The Geometry of Low-Resource Language Representations

Francois Meyer and Jan Buys
Department of Computer Science
University of Cape Town
francois.meyer@uct.ac.za, jbuys@cs.uct.ac.za

Abstract

The performance gap between low- and high-resource languages in LLMs is widely known, but it remains unclear which internal model factors drive these disparities. In this paper, we characterise this gap through the lens of representational geometry. Comparing the geometric properties of hidden representations across 30 languages reveals that LLM geometry is systematically related to language data availability. The most consistent effect is in final layers, where low-resource languages exhibit representational degeneration. To counter this, we investigate the effectiveness of regularisation terms to penalise degeneration during continued pretraining (CPT). Experiments monolingually adapting 9 base LLMs to 10 African languages show that geometric regularisation successfully reduces representational degeneration during CPT. For larger models, cosine similarity-based regularisation marginally improves performance over vanilla CPT, with more consistent gains on the most challenging tasks. We establish that the representational geometry of low- and high-resource languages in LLMs is measurably distinct, and that targeted geometric intervention is a viable strategy for improving CPT for low-resource languages.

1 Introduction

LLM capabilities have developed unevenly across languages, with performance in low-resource languages lagging far behind that of high-resource languages (Adelani et al., 2025). This is primarily attributed to data scarcity (Joshi et al., 2020), with low-resource languages lacking sufficient pretraining data to enable generalisation across tasks. From an interpretability perspective, it is unclear which internal model factors contribute to performance degradation for low-resource languages. One way to study this is to compare how multilingual LLMs represent and process low- versus high-resource languages, with the goal of identifying the underlying causes of cross-lingual performance disparities.

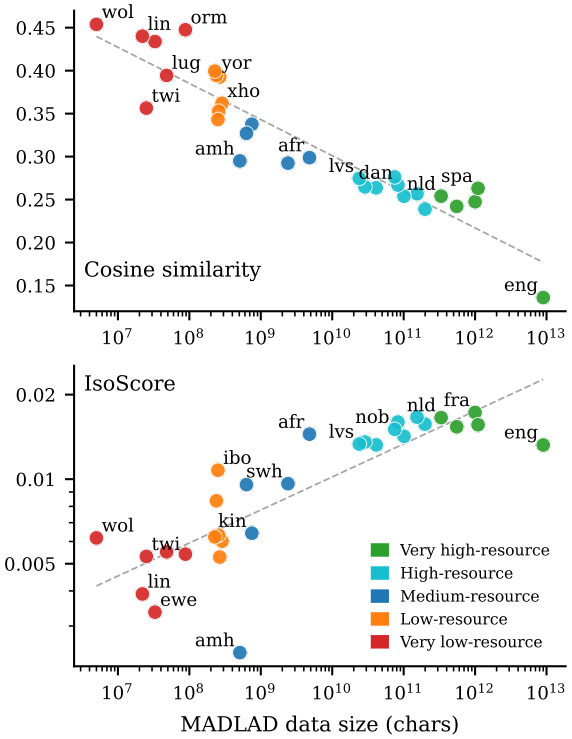

Figure 1: Geometric properties of final-layer token representations in Llama 3.1 8B (see Figures 4 and 5 in the appendix for all models). Lower-resource languages exhibit greater representational degeneration.

In this work, we study the role of LLM geometry. We analyse how the geometric properties of language vector spaces (token embeddings across model layers) relate to data size. Our hypothesis is that vector spaces of low-resource languages are geometrically degenerate, that is, they fail to utilise the full representational capacity. To test this, we use two metrics that characterise representational geometry from complementary perspectives: (1) *Cosine similarity* measures how directionally distinct individual vectors are, with high similarities indicating representational collapse (Gao et al., 2019). (2) *Isotropy* measures how uniformly the dimensions are utilised – anisotropic spaces have reduced effective dimensionality, which acts as a representational bottleneck (Rudman et al., 2022).

Cosine similarity is a local, relational metric, while isotropy measures global space utilisation.

First, we conduct a multilingual analysis of LLM geometry. Using cosine similarity and isotropy, we compare how LLMs represent low- and high-resource languages. Across 30 languages and 9 open models, we observe a clear relationship between resource level and layer-wise geometry. The most consistent trend across metrics is in final layer representations: low-resource languages exhibit higher cosine similarity and anisotropy (as shown in Figure 1), both of which reflect representational degeneration. This shows that LLMs fail to learn geometrically rich representations for low-resource languages, limiting their ability to fully leverage their representational capacity.

Next, we use these findings to augment continued pretraining (CPT), wherein a base LLM is adapted for a low-resource language. We employ two regularisation terms that augment the training objective to explicitly steer LLM representations toward desired geometric properties: CosReg (Gao et al., 2019), which reduces pairwise cosine similarity to encourage more separable representations, and I-STAR (Rudman and Eickhoff, 2024), which increases isotropy to encourage more uniform use of the embedding space. These regularisation techniques have not previously been applied to CPT.

We adapt 9 base LLMs through monolingual CPT for 10 African languages, comparing vanilla CPT to geometrically regularised CPT. Geometric regularisation effectively reduces representational degeneration, offering a simple method for targeted geometric intervention. Evaluating on IrokoBench (Adelani et al., 2025), CosReg marginally improves CPT for larger models (Llama 3.1 8B, Qwen 3 8B, and Gemma 3 4/12B), producing its largest gains on AfriMGSM, a highly challenging task for under-resourced languages. CosReg outperforms I-STAR, suggesting that local separability is a more influential factor in low-resource language modelling than overcoming global representational bottlenecks.

In summary: we analyse how LLM geometry differs between low- and high-resource languages, observe clear differences in final-layer representations, and use geometrically regularised CPT to reduce these differences. Our findings highlight representational degeneration as a factor in the underperformance of low-resource languages, and present geometrically regularised CPT as a practical option for countering it.

2 Related Work

Previous studies on model geometry have repeatedly identified representational degeneration: embeddings fail to utilise the full capacity of the space. The most common metric for quantifying this phenomenon is cosine similarity, with high similarities reflecting embeddings clustered in a narrow region (Gao et al., 2019; Godey et al., 2024). LM representation spaces have also been shown to be anisotropic: variance is dominated by a few axes and available dimensions are not utilised uniformly (Rajaee and Pilehvar, 2022; Rudman et al., 2022). Godey et al. (2024) studied the connection between representational degeneration and model capacity, showing that final-layer representations of smaller LMs collapse onto a lower-dimensional subspace that acts as a representational bottleneck.

Several methods have been proposed to address representational collapse. A common strategy is to penalise degeneration during training with regularisation terms, based on cosine similarity (Gao et al., 2019), isotropic metrics (Zhang et al., 2022; Ji et al., 2023; Rudman and Eickhoff, 2024), or contrastive learning (Zhang et al., 2022; Xiao et al., 2023). Findings have been mixed and there is no clear consensus about the relationship between geometric properties and downstream performance. Reducing representational degeneration improved performance in some contexts (Gao et al., 2019; Rajaee and Pilehvar, 2022; Zhang et al., 2022; Hämmerl et al., 2023; Ji et al., 2023), but in others it had no effect or even degraded it (Rajaee and Pilehvar, 2021; Kovaleva et al., 2021; Zhang et al., 2022; Ding et al., 2022; Rudman and Eickhoff, 2024).

Some of the works outlined above have studied multilingual models (Rajaee and Pilehvar, 2022; Hämmerl et al., 2023; Ji et al., 2023) and designed methods to counter representational collapse in multilingual settings, such as training on parallel corpora (Hämmerl et al., 2023) or code-switched data (Ji et al., 2023). However, these studies focussed on masked LMs and sentence-level semantic tasks. In modern decoder-only LLMs, the relationship between data scarcity and representational geometry has not been studied. The question of whether low-resource languages have systematically different representational geometry from high-resource ones, and whether this can be addressed through geometric regularisation, remains unexplored. In this work we investigate both.

3 Methodology

This paper presents two complementary studies. First, we systematically compare the geometry of LLM representation spaces across low- and high-resource languages (Section 4). Second, based on these findings, we test regularisation terms that counteract the geometric deficiencies observed for low-resource languages (Section 5). Our overall approach follows the paradigm of actionable interpretability (Mosbach et al., 2024; Orgad et al., 2026): using insights from interpretability analysis to inform modelling decisions.[1] In this section, we describe the experimental setup of our geometric analysis (Section 3.1) and our CPT experiments with geometric regularisation (Sections 3.2–3.3).

3.1 Geometric Analysis

Our starting point is to compare the geometric properties of representation spaces across different languages, models, and layers. We study 30 languages, selected to span a wide range of data availability. To quantify per-language data scarcity, we use corpus size in the MADLAD-400 dataset (Kudugunta et al., 2023) as a proxy. MADLAD is a popular multilingual pretraining corpus based on CommonCrawl. Although it does not reflect the exact training mixtures of the models we analyse (their pretraining corpora are not publicly disclosed), it provides a reasonable approximation of relative web-scale data availability across languages. Based on MADLAD data size, we group languages into five tiers of "resourcedness":

- **Very high** (>200B chars): English, Spanish, French, Italian, Dutch
- **High** (>5B chars): Estonian, Turkish, Swedish, Finnish, Danish, Norwegian, Lithuanian, Latvian
- **Medium** (500M–5B chars): Afrikaans, Kiswahili, Kinyarwanda, Hausa, Amharic
- **Low** (100M–500M chars): isiXhosa, chiShona, isiZulu, Igbo, Yorùbá, Sesotho
- **Very low** (<100M chars): Oromo, Luganda, Ewe, Twi, Lingala, Wolof

We compare the representation spaces of these languages across 9 open models: Llama 3.2 (1B, 3B), 3.1 (8B) (Llama Team, 2024), Gemma 3 (1B, 4B, 12B) (Gemma Team, 2025), and Qwen 3 (1.7B, 4B, 8B) (Qwen Team, 2025). Our setup spans languages with varying levels of data availability, models with varying levels of multilingual competence, and models ranging from 1B to 12B parameters.

We use two metrics to quantify vector space geometry: cosine similarity and isotropy. For each language, we pass sentences from FLORES (Goyal et al., 2022) through a model and compute both metrics based on token embeddings from each layer.

3.1.1 Cosine similarity

For layer l, we compute the average pairwise cosine similarity as

$$\text{CosSim}(l) = \frac{1}{N} \sum_{(i,j)} \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{|\mathbf{h}_i||\mathbf{h}_j|}, \quad (1)$$

where i, j correspond to randomly sampled pairs of tokens from FLORES sentences, and $\mathbf{h}_i, \mathbf{h}_j$ are hidden representations of these tokens in layer l. We sample $N = 1000$ token pairs per layer.

Cosine similarity is a local, relational metric, quantifying how directionally distinct vectors are. Averaged across a sample of embeddings, it estimates how geometrically concentrated or dispersed the representations of a language are. For LLM representations, cosine similarity typically ranges from 0 to 1. Low values indicate well-separated embeddings, while high values indicate embeddings collapsed onto a narrow, high-dimensional cone.

3.1.2 Isotropy

Average pairwise cosine similarity is often used as a measure of isotropy. However, Rudman et al. (2022) argue that it fails to capture the true definition of isotropy: how uniformly variance is distributed across the dimensions of a vector space.

They propose IsoScore, a metric that quantifies the degree to which the variance of a set of embeddings is uniformly distributed across all dimensions. Based on eigenvalues of the covariance matrix, it quantifies how evenly variance is spread across dimensions. Unlike cosine similarity, it is invariant to transformations that do not change isotropy (e.g. mean-based shifting). IsoScore measures global space utilisation, while cosine similarity captures local separability between individual vectors.

