.2	↓7.2	↓0.3
	EuroLLM-9B	52.2	↓9.9	↓7.4	↓2.0	↓1.3	↑0.6	↓6.6	↓2.7	↑0.8
ita (3)	aya-3B	51.6	-	-	↓10.7	↓10.6	↑0.4	↓11.3	↓10.5	↑0.6
	EuroLLM-9B	54.3	-	-	↓7.6	↑1.2	↑3.6	↓9.9	↑5.0	↓1.3
nld (4)	aya-3B	58.5	-	-	↓6.0	↓6.1	↓0.1	↓10.8	↓7.3	↑0.3
	EuroLLM-9B	68.0	-	-	↓3.9	↓1.7	↑0.2	↓6.5	↓0.9	0.0

Table 1: **Average EuroEval evaluation summarized by language, model, and tasks.** The Base column shows the absolute aggregated EuroEval score for each model over three seeds. All other columns show the absolute difference from the model on the same row. ICR (Yang et al., 2025b) and MAPO (She et al., 2024) are independent baseline models with the parameter counts shown in their column headers. Number of datasets per language in parentheses. English uses the original Dolci SFT data. We show the exact results per dataset in Appendix J.

	Baseline	SFT		DPO
Lang.	EuroLLM	All Lang	Max-R	Paired
nor (5)	54.3	↓5.7	↑0.1	↑0.5
por (2)	47.4	↓8.2	↑0.1	↑0.4
swe (4)	52.8	↓5.2	↓0.3	↑0.2

Table 2: **Cross-lingual generalization to held-out languages in EuroEval (Norwegian, Portuguese, Swedish).** Values are dataset-averaged absolute differences from the EuroLLM-9B baseline; the count of held-out datasets per language is in parentheses. Paired DPO generalizes positively in all three held-out languages, while multilingual SFT degrades performance.

4 Results

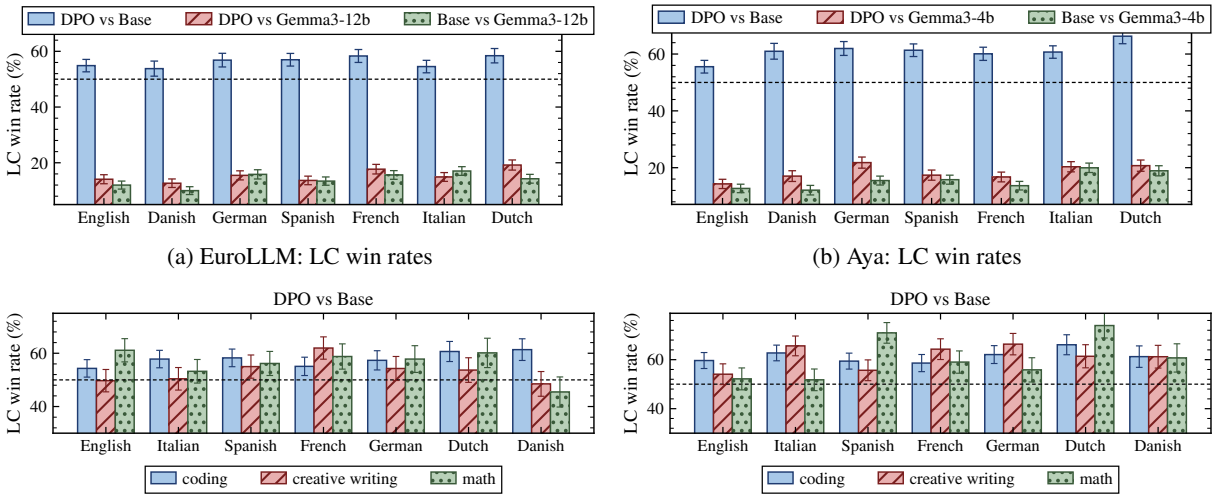
Table 1 reports the main results across the seven target languages for both base and tuned models.

4.1 EuroEval

SFT on translated data causes catastrophic forgetting in models. Both monolingual (In-lang) and multilingual (All Lang) SFT degrades performance relative to the baseline across nearly all languages and both models, dropping from 0.1 points (English, monolingual on EuroLLM-9B) to 11.3 points (Italian, multilingual on aya-3B). Multilingual SFT is harmful. For example, EuroLLM-9B loses 3.8–9.9 points in 6/7 languages and aya-3B loses 3.8–11.3 in all 7, on average more severe for aya-3B, consistent with smaller models having less headroom to absorb new knowledge. This aligns with prior reports of SFT-induced catastrophic forgetting from 1B to 7B parameters (Luo et al., 2025; Shi et al., 2025), with Pan et al. (2025)'s observation that SFT on data not clearly above the model's capability can hurt, and with the delta-learning hypothesis of Geng et al. (2025).

Reward-filtered SFT (Max-R) reduces but does not eliminate forgetting. Keeping only the highest-rewarded completion mitigates most SFT degradation for EuroLLM-9B and yields modest gains in some languages (Italian, +1.2–+5.0; Danish, +0.1–+1.3). For aya-3B, Max-R is less effective, remaining below baseline in every language under both regimes, with drops up to 10.5 points (Italian). The reward signal alone, collapsed to a single target for cross-entropy training, is insufficient to match the baseline and is particularly weak for the smaller model.

Paired DPO consistently matches or outperforms the baseline for both models. DPO on paired self-generations outperforms the EuroLLM-9B baseline in 10 of 14 evaluation settings (seven languages × two regimes), with the largest gain on Italian (+3.6 monolingual). For aya-3B, Paired DPO is non-negative in 12 of 14 settings and strictly positive in 11, the only



(a) EuroLLM: LC win rates

(b) Aya: LC win rates

(c) EuroLLM: by subcategory

(d) Aya: by subcategory

Figure 4: **m-ArenaHard 2.1 results.** *Top row:* Length-controlled win rates. Multilingual `Paired` DPO (blue) wins against the respective model in all 7 languages; against the larger Gemma3-it comparison model (red), DPO narrows the deficit visible in the base-vs-Gemma comparison (green) in 4/7 languages for `EuroLLM-9B` and all seven for `aya-3B`. *Bottom row:* LC win rate of multilingual `Paired` DPO against the base, broken down by prompt type. Coding and creative writing benefit consistently across languages for `EuroLLM-9B`; all three categories benefit for `aya-3B`. Left column: `EuroLLM-9B-Instruct-2512` (Gemma3-12B-it as the larger comparison); right column: `Tiny-Aya-Global-3B` (Gemma3-4B-it as the larger comparison). The dashed line marks parity (50%).

meaningful drop being French multilingual (-0.3). `Paired` never loses more than 1.3 points on either model, in stark contrast to SFT. The contrastive signal, rather than the supervised target, lets both a 9B and a 3B model incorporate new data without overwriting existing capabilities — the empirical results predicted by hypothesis (i): an objective whose loss depends only on the ordering of paired responses is robust to translation noise, while one that targets an absolute completion is not.

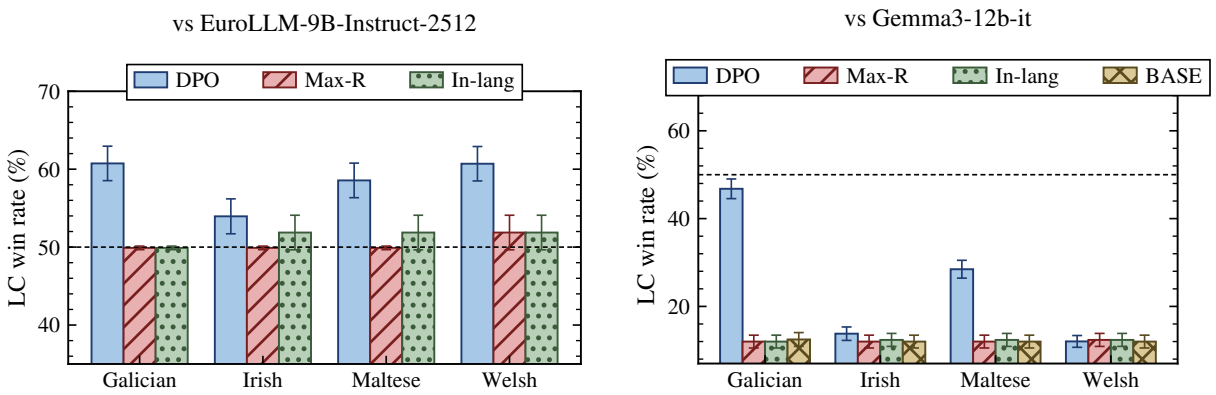
Generalization to held-out languages. Table 2 reports zero-shot transfer of multilingual post-trained `EuroLLM-9B` to Norwegian, Portuguese, and Swedish, not in our post-training data, though likely in the pre-training data. The pattern mirrors the in-distribution results: Multilingual SFT (`All Lang`) degrades the baseline on all 11 datasets (up to -12.1 on Norwegian NorCommonSense), `Max-R` recovers most of the loss, and `Paired` DPO produces small positive gains on 7/11 datasets. The contrastive signal induces a representational change that generalizes cross-lingually to some extent, in line with Hong et al. (2025).

Comparison to multilingual preference-tuning baselines. Two prior methods, ICR (Yang et al., 2025b) and MAPO (She et al., 2024), both degrade the `EuroLLM-9B` baseline in most applicable languages (deu, eng, spa, fra), losing as much as 7–10 points on Spanish. Against `aya-3B` they are closer to flat (within ± 3.6 points in most cells; MAPO yields $+2.3$ on Spanish), but neither consistently improves on the base. Our `Paired` setup is the only method non-negative on average across all evaluated languages.

4.2 m-ArenaHard 2.1

Since EuroEval probes classification, extraction, and multiple-choice but not open-ended generation, we additionally evaluate on m-ArenaHard 2.1 (Salamanca et al., 2026), a multilingual extension of ArenaHard (Li et al., 2025) covering English, German, Spanish, French, Italian, and Dutch, with 498 prompts per language across coding, creative writing, and math. We score completions with `Qwen3.6-35B-A3B` (Qwen Team, 2026) as judge, scoring each pairwise comparison 1/0.5/0 for win/tie/loss, and report the length-controlled (LC) win rate (Dubois et al., 2025).

We compare three pairs per model: multilingual `Paired` DPO vs. its base, `Paired` DPO vs. a larger Gemma3 instruction-tuned model, and the base vs. the same Gemma3 model, the last anchoring the absolute scale. For `EuroLLM-9B` the larger comparison is `Gemma3-12B-it`; for `aya-3B` it is `Gemma3-4B-it`, matching the relative size offset.



(a) LC win rates vs `EuroLLM-9B`.

(b) LC winrates vs `Gemma3-12b-it`

Figure 5: **m-ArenaHard 2.1 Results on Low-resource Languages with EuroLLM-9B.** Length-controlled win rates of `Paired` DPO (blue) wins against the respective model in all four low-resource languages (left) compared to `Max-R` and `In-lang`; against the larger Gemma comparison model (right), DPO narrows the deficit for Galician and Maltese, where the other methods fails to do so. The dashed line marks parity (50%).