Following Rudman et al. (2022), we compute IsoScore for layer l, over a point cloud X of $N = 5000$ randomly sampled token embeddings from FLORES sentences, as

$$\text{IsoScore}(l) = \frac{\left(d - \delta(X)^2(d - \sqrt{d})\right)^2 - d}{d(d-1)}, \quad (2)$$

where d is the embedding dimension and $\delta(X) \in$

[1] Code for our geometric analysis and regularised CPT is available at github.com/francois-meyer/cpt-geometry.

$[0, 1]$ is the isotropy defect, measuring how far the variance distribution of X deviates from uniformity (we refer the reader to Rudman et al. (2022) for a more detailed presentation of the IsoScore metric). A score of 1 indicates perfectly isotropic representations (all dimensions utilised evenly), while lower values reflect increasing anisotropy.

3.2 CPT with Geometric Regularisation

Our geometric analysis, outlined above, is an interpretability study that aims to highlight differences in LLM geometry across languages. Next, we consider a practical application where these differences might provide actionable insights: continued pretraining (CPT). In CPT, a pretrained checkpoint is further trained on a target language corpus under the model's original pretraining objective. It is used to adapt pretrained models to a specific language or limited set of languages.

CPT is the standard approach for developing high-quality LMs for low-resource languages, particularly in settings where training from scratch is computationally infeasible. For African languages, CPT has produced state-of-the-art results across different architectures, from masked language models (Alabi et al., 2022), to encoder–decoder models (Oladipo et al., 2023), and more recently decoder-based LLMs (Buzaaba et al., 2025; Yu et al., 2026).

We apply monolingual CPT to all 9 LLMs from our geometric analysis, on corpora from 10 African languages: Kinyarwanda, Hausa, Amharic, isiXhosa, chiShona, isiZulu, Igbo, Yorùbá, Sesotho, and Oromo. For CPT we use the WURA corpus (Oladipo et al., 2023), a high-quality filtered subset of mC4. We select languages in WURA that are in our very low-, low-, and medium-resource tiers, as this is where CPT offers most benefit (see Table 4 in the Appendix for corpus statistics). For each model and language, we continue pretraining the base LLM[2] for 1 epoch on the target-language corpus. Our baseline models are adapted with vanilla CPT, in which the standard LM loss is minimised over the target-language corpus. Our hyperparameter tuning process and final CPT configuration is detailed in Appendix A.1.

Vanilla CPT is a strong baseline for low-resource language adaptation. Because CPT updates model parameters on target-language data, it provides an opportunity to correct the geometric deficiencies inherited from pretraining. We explicitly encourage this by augmenting the CPT objective with regularisation terms that penalise representational degeneration. We test two regularisation strategies, each targeting one of the geometric metrics from our analysis. We tune hyperparameters for both regularisers, as detailed in Appendix A.2.

CosReg Following Gao et al. (2019), we augment the training objective with a penalty based on the mean pairwise cosine similarity of all M token representations at layer l in a batch:

$$\mathcal{L}_{\text{CosReg}} = \mathcal{L}_{\text{NLL}} + \lambda \cdot \frac{1}{M^2} \sum_i \sum_{j \neq i} \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{|\mathbf{h}_i||\mathbf{h}_j|}, \quad (3)$$

where $\mathbf{h}_1, \ldots, \mathbf{h}_M$ are token representations at layer l in a training mini-batch. We set $\lambda = 1$ to penalise high pairwise cosine similarity and encourage more separable representations.

I-STAR Rudman and Eickhoff (2024) propose I-STAR, a regularisation term that estimates isotropy from the representations of a batch. It is based on IsoScore*, a differentiable variant of IsoScore that can be optimised during training. They apply it during masked LM finetuning, but suggest its use in pretraining as a promising direction for future work. We adopt I-STAR for CPT of decoder LLMs.

At each training step, we compute IsoScore* for a randomly sampled batch subset of $M' = 1{,}000$ token embeddings at layer l, and augment the training objective as follows:

$$\mathcal{L}_{\text{I-STAR}} = \mathcal{L}_{\text{NLL}} + \lambda \cdot (1 - \text{IsoScore}^\star). \quad (4)$$

We set $\lambda = 1$ to penalise anisotropy and encourage more even utilisation of dimensions. Rudman and Eickhoff (2024) found that *encouraging* anisotropy with I-STAR ($\lambda < 0$) produced better results in the context of masked LM finetuning. We did test negative values of λ during hyperparameter tuning (Appendix A.2) but found no benefit for CPT.

3.3 Evaluation

To test the impact of geometric intervention, we compare target-language performance after vanilla CPT and geometrically regularised CPT. We evaluate adapted models on target-language data sets in IrokoBench (Adelani et al., 2025), a human-translated benchmark for African languages that includes natural language inference (AfriXNLI),

[2]We use base models, instead of instruction-tuned versions, in all our experiments. CPT extends pretraining, so it is more naturally applied to base models that have not undergone additional post-training.

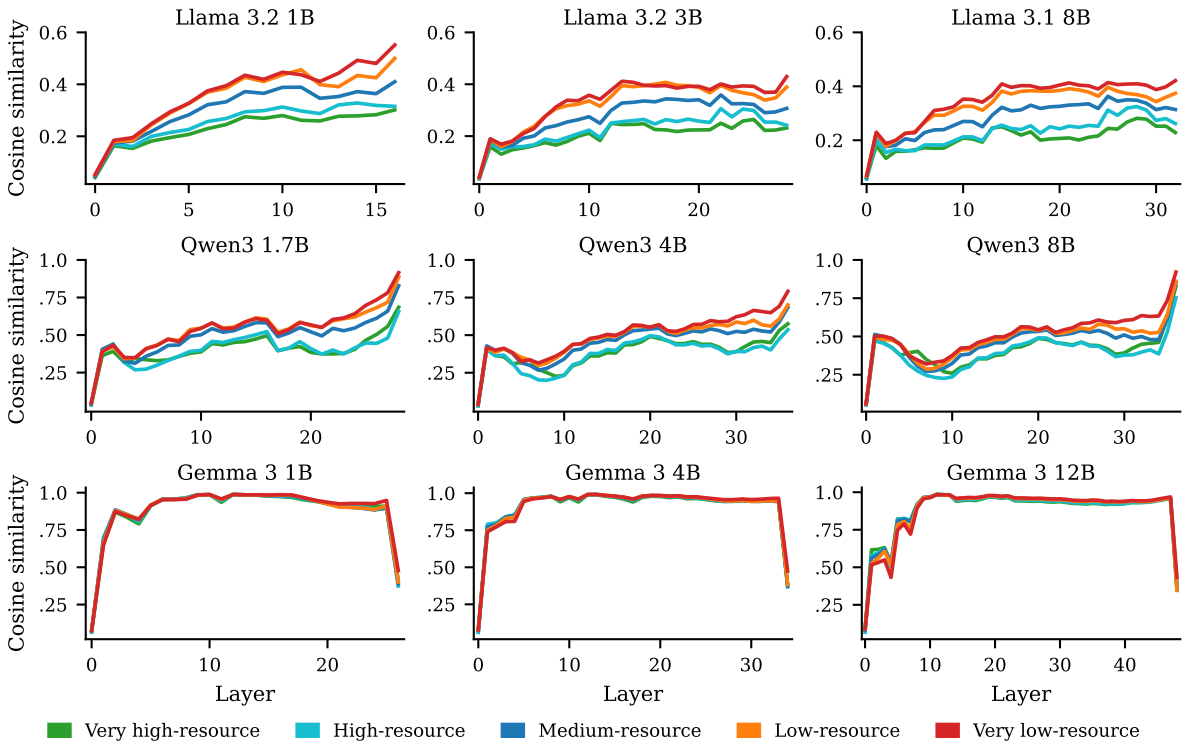

Figure 2: Pairwise cosine similarity of representations across model layers, calculated as specified in Section 3.1.1 and averaged across languages within each tier of data availability.

multi-choice knowledge-based QA (AfriMMLU), and mathematical reasoning (AfriMGSM). We evaluate with LM Evaluation Harness (Gao et al., 2024) in zero-shot and few-shot settings (5-shot for AfriMMLU and AfriXNLI, 8-shot for AfriMGSM), using prompt templates from Adelani et al. (2025).

4 Multilingual Geometric Analysis

We present our geometric analysis (Section 3.1) of nine models across 30 languages.

4.1 Geometry Across Languages and Layers

We observe clear differences between the geometric properties of high- and low-resource languages. Figure 2 plots average cosine similarity, grouped by data tier. Lower-resource languages have higher cosine similarity across most layers, while higher-resource language representations are more dispersed, indicating better local separability.

This trend is not only a binary distinction between low- and high-resource languages. In many layers, language data size and cosine similarity are strongly correlated: the lower-resource a language is, the higher its cosine similarity. This relationship can be seen in the Llama and Qwen plots of Figure 2, where cosine similarity decreases monotonically with data tier for many layers (from very low-resource to very high-resource). To quantify this relationship, we compute Spearman ρ correlation coefficients between MADLAD data size and cosine similarity for each layer. Table 1 reports ρ for evenly-spaced layers across all models. Deeper layers have strong negative correlations (similarity decreases as language "resourcedness" increases).

We also observe clear differences between the isotropy of high- and low-resource languages in some layers. Low-resource language representation spaces tend to be more anisotropic. This is hard to visualise because IsoScore differences are small (see Figure 3 in the appendix), but the correlation analysis between data size and IsoScore in Table 1 shows a clear pattern. Input layers have weak positive correlations, middle-layers have inconsistent correlations, and final layers have the strongest, positive correlations (IsoScore increases as language "resourcedness" increases).

One language emerged as an outlier: Amharic consistently deviated from what its data availability would predict (as shown in Figure 1). The layer-wise geometric profiles of Amharic were also markedly different from other languages in its resource tier. We attribute this to Amharic's use of the Ge'ez script, which sets it apart from the remaining languages, all of which use Latin script.

Model	0%	25%	50%	75%	100%
Cosine similarity					
Llama 3.2 1B	**-0.52**	**-0.84**	**-0.89**	**-0.88**	**-0.88**
Llama 3.2 3B	-0.24	**-0.87**	**-0.89**	**-0.85**	**-0.90**
Llama 3.1 8B	**-0.52**	**-0.93**	**-0.93**	**-0.90**	**-0.95**
Qwen3 1.7B	**-0.56**	**-0.76**	**-0.85**	**-0.82**	**-0.73**
Qwen3 4B	**-0.56**	**-0.78**	**-0.91**	**-0.85**	**-0.74**
Qwen3 8B	**-0.56**	**-0.73**	**-0.89**	**-0.89**	**-0.58**
Gemma 3 1B	-0.27	0.58	0.47	-0.17	-0.42
Gemma 3 4B	**-0.54**	0.80	**-0.75**	**-0.83**	-0.40
Gemma 3 12B	**-0.52**	0.38	**-0.71**	**-0.81**	-0.25
IsoScore					
Llama 3.2 1B	0.20	0.04	0.20	0.29	**0.85**
Llama 3.2 3B	0.15	0.10	0.13	0.36	**0.88**
Llama 3.1 8B	**0.72**	-0.26	0.06	0.07	**0.85**
Qwen3 1.7B	0.39	0.09	0.26	0.30	0.48
Qwen3 4B	0.47	-0.10	-0.11	-0.19	**0.73**
Qwen3 8B	0.45	-0.33	-0.17	-0.39	**0.70**
Gemma 3 1B	-0.04	-0.06	**0.83**	**0.76**	**0.53**
Gemma 3 4B	0.03	-0.29	**0.53**	**0.68**	0.47
Gemma 3 12B	0.30	-0.27	**0.55**	**0.71**	0.25

Table 1: Spearman ρ between MADLAD size and geometric metrics across layers (0% = input embeddings, 100% = final layer, remaining columns are evenly-spaced hidden layers). We **boldface** $\rho < -0.5$ for cosine similarity and $\rho > 0.5$ for IsoScore.