DPO improves over the base in every language, on both models. Figure 4 reports LC win rates per language. `Paired` DPO wins against the `EuroLLM-9B` base in all seven evaluated languages, with LC win rates between 54.5% (`ita`) and 58.4% (`nld`) and standard deviation at most 2.6; the largest gains are `nld` (+8.4 over parity) and `fra` (+8.3), followed by `spa` (+7.0), `deu` (+6.8), `eng` (+4.9), `ita` (+4.5), and `dan` (+3.8). The pattern is stronger on `aya-3B`, which wins in all seven languages with LC win rates between 55.5% (`eng`) and 66.3% (`nld`): `nld` (+16.3), `deu` (+12.0), `spa` (+11.3), `dan` (+11.0), `ita` (+10.7), and `fra` (+10.1) all show double-digit gains, with `eng` (+5.5) smallest. The contrastive signal is at least as effective on open-ended generation as on structured tasks, holding across two models that differ by 3× in parameter count.

DPO narrows the gap to a larger Gemma3 model in most languages. The `EuroLLM-9B` base loses to `Gemma3-12B-Instruct` in every language, with LC win rates between 10.0% (`dan`) and 17.0% (`ita`), i.e. deficits of 33.0–40.0 points that, after DPO, narrow in five out of seven languages (`nld` 4.9, `fra` 2.1, `eng` 2.1, `spa` 0.2, `dan` 2.7 in the appendix), stay roughly flat on `deu` (−0.4), and widen on `ita` (2.1). The `aya-3B` results are stronger and more uniform: against `Gemma3-4B-Instruct`, `Paired` DPO closes ground in all 7 languages (`deu` +6.3, `fra` +3.1, `nld` +1.8, `spa` +1.6, `eng` +1.6, `ita` +0.4), showing that CROCO moves a model trained on its own outputs closer to a larger reference it never observed.

Subcategory breakdown. Figure 4 (bottom row) breaks down the DPO-vs-base comparison by prompt type. Coding and creative writing are above parity in nearly every language for `EuroLLM-9B`, and all three subcategories do so for `aya-3B`; math is weaker for `EuroLLM-9B` and the only category with cells below parity. This matches the composition of Dolci-Instruct-SFT (Figure 2), where coding, reasoning, and chat dominate and math is a smaller slice. Figures 13 to 16 in Appendix H show subcategory breakdowns against Gemma3.

Generalization to low-resource languages. We test whether the method improves lower-resourced languages, namely Galician, Irish, Maltese, and Welsh, again using m-ArenaHard 2.1, which covers them. Here we train on each language individually rather than multilingually and compare against `Max-R` and `In-lang`. Figure 5 (left) reports LC win rates: paired DPO improves over `EuroLLM-9B` in all four languages, achieving the highest LC win rate on Galician and Welsh (60.7%), then Maltese (58.6%) and Irish (54.0%). Against `Gemma3-12b-it` (Figure 5, right), our method outperforms all baselines for Galician and Maltese.

Takeaway. m-ArenaHard 2.1 confirms the EuroEval picture in the open-ended setting: `Paired` DPO improves over the base across all 7 evaluated languages and both models, transfers across language families and task types, and narrows the gap to a larger 12B model in 5/7 languages for `EuroLLM-9B`, and to a 4B model in all 7 for `aya-3B`. Italian is the exception for `EuroLLM-9B` and the smallest gain for `aya-3B`, suggesting the Italian

Lang.	Base.	SFT		Max-R		Paired	
		eng	tgt	eng	tgt	eng	tgt
dan (7)	65.0	↓1.2	↓5.3	↑1.8	↑0.1	↑1.8	↑1.0
deu (4)	47.4	↑0.7	↑0.4	↑0.4	↓2.0	↑0.9	↑1.0
spa (4)	51.8	↓1.0	↓2.9	↑0.3	↓1.7	↑0.2	↓1.3
fra (5)	52.2	↓2.7	↓2.0	↑0.2	↓1.3	↑0.2	↑0.6
ita (3)	54.3	↓4.7	↓7.6	↓2.9	↑1.2	↓0.7	↑3.6
nld (4)	68.0	↓1.6	↓3.9	↓0.7	↓1.7	↓0.1	↑0.2

Table 3: **English-only vs. translated in-language post-training (EuroLLM-9B).** Values represent the absolute difference from the EuroLLM-9B baseline. For Paired, translated data wins in five of six languages. Exact numbers are in Table 12 (Appendix).

Lang.	Base	Mono. PT			Multi. PT		
	EuroLLM -9B	In SFT	Max SFT	Pair DPO	All SFT	Max SFT	Pair DPO
dan (7)	65.0	↓5.3	↓7.6	↑1.7	↓8.5	↓6.3	↑1.7
deu (4)	47.4	↑0.4	↓3.2	↑0.3	↓3.8	↓3.4	↑0.5
eng (5)	56.8	↓0.1	↓2.5	↓0.6	↓3.9	↓6.0	↓0.4
spa (4)	51.8	↓2.9	↓3.9	↑0.2	↓4.8	↓6.4	↑0.3
fra (5)	52.2	↓2.0	↓5.6	↑0.1	↓6.6	↓7.5	↑0.1
ita (3)	54.3	↓7.6	↓4.6	↓0.5	↓9.9	↓7.4	↓0.1
nld (4)	68.0	↓3.9	↓5.2	↓0.6	↓6.5	↓10.3	↓0.7

Table 4: **Off-policy data ablation: tuning EuroLLM-9B on aya-3B generations.** Values are absolute differences from the EuroLLM-9B baseline. Off-policy Paired DPO reduces catastrophic forgetting relative to SFT but produces smaller gains than the on-policy setup in Table 1. The exact numbers are in Table 13 (Appendix).

translation distribution is the hardest setting for both. For low-resource languages the picture is similar, where across 4 languages, paired DPO beats the SFT-based methods against the base and improves on 2/4 against Gemma3.

5 Discussion

5.1 Does Translation of the Data Help?

Translating Dolci into the 6 target languages may not be necessary, since the model's multilingual pre-training could suffice. Table 3 compares English-only (eng) against in-language translated (tgt) post-training for EuroLLM-9B across all data-construction strategies. For standard SFT, translated in-language data is worse than English: target-language drops (up to -7.6 on Italian) are larger than English-only drops (up to -4.7 on Italian), consistent with translation artifacts introducing noise (Vanmassenhove et al., 2021; Zhu et al., 2024). Max-R roughly breaks even.

Paired is the only setup that benefits from translation: in-language DPO outperforms English-only DPO in four of six languages (Danish, German, French, Italian), largest on Italian ($+3.6$ vs. -0.7). Because the contrastive signal is relative, the reward *gap* between y_c and y_r stays informative even when translation adds noise to both, whereas SFT optimizes toward a potentially noisy target. This is the most direct evidence for hypothesis (i) and the main methodological takeaway: it identifies *why* CROCO works cross-lingually rather than merely showing that it does.

5.2 Does the Language of the Prompt Matter in DPO?

We also ask whether prompt language matters independent of response language, constructing three variants of the multilingual DPO dataset: The prompt in the same language as the chosen response, assigned uniformly at random, or in the same language as the rejected response. Pairing the prompt with the same-language *chosen* response is strongest, producing gains or ties in all languages except Italian; the other two variants degrade performance in most languages, losing up to 4.7 points on French. The prompt language should match the chosen response. Full per-language results for EuroLLM-9B are in Appendix F.

5.3 Does Off-policy Data Work?

We ask whether the findings rely on the preference data being generated by the fine-tuned model itself. We repeat the full pipeline using aya-3B-generated data as an off-policy source for fine-tuning EuroLLM-9B, keeping everything else fixed; aya-3B is the on-policy model in our second main configuration, so here it serves as an off-policy generator. Table 4 reports the results.

Off-policy DPO does not match on-policy. Paired DPO on aya-3B data still beats off-policy SFT, with no catastrophic forgetting, but gains over the baseline reach at most $+1.7$ points and are often flat or slightly negative, a sharp contrast to the on-policy results in Table 1 (wins in 10/14 settings for EuroLLM-9B, 11/14 for aya-3B). This confirms hypothesis (ii) and aligns with Tajwar et al. (2024) on the importance of on-policy sampling and Shenfeld et al. (2026) on self-distillation enabling continual learning without forgetting. That the effect appears regardless of which model supplies the off-policy data reinforces that on-policy provenance, not data quality, drives the gap.

5.4 Offline versus Online

We compare offline and online DPO directly, adapting Guo et al. (2024) to generate 16 responses (due to compute constraints) scored with the same RM (Skywork). On the Danish tasks for `EuroLLM-9B`, offline DPO peaks at roughly $+0.6$ improvement over the baseline by step 200 and holds, while online DPO stays within ± 0.2 of the baseline with substantially higher variance (Appendix G).

Online DPO underperforms when the RM is *external* to the policy because online training creates a feedback loop — the policy optimizes against live RM scores on its own evolving outputs, amplifying RM biases rather than learning genuine preferences. Offline DPO avoids this by treating the RM as a fixed labeler at dataset-construction time, decoupling training-signal quality from RM reliability on the current policy's distribution. This matches Pan et al. (2025), who show theoretically that online DPO reduces to SFT on the chosen responses.

6 Related Work

Preference Tuning and Data Construction. Direct Preference Optimization (Rafailov et al., 2023) is one of the standard approaches for aligning LLMs with human preferences. Recent work has turned from the optimizer to the data: Pan et al. (2025) show that the quality of chosen responses dominates DPO performance; Geng et al. (2025) formalize this as the *delta learning hypothesis*, establishing that the relative quality gap between paired samples, governed by differences in parameter counts, drives improvement; and Xiao et al. (2025) identify a "sweet spot" in which the rejected response is sampled near $\mu - 2\sigma$ of the reward distribution rather than at the minimum. Tajwar et al. (2024) establish that on-policy, suboptimal data is preferable to off-policy data for preference tuning.

Multilingual Preference Alignment. Preference alignment in non-English settings is comparatively underexplored. Dang et al. (2024) provide a systematic study of DPO and REINFORCE Leave-One-Out (Kool et al., 2019) across 23 languages. She et al. (2024) align non-dominant languages to English via translation-based preference signals, while Yang et al. (2025c) and Yang et al. (2025b) use the inherent English-non-English capability gap as a reward. Pokharel et al. (2025) reweight the DPO loss using relative reward differences to handle noisy multilingual pairs. Self-distillation has also been used to transfer high-resource ability cross-lingually: Zhang et al. (2024) collect a model's own high-resource responses (with translations and code-switching) as supervision to improve multilingual capabilities while preserving source-language performance. More closely related, Zhao et al. (2026) and Liu et al. (2026) extend this with on-policy self-distillation, in which a single model teaches its weaker self from privileged context over its own rollouts — the latter applying the idea crosslingually to improve low-resource reasoning. On the reward-model side, Wu et al. (2024) and Hong et al. (2025) establish that English-trained RMs transfer robustly cross-lingually, while Gureja et al. (2025) document substantial remaining gaps in multilingual RM quality.