Amharic shares no subword tokens with our other languages. Its representations originate from a geometrically isolated set of input embeddings and remain geometrically distinct throughout model layers. This highlights that resource level is not the only factor influencing LLM geometry and that linguistic typology can also play a role.

4.2 Final Layer Degeneration

The most consistent trend across metrics and models concerns final layer representations, where low-resource languages reliably exhibit greater representational degeneration than high-resource ones. As shown in Figure 1 for Llama 3.1 8B (and Figures 4 and 5 in the appendix for all models), final-layer degeneration worsens as language data decreases. While absolute measures of degeneration actually improve for some models/languages in final layers (e.g. cosine similarity of Gemma models in Figure 2, IsoScore of most models in Figure 3), cross-lingual geometric differences persist and, in many instances, become more pronounced. The last column of Table 1 shows that the final layer is the only layer where representational degeneration correlates with data scarcity across all models and both metrics (ρ is consistently negative for cosine similarity and positive for IsoScore).

	Δ Cosine similarity			Δ IsoScore		
Model	CPT	+CR	+IS	CPT	+CR	+IS
Llama 3.2 1B	-.096	-.278	-.089	.000	-.004	+.014
Llama 3.2 3B	-.023	-.176	-.024	.000	-.003	+.022
Llama 3.1 8B	-.020	-.202	-.049	.000	-.003	+.027
Qwen3 1.7B	-.105	-.620	-.048	.000	-.002	+.004
Qwen3 4B	-.095	-.491	-.032	.000	-.002	+.006
Qwen3 8B	-.004	-.691	+.023	.000	.000	+.001
Gemma 3 1B	-.016	-.222	-.012	-.003	-.011	+.029
Gemma 3 4B	+.020	-.204	+.001	-.003	-.009	+.014
Gemma 3 12B	+.054	-.180	+.035	-.004	-.009	+.005

Table 2: Final-layer geometry change from base models to CPT and regularised CPT (+CR: cosine regularisation, +IS: I-STAR), averaged over 10 CPT languages.

4.3 Geometry Across Models

Figures 2 and 3 also show geometric differences between model families. Yu et al. (2026) showed that Gemma 3 has strong multilingual capabilities, followed by Qwen 3 and Llama 3. This is reflected in LLM geometry – differences between lower and higher-resource languages shrink as multilingual coverage expands from Llama, to Qwen, and Gemma. Table 1 also shows that the correlation between data size and final-layer degeneration is strongest in Llama, weaker in Qwen, and weakest in Gemma (this holds for both cosine similarity and IsoScore). However, broader multilingual coverage does not eliminate cross-lingual disparities completely. Even in Gemma, where geometric differences are hard to discern in our plots, data scarcity still consistently correlates with representational degeneration in deeper layers.

A notable feature of our plots is the distinct geometric profile of Gemma compared to Llama and Qwen. In Figure 2, Gemma has high cosine similarity across most layers and a dramatic drop in the final layer. This likely reflects differences in architecture and pretraining between the model families. For example, unlike Llama and Qwen, Gemma applies layer normalisation both before and after each sublayer, which may constrain angular variation across layers. Gemma also uses a larger vocabulary, which requires more discriminative final-layer representations to predict next tokens. Other factors in Gemma pretraining, such as knowledge distillation and multi-modal pretraining, might also influence model geometry. Given architectural differences and varying pretraining choices, it is unsurprising that models have distinct geometric profiles. A detailed investigation of these effects is outside the scope of this work. Instead, we focus on what

Model	Variant	AfriXNLI		AfriMMLU		AfriMGSM		Overall		
		0-shot	5-shot	0-shot	5-shot	0-shot	8-shot	Avg	Δ	Δ%
Llama 3.2 1B	Base	32.9	32.4	**27.1**	25.8	2.1	**2.7**	20.5	–	–
	CPT	**34.0**	**33.5**	26.1	**26.6**	2.2	2.5	**20.8**	**+0.3**	**+1.5%**
	+ CosReg	33.4	33.2	26.5	**26.6**	**2.3**	2.6	**20.8**	**+0.3**	**+1.5%**
	+ I-STAR	33.9	33.2	26.0	26.5	2.2	2.6	20.7	+0.2	+1.0%
Gemma 3 1B	Base	33.4	32.7	**24.9**	25.8	1.7	2.7	20.2	–	–
	CPT	34.5	**33.4**	24.2	**25.9**	**2.2**	3.7	**20.7**	**+0.5**	**+2.5%**
	+ CosReg	34.4	33.0	24.4	25.2	1.9	**4.0**	20.5	+0.3	+1.5%
	+ I-STAR	**34.9**	33.0	24.4	25.8	2.0	3.7	20.6	+0.4	+2.0%
Qwen3 1.7B	Base	33.3	31.4	**29.2**	29.7	**3.3**	**4.5**	**21.9**	–	–
	CPT	33.4	**32.6**	28.0	29.8	2.4	3.3	21.6	-0.3	-1.4%
	+ CosReg	**33.9**	32.3	28.2	**30.0**	2.6	3.2	21.7	**-0.2**	**-0.9%**
	+ I-STAR	33.4	**32.6**	28.4	29.9	2.4	3.3	21.7	**-0.2**	**-0.9%**
Llama 3.2 3B	Base	33.6	32.3	26.4	27.5	**2.7**	**4.1**	21.1	–	–
	CPT	33.8	**32.7**	26.5	**28.4**	**2.7**	4.0	**21.3**	**+0.2**	**+0.9%**
	+ CosReg	**33.9**	32.4	**27.0**	28.0	**2.7**	3.9	**21.3**	**+0.2**	**+0.9%**
	+ I-STAR	**33.9**	32.6	26.6	28.2	2.6	3.7	**21.3**	**+0.2**	**+0.9%**
Qwen3 4B	Base	33.6	32.1	**31.8**	34.4	**4.7**	**6.6**	**23.9**	–	–
	CPT	33.8	33.0	30.9	**34.8**	3.1	6.5	23.7	-0.2	-0.8%
	+ CosReg	33.1	**33.3**	30.6	34.5	3.5	6.5	23.6	-0.3	-1.3%
	+ I-STAR	**34.0**	**33.3**	30.9	34.6	3.3	6.5	23.8	**-0.1**	**-0.4%**
Gemma 3 4B	Base	34.9	33.6	30.4	34.0	4.0	7.9	24.2	–	–
	CPT	36.1	35.7	31.9	34.9	5.7	11.4	25.9	+1.7	+7.0%
	+ CosReg	**36.9**	35.6	**32.7**	**35.1**	**6.1**	**11.5**	**26.3**	**+2.1**	**+8.7%**
	+ I-STAR	36.0	**35.9**	32.1	34.4	5.6	11.3	25.9	+1.7	+7.0%
Llama 3.1 8B	Base	33.4	32.9	28.7	32.0	4.8	6.9	23.1	–	–
	CPT	35.0	35.3	**30.2**	**33.6**	6.7	**9.4**	25.0	+1.9	+8.2%
	+ CosReg	35.3	**36.1**	29.9	33.4	**6.8**	9.2	**25.1**	**+2.0**	**+8.7%**
	+ I-STAR	**35.5**	35.3	29.9	32.8	6.6	8.8	24.8	+1.7	+7.4%
Qwen3 8B	Base	**34.3**	32.5	31.8	37.0	**6.2**	8.2	25.0	–	–
	CPT	33.9	32.3	**33.4**	36.8	5.9	9.0	25.2	+0.2	+0.8%
	+ CosReg	33.6	**32.6**	**33.4**	37.0	5.9	**9.6**	**25.3**	**+0.3**	**+1.2%**
	+ I-STAR	33.9	32.5	33.1	**37.3**	5.4	9.4	**25.3**	**+0.3**	**+1.2%**
Gemma 3 12B	Base	38.6	38.0	38.3	48.0	13.1	22.1	33.0	–	–
	CPT	40.0	**40.5**	**43.8**	**49.6**	15.6	**28.2**	36.3	+3.3	+10.0%
	+ CosReg	**40.1**	40.3	43.7	49.5	**16.5**	**28.2**	**36.4**	**+3.4**	**+10.3%**
	+ I-STAR	**40.1**	**40.5**	43.7	49.2	16.2	27.8	36.2	+3.2	+9.7%

Table 3: Average performance (%) across CPT languages (Base = pretrained model, CPT = vanilla continued pretraining, +CosReg / +I-STAR = regularised CPT). AfriMGSM scores are the averages of direct and chain-of-thought evaluations. We **boldface** the best-performing variant per model. Δ and Δ% show, respectively, absolute and relative performance gains over base models.

is consistent across all models: low-resource languages exhibit greater representational degeneration, regardless of architecture.

5 CPT with Geometric Regularisation

In the previous section, we found that final-layer degeneration reliably worsens with data scarcity. We therefore apply regularisation (Equations 3 and 4) to final-layer token representations during CPT. We present our CPT experiments (Sections 3.2–3.3), comparing the impact of vanilla CPT and geometrically regularised CPT on LLM geometry and downstream performance across 10 African languages.

5.1 CPT and Representational Degeneration

First, we test whether geometric regularisation successfully changes LLM geometry as intended. Table 2 confirms that it does. CosReg consistently decreases average pairwise cosine similarity in the final layer, so token representations become more locally separable. I-STAR consistently increases IsoScore in the final layer, so representation spaces utilise available dimensions more evenly. The effect of cosine similarity-based regularisation on isotropy, and vice versa, is negligible. This reaffirms that the two metrics measure different aspects of LLM geometry and that intervening on one does

not affect the other. The effect of final-layer regularisation on other layers is also negligible (see Figures 6 and 7 in the appendix), so our intervention is localised. We alter the geometry of a specific layer as intended, without disrupting the representational geometry of other layers.

Without regularisation, vanilla CPT already reduces final-layer degeneration, but only marginally. Table 2 shows that final-layer cosine similarity decreases after vanilla CPT for most models, more so for the less multilingual models (Llama and Qwen) than for Gemma. Models learn more locally separable representations for low-resource languages, which might partially explain the success of CPT. Next, we study whether explicitly encouraging this geometric shift through regularisation can improve downstream performance over vanilla CPT.

5.2 IrokoBench Results

Table 3 summarises the performance of our CPT models on IrokoBench, while Tables 5 and 6 in the appendix report the full set of results. Results are mixed across individual tasks and languages, but a clearer pattern emerges when results are viewed by model size, which is why Table 3 orders models from smallest to largest. The final three columns of Table 3 report average performance level and average gains over base models (in absolute and relative terms), across all tasks and languages.

CPT and geometric regularisation performance scales with model size. For smaller models (3B and below), geometrically regularised CPT offers no gains beyond vanilla CPT. However, at this scale, all forms of CPT achieve only marginal (or no) improvements over base models. With smaller models and limited pretraining data, CPT is generally less effective for language adaptation, and geometric regularisation fails to overcome this.

For larger models (4B and above), CPT more reliably improves over base models, and CosReg marginally improves average performance over vanilla CPT. While average gains are small, CosReg achieves notable improvements over vanilla CPT in several instances: AfriXNLI 5-shot for Llama 3.1 8B ($35.3 \rightarrow 36.1$), and AfriXNLI and AfriMMLU 0-shot for Gemma 3 4B ($36.1 \rightarrow 36.9$ and $31.9 \rightarrow 32.7$). On AfriMGSM, the most challenging task in IrokoBench, CosReg outperforms vanilla CPT across nearly all larger models, with notable gains in some cases ($9.0 \rightarrow 9.6$ for Qwen3 8B and $15.6 \rightarrow 16.5$ for Gemma 3 12B). At these low absolute performance levels, such improvements are proportionally substantial and reflect the average gain across 10 languages.