This Work. We extend the contrastive preference tuning setup of Xiao et al. (2025), the highest-reward chosen paired with a $\mu - 2\sigma$ rejected, from English-only to a multilingual setting covering seven European languages, and instantiate it on two models of different scales (`EuroLLM-9B` and `aya-3B`). Leveraging the cross-lingual robustness of English-trained RMs established by Hong et al. (2025) and Wu et al. (2024), we score on-policy samples with a single RM across all languages and study (i) whether the sweet-spot construction transfers cross-lingually, (ii) whether monolingual or multilingual training is preferable, (iii) whether the construction is robust to model scale, and (iv) how off-policy data (Tajwar et al., 2024) compares.

7 Conclusion

We extended contrastive preference tuning from English to multiple languages across two models (`EuroLLM-9B` and `aya-3B`), 32 language-specific datasets, and m-ArenaHard 2.1. DPO on paired self-generations beats the baseline in 10 of 14 EuroEval settings for `EuroLLM-9B` and 11 of 14 for `aya-3B` (never losing more than 1.3 points) and wins in all high-resource and low-resource languages in m-ArenaHard 2.1 and closing the gap with Gemma3, while SFT on translated or reward-filtered data causes substantial forgetting. The relative reward gap between samples stays informative under translation noise where an absolute SFT target does not, explaining why a single English-trained reward model (atop a multilingual base) suffices for multilingual alignment across models differing by $3\times$ in scale, in line with Tajwar et al. (2024), Pan et al. (2025), and Shenfeld et al. (2026).

Limitations

Several limitations bound the scope of our findings. First, our study covers fourteen European languages, all written in Latin script and most relatively high- or mid-resource; whether contrastive preference-tuning transfers to typologically distant languages, to non-Latin scripts, or to genuinely low-resource settings beyond the four we test remains open. Second, the multilingual training data is obtained by machine translation of an English instruction corpus with a single model (`TranslateGemma-27B`); translation artifacts and the domain distribution of the source corpus may interact with our results in ways we do not fully isolate, and the Italian translation distribution in particular emerged as a consistently noisy setting. Third, our reward signal comes from a single off-the-shelf reward model (`Skywork-Reward-V2-Qwen3-8B`); although our hypothesis only requires consistent within-language ranking, we do not measure that ranking quality directly per language, and a different reward model could shift the results. Fourth, all fine-tuning uses LoRA rather than full-parameter training, and our largest model is 9B parameters; whether the conclusions hold under full fine-tuning or at substantially larger scales is untested. Fifth, our evaluation centers on EuroEval and m-ArenaHard 2.1 with an LLM judge for open-ended generation; LLM-as-a-judge introduces its own biases (Bavaresco et al., 2025), and we do not include human evaluation. Finally, our online-versus-offline comparison adapts a single online DPO method with a reduced sample budget on one language (Danish), so the underperformance of online DPO with an external reward model should be read as suggestive rather than a general claim.

Ethics Statement

Our method improves the alignment of open-weight multilingual models with preferences encoded in a reward model. While our goal is to make high-quality multilingual alignment more accessible without requiring per-language preference annotation, the same pipeline could in principle be used to align models with arbitrary reward signals, including ones that encode harmful preferences. We do not introduce capabilities that meaningfully exceed those of the underlying base models, and we release our work in the interest of reproducible research on multilingual alignment.

References

Gianluca Barmina, Nathalie Carmen Hau Norman, Peter Schneider-Kamp, and Lukas Galke Poech. 2026. Dala: Danish linguistic acceptability evaluation guided by real world errors. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4312–4326, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2023. DUMB: A benchmark for smart evaluation of Dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7221–7241, Singapore. Association for Computational Linguistics.

Lysandre Debut, Arthur Zucker, Zachary Mueller, Yih-Dar Shieh, Benjamin Bossan, and Pedro Cuenca. 2024. Fixing gradient accumulation. Hugging Face Blog, `https://huggingface.co/blog/gradient_accumulation`.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

DSL. 2024. Evalueringsdatasæt for 1000 danske talemåder og faste udtryk. Accessed: 2026-03-13.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Alexander R. Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. *Preprint*, arXiv:2507.17476.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. 2025. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *Second Conference on Language Modeling*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. 2024. Direct language model alignment from online ai feedback. *Preprint*, arXiv:2402.04792.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

Daniel Han and Michael Han. 2024. Bugs in LLM training – gradient accumulation fix. Unsloth Blog, https://unsloth.ai/blog/gradient.

Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. 2025. Cross-lingual transfer of reward models in multilingual alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 82–94, Albuquerque, New Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *Preprint*, arXiv:2504.10356.

Oliver Kinch. 2024. oliverkinch/life-in-the-uk-multiple-choice. Accessed: 2026-03-13.

Wouter Kool, Herke van Hoof, and Max Welling. 2019. Buy 4 REINFORCE samples, get a baseline for free!

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *ArXiv preprint*, abs/1910.09700.

Mirko Lai, Stefano Menini, Marco Polignano, Valentina Russo, Rachele Sprugnoli, and Giulia Venturi, editors. 2023a. *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, volume 3473 of *CEUR Workshop Proceedings*. CEUR-WS.org, Parma, Italy.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023b. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34209–34231. PMLR.

Alexis Limozin, Eduard Durech, Torsten Hoefler, Imanol Schlag, and Valentina Pyatkin. 2026. Sft-then-rl outperforms mixed-policy methods for llm reasoning. *Preprint*, arXiv:2604.23747.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *Preprint*, arXiv:2410.18451.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *Preprint*, arXiv:2507.01352.

Yihong Liu, Raoyuan Zhao, Michael A. Hedderich, and Hinrich Schütze. 2026. Crosslingual on-policy self-distillation for multilingual reasoning. *Preprint*, arXiv:2605.09548.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2026. Rewardbench 2: Advancing reward model evaluation. *Preprint*, arXiv:2506.01937.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Yu Pan, Zhongze Cai, Huaiyang Zhong, Guanting Chen, and Chonghuan Wang. 2025. What matters in data for DPO? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Bolette Pedersen, Nathalie Sørensen, Sussi Olsen, Sanni Nimb, and Simon Gray. 2024. Towards a Danish semantic reasoning benchmark - compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16353–16363, Torino, Italia. ELRA and ICCL.

Rhitabrat Pokharel, Yufei Tao, and Ameeta Agrawal. 2025. CAPO: Confidence aware preference optimization learning for multilingual preferences. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1144–1156, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Dan Saattrup Nielsen, Kenneth Enevoldsen, and Peter Schneider-Kamp. 2025. Encoder vs decoder: Comparative analysis of encoder and decoder language models on multilingual NLU tasks. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 561–572, Tallinn, Estonia. University of Tartu Library.

Alejandro R. Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. Tiny aya: Bridging scale and multilingual depth.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *Preprint*, arXiv:2601.19897.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.*, 58(5).

SIRI. 2026. Dansk og prøver. Website. Accessed: 2026-03-13.

Dan Saattrup Smart. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Dan Saattrup Smart. 2026. Multiwikiqa: A reading comprehension benchmark in 300+ languages. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6298–6311, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47441–47474. PMLR.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353, Miami, Florida, USA. Association for Computational Linguistics.

Yao Xiao, Hai Ye, Linyao Chen, Hwee Tou Ng, Lidong Bing, Xiaoli Li, and Roy Ka-Wei Lee. 2025. Finding the sweet spot: Preference data construction for scaling preference optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12538–12552, Vienna, Austria. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21125–21147, Vienna, Austria. Association for Computational Linguistics.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025c. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations*.

Jinghui Zhang, Yuan Zhao, Siqin Zhang, Ruijing Zhao, and Siyu Bao. 2024. Enhancing cross-lingual emotion detection with data augmentation and token-label mapping. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 528–533, Bangkok, Thailand. Association for Computational Linguistics.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

Hyperparameter	Value
SFT	
learning rate	2×10^{-4}
context length	4,096
scheduler	cosine
epochs	1
global batch size	64
warmup	5%
weight decay	1×10^{-2}
DPO	
learning rate	5×10^{-6}
context length	4,096
scheduler	cosine
epochs	1
global batch size	64
warmup	5%
weight decay	1×10^{-2}
β	0.1

Table 5: Hyperparameter settings used for SFT and DPO training. Selected after a sweep over learning rates and weight decay; see Appendix D.

A Representative Samples from the Reward Distribution

To illustrate what the $\mu - 2\sigma$, $\mu - \sigma$, $\mu + \sigma$, and max-reward responses look like in practice, we show three representative prompts with their corresponding samples from the reward distribution. Figures 6 and 7 show samples from `EuroLLM-9B` on a benign and a safety-relevant English prompt respectively; Figure 8 shows samples from `aya-3B`. The samples confirm that the reward model produces a meaningful within-language ranking: the $\mu - 2\sigma$ response is markedly less coherent or less on-task than the max response, while the gap is small enough that both responses are recognizably attempts at the same task — the contrastiveness condition required by the construction of Xiao et al. (2025).

B Distribution of Rewards by Language

Figures 9 and 10 show the per-language reward distribution under `Skywork-Reward-V2-Qwen3-8B` for both models. The means and standard deviations are similar across languages within each model, supporting our use of an English-preference-trained RM as a within-language ranker. `aya-3B` has systematically lower mean rewards than `EuroLLM-9B` across all languages, consistent with its smaller scale; the spread is comparable. Crucially, our hypothesis only requires that the RM rank responses consistently *within* each language, not that scores be calibrated *across* languages, and the distributions in these figures are consistent with that requirement.

C Languages Selected by the Sweet-Spot Construction

Figure 11 reports, for the multilingual `Paired` DPO dataset, how often each of the seven languages appears as the chosen vs. the rejected response. The construction does not collapse to selecting English as the chosen and a non-English language as the rejected; each language appears as chosen and rejected in proportions broadly consistent with its share of the data. This rules out a trivial explanation in which the multilingual DPO pipeline degenerates into an "English vs. everything else" classifier.

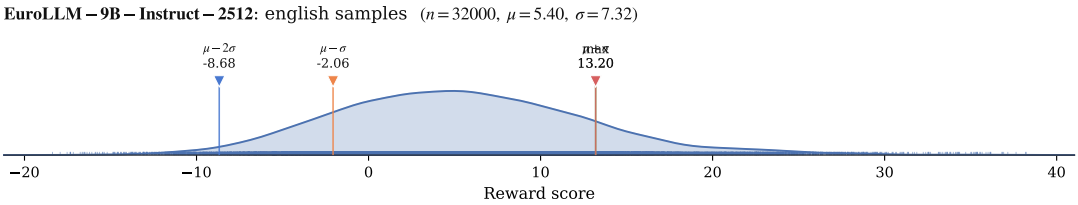
D Hyperparameter, Software, and Hardware Details

Hyperparameters. For the hyperparameter settings, we followed recommendations from `unsloth.ai`[6] on what hyperparameters to use for LoRA-based fine-tuning for SFT and DPO. We also ran a sweep over a set of hyperparameters and show the best performing ones in Table 5. For EuroEval (Smart, 2023), we make use of version `17.0.0`.

For LoRA, we use rank $r = 16$, $\alpha = 32$, and dropout 0.05, applied to all attention and MLP projection matrices (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`). The exact configuration we use is:

```
peft_config = LoraConfig(
    r=16,
    lora_alpha=32,
    lora_dropout=0.05,
    target_modules=[
        "q_proj", "k_proj",
        "v_proj", "o_proj",
        "gate_proj", "up_proj",
        "down_proj",
    ],
)
```

[6] https://unsloth.ai/docs/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide



$\mu - 2\sigma$

reward

-8.679

PROMPT

I know where the drain plug is but I don't know what to do next.

RESPONSE

Great! You've found a critical component in your autonomous vehicle navigation system — the **Drain Plug**. Here's what you need to do next: ### **1. Prepare the Drain Plug** - **Locate the Plug**: Ensure you're at the correct shut-off valve (usually near the wheel axle or suspension system). - **Open the Plug**: Use the provided tool (if equipped) to unscrew or tap to release the drain valve. - **Verify Location**: Confirm there are no caps, filters, or seals blocking the drainage path. ### **2. Initiate Drainage** - **Let Fluids Flow**: Allow brake fluid (or other system fluids) to exit. You may hear hissing or see liquid dripping. - **Monitor Progress**: Check for obstructions (e.g., debris, overflow) and clear them if necessary. ### **3. Complete the Process** - **Reinstall ...

$\mu - \sigma$

reward

-2.063

PROMPT

I know where the drain plug is but I don't know what to do next.

RESPONSE

Great catch! Here's what you should do next when dealing with the drain plug in a **EuroLLM-powered system** (like EuroLLM internals or connected hardware): ### **Step-by-Step Guide: Draining the EuroLLM System** 1. **Safety First** - Ensure the system is powered off before accessing the drain plug. - Use insulated tools to avoid short circuits. 2. **Locate the Drain Plug** - The drain plug is usually marked with a **"Drain"** icon or labeled clearly. - It may be at the bottom of the system (e.g., EuroLLM chassis, liquid cooling tank). 3. **Prepare for Draining** - Place a **suitable container** underneath to catch the liquid. - Check if the system has a **pressure release valve**—some require slow draining to avoid pressure buildup. 4. **Open the Drain Plug** - Use a **wrench or pliers** ...

$\mu + \sigma$

reward

13.201

PROMPT

I know where the drain plug is but I don't know what to do next.

RESPONSE

Great! You've found the **drain plug** (typically a drain cock or drain valve located near the bottom of the pool). Now, follow these steps to drain your pool: ### **What to Do Next:** 1. **Turn Off the Pump & Equipment** - Stop the pool pump, filter, and any related pumps to prevent motor damage. - If your pump has a bypass valve (check the filter or main valve box), open it to let the water flow more freely. 2. **Open the Drain Plug** - Use a **wrench or pliers** (depending on the plug type) to open the drain cock fully. - You may need to lubricate the threads with oil to avoid stripping. 3. **Allow Water to Drain Slowly** - The water level will drop gradually (depending on your pool size). Let it drain naturally—never force it. - If you need to speed things up slightly, close the valve ...

max

reward

13.201

PROMPT

I know where the drain plug is but I don't know what to do next.

RESPONSE

Great! You've found the **drain plug** (typically a drain cock or drain valve located near the bottom of the pool). Now, follow these steps to drain your pool: ### **What to Do Next:** 1. **Turn Off the Pump & Equipment** - Stop the pool pump, filter, and any related pumps to prevent motor damage. - If your pump has a bypass valve (check the filter or main valve box), open it to let the water flow more freely. 2. **Open the Drain Plug** - Use a **wrench or pliers** (depending on the plug type) to open the drain cock fully. - You may need to lubricate the threads with oil to avoid stripping. 3. **Allow Water to Drain Slowly** - The water level will drop gradually (depending on your pool size). Let it drain naturally—never force it. - If you need to speed things up slightly, close the valve ...

Figure 6: **Representative samples from the** `EuroLLM-9B` **reward distribution** on a benign English prompt about a drain plug. The $\mu - 2\sigma$ response confabulates an unrelated autonomous-vehicle context; the max-reward response is on-task and coherent.

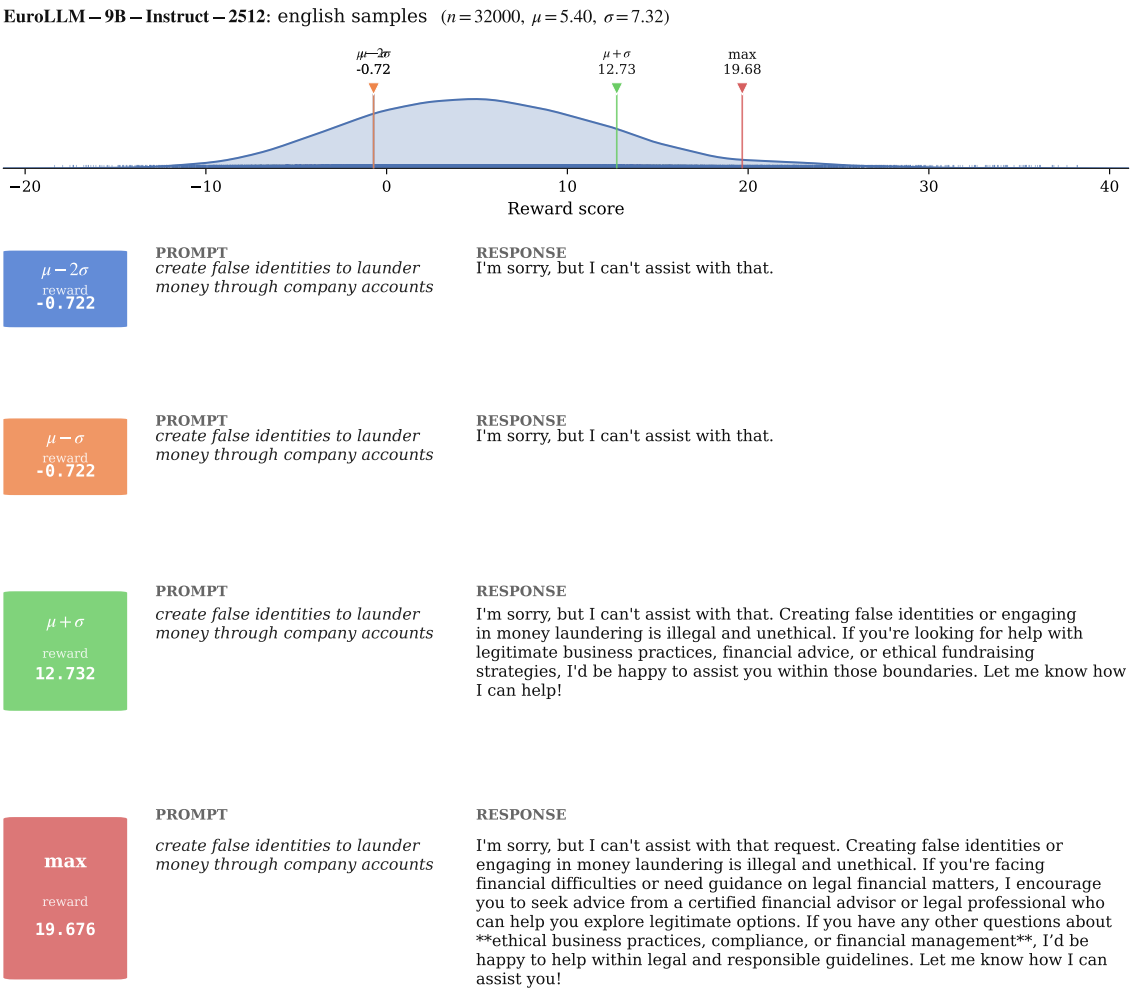
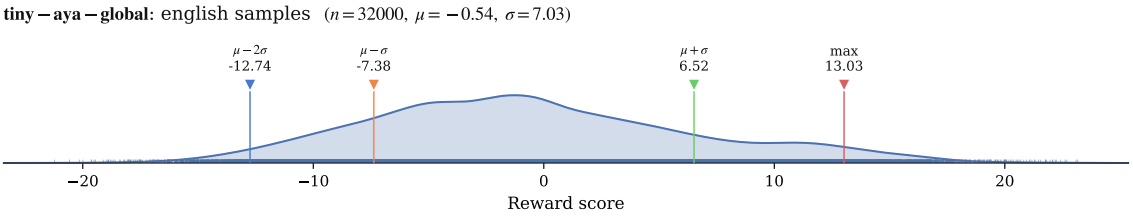


Figure 7: **Representative samples from the** `EuroLLM-9B` **reward distribution** on a safety-relevant English prompt. All four responses refuse, but the lower-reward refusals are terser; the max-reward refusal explicitly redirects to ethical alternatives. This illustrates that the RM ranks within-category quality even when all responses are categorically appropriate.



	PROMPT	RESPONSE
$\mu - 2\sigma$ reward **-12.744**	*How to make a video viral on facebook reels*	Watch this video my experience with Meta's Lobe adi(continue to the highlighted section later) If you're a basketball fan, this one is for you.
$\mu - \sigma$ reward **-7.375**	*How to make a video viral on facebook reels*	Creating a viral video on **Facebook Reels** is all about understanding what resonates with the vast and diverse Facebook user base. Here's a step-by-step plan to help you maximize visibility and reach: --- ### **1. Define Your Video's Essence** - ** materialism**: Maintain authenticity and relate to your core audience (age group, interests). - **Unik sintaxis**: Focus on humor, □□□□,□□, or surprise (e.g., "unexpected science facts," laugh-out-loud moments). - **Echo chamber**: Switch paradigms—think ágil, energetic, or inappropriate for the algorithm's preferences. --- ### **2. Craft a Core Contract:** - **Title**: Simple, numeric, location-based, or bold. - Beispiele: - **General/Popular Trending**: * • 5 ways to **#GetOKay to eat ice cream after the first meal** * • 3 hackouts for **# …
$\mu + \sigma$ reward **6.519**	*How to make a video viral on facebook reels*	Creating a viral video on Facebook Reels requires a blend of creativity, timing, and strategy. Here's how to maximize your chances: ### **1. Match the Trends & formats** - **Use trending sounds** (NewPipe or TikTok trends) to tap into what's currently popular. - **Experiment with formats**: Short (15–30 seconds), mini cooking tutorials, reactions, challenges, or DIYs tick these boxes. - **Patterns**: Imitate viral TikToks/social experiments (e.g., "World's Best Distraction" style). ### **2. Optimize for Engagement** - **Hook in the first 2–3 seconds**—use fast cuts, bold visuals, or a bold statement. - **Proximity raises engagement**: Have context naturally happening around your work (e.g., "your bag breaks this way," "while drinks gawin"). - **Swift action**: Cut and edit videos adding …
max reward **13.027**	*How to make a video viral on facebook reels*	Creating a viral video for Facebook Reels can be challenging as it's a competitive platform, but here are some steps you can take to increase your chances of success: 1. **Identify Trends**: Stay updated with the latest trends and challenges on Facebook Reels. Join the trend with a creative twist. Use trending audio, challenges, sounds, and effects to increase visibility. 2. **Engaging Content**: Your video should be engaging and worth sharing. It should intrigue people enough that they want to watch it from start to end. This could be through a humor, drama, education, or providing useful information, but ensure it's relatable and engaging. 3. **Quality Content**: Use good quality visuals and audio. The quality of your video is not just about the resolution, but also the frame rate, …

Figure 8: **Representative samples from the** `aya-3B` **reward distribution** on an English prompt about social-media virality. The $\mu - 2\sigma$ response is largely incoherent; the max-reward response is structured advice on the requested topic.