CosReg outperforms I-STAR. Across models and tasks, CosReg tends to perform better than I-STAR. We cannot rule out that I-STAR could prove effective for CPT under different conditions, such as larger target-language corpora or if applied across multiple layers (the latter would require reducing the computational overhead of I-STAR). However, the comparison between CosReg and I-STAR is informative because, as shown in Table 2, both regularisers alter model geometry as intended. Our results show that decreasing final-layer cosine similarity is a more reliable way to improve downstream performance than increasing final-layer isotropy. This suggests that, for low-resource language adaptation, local separability between token representations is more influential than global utilisation of the embedding space.

6 Conclusion

This work conducts a systematic analysis of LLM geometry, highlighting clear differences in how models construct representation spaces for low- and high-resource languages. Our results reveal a direct link between data scarcity and LLM geometry: lower-resource languages exhibit progressively worse representational degeneration, especially in final layers. Based on these findings, we adopt geometric regularisation techniques and apply them during CPT, where they successfully reduce final-layer degeneration for target languages. Cosine similarity-based regularisation frequently improves downstream performance for larger models, offering a simple technique to enhance CPT.

Our approach is based on actionable interpretability (Mosbach et al., 2024; Orgad et al., 2026), which aims to use insights from interpretability experiments to inform modelling decisions. Our work shows how the perspective of LLM geometry is suited to this paradigm: geometric analyses can reveal internal deficiencies, while targeted geometric intervention can shift representational geometry towards desirable properties. More broadly, our work shows that actionable interpretability offers a promising approach to low-resource NLP. By first understanding *why* models underperform in low-resource settings, we can move beyond trial-and-error towards principled modelling decisions for low-resource languages.

7 Limitations

Analysis scope Our geometric analysis focusses on the relationship between language resource level and representational degeneration. While our results highlight the impact of other factors on LLM geometry, such as Amharic typography (Section 4.1) and model architecture (Section 4.3), we do not investigate these effects in detail. The role of linguistic typology and architectural choices in shaping LLM geometry warrants further study, as such insights could similarly inform strategies for improving low-resource language representations. We do not claim that data availability is the primary determinant of representational geometry, only that it is a strong and consistent one. Our main finding – that cross-lingual geometric disparities scale with data availability – holds across all 9 models and is robust when restricted to the Latin-script languages in our set, where script effects are controlled for.

Training scope The main practical limitation is that we focus on a single phase of LLM training, namely CPT. We cannot guarantee that geometric regularisation will effectively steer LLM geometry in other phases, such as training from scratch or instruction tuning. We initially applied geometric regularisation during instruction tuning of base models, but all forms of instruction tuning failed to reliably improve IrokoBench performance. Given the lack of high-quality instruction following data for African languages, this remains challenging for our target languages. We therefore focus on CPT, a reliable method for improving low-resource language performance and a natural setting for evaluating geometric intervention.

Performance gains The impact of geometric regularisation on performance is inconsistent across tasks, and most gains are small in absolute terms. CosReg marginally improves average performance over vanilla CPT, but only for larger models. Our primary practical contribution is not a new method for state-of-the-art CPT. Rather, we demonstrate that representational geometry is a measurable and actionable factor in low-resource language modelling. Our approach requires no additional data. Yet it reliably shifts final-layer geometry as intended, without degrading performance, and in some cases leads to meaningful gains on challenging tasks like AfriMGSM. This is sufficient to establish geometric regularisation as a promising direction for low-resource language adaptation.

8 Acknowledgements

This research was supported by Google.org and the Google Cloud Research Credits program for the Gemma Academic Program. Computations were also performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za. This work is also based on research supported in part by the National Research Foundation of South Africa (Grant Number: 151601).

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Salomey Osei, Shamsuddeen Hassan Muhammad, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lugha-llama: Adapting large language models for african languages. *Preprint*, arXiv:2504.06536.

Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. 2022. On isotropy calibration of transformer models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence

Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

Google Gemma Team. 2025. Gemma 3.

Nathan Godey, Éric Villemonte de la Clergerie, and Benoît Sagot. 2024. Why do small language models underperform? studying language model saturation via the softmax bottleneck. In *First Conference on Language Modeling*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Yixin Ji, Jikai Wang, Juntao Li, Hai Ye, and Min Zhang. 2023. Isotropic representation can improve zero-shot cross-lingual transfer on multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8104–8118, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105, Miami, Florida, USA. Association for Computational Linguistics.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron Wallace, Sarah Wiegreffe, Eric Wong, Ian Tenney, and Mor Geva. 2026. Interpretability can be actionable. *Preprint*, arXiv:2605.11161.

Alibaba Group Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

William Rudman and Carsten Eickhoff. 2024. Stable anisotropic regularization. In *The Twelfth International Conference on Learning Representations*.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. 2023. On isotropy, contextualization and learning dynamics of contrastive-based sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12266–12283, Toronto, Canada. Association for Computational Linguistics.

Hao Yu, Tianyi Xu, Michael A. Hedderich, Wassim Hamidouche, Syed Waqas Zamir, and David Ifeoluwa Adelani. 2026. Afriquellm: How data mixing and model architecture impact continued pre-training for african languages. *Preprint*, arXiv:2601.06395.

Haode Zhang, Haowen Liang, Yuwei Zhang, Liming Zhan, Xiaolei Lu, Albert Lam, and Xiao-Ming Wu. 2022. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 532–542, Seattle, United States. Association for Computational Linguistics.

A Hyperparameters

A.1 Vanilla CPT

We tune hyperparameters for vanilla CPT based on three pilot languages (Oromo, isiXhosa, isiZulu) using Llama 3.1 8B and Gemma 3 4B as representative models, selecting final settings based on AfriXNLI validation performance. We test two learning rates, 1×10^{-5} and 1×10^{-4}. The higher rate performed worse, so we adopt 1×10^{-5} with cosine decay to 1×10^{-6}. We train for 1, 2, and 3 epochs and find that performance plateaus after 1 epoch, so we train all subsequent models for 1 epoch only. We use a maximum sequence length of 512 tokens. These settings carry over unchanged to geometrically regularised CPT.

A.2 Geometric regularisation

For both regularisers, we tune on the same three pilot languages, carrying over the CPT settings above. We select final hyperparameter values based on AfriXNLI validation performance and apply them uniformly to all other languages and models.

CosReg. We tune the regularisation coefficient $\lambda \in \{-10, -1, 0.1, 1, 10\}$, where positive values penalise high cosine similarity (encouraging separation) and negative values do the opposite. We find $\lambda = 1$ performs best, which aligns with our expectation that penalising representational degeneration is beneficial. We also compare applying CosReg to the final layer only, to all layers, and to subsets of the final 3 and 5 layers, finding that regularising more layers fails to improve over final-layer only. We use $\lambda = 1$ applied to the final layer for all CosReg experiments.

I-STAR. I-STAR introduces several additional hyperparameters, which we describe before detailing our tuning procedure.

Rudman and Eickhoff (2024) propose I-STAR, a regularisation term based on IsoScore*, which is a differentiable variant of IsoScore (Rudman et al., 2022). Estimating covariance from one batch underestimates the true IsoScore of a representation space, which I-STAR addresses via shrinkage: interpolating the batch covariance with a stable reference covariance. At regular intervals, we compute a reference covariance matrix C_0 from a sample of N token embeddings at layer l drawn from WURA training data. At each training step, we compute the empirical covariance C over a randomly sampled batch subset of $M' = 1{,}000$ token embeddings at layer l, and form the shrinkage estimate:

$$\Sigma = (1 - \zeta)C + \zeta C_0, \tag{5}$$

where $\zeta \in [0, 1]$ controls the balance between the batch and the global reference estimate. IsoScore* is computed from Σ, giving the training objective:

$$\mathcal{L}_{\text{I-STAR}} = \mathcal{L}_{\text{NLL}} + \lambda \cdot (1 - \text{IsoScore}^\star). \tag{6}$$

The sign of λ determines the direction of the intervention: positive values encourage greater isotropy, while negative values encourage greater anisotropy. As with CosReg, the regularisation can be applied to a single layer or to all layers simultaneously.

In all our experiments, we apply I-STAR regularisation to final-layer representations. Applying I-STAR to all layers requires computing IsoScore* at every layer at each training step, which is prohibitively slow for the scale of our experiments. Selecting the final layer is also motivated by our analysis showing that the strongest representational degeneration effect, in terms of IsoScore, is found in the final layer (Table 1).

We tune the regularisation coefficient $\lambda \in \{-10, -1, 1, 10\}$, testing both signs to assess whether encouraging or discouraging isotropy is more beneficial. We also tune the shrinkage parameter $\zeta \in \{0.1, 0.2, 0.5, 0.8\}$, which controls how strongly the batch covariance is regularised towards the global reference. Lastly, we tune the reference sample size $N \in$

Code	Language	Words	Characters
hau	Hausa	151,868,683	858,393,671
amh	Amharic	86,144,742	453,397,345
sot	Sesotho	35,331,080	193,620,015
zul	isiZulu	31,584,011	271,658,420
yor	Yorùbá	30,854,371	155,028,472
ibo	Igbo	29,890,153	161,814,604
sna	chiShona	28,464,659	216,818,585
kin	Kinyarwanda	26,327,273	194,417,463
xho	isiXhosa	13,082,332	111,534,373
orm	Oromo	6,296,883	49,322,425

Table 4: Corpus sizes for CPT (word counts are based on white-space separated tokens in the raw text).

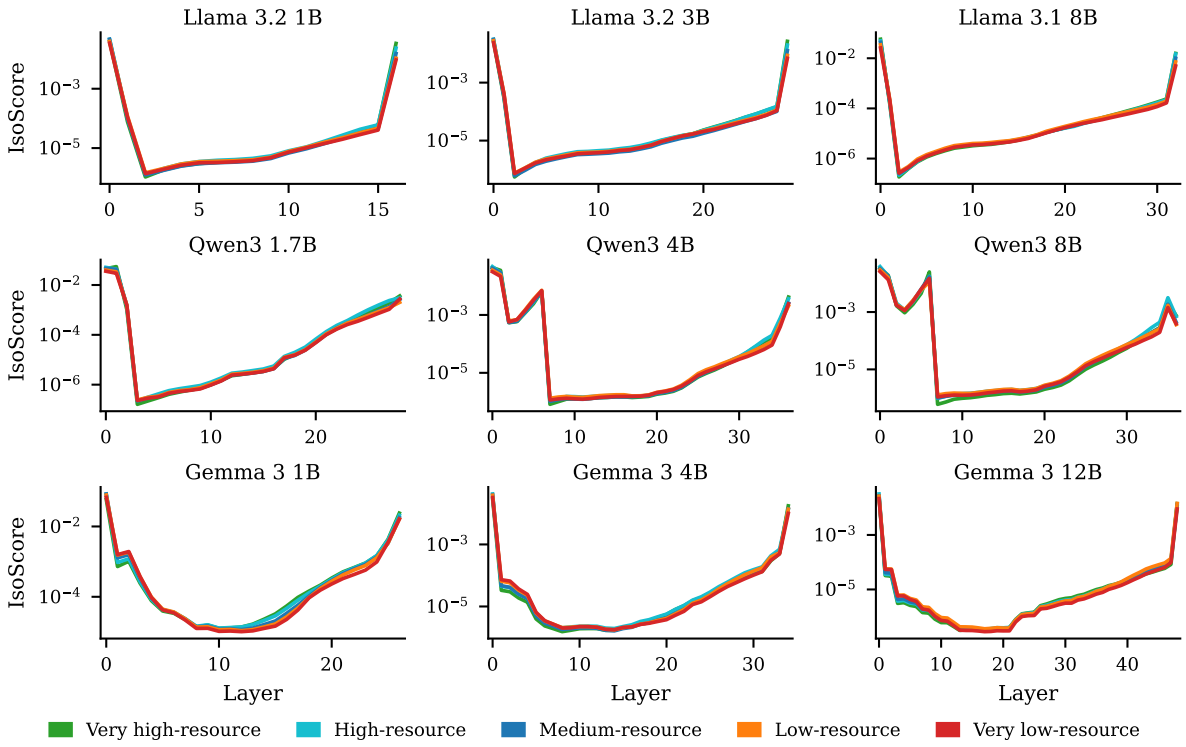

Figure 3: Isotropy (IsoScore) of representations across model layers, calculated as specified in Section 3.1.2 and averaged across languages within each tier of data availability.