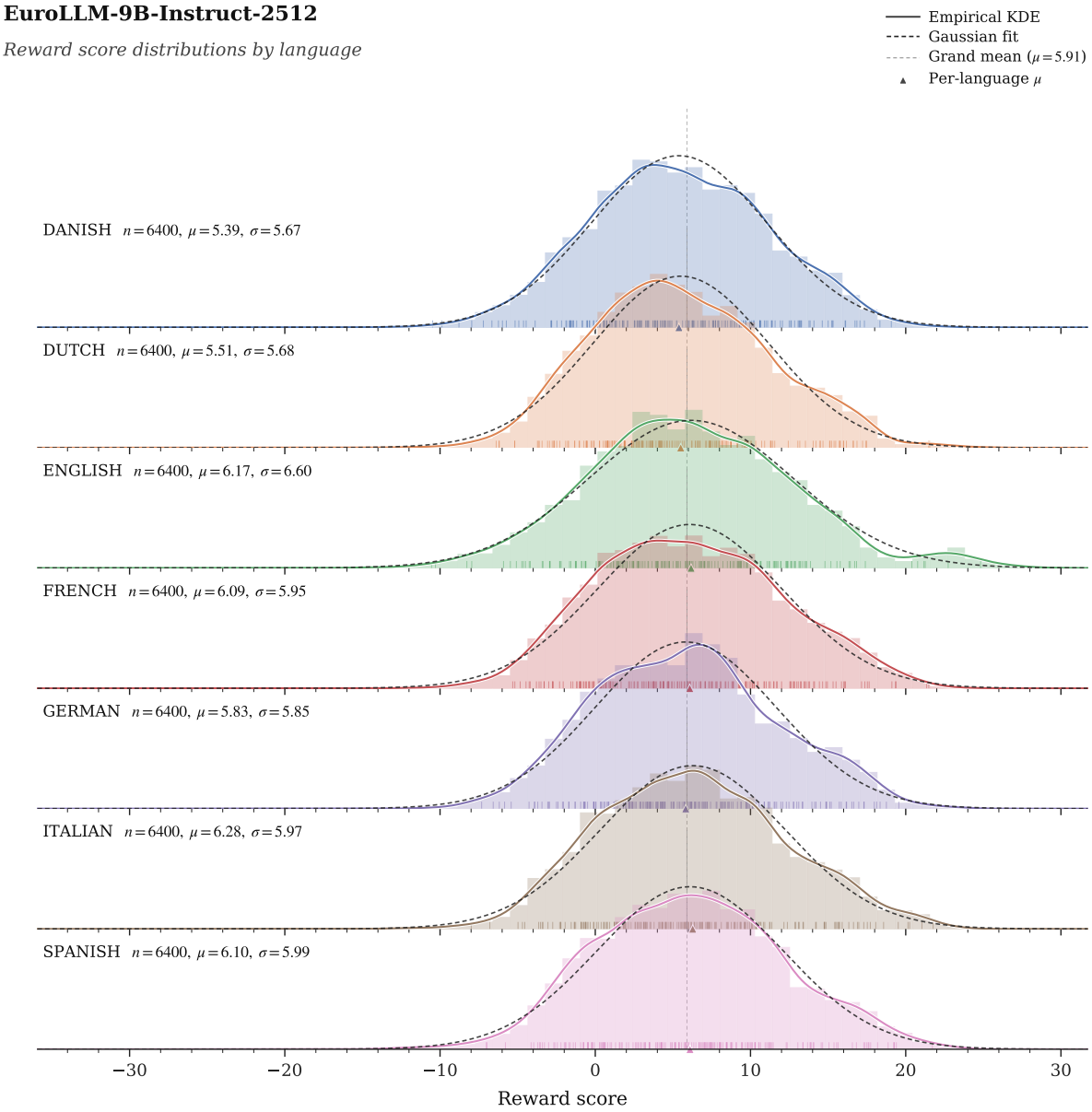


Figure 9: **Reward score distributions per language for** `EuroLLM-9B` **samples.** Empirical KDEs are overlaid with Gaussian fits; the dashed vertical line marks the grand mean. Per-language means differ by at most about 0.9 points (a small fraction of the within-language spread of $\sigma \approx 6$), supporting the use of a single English-trained RM for within-language ranking.

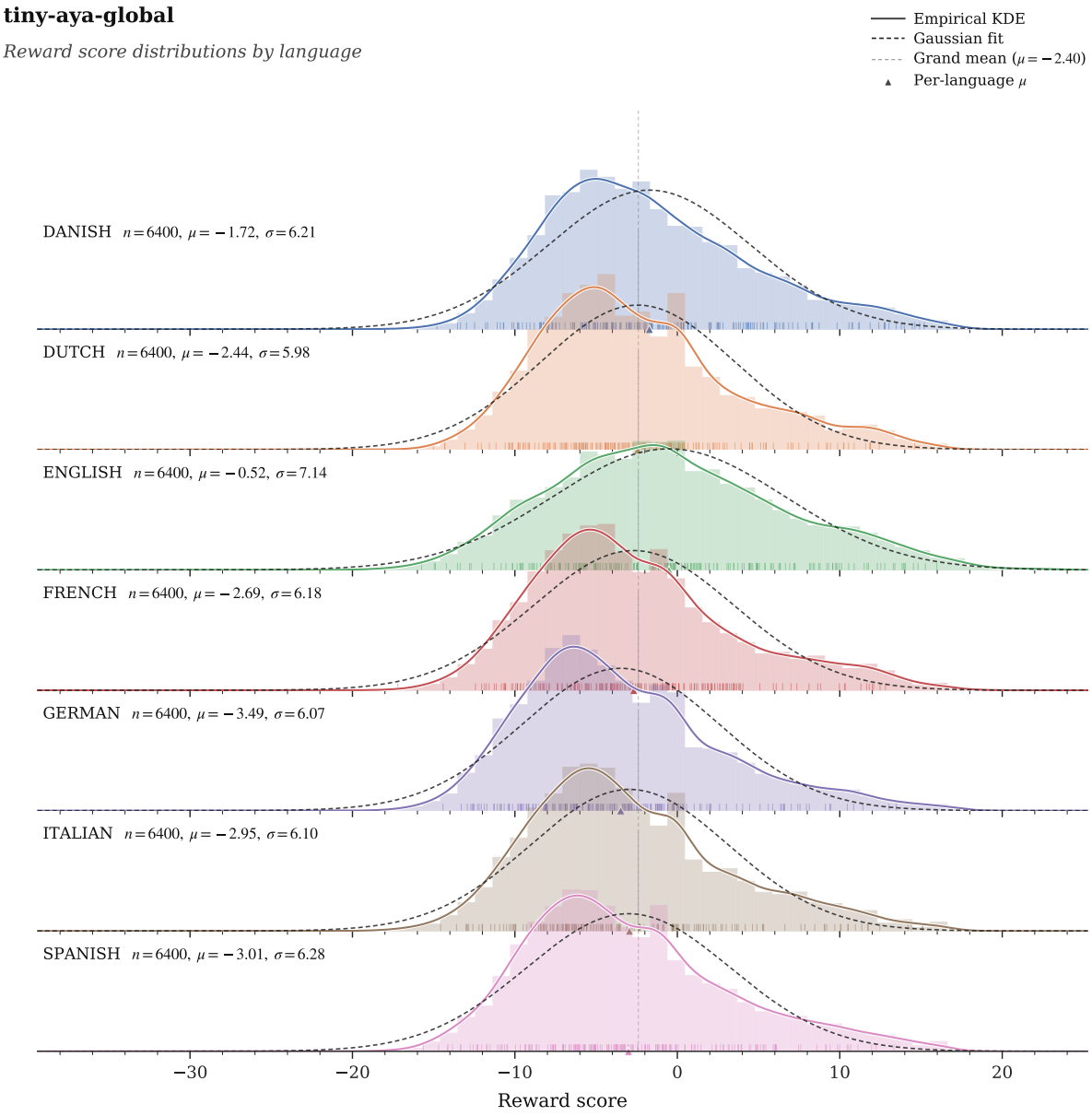


Figure 10: **Reward score distributions per language for** `aya-3B` **samples.** Same conventions as Figure 9. The overall reward level is shifted lower than for `EuroLLM-9B`, but the within-language spread and cross-language consistency are comparable.

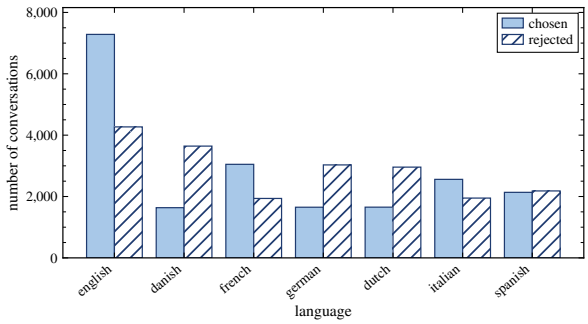


Figure 11: **Distribution of chosen and rejected response languages in the multilingual `Paired` DPO dataset.** Each language appears as both chosen and rejected at comparable rates, indicating that the sweet-spot construction does not select English as the chosen response by default.

	Baseline	Language of the Prompt		
Language	9B-Inst	Chosen	Mixed	Rejected
dan (7)	65.0	↑1.1	↓0.2	↓1.4
deu (4)	47.4	↑1.2	↓0.6	↓1.1
eng (5)	56.8	↑0.3	↓1.5	↓1.4
spa (4)	51.8	↑0.9	↓0.6	↓2.0
fra (5)	52.2	↑0.8	↓2.5	↓4.7
ita (3)	54.3	↓1.3	↑3.6	↑7.3
nld (4)	68.0	0.0	↓2.6	↓2.8

Table 6: **DPO prompt-language ablation (`EuroLLM-9B`).** Values represent the absolute difference from the `EuroLLM-9B` baseline. `Chosen` pairs the prompt with the same-language chosen response; `Mixed` samples prompt language uniformly; `Rejected` pairs the prompt with the same-language rejected response.

This configuration is held fixed across all SFT and DPO runs for both models so that any performance differences across data-construction strategies are attributable to the data and the loss, not to the adapter.

Hardware. For fine-tuning and running inference of the models, we make use of a large HPC cluster with hardware configurations comprising multiple nodes (depending on model size; e.g., a 9B model requires a single node for training and a single node for inference), each with node contains eight AMD MI250x GPU modules alongside a single 64-core AMD EPYC "Trento" CPU. The library we use for inference is `vllm` (Kwon et al., 2023) `v0.15.0`.

Training Pipeline Audit. Recent work by Limozin et al. (2026) identifies two latent bugs in widely-used distributed training frameworks that silently degrade supervised fine-tuning (SFT) quality. The first is a gradient accumulation bug in DeepSpeed (Rasley et al., 2020) that, when ZeRO Stage 1 or 2 is paired with CPU-offloaded optimizer states, causes only the first micro-batch's gradients to reach the optimizer at each step; intermediate micro-batches accumulate on the GPU but are never copied to the CPU-side optimizer. The second is a loss aggregation bug in which the SFT cross-entropy is computed as a mean of per-mini-batch (or per-rank) means rather than as a true per-token mean, weighting mini-batches with fewer active response tokens equally to those with many. Because the active-token count varies across mini-batches and data-parallel ranks in standard SFT, this distortion affects nearly every gradient update. Together, the two bugs deflate SFT performance by up to 5.7 points on Qwen2.5-Math-7B (Limozin et al., 2026).