$\{10{,}000, 50{,}000, 100{,}000, 250{,}000\}$, which determines how many token embeddings are used to compute the stable reference covariance C_0.

We find that $\lambda = 1$, $\zeta = 0.2$, and $N = 250{,}000$ perform best and use these values across all models and languages.

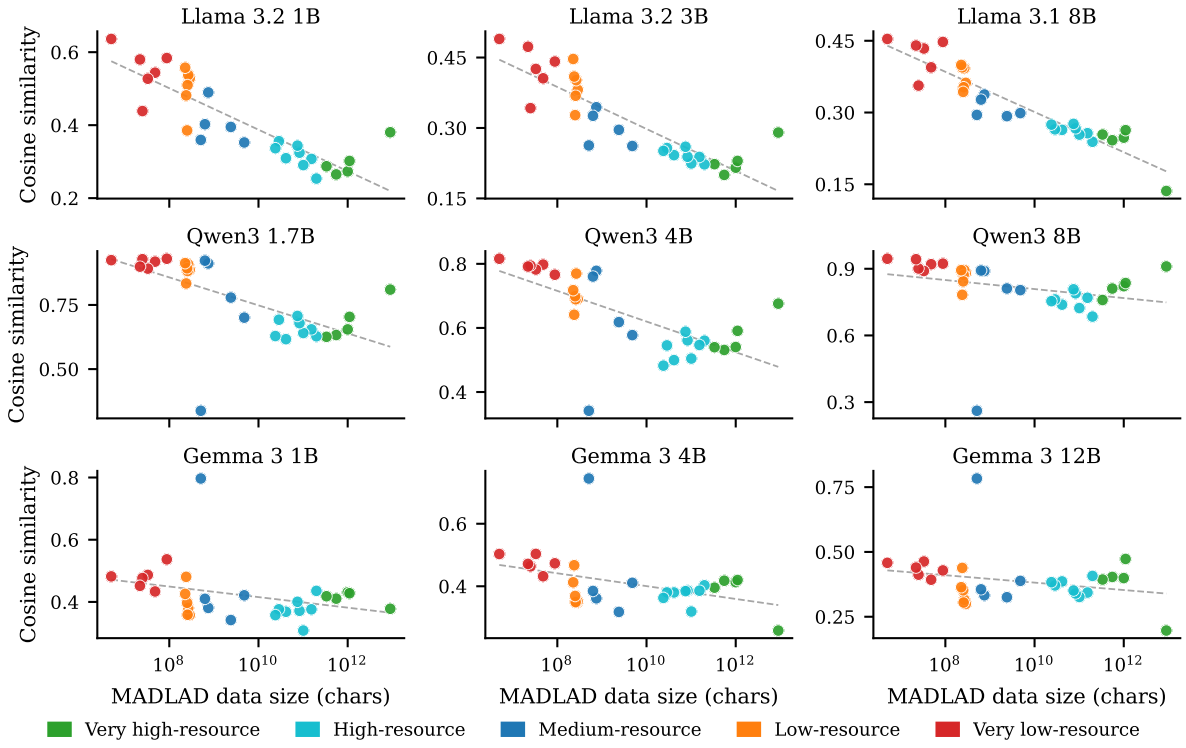

Figure 4: The relationship between pretraining data availability and average pairwise cosine similarity of final-layer representations. Each dot represents a single language.

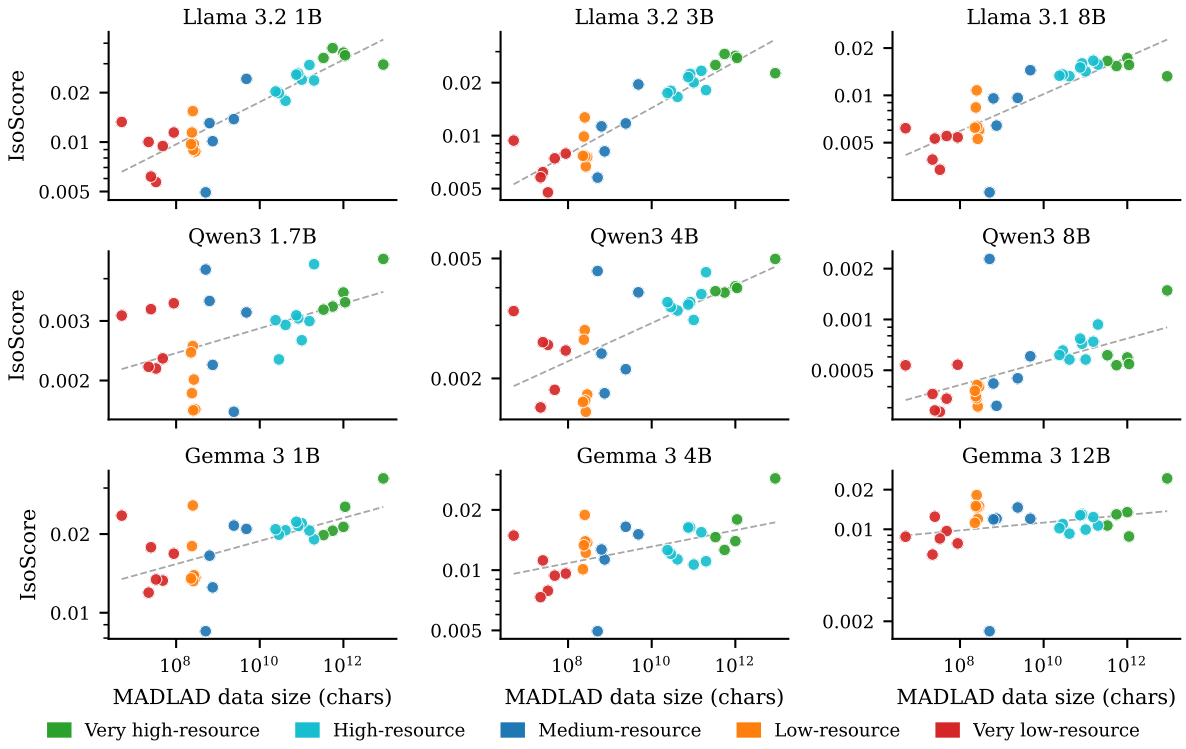

Figure 5: The relationship between pretraining data availability and isotropy (IsoScore) of final-layer representations. Each dot represents a single language.

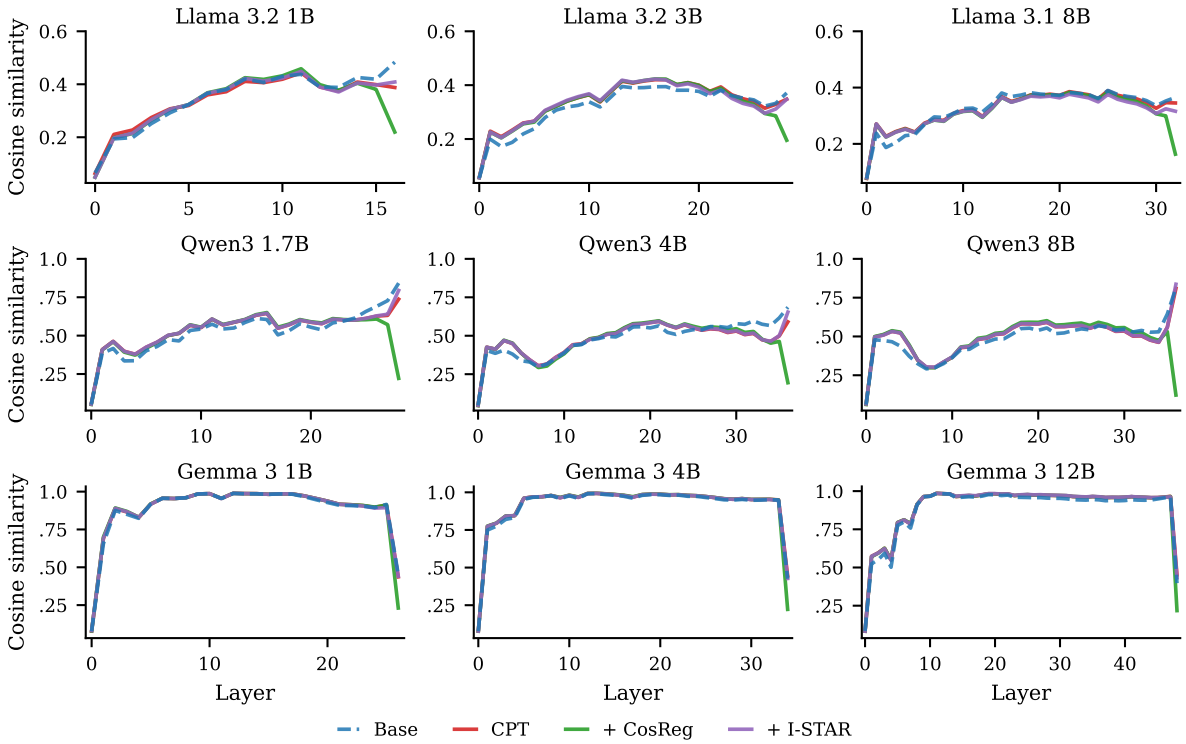

Figure 6: How do CPT and geometric regularisation change cosine similarity? We plot layer-wise cosine similarity, averaged across all 10 languages from our CPT experiments, comparing base models to vanilla CPT models (which only marginally affect cosine similarity), and geometrically regularised CPT models (CosReg successfully reduces final-layer cosine similarity, while leaving other layers unaffected).

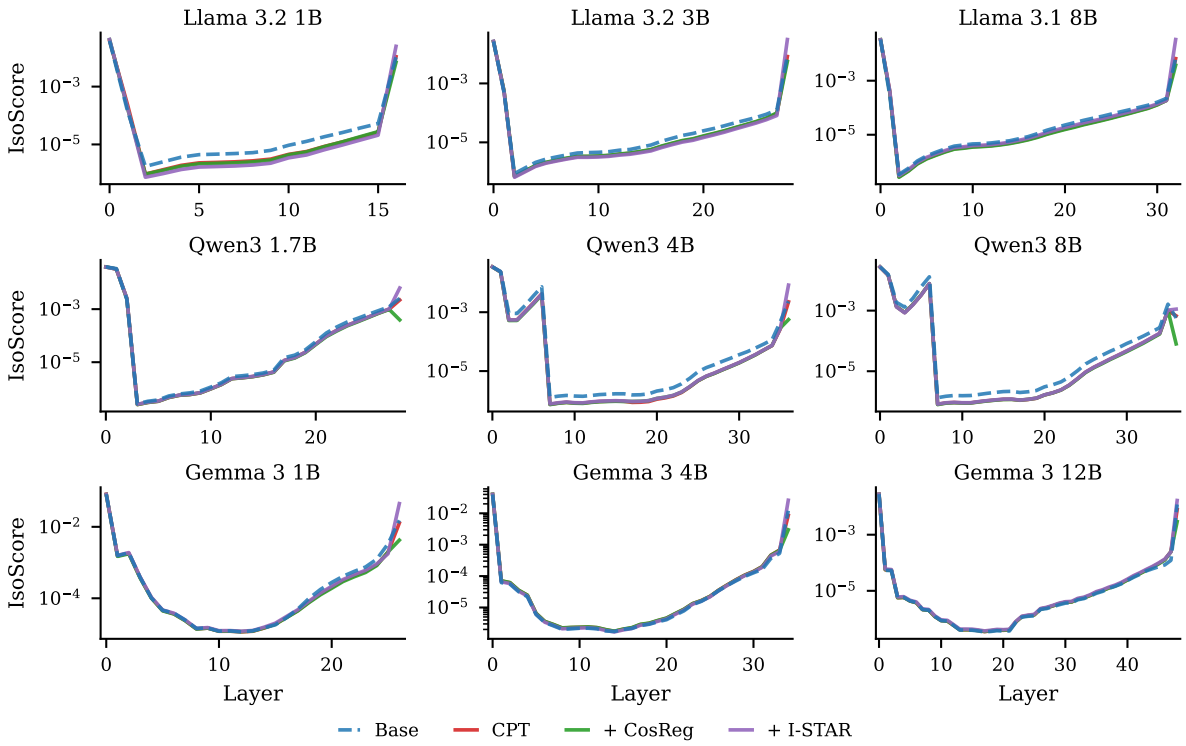

Figure 7: How do CPT and geometric regularisation change isotropy? We plot layer-wise IsoScore, averaged across all 10 languages from our CPT experiments, comparing base models to vanilla CPT models (which only marginally affect isotropy), and geometrically regularised CPT models (I-STAR successfully increases final-layer IsoScore, while leaving other layers unaffected).