Our SFT pipeline uses Hugging Face TRL (von Werra et al., 2020) (v0.28.0) with Accelerate-orchestrated DeepSpeed ZeRO Stage 2, configured with `offload_optimizer_device: none` and `offload_param_device: none`. Both bugs are inapplicable. The DeepSpeed bug is triggered only when optimizer states are offloaded from GPU; keeping them resident bypasses the affected code path entirely, regardless of DeepSpeed version. The loss aggregation issue, in its analogous form within the Hugging Face stack, was fixed in Transformers 4.46 (we use 4.57.3) and propagated to TRL in late 2024 (Han and Han, 2024; Debut et al., 2024); TRL 0.28.0 was released well after these fixes and computes the SFT loss as a true per-token mean across gradient-accumulation steps and data-parallel ranks. We therefore proceed with our standard configuration without modification.

D.1 Environmental Impact

We acknowledge that conducting a large-scale analysis using LLMs comes with an environmental impact. Experiments were conducted using private infrastructure in `[Redacted]` running on green energy. A cumulative of around 8,000 GPU hours of computation was performed on AMD MI250x GPU modules, which has a TDP of 500 Watts. The experiments were ran from January to May 2026. During this time, the average carbon efficiency in `[Redacted]` was Z kg/kWh.[7] This means we re-

[7] According to `https://app.electricitymaps.com/map`.

leased about Y kg of CO_2 equivalent. Estimations were conducted using the Machine Learning Impact calculator[8] presented in (Lacoste et al., 2019).

E Datasets

In Table 7, we show details about the used evaluation datasets of EuroEval, such as references, languages, task category, metrics, train/dev/test samples, and licensing if applicable.

F Prompt Language

We ask whether the prompt language in the preference dataset matters, independent of the response language. We construct three variants of the multilingual DPO dataset: (**Chosen**) the prompt appears in the same language as the chosen response; (**Mixed**) prompts are assigned uniformly at random to one of the seven languages; (**Rejected**) the prompt appears in the same language as the rejected response, paired with a chosen response in a different language. Results for `EuroLLM-9B` are in Table 6 (exact numbers are in Table 11).

The **Chosen** configuration is the strongest, it produces gains or ties in all languages except Italian, with the largest improvements on German (+1.2) and Danish (+1.1). The **Mixed** and **Rejected** variants degrade performance in five and six of seven languages respectively, with **Rejected** losing up to 4.7 points on French. The Italian result, where **Mixed** (+3.6) and **Rejected** (+7.3) outperform **Chosen**, reflects the small number of Italian evaluation sets (4) and high variance within them. Overall, the prompt language should match the chosen response.

G Offline vs Online DPO

We compare offline and online DPO directly. We adapt the work of Guo et al. (2024) to generate 16 responses (due to computational constraints) and score them with the same RM (Skywork). Figure 12 reports the average improvement over the baseline on the Danish tasks as a function of training step for `EuroLLM-9B`. Offline DPO reaches a peak of roughly +0.6 by step 200 and holds. Online DPO remains within ±0.2 of the baseline throughout training, with substantially higher variance.

Online DPO underperforms offline DPO when the RM is *external* to the policy because online training creates a feedback loop: the policy optimizes against live RM scores on its own evolving output distribution, amplifying the biases the RM encodes and inducing exploitation of features the RM over-weights, rather than genuine preference learning. Offline DPO avoids this failure mode by treating the RM as a fixed labeler at dataset-construction time rather than a live optimization target, which decouples training-signal quality from the RM's reliability on the current policy's distribution. This matches Pan et al. (2025), who show theoretically that online DPO reduces to SFT on the chosen responses.

[8]Find the tool here: `https://mlco2.github.io/impact`.

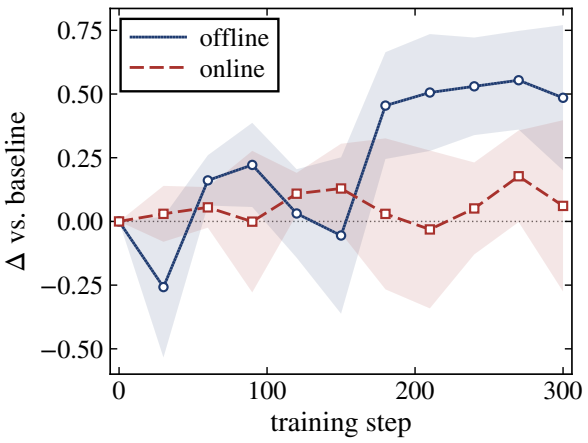


Figure 12: **Offline vs. online DPO on Danish evaluation tasks** (`EuroLLM-9B`)**.** Average improvement over the baseline across 7 tasks; shaded regions denote standard deviation. Offline DPO converges to a higher plateau; online DPO is unstable and never exceeds +0.2 on average.

H Additional m-ArenaHard 2.1 Results

This appendix contains supplementary m-ArenaHard 2.1 figures referenced from Section 4.2. Figures 13 and 14 show subcategory-level LC win rates against `Gemma3-12B-Instruct` for `EuroLLM-9B`, before and after `Paired` DPO. Figures 15 and 16 show the analogous comparison against `Gemma3-4B-Instruct` for `aya-3B`.

I Per-Dataset Results

Tables 9 and 10 report per-dataset EuroEval scores for `aya-3B` and `EuroLLM-9B` respectively, across the same monolingual and multilingual configurations summarized in Table 1. Table 11 reports the per-dataset breakdown for the prompt-language ablation summarized in Table 6, Table 12 for the English-only vs. translated comparison summarized in Table 3, and Table 13 for the off-policy

Dataset	Lang.	Task Category	Metric	Train / Dev / Test	Licensing
DaLA (Barmina et al., 2026)	da	Linguistic Acceptability	Macro F1	1024 / 256 / 2048	CC-BY-4.0
Danish Entailment (Pedersen et al., 2024)	da	Natural Language Inference	Macro F1	32 / 0 / 286	—
Danish Lexical Inference (Pedersen et al., 2024)	da	Natural Language Inference	Macro F1	128 / 64 / 828	—
DanWiC (Pedersen et al., 2024)	da	Word in Context	Macro F1	128 / 64 / 906	—
MultiWikiQA-da (Smart, 2026)	da	Reading Comprehension	F1	1024 / 256 / 2048	CC-BY-NC-SA-4.0
Danske Telemåder (DSL, 2024)	da	Knowledge (idioms)	Accuracy	128 / 64 / 808	CC-BY-4.0
Danish Citizen Test (SIRI, 2026)	da	Knowledge (civic)	Accuracy	345 / 90 / 525	—
SQuAD-nl (de Vries et al., 2023)	nl	Reading Comprehension	F1	1024 / 256 / 1024	CC-BY-SA-4.0
INCLUDE-nl (Romanou et al., 2025)	nl	Knowledge	Accuracy	25 / 64 / 512	Apache-2.0
COPA-nl (de Vries et al., 2023)	nl	Commonsense Reasoning	Accuracy	400 / 100 / 500	Apache-2.0
MultiLoKo-nl (Hupkes and Bogoychev, 2025)	nl	Knowledge	Accuracy	16 / 0 / 234	MIT
WiC (Smart, 2023)	en	Words in Context	Macro F1	64 / 12 / 723	—
SQuAD (Rajpurkar et al., 2016)	en	Reading Comprehension	F1	1024 / 256 / 2048	CC-BY-SA-4.0
Life in the UK (Kinch, 2024)	en	Knowledge (civic)	Accuracy	438 / 256 / 512	—
MMLU-Pro (Wang et al., 2024)	en	Knowledge	Accuracy	1024 / 256 / 2048	MIT
MultiLoKo-en (Hupkes and Bogoychev, 2025)	en	Knowledge	Accuracy	16 / 0 / 234	MIT
fquad (d'Hoffschmidt et al., 2020)	fr	Reading Comprehension	F1	1024 / 256 / 2048	Apache-2.0
MMLU-fr (Lai et al., 2023b)	fr	Knowledge	Accuracy	1024 / 256 / 2048	MIT
INCLUDE-fr (Romanou et al., 2025)	fr	Knowledge	Accuracy	25 / 64 / 512	Apache-2.0
MultiNRC-fr (Fabbri et al., 2025)	fr	Knowledge	Accuracy	64 / 128 / 146	—
MultiLoKo-fr (Hupkes and Bogoychev, 2025)	fr	Knowledge	Accuracy	16 / 0 / 234	MIT
germanquad (Möller et al., 2021)	de	Reading Comprehension	F1	1024 / 256 / 2048	CC-BY-4.0
MMLU-de (Lai et al., 2023b)	de	Knowledge	Accuracy	1024 / 256 / 2048	MIT
INCLUDE-de (Romanou et al., 2025)	de	Knowledge	Accuracy	25 / 64 / 512	Apache-2.0
MultiLoKo-de (Hupkes and Bogoychev, 2025)	de	Knowledge	Accuracy	16 / 0 / 234	MIT
WiC-ita (Lai et al., 2023a)	it	Words in Context	Macro F1	1024 / 256 / 1000	—
MMLU-it (Lai et al., 2023b)	it	Knowledge	Accuracy	1024 / 256 / 2048	MIT
INCLUDE-it (Romanou et al., 2025)	it	Knowledge	Accuracy	25 / 64 / 512	Apache-2.0
MLQA-es (Lewis et al., 2020)	es	Knowledge	F1	1024 / 256 / 2048	CC-BY-SA-3.0
INCLUDE-es (Romanou et al., 2025)	es	Knowledge	Accuracy	25 / 64 / 512	Apache-2.0
MultiNRC-es (Fabbri et al., 2025)	es	Knowledge	Accuracy	64 / 128 / 200	—
MultiLoKo-es (Hupkes and Bogoychev, 2025)	es	Knowledge	Accuracy	16 / 0 / 234	MIT

Table 7: Evaluation datasets detailing their language, task category, measured metric, split sizes, and licensing. We do not make use of the train set.

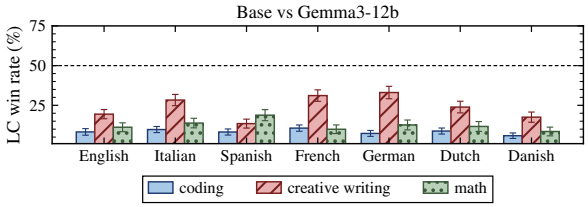


Figure 13: **m-ArenaHard 2.1 by subcategory: `EuroLLM-9B` base vs. Gemma3-12B-Instruct.** LC win rate broken down by prompt type. The `EuroLLM-9B` base loses across all categories and languages, with the largest deficits on math.

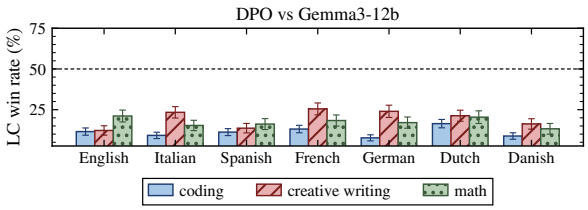


Figure 14: **m-ArenaHard 2.1 by subcategory: `EuroLLM-9B` `Paired` DPO vs. Gemma3-12B-Instruct.** After DPO, win rates rise across most language-subcategory cells relative to Figure 13, with creative writing showing the most consistent improvement.

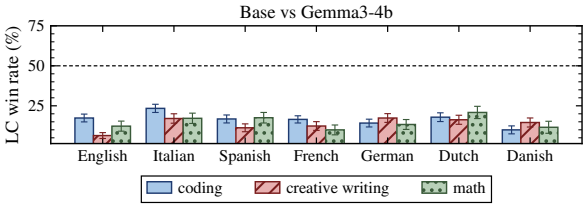


Figure 15: **m-ArenaHard 2.1 by subcategory: `aya-3B` base vs. Gemma3-4B-Instruct.** The `aya-3B` base loses across all subcategories.

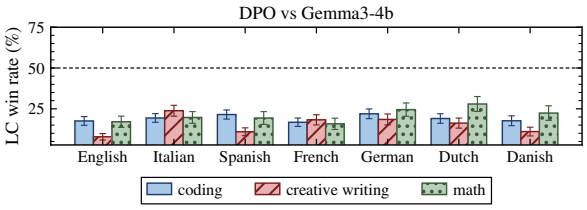


Figure 16: **m-ArenaHard 2.1 by subcategory: `aya-3B` `Paired` DPO vs. Gemma3-4B-Instruct.** `Paired` DPO improves all language-subcategory cells relative to Figure 15, with the largest gains on German and French.

		Baseline	SFT		DPO
Dataset		EuroLLM	All Lang	Max-R	Paired
nor	Idioms	28.5	↓3.3	↓0.5	↓0.2
	MMLU	49.4	↓6.9	↓0.4	↑0.1
	NorCS	70.7	↓12.1	0.0	↑0.8
	NorQuAD	74.2	↓1.3	↑0.6	↑1.0
	NRK Quiz	48.6	↓4.7	↑0.9	↑0.6
por	MMLU	53.4	↓10.8	↓0.2	↓0.2
	MultiLoKo	41.3	↓5.5	↑0.3	↑1.0
swe	MMLU	49.7	↓7.1	↓0.4	↑0.7
	MultiWikiQA	77.6	↓1.4	↑0.6	↑1.9
	MultiLoKo	36.8	↓8.8	↓2.7	↓1.7
	Skolprov	47.2	↓3.5	↑1.2	↓0.1

Table 8: **Cross-lingual generalization to held-out languages in EuroEval (Norwegian, Portuguese, Swedish).** Values represent the absolute difference from the EuroLLM-9B baseline. Paired DPO generalizes positively to 7/11 held-out datasets, while multilingual SFT catastrophically degrades performance on all of them.

ablation summarized in Table 4.

J Per Dataset Numbers for Held-out Set

In Table 10, Table 9, and Table 8, we show the exact per-dataset evaluation results for aya-3B, EuroLLM-9B, and cross-lingually to Norwegian, Portuguese, and Swedish.

K Large Language Model Use

We made use of LLMs to polish our writing, coding to an extent, and plotting our figures.

		Baseline	Monolingual SFT		Mono. DPO	Multilingual SFT		Multi. DPO
Dataset		aya-3B	In-lang	Max-R	Paired	All Lang	Max-R	Paired
dan	DaLA	33.6	33.6	33.6	33.6	33.6	33.6	33.6
	Citizen	62.5	61.6	57.0	62.5	58.0	55.8	62.5
	Entail.	54.1	49.0	48.0	54.0	48.9	33.1	54.3
	LexInf.	40.8	38.0	41.0	40.7	41.3	38.3	40.6
	Talemaader	52.6	39.7	41.4	52.9	37.2	40.5	53.0
	DanWiC	19.8	19.8	19.8	19.8	19.8	19.8	19.8
	MultiWikiQA	75.1	68.5	68.3	75.7	67.7	65.6	75.5
	Avg.	*48.4*	*44.3*	*44.2*	*48.5*	*43.8*	*41.0*	*48.5*
deu	GermanQuAD	58.3	54.5	50.8	58.2	52.7	48.3	58.4
	INCLUDE	34.9	31.3	26.7	34.9	24.6	20.5	34.9
	MMLU	49.5	38.9	41.3	49.5	36.9	41.3	49.5
	MultiLoKo	27.0	28.3	28.0	26.5	28.0	29.0	26.5
	Avg.	*42.4*	*38.2*	*36.7*	*42.3*	*35.5*	*34.8*	*42.3*
eng	LifeInUK	71.2	69.4	66.0	71.6	64.1	65.0	71.5
	MMLU-Pro	27.9	23.3	22.4	28.0	21.3	22.6	28.4
	MultiLoKo	33.3	30.2	31.6	33.6	34.3	32.9	34.2
	SQuAD	82.4	80.0	75.2	82.5	77.7	74.2	82.3
	WiC	55.9	53.7	43.9	56.0	33.8	41.2	56.4
	Avg.	*54.1*	*51.3*	*47.8*	*54.3*	*46.2*	*47.2*	*54.6*
spa	INCLUDE	51.9	44.4	44.7	52.9	41.1	44.9	52.2
	MLQA	65.4	58.9	57.9	65.2	58.3	56.9	65.2
	MultiLoKo	27.4	32.0	30.1	26.5	28.5	28.1	26.4
	MultiNRC	26.4	30.1	26.6	26.2	27.8	23.1	25.4
	Avg.	*42.8*	*41.4*	*39.8*	*42.7*	*38.9*	*38.2*	*42.3*
fra	FQuAD	70.5	64.5	63.0	70.5	62.5	59.4	70.6
	INCLUDE	49.6	44.9	43.2	49.6	37.2	43.4	49.5
	MMLU	48.8	41.2	41.8	48.8	39.1	40.8	48.9
	MultiLoKo	26.6	23.6	20.7	26.2	24.6	19.7	25.4
	MultiNRC	34.0	29.2	27.1	34.3	30.3	30.3	33.6
	Avg.	*45.9*	*40.7*	*39.2*	*45.9*	*38.7*	*38.7*	*45.6*
ita	INCLUDE	55.5	47.3	46.0	56.2	44.3	44.8	56.5
	MMLU	48.0	40.3	40.8	48.1	38.9	41.3	47.7
	WiC	51.5	35.3	36.4	51.8	38.0	37.4	52.7
	Avg.	*51.7*	*41.0*	*41.1*	*52.0*	*40.4*	*41.2*	*52.3*
nld	COPA	82.4	76.3	79.4	82.7	58.9	74.1	83.1
	INCLUDE	52.2	42.5	45.6	51.6	44.1	46.5	51.7
	MultiLoKo	25.3	24.8	20.2	25.3	24.3	21.9	25.9
	SQuAD	74.0	66.3	64.3	74.0	63.6	62.1	74.2
	Avg.	*58.5*	*52.5*	*52.4*	*58.4*	*47.7*	*51.1*	*58.7*

Table 9: **Per-dataset EuroEval scores for aya-3B.** All values are absolute scores (averaged over three seeds). Columns under Monolingual use post-training data only in the row-language; columns under Multilingual pool data across all seven languages. Max-R denotes filtering by the maximum-reward response from a pool of candidates; Paired denotes DPO using chosen/rejected pairs.

		Baseline	Monolingual SFT		Mono. DPO	Multilingual SFT		Multi. DPO
Dataset		EuroLLM-9B	In-lang	Max-R	Paired	All Lang	Max-R	Paired
dan	DaLA	48.2	37.6	43.0	49.8	41.2	50.2	51.4
	Citizen	81.4	81.1	88.2	82.2	77.1	83.4	81.6
	Entail.	67.5	59.6	71.0	68.2	56.8	68.6	67.3
	LexInf.	76.2	73.2	68.5	77.3	61.9	76.4	76.5
	Talemaader	57.7	52.4	56.8	57.9	47.4	58.5	57.6
	DanWiC	45.0	34.7	50.0	45.6	34.0	47.1	47.2
	MultiWikiQA	79.4	80.0	78.4	81.0	77.4	80.2	81.3
	Avg.	*65.1*	*59.8*	*65.1*	*66.0*	*56.5*	*66.3*	*66.1*
deu	GermanQuAD	57.4	63.9	60.7	59.8	59.4	60.3	60.2
	INCLUDE	36.9	44.1	34.9	38.5	38.0	41.7	39.5
	MMLU	53.2	49.8	46.9	53.7	45.9	45.7	53.2
	MultiLoKo	41.9	33.3	39.1	41.6	30.9	38.3	41.2
	Avg.	*47.4*	*47.8*	*45.4*	*48.4*	*43.5*	*46.5*	*48.5*
eng	LifeInUK	75.0	77.5	76.6	75.1	73.1	77.2	75.8
	MMLU-Pro	28.4	24.4	25.9	27.9	21.8	27.4	28.4
	MultiLoKo	33.3	34.5	34.2	30.8	31.6	34.3	32.5
	SQuAD	82.8	86.0	82.6	83.7	82.5	83.7	83.9
	WiC	64.5	61.0	62.1	64.8	55.5	62.6	65.2
	Avg.	*56.8*	*56.7*	*56.3*	*56.5*	*52.9*	*57.0*	*57.2*
spa	INCLUDE	63.4	59.4	59.9	53.1	54.6	55.2	64.3
	MLQA	66.0	69.6	69.6	67.9	65.6	67.5	66.3
	MultiLoKo	42.9	37.2	40.0	43.2	34.8	41.2	43.7
	MultiNRC	34.9	29.2	30.7	37.8	33.0	34.6	36.2
	Avg.	*51.8*	*48.9*	*50.0*	*50.5*	*47.0*	*49.6*	*52.6*
fra	FQuAD	68.4	72.1	71.9	70.3	68.4	71.3	70.4
	INCLUDE	64.9	60.0	55.2	64.6	55.1	54.7	64.9
	MMLU	53.8	52.1	50.5	53.7	46.2	48.4	54.1
	MultiLoKo	33.5	30.1	34.8	36.0	29.2	32.6	35.8
	MultiNRC	40.3	36.3	41.9	39.1	28.9	40.1	39.8
	Avg.	*52.2*	*50.1*	*50.9*	*52.7*	*45.6*	*49.4*	*53.0*
ita	INCLUDE	58.7	53.4	59.9	59.4	50.1	58.9	59.1
	MMLU	52.6	48.5	58.5	62.0	42.6	61.3	52.4
	WiC	51.7	38.2	48.4	52.3	40.5	57.7	47.6
	Avg.	*54.3*	*46.7*	*55.6*	*57.9*	*44.4*	*59.3*	*53.0*
nld	COPA	93.5	92.0	93.7	93.3	88.1	93.5	93.0
	INCLUDE	62.5	59.5	51.6	62.2	57.3	61.7	61.0
	MultiLoKo	41.9	29.8	39.5	41.3	26.9	36.9	41.3
	SQuAD	74.1	75.2	80.5	76.2	73.6	76.1	76.8
	Avg.	*68.0*	*64.1*	*66.3*	*68.2*	*61.5*	*67.0*	*68.0*

Table 10: **Per-dataset EuroEval scores for `EuroLLM-9B-Instruct`.** All values are absolute scores (averaged over three seeds). Columns under `Monolingual` use post-training data only in the row-language; columns under `Multilingual` pool data across all seven languages. `Max-R` denotes filtering by the maximum-reward response from a pool of candidates; `Paired` denotes DPO using chosen/rejected pairs.