Model	Variant	kin	hau	amh	xho	sna	zul	ibo	yor	sot	orm	Avg	kin	hau	amh	xho	sna	zul	ibo	yor	sot	orm	Avg
		AfriXNLI 0-shot (%)											*AfriXNLI* 5-shot (%)										
Llama 3.2 1B	Base	33.3	34.5	32.8	32.3	31.3	33.0	32.7	**34.2**	33.7	30.8	32.9	31.7	33.7	30.7	31.7	29.2	32.3	32.7	**34.7**	32.8	32.5	32.4
	CPT	33.3	**34.8**	**33.0**	**33.5**	33.7	**36.2**	**33.8**	33.8	33.3	**33.5**	**34.0**	**34.3**	33.7	33.8	**34.2**	31.3	35.5	33.2	32.8	**33.8**	**33.0**	**33.5**
	+ CosReg	**33.7**	34.5	32.7	33.3	**34.5**	33.0	33.5	33.0	31.5	33.3	33.4	33.2	**35.0**	**34.2**	33.7	31.2	**35.7**	32.5	32.2	32.7	32.5	33.2
	+ I-STAR	33.3	34.7	32.5	33.2	33.5	34.2	33.7	33.5	**35.2**	**33.5**	33.9	33.0	34.7	33.5	33.2	**31.5**	34.7	**33.8**	33.3	32.7	31.7	33.2
Llama 3.2 3B	Base	33.2	34.2	33.2	33.3	33.3	34.5	33.8	**33.0**	33.5	**34.0**	33.6	**33.3**	**33.7**	34.3	31.8	30.8	33.7	31.8	29.2	**33.0**	33.2	32.3
	CPT	**33.3**	34.2	35.0	**33.7**	**34.3**	34.8	**34.2**	32.5	33.3	33.5	33.8	31.5	33.0	34.5	32.7	33.5	34.7	31.7	30.5	32.8	**34.3**	**32.7**
	+ CosReg	**33.3**	**35.3**	**35.2**	33.5	34.2	**35.7**	**34.2**	32.0	33.3	33.3	**33.9**	30.2	33.0	33.3	32.7	**34.3**	**35.8**	31.8	29.7	30.2	33.8	32.4
	+ I-STAR	**33.3**	34.5	35.0	**33.7**	33.7	35.0	**34.2**	32.5	**35.0**	33.2	**33.9**	31.2	32.5	**34.8**	**33.0**	33.5	33.8	**33.5**	**30.7**	32.7	32.7	32.6
Llama 3.1 8B	Base	32.5	35.2	34.3	33.7	32.3	**35.5**	33.5	30.5	34.2	33.3	33.4	32.3	32.7	31.2	33.7	33.3	33.2	32.8	29.8	32.8	35.7	32.9
	CPT	32.7	40.0	**34.5**	**35.0**	36.2	32.7	34.0	37.0	35.0	32.8	35.0	31.5	39.0	**37.3**	34.8	35.3	37.2	**33.5**	34.7	**36.5**	35.2	35.3
	+ CosReg	34.2	**40.5**	34.2	33.2	**37.3**	34.3	34.0	35.8	**35.3**	33.3	35.3	32.0	**40.0**	36.3	**36.8**	**35.7**	**38.0**	32.8	**37.3**	35.7	**36.7**	**36.1**
	+ I-STAR	**34.7**	39.2	34.2	34.0	35.8	33.7	**34.3**	**39.5**	34.3	**34.0**	**35.5**	**32.7**	39.7	35.7	36.7	33.7	35.8	33.3	35.7	34.8	35.3	35.3
Qwen3 1.7B	Base	33.5	33.3	35.2	33.2	32.8	33.2	33.2	32.5	33.7	**34.0**	33.3	28.7	32.3	32.3	**33.2**	30.2	31.8	31.8	**33.0**	29.3	**32.5**	31.4
	CPT	33.3	32.3	**36.3**	33.3	**34.7**	32.7	**33.5**	**35.2**	32.8	33.0	33.4	**32.2**	**33.2**	34.3	32.7	34.7	32.5	31.8	32.3	**32.5**	31.2	**32.6**
	+ CosReg	**33.8**	33.5	35.2	33.3	34.0	**34.8**	**33.5**	33.5	**35.3**	33.2	**33.9**	**32.2**	32.5	34.0	33.0	34.5	**34.3**	31.5	31.3	30.3	30.8	32.3
	+ I-STAR	33.5	**35.0**	35.0	**33.5**	34.5	31.3	33.3	32.8	33.2	33.3	33.4	32.0	31.5	**34.5**	32.8	**35.0**	33.7	**32.0**	32.7	**32.5**	31.5	**32.6**
Qwen3 4B	Base	32.7	**33.3**	**37.5**	33.2	34.0	33.7	33.7	34.5	33.5	33.5	33.6	31.2	30.3	34.3	32.7	32.3	32.8	31.2	33.2	31.5	**34.0**	32.1
	CPT	**33.5**	32.0	34.8	33.5	34.3	**34.7**	33.5	34.5	**34.5**	33.3	33.8	33.5	31.3	35.7	34.0	31.8	**35.0**	30.7	34.5	33.7	32.7	33.0
	+ CosReg	**33.5**	32.5	34.2	**34.0**	31.3	33.2	**33.8**	32.7	33.5	**33.8**	33.1	**34.2**	**33.0**	**38.5**	34.2	31.8	32.7	32.5	**34.7**	33.5	33.2	**33.3**
	+ I-STAR	**33.5**	32.2	34.3	33.3	**35.3**	34.5	33.5	**36.0**	34.3	33.3	**34.0**	33.0	32.3	36.3	**34.3**	**33.7**	34.3	**32.8**	33.0	**34.0**	32.5	**33.3**
Qwen3 8B	Base	32.2	34.8	33.2	**33.8**	32.5	**35.8**	**34.2**	34.5	34.0	**36.5**	**34.3**	32.3	32.3	37.3	31.8	31.7	34.3	**31.7**	31.3	33.3	**33.5**	32.5
	CPT	**33.2**	34.2	33.7	33.5	33.3	33.8	**34.2**	32.3	34.2	36.3	33.9	32.2	**32.5**	**39.0**	33.8	**32.5**	33.8	29.8	**33.2**	31.7	31.0	32.3
	+ CosReg	30.3	**38.2**	**34.2**	32.3	**33.8**	34.0	31.8	32.7	34.3	35.2	33.6	**34.0**	32.2	38.3	33.3	31.7	**35.0**	31.5	30.5	33.7	31.5	**32.6**
	+ I-STAR	33.0	34.7	33.2	32.2	33.2	33.7	33.2	**35.5**	**35.0**	35.0	33.9	31.3	32.3	38.8	**34.3**	31.7	34.8	31.0	29.2	**35.3**	32.3	32.5
Gemma 3 1B	Base	33.8	35.0	**36.0**	33.3	32.0	33.8	32.8	33.0	33.0	33.8	33.4	32.7	34.0	36.2	33.3	32.3	32.5	**34.5**	32.0	31.0	31.8	32.7
	CPT	**35.2**	37.0	34.7	33.3	35.5	33.3	**33.8**	33.2	34.0	35.2	34.5	**32.8**	35.5	37.0	**34.0**	32.0	33.5	33.5	**34.0**	**32.7**	**32.8**	**33.4**
	+ CosReg	33.0	36.5	34.3	**34.5**	35.2	33.8	**33.8**	**33.3**	34.0	**35.8**	34.4	31.2	**36.2**	**38.2**	**34.0**	30.8	33.5	33.3	33.8	31.8	32.2	33.0
	+ I-STAR	33.7	**38.2**	35.0	34.0	**36.2**	**34.2**	**33.8**	**33.3**	**36.0**	34.8	**34.9**	32.2	35.8	36.3	32.8	**33.0**	**33.8**	32.7	33.2	31.3	32.3	33.0
Gemma 3 4B	Base	35.2	38.0	**38.7**	34.3	36.5	33.2	31.8	35.8	35.8	33.8	34.9	32.7	36.0	39.8	35.8	34.8	33.7	31.2	33.7	34.2	30.8	33.6
	CPT	34.8	39.2	37.0	37.3	**42.5**	33.3	31.3	34.8	37.0	34.5	36.1	35.2	**41.5**	40.5	36.8	34.7	35.5	**35.7**	33.5	**35.8**	32.7	35.7
	+ CosReg	**37.5**	39.8	36.5	**37.8**	41.0	**34.3**	**35.0**	**36.7**	**37.2**	33.0	**36.9**	**35.8**	41.3	**41.3**	35.8	35.2	**35.8**	34.5	**35.2**	34.5	32.5	35.6
	+ I-STAR	35.5	**40.2**	37.7	36.5	39.2	33.2	33.2	34.5	36.2	**36.0**	36.0	35.5	41.0	40.5	**37.3**	**36.2**	34.7	35.3	34.3	35.0	**33.8**	**35.9**
Gemma 3 12B	Base	39.2	**44.5**	38.7	36.3	43.3	**39.2**	38.2	34.2	35.7	**36.7**	38.6	37.2	42.7	43.0	39.3	40.0	**43.2**	35.0	37.2	35.0	32.5	38.0
	CPT	40.8	43.0	41.5	36.2	43.2	38.8	**42.2**	42.3	39.7	34.2	40.0	39.5	43.3	**45.8**	45.8	**42.8**	41.8	38.0	**41.7**	**37.0**	34.2	**40.5**
	+ CosReg	40.0	43.7	42.0	**36.8**	43.5	37.3	40.2	**45.0**	**40.7**	34.0	**40.1**	38.3	44.3	44.7	**46.3**	41.8	42.3	36.8	41.3	36.3	**34.8**	40.3
	+ I-STAR	**41.2**	43.3	**42.2**	36.0	**44.7**	38.5	41.0	42.2	39.0	34.8	**40.1**	**40.0**	**45.3**	44.0	45.2	42.5	**43.2**	**38.2**	39.7	**37.0**	33.2	**40.5**