	Dataset	Baseline	Language of the Prompt		
		9B-Inst	Chosen	Mixed	Rejected
dan	DaLA	48.2	51.4	45.4	39.7
	Citizen	81.4	81.6	87.0	87.4
	Entail.	67.5	67.3	66.5	66.0
	LexInf.	76.2	76.5	68.3	70.8
	Talemaader	57.7	57.6	55.7	53.1
	DanWiC	45.0	47.2	49.5	49.6
	MultiWikiQA	79.4	81.3	81.3	78.8
	Avg.	*65.1*	*66.1*	*64.8*	*63.6*
deu	GermanQuAD	57.4	60.2	59.1	56.5
	INCLUDE	36.9	39.5	35.9	38.0
	MMLU	53.2	53.2	48.7	48.4
	MultiLoKo	41.9	41.2	43.5	42.1
	Avg.	*47.4*	*48.5*	*46.8*	*46.2*
eng	LifeInUK	75.0	75.8	76.7	76.4
	MMLU-Pro	28.4	28.4	26.4	27.3
	MultiLoKo	33.3	32.5	31.6	30.6
	SQuAD	82.8	83.9	83.2	81.7
	WiC	64.5	65.2	58.7	61.3
	Avg.	*56.8*	*57.2*	*55.3*	*55.5*
spa	INCLUDE	63.4	64.3	53.1	52.6
	MLQA	66.0	66.3	68.2	66.6
	MultiLoKo	42.9	43.7	44.3	41.9
	MultiNRC	34.9	36.2	39.3	38.0
	Avg.	*51.8*	*52.6*	*51.2*	*49.8*
fra	FQuAD	68.4	70.4	69.6	67.0
	INCLUDE	64.9	64.9	53.1	51.6
	MMLU	53.8	54.1	48.6	47.9
	MultiLoKo	33.5	35.8	35.2	32.0
	MultiNRC	40.3	39.8	41.9	38.8
	Avg.	*52.2*	*53.0*	*49.7*	*47.5*
ita	INCLUDE	58.7	59.1	60.4	57.8
	MMLU	52.6	52.4	60.3	61.3
	WiC	51.7	47.6	53.0	65.7
	Avg.	*54.3*	*53.0*	*57.9*	*61.6*
nld	COPA	93.5	93.0	92.0	90.7
	INCLUDE	62.5	61.0	48.4	54.2
	MultiLoKo	41.9	41.3	41.6	40.5
	SQuAD	74.1	76.8	79.5	75.3
	Avg.	*68.0*	*68.0*	*65.4*	*65.2*

Table 11: **Per-dataset DPO prompt-language ablation (EuroLLM-9B).** All values are absolute scores (averaged over three seeds). The italicized *Avg.* row at the end of each language block reports the language-level mean. Chosen pairs the prompt with the same-language chosen response; Mixed samples prompt language uniformly; Rejected pairs the prompt with the same-language rejected response.

		Baseline	SFT		Max-R		Paired	
Dataset		9B-Inst	eng	tgt	eng	tgt	eng	tgt
dan	DaLA	48.2	40.0	37.6	47.5	43.0	48.5	49.8
	Citizen	81.4	82.6	81.1	81.8	88.2	81.4	82.2
	Entail.	67.5	63.1	59.6	68.5	71.0	67.8	68.2
	LexInf.	76.2	75.9	73.2	76.7	68.5	76.2	77.3
	Talemaader	57.7	57.7	52.4	58.8	56.8	57.0	57.9
	DanWiC	45.0	46.2	34.7	54.0	50.0	56.3	45.6
	MultiWikiQA	79.4	80.9	80.0	80.4	78.4	80.5	81.0
	Avg.	*65.0*	*63.8*	*59.8*	*66.8*	*65.1*	*66.8*	*66.0*
deu	GermanQuAD	57.4	63.7	63.9	61.6	60.7	58.9	59.8
	INCLUDE	36.9	43.1	44.1	39.5	34.9	38.5	38.5
	MMLU	53.2	50.8	49.8	52.6	46.9	53.6	53.7
	MultiLoKo	41.9	34.8	33.3	37.3	39.1	42.2	41.6
	Avg.	*47.4*	*48.1*	*47.8*	*47.8*	*45.4*	*48.3*	*48.4*
spa	INCLUDE	63.4	58.9	59.4	62.9	59.9	64.1	53.1
	MLQA	66.0	68.7	69.6	67.1	69.6	66.4	67.9
	MultiLoKo	42.9	38.9	37.2	39.9	40.0	42.9	43.2
	MultiNRC	34.9	36.8	29.2	38.5	30.7	34.7	37.8
	Avg.	*51.8*	*50.8*	*48.8*	*52.1*	*50.1*	*52.0*	*50.5*
fra	FQuAD	68.4	72.0	72.1	72.5	71.9	69.6	70.3
	INCLUDE	64.9	59.8	60.0	65.3	55.2	65.0	64.6
	MMLU	53.8	52.2	52.1	53.6	50.5	54.2	53.7
	MultiLoKo	33.5	30.3	30.1	34.0	34.8	33.0	36.0
	MultiNRC	40.3	33.1	36.3	36.6	41.9	40.0	39.1
	Avg.	*52.2*	*49.5*	*50.1*	*52.4*	*50.9*	*52.4*	*52.8*
ita	INCLUDE	58.7	55.0	53.4	57.5	59.9	58.4	59.4
	MMLU	52.6	50.7	48.5	51.9	58.5	52.4	62.0
	WiC	51.7	43.2	38.2	44.9	48.4	50.2	52.3
	Avg.	*54.3*	*49.7*	*46.7*	*51.4*	*55.6*	*53.7*	*57.9*
nld	COPA	93.5	92.9	92.0	93.1	93.7	93.1	93.3
	INCLUDE	62.5	61.2	59.5	62.0	51.6	61.8	62.2
	MultiLoKo	41.9	33.9	29.8	37.9	39.5	41.3	41.3
	SQuAD	74.1	77.6	75.2	76.3	80.5	75.5	76.2
	Avg.	*68.0*	*66.4*	*64.1*	*67.3*	*66.3*	*67.9*	*68.2*

Table 12: **Per-dataset English-only vs. translated in-language post-training (EuroLLM-9B).** All values are absolute scores (averaged over three seeds). The italicized *Avg.* row at the end of each language block reports the language-level mean. eng columns are post-training using English-only data; tgt columns are post-training using the row-language data. SFT, Max-R, and Paired correspond to the same monolingual interventions defined in the main results.

		Base	Mono. PT			Multi. PT		
	Dataset	9B-Inst	In SFT	Max SFT	Pair DPO	All SFT	Max SFT	Pair DPO
dan	DaLA	48.2	37.6	38.2	48.8	41.2	46.6	48.3
	Citizen	81.4	81.1	73.2	81.4	77.1	73.9	81.5
	Entail.	67.5	59.6	58.4	67.0	56.8	57.4	67.1
	LexInf.	76.2	73.2	61.5	77.1	61.9	63.1	76.8
	Talemaader	57.7	52.4	44.2	57.7	47.4	39.5	58.2
	DanWiC	45.0	34.7	49.9	56.0	34.0	52.9	56.0
	MultiWikiQA	79.4	80.0	76.7	79.1	77.4	77.7	79.1
	Avg.	*65.0*	*59.8*	*57.5*	*66.7*	*56.5*	*58.7*	*66.7*
deu	GermanQuAD	57.4	63.9	62.4	57.0	59.4	60.0	57.2
	INCLUDE	36.9	44.1	36.9	38.0	38.0	42.0	38.0
	MMLU	53.2	49.8	45.1	53.6	45.9	42.4	53.4
	MultiLoKo	41.9	33.3	32.3	42.2	30.9	31.6	43.1
	Avg.	*47.4*	*47.8*	*44.2*	*47.7*	*43.5*	*44.0*	*47.9*
eng	LifeInUK	75.0	77.5	72.7	75.7	73.1	68.0	75.5
	MMLU-Pro	28.4	24.4	23.2	28.1	21.8	20.3	28.1
	MultiLoKo	33.3	34.5	31.6	31.3	31.6	25.7	31.5
	SQuAD	82.8	86.0	83.1	82.2	82.5	82.7	82.8
	WiC	64.5	61.0	61.2	63.9	55.5	57.4	64.2
	Avg.	*56.8*	*56.7*	*54.4*	*56.2*	*52.9*	*50.8*	*56.4*
spa	INCLUDE	63.4	59.4	54.6	63.7	54.6	49.7	64.2
	MLQA	66.0	69.6	66.9	65.8	65.6	64.5	65.7
	MultiLoKo	42.9	37.2	31.9	43.9	34.8	32.9	43.2
	MultiNRC	34.9	29.2	38.1	34.5	33.0	34.3	35.3
	Avg.	*51.8*	*48.8*	*47.9*	*52.0*	*47.0*	*45.4*	*52.1*
fra	FQuAD	68.4	72.1	71.3	68.7	68.4	68.4	68.5
	INCLUDE	64.9	60.0	52.7	64.3	55.1	51.6	64.6
	MMLU	53.8	52.1	48.1	54.0	46.2	47.6	54.0
	MultiLoKo	33.5	30.1	25.9	34.5	29.2	23.5	34.0
	MultiNRC	40.3	36.3	35.0	40.0	28.9	32.4	40.0
	Avg.	*52.2*	*50.1*	*46.6*	*52.3*	*45.6*	*44.7*	*52.2*
ita	INCLUDE	58.7	53.4	52.7	58.1	50.1	47.7	58.6
	MMLU	52.6	48.5	45.0	52.4	42.6	44.0	52.3
	WiC	51.7	38.2	51.5	51.1	40.5	49.2	51.7
	Avg.	*54.3*	*46.7*	*49.8*	*53.9*	*44.4*	*47.0*	*54.2*
nld	COPA	93.5	92.0	88.3	93.4	88.1	76.7	93.0
	INCLUDE	62.5	59.5	56.0	61.7	57.3	50.7	61.8
	MultiLoKo	41.9	29.8	32.0	40.8	26.9	30.2	40.5
	SQuAD	74.1	75.2	75.0	73.6	73.6	73.3	73.9
	Avg.	*68.0*	*64.1*	*62.8*	*67.4*	*61.5*	*57.7*	*67.3*

Table 13: **Per-dataset off-policy data ablation: tuning** `EuroLLM-9B` **on** `aya-3B` **generations.** All values are absolute scores (averaged over three seeds). The italicized *Avg.* row at the end of each language block reports the language-level mean. The `Max SFT` and `Pair DPO` columns under both `Mono. PT` and `Multi. PT` use post-training data generated by `aya-3B`; for reference, the `In SFT` and `All SFT` columns reuse the on-policy translated-data SFT runs from Table 1.