		AfriMMLU 0-shot (%)											*AfriMMLU* 5-shot (%)										
Llama 3.2 1B	Base	24.4	**27.2**	**25.6**	24.2	**29.8**	26.2	**28.4**	28.2	**28.8**	**26.6**	**27.1**	24.2	26.2	26.8	25.0	26.4	26.2	24.4	26.0	25.4	**28.8**	25.8
	CPT	**25.0**	25.6	25.2	24.0	27.4	25.2	26.2	**28.4**	26.6	26.4	26.1	**26.0**	26.8	26.8	**28.2**	26.8	26.0	**25.2**	**27.2**	25.8	27.8	**26.6**
	+ CosReg	24.8	25.4	**25.6**	**25.0**	29.2	**26.6**	27.0	27.8	26.6	26.0	26.5	24.8	26.2	**27.6**	27.2	**27.4**	26.6	24.8	26.2	**27.4**	**28.8**	**26.6**
	+ I-STAR	**25.0**	25.0	25.2	24.0	28.2	25.8	25.6	27.8	26.4	26.0	26.0	25.2	**27.4**	25.4	27.0	26.6	**27.0**	24.2	26.8	25.2	**28.8**	26.5
Llama 3.2 3B	Base	23.6	24.6	25.8	23.6	27.6	28.4	26.4	**28.0**	27.2	27.8	26.4	26.6	28.8	24.2	24.4	27.6	**28.2**	25.8	**29.6**	27.2	29.0	27.5
	CPT	24.8	25.6	**27.2**	24.0	27.0	27.4	**26.6**	26.4	28.8	27.8	26.5	**27.8**	29.4	**25.8**	**25.4**	**29.2**	27.4	**27.4**	29.4	**29.0**	30.6	**28.4**
	+ CosReg	24.8	**27.0**	26.8	**24.8**	27.0	**28.6**	26.4	26.4	**29.6**	28.4	**27.0**	27.4	**30.0**	**25.8**	24.4	27.6	27.0	27.2	29.0	27.8	**31.6**	28.0
	+ I-STAR	**25.0**	26.4	26.2	24.6	**27.8**	26.4	**26.6**	26.2	27.8	**28.8**	26.6	27.2	29.6	25.4	**25.4**	28.8	26.6	**27.4**	28.6	**29.0**	31.0	28.2
Llama 3.1 8B	Base	29.0	27.2	30.2	27.0	28.2	29.8	29.4	29.4	28.8	29.6	28.7	34.0	35.4	34.6	28.0	30.2	33.2	32.8	32.0	29.0	33.8	32.0
	CPT	30.4	30.2	31.0	28.2	30.2	**30.4**	**30.8**	**30.4**	30.0	**30.8**	**30.2**	34.0	**37.0**	34.8	28.4	**35.0**	**35.2**	**34.8**	**33.2**	31.2	33.2	**33.6**
	+ CosReg	30.0	**30.6**	**31.2**	28.2	30.6	29.2	30.0	29.6	**30.6**	30.0	29.9	**34.4**	36.0	**35.6**	**28.8**	34.0	32.8	34.6	32.8	**32.8**	**34.8**	33.4
	+ I-STAR	**31.2**	30.0	29.4	**28.6**	**30.8**	30.0	**30.8**	29.2	29.4	28.8	29.9	33.6	36.0	33.0	27.6	33.8	33.6	33.6	31.8	30.8	34.0	32.8
Qwen3 1.7B	Base	**27.0**	29.0	30.0	28.2	**30.2**	**27.8**	**28.4**	**32.4**	**29.6**	**30.0**	**29.2**	**29.0**	**27.8**	32.0	**30.2**	29.8	27.2	31.0	30.4	29.6	**32.4**	29.7
	CPT	26.4	28.2	30.6	27.4	29.4	26.8	27.8	30.4	28.2	27.6	28.0	28.8	27.4	32.2	28.6	**30.8**	30.2	30.2	32.0	30.0	30.2	29.8
	+ CosReg	26.4	**29.6**	30.4	26.6	28.6	26.8	28.2	30.4	29.2	27.6	28.2	28.0	**27.8**	**33.4**	28.8	30.4	29.0	**32.0**	32.6	**31.2**	30.2	**30.0**
	+ I-STAR	26.4	**29.6**	**30.8**	**28.4**	29.2	27.2	27.0	30.8	28.4	28.4	28.4	28.6	**27.8**	32.2	29.6	29.8	**31.2**	30.2	**33.0**	29.6	29.6	29.9
Qwen3 4B	Base	**30.6**	**32.8**	**31.6**	**31.2**	31.2	31.8	**31.8**	30.8	**30.0**	**35.8**	**31.8**	**31.4**	35.0	34.8	**34.0**	33.0	33.4	35.6	34.4	35.8	**37.2**	34.4
	CPT	28.8	30.4	31.2	29.4	**32.0**	**32.2**	29.4	**31.6**	**30.0**	34.6	30.9	30.2	**36.6**	38.0	32.6	32.2	34.8	**37.8**	36.0	**36.8**	36.4	**34.8**
	+ CosReg	28.8	31.2	**31.6**	29.8	31.2	31.6	28.8	30.4	29.2	34.6	30.6	30.8	36.0	**38.8**	32.8	32.4	**35.2**	**37.8**	35.6	34.6	35.4	34.5
	+ I-STAR	28.8	30.2	30.4	29.2	**32.0**	31.6	30.0	31.2	**30.0**	35.0	30.9	29.8	36.0	38.2	32.8	**33.2**	34.4	36.6	**36.4**	36.0	36.4	34.6
Qwen3 8B	Base	27.4	33.2	34.4	32.6	31.8	32.0	29.0	34.6	31.4	34.2	31.8	**33.2**	37.2	41.2	**36.0**	**38.4**	39.0	37.4	**37.4**	35.6	38.4	37.0
	CPT	28.6	**35.8**	35.6	**34.2**	32.8	34.8	30.6	34.6	33.4	**36.0**	**33.4**	**33.2**	37.8	43.4	35.8	34.6	39.8	36.2	36.4	38.0	39.4	36.8
	+ CosReg	**29.0**	35.0	**36.2**	33.6	33.2	**36.2**	**30.8**	34.2	**33.6**	34.8	**33.4**	32.8	**38.6**	43.2	34.8	36.0	**40.2**	37.0	37.0	37.4	38.8	37.0
	+ I-STAR	28.0	34.8	**36.2**	33.4	**33.4**	35.2	30.2	**35.0**	**33.6**	34.2	33.1	32.4	37.0	**43.8**	35.6	35.8	**40.2**	**38.4**	37.0	**39.4**	**40.0**	**37.3**
Gemma 3 1B	Base	22.4	22.8	21.2	**26.0**	23.6	**25.2**	24.2	**27.6**	27.6	**25.0**	**24.9**	24.6	**24.6**	22.8	27.8	24.6	26.2	**25.4**	**25.6**	27.6	26.2	25.8
	CPT	22.6	23.0	**22.6**	24.4	24.8	23.6	24.8	24.8	27.2	22.6	24.2	**25.6**	23.8	**23.2**	**28.0**	25.2	**27.6**	23.8	24.6	**28.6**	26.0	**25.9**
	+ CosReg	22.6	23.0	20.8	25.6	24.6	23.8	**25.2**	24.8	**27.8**	21.8	24.4	24.2	22.8	22.4	25.6	26.0	26.0	24.8	24.2	27.4	26.2	25.2
	+ I-STAR	**24.2**	**23.2**	**22.6**	24.6	**25.2**	23.6	24.8	24.4	27.4	22.6	24.4	24.6	23.4	22.6	27.8	**26.2**	26.6	24.6	25.0	27.0	**26.6**	25.8
Gemma 3 4B	Base	27.8	31.4	30.6	32.6	30.4	30.2	30.8	33.4	28.4	29.0	30.4	34.2	36.2	36.8	35.0	35.4	**35.6**	36.4	30.0	33.6	29.4	34.0
	CPT	**28.8**	**35.4**	32.6	32.4	33.0	31.2	**33.4**	30.6	32.8	29.2	31.9	**35.4**	**37.8**	**38.6**	36.2	36.6	34.2	37.0	32.6	34.6	29.8	34.9
	+ CosReg	28.4	35.0	33.2	**33.6**	**34.6**	31.6	32.8	**33.8**	**33.8**	**30.4**	**32.7**	34.8	**37.8**	38.4	**36.4**	**37.6**	32.4	36.2	**32.8**	**35.4**	**32.8**	**35.1**
	+ I-STAR	**28.8**	35.0	**34.2**	**33.6**	32.8	**32.2**	33.0	31.2	32.4	30.0	32.1	35.2	37.2	37.8	35.4	36.4	32.0	**37.4**	32.2	34.2	29.8	34.4
Gemma 3 12B	Base	39.8	42.2	44.2	37.8	36.2	39.2	42.2	35.6	38.0	33.6	38.3	**47.4**	**52.8**	53.4	46.6	52.2	49.2	48.2	44.8	50.0	41.0	48.0
	CPT	**42.4**	48.8	48.6	**46.0**	**44.6**	41.4	45.4	42.2	44.8	38.6	**43.8**	46.8	51.2	**56.4**	**50.2**	**55.4**	**50.8**	48.6	45.0	53.2	**44.8**	**49.6**
	+ CosReg	**42.4**	48.4	48.2	45.2	44.0	40.8	**46.4**	**43.8**	44.4	37.8	43.7	45.4	52.0	55.2	49.2	53.8	50.2	**49.8**	**46.4**	**53.8**	**44.8**	49.5
	+ I-STAR	41.0	**49.6**	**49.0**	45.6	44.0	**41.8**	45.8	41.8	**45.0**	**38.8**	43.7	46.4	52.0	55.2	49.8	53.4	**50.8**	48.6	46.2	51.0	**44.8**	49.2

Table 5: AfriXNLI and AfriMMLU performance (%) across CPT languages. Best variants **boldfaced**.

Model	Variant	kin	hau	amh	xho	sna	zul	ibo	yor	sot	orm	Avg	kin	hau	amh	xho	sna	zul	ibo	yor	sot	orm	Avg
		AfriMGSM 0-shot (%)											*AfriMGSM* 8-shot (%)										
Llama 3.2 1B	Base	2.0	2.4	2.8	3.2	**2.0**	2.0	**1.6**	**2.8**	2.0	**2.0**	2.2	**2.4**	2.8	**1.6**	1.6	2.0	2.4	**2.8**	**3.6**	2.4	**2.4**	2.5
	CPT	**3.2**	**3.6**	2.4	2.4	**2.0**	2.0	0.8	1.2	2.0	**2.0**	2.1	2.0	**4.4**	1.2	**3.2**	**2.8**	2.8	2.4	2.0	2.8	1.6	2.7
	+ CosReg	2.8	3.2	2.8	**3.6**	1.6	**2.4**	1.2	1.2	2.0	**2.0**	2.2	**2.4**	4.0	1.2	**3.2**	**2.8**	**3.6**	2.4	1.6	**3.2**	2.0	**2.8**
	+ I-STAR	**3.2**	**3.6**	**3.2**	2.8	1.6	**2.4**	0.8	1.6	**2.4**	**2.0**	**2.3**	2.0	4.0	1.2	**3.2**	2.4	2.8	2.4	1.6	**3.2**	2.0	2.6
Llama 3.2 3B	Base	2.4	**4.0**	**3.2**	**3.2**	3.2	**3.2**	**3.6**	**4.4**	**2.4**	2.8	**3.2**	**4.8**	**5.6**	2.4	**5.6**	3.6	4.4	2.8	**4.8**	**5.6**	2.8	**4.4**
	CPT	**3.2**	**4.0**	2.8	2.8	**3.6**	1.6	1.6	2.4	**2.4**	3.6	2.8	**4.8**	4.4	**3.6**	4.8	4.0	4.4	2.8	3.6	3.6	3.2	4.0
	+ CosReg	**3.2**	3.6	2.4	2.0	2.4	2.8	2.4	2.0	**2.4**	**4.0**	2.8	**4.8**	4.0	2.8	4.4	**4.4**	**4.8**	**3.2**	3.6	4.0	3.2	4.0
	+ I-STAR	2.4	**4.0**	2.8	2.4	**3.6**	2.4	2.0	2.4	**2.4**	3.6	2.8	**4.8**	4.0	**3.6**	4.0	4.0	4.4	2.8	3.6	4.0	**3.6**	3.9
Llama 3.1 8B	Base	6.4	7.2	2.8	5.2	5.6	4.4	6.4	5.2	5.2	4.0	5.5	**6.8**	5.6	4.8	**5.6**	**5.6**	4.0	**5.6**	**8.0**	**5.2**	**5.6**	**5.8**
	CPT	**9.2**	**9.2**	3.2	5.6	**8.8**	**8.4**	5.6	6.4	8.8	**4.4**	7.4	6.4	**8.8**	5.2	5.2	4.8	**6.4**	4.8	5.6	**5.2**	4.4	5.7
	+ CosReg	**9.2**	8.8	3.2	5.2	8.0	**8.4**	5.2	**8.8**	**10.0**	**4.4**	**7.6**	5.6	**8.8**	4.4	**5.6**	4.4	5.6	4.8	5.6	4.8	4.0	5.5
	+ I-STAR	8.8	7.6	**4.8**	**6.0**	**8.8**	**8.4**	**7.2**	8.0	8.4	**4.4**	7.5	5.6	7.6	**6.0**	5.2	5.2	6.0	4.0	5.2	4.8	4.4	5.3
Qwen3 1.7B	Base	**4.0**	**4.4**	**4.4**	**3.2**	**5.2**	3.2	**3.2**	**3.6**	**4.0**	**2.4**	**3.7**	**4.0**	**5.2**	5.6	**3.2**	**5.2**	**4.4**	**3.6**	**4.8**	**4.0**	**3.6**	**4.2**
	CPT	2.4	2.8	**4.4**	2.8	2.8	2.8	1.2	1.6	3.2	**2.4**	2.4	3.2	2.4	5.2	**3.2**	2.0	1.6	2.0	1.2	2.8	**3.6**	2.4
	+ CosReg	2.0	2.8	3.6	2.4	2.0	2.4	1.6	2.4	3.2	**2.4**	2.4	3.6	2.4	2.8	1.6	2.4	1.6	2.8	0.8	2.8	2.8	2.3
	+ I-STAR	2.4	2.8	4.0	2.8	2.4	**4.0**	1.2	1.6	3.2	**2.4**	2.5	3.2	2.8	**6.0**	2.4	2.4	2.0	2.0	1.2	2.8	**3.6**	2.5
Qwen3 4B	Base	**4.0**	**4.8**	**10.4**	**4.8**	**4.4**	**4.8**	**3.6**	**3.6**	**4.8**	**4.8**	**4.4**	3.6	5.2	10.0	**6.0**	**5.6**	5.2	**4.0**	**4.4**	**6.8**	**5.6**	**5.2**
	CPT	3.2	4.0	6.8	2.0	2.8	2.0	2.0	2.0	3.2	2.0	2.6	**4.8**	**6.4**	**12.4**	4.8	5.2	**6.4**	3.2	**4.4**	6.0	5.2	**5.2**
	+ CosReg	2.8	4.0	6.0	2.0	2.8	2.0	1.6	2.4	3.2	2.0	2.5	3.6	5.6	11.2	4.4	**5.6**	4.8	3.2	3.2	5.2	4.8	4.5
	+ I-STAR	2.8	2.4	6.8	2.4	2.8	2.4	1.6	2.0	4.0	2.0	2.5	**4.8**	5.6	12.0	4.8	5.2	5.2	3.2	3.6	6.4	4.8	4.8
Qwen3 8B	Base	**6.0**	4.8	**16.8**	**4.8**	**5.6**	**7.6**	**4.0**	**6.0**	**5.6**	5.6	**5.6**	**7.2**	6.0	**16.0**	7.2	**7.2**	5.6	**4.8**	**6.8**	8.4	**7.2**	**6.7**
	CPT	5.2	**9.2**	9.6	**4.8**	3.6	5.2	2.8	3.6	4.4	**6.0**	5.0	4.4	8.0	12.8	7.2	6.0	5.6	4.4	4.0	8.0	6.4	6.0
	+ CosReg	4.8	8.0	10.0	**4.8**	4.8	5.2	2.4	3.2	4.4	5.6	4.8	3.6	**9.6**	10.8	7.2	4.8	**6.8**	4.4	4.4	8.8	6.4	6.2
	+ I-STAR	5.6	7.2	8.8	4.4	4.0	4.0	2.8	3.6	**5.6**	4.8	4.7	4.0	8.4	13.2	**8.0**	5.6	**6.8**	4.0	4.8	**9.2**	6.0	6.3
Gemma 3 1B	Base	1.6	2.0	**2.0**	1.2	2.8	1.2	1.2	0.8	1.2	1.6	1.5	2.4	3.6	3.6	2.8	3.2	3.2	2.4	3.6	3.6	3.2	3.1
	CPT	1.6	**2.8**	1.6	**2.0**	**3.6**	**2.0**	**1.6**	1.2	1.6	2.0	**2.0**	5.2	**5.2**	6.0	**4.8**	**4.8**	**4.8**	**5.2**	**5.6**	4.4	**6.0**	5.1
	+ CosReg	1.6	1.2	1.2	1.2	1.6	1.6	**1.6**	1.2	**2.8**	2.0	1.6	**6.0**	4.8	4.8	4.4	**4.8**	**4.8**	**5.2**	**5.6**	**4.8**	**6.0**	**5.2**
	+ I-STAR	**2.0**	2.4	1.6	1.6	1.6	1.6	**1.6**	**1.6**	2.0	**2.4**	1.9	5.6	**5.2**	**6.4**	4.4	**4.8**	**4.8**	4.8	**5.6**	4.0	**6.0**	5.0
Gemma 3 4B	Base	2.4	4.4	**10.0**	3.2	2.4	4.0	2.4	2.8	3.2	1.2	2.9	7.2	7.6	9.2	6.4	6.8	7.2	5.6	3.6	7.2	3.6	6.1
	CPT	5.2	6.0	8.0	5.2	4.4	4.4	**3.2**	3.6	**5.6**	4.4	4.7	10.8	10.4	**12.8**	8.8	10.0	9.6	**8.8**	**7.6**	8.8	8.0	9.2
	+ CosReg	**5.6**	**7.2**	8.4	**5.6**	**5.2**	**5.6**	**3.2**	**5.2**	4.4	**5.6**	**5.3**	10.4	10.4	12.0	**9.2**	**10.8**	**10.4**	**8.8**	**7.6**	**9.2**	**8.8**	**9.5**
	+ I-STAR	4.0	6.8	7.2	4.8	4.8	4.8	2.4	4.0	4.0	4.4	4.4	**11.2**	**10.8**	12.4	8.0	10.0	9.6	**8.8**	6.8	**9.2**	8.4	9.2
Gemma 3 12B	Base	14.0	15.6	**19.2**	11.2	11.2	12.4	8.8	7.6	11.6	8.0	11.2	20.8	20.8	21.6	18.4	**23.2**	20.0	14.4	12.8	19.6	14.4	18.3
	CPT	15.6	16.0	14.8	13.2	15.2	17.6	**11.2**	12.4	16.8	15.2	14.8	**24.4**	**23.2**	**24.4**	**20.4**	22.4	**24.4**	17.6	17.2	**23.6**	20.8	**21.6**
	+ CosReg	**17.6**	**16.4**	13.6	12.8	16.4	**18.0**	10.4	13.2	**17.6**	**16.8**	15.5	23.2	**23.2**	22.0	17.6	22.4	21.2	**18.0**	17.2	**23.6**	**22.4**	21.0
	+ I-STAR	16.0	**16.4**	14.0	**14.4**	**17.6**	**18.0**	10.4	**13.6**	**17.6**	**16.8**	**15.6**	24.0	22.4	24.0	18.8	22.0	23.6	17.6	**18.8**	22.4	**22.4**	21.3
		AfriMGSM CoT 0-shot (%)											*AfriMGSM CoT* 8-shot (%)										
Llama 3.2 1B	Base	**2.8**	1.2	**2.0**	2.0	2.0	**2.4**	**2.0**	**2.4**	1.6	**2.0**	2.0	3.2	2.0	**2.0**	**3.2**	**4.8**	2.0	**2.0**	**3.2**	3.2	**2.4**	**2.9**
	CPT	2.0	1.6	1.6	**3.2**	**2.8**	2.0	**2.0**	2.0	2.0	**2.0**	2.2	2.8	**2.8**	1.6	2.8	2.8	2.4	1.2	2.4	2.4	1.6	2.4
	+ CosReg	**2.8**	**2.0**	1.6	**3.2**	2.0	**2.4**	**2.0**	**2.4**	**2.4**	**2.0**	**2.4**	2.4	2.4	1.6	**3.2**	2.8	**2.8**	1.2	1.2	**3.6**	2.0	2.4
	+ I-STAR	2.4	**2.0**	**2.0**	2.4	2.0	1.6	1.6	**2.4**	2.0	**2.0**	2.0	**3.6**	**2.8**	0.8	**3.2**	2.8	2.0	1.2	2.8	2.4	1.6	2.5
Llama 3.2 3B	Base	2.0	2.0	2.0	**3.2**	2.0	**3.2**	1.2	1.6	**2.8**	2.0	2.2	2.8	**6.0**	**3.6**	3.2	3.6	2.4	**2.4**	3.6	**4.8**	4.4	3.7
	CPT	2.8	2.8	**2.8**	2.4	**2.8**	1.6	2.4	2.4	2.4	**3.2**	2.5	4.0	3.2	2.4	**5.2**	**4.8**	**4.4**	1.6	4.0	3.6	**5.2**	**4.0**
	+ CosReg	**3.2**	2.4	2.0	2.8	2.4	2.0	**3.2**	**2.8**	2.0	2.4	**2.6**	3.6	2.8	2.8	3.2	4.0	4.0	**2.4**	**4.8**	4.0	4.8	3.7
	+ I-STAR	2.0	**3.2**	2.4	2.8	**2.8**	1.2	1.6	2.0	2.0	**3.2**	2.3	**4.4**	3.2	2.4	2.8	4.4	4.0	2.0	4.0	2.0	4.4	3.5
Llama 3.1 8B	Base	4.4	7.6	**3.6**	3.6	3.6	2.0	**4.4**	3.6	4.8	2.0	4.0	7.6	15.6	3.6	4.8	8.0	6.8	8.4	6.8	7.2	6.4	8.0
	CPT	5.6	9.2	**3.6**	4.4	8.8	**7.6**	**4.4**	4.4	6.4	4.0	**6.1**	14.0	**16.8**	**5.6**	10.8	**15.2**	12.0	**14.0**	**11.2**	14.4	8.8	**13.0**
	+ CosReg	**6.4**	**10.0**	2.8	**4.8**	**9.6**	4.8	**4.4**	**4.8**	**7.6**	2.4	**6.1**	14.0	14.0	4.8	**12.4**	14.4	**12.8**	13.2	**11.2**	**15.2**	8.8	12.9
	+ I-STAR	**6.4**	8.0	2.8	**4.8**	8.4	5.2	4.0	**4.8**	5.6	**4.4**	5.7	**14.4**	12.8	4.4	9.6	14.0	10.4	12.8	**11.2**	14.4	**10.0**	12.2
Qwen3 1.7B	Base	**3.6**	2.8	**7.2**	3.2	**3.6**	**2.8**	3.6	2.4	**2.0**	**2.4**	**2.9**	**4.0**	5.2	3.6	3.2	**6.4**	**6.4**	**3.2**	5.2	**4.8**	4.4	**4.8**
	CPT	2.0	**3.2**	4.4	3.2	3.2	2.4	2.4	2.0	1.2	**2.4**	2.4	3.2	**5.6**	4.0	**4.4**	4.4	3.2	2.4	**6.4**	2.8	**5.2**	4.2
	+ CosReg	2.4	2.8	3.6	**4.0**	2.0	**2.8**	3.6	**3.2**	**2.0**	**2.4**	2.8	3.2	**5.6**	2.4	3.2	5.2	4.8	2.4	5.6	3.6	3.6	4.1
	+ I-STAR	2.4	2.0	4.0	2.0	2.4	1.6	**4.4**	2.4	1.6	2.0	2.3	**4.0**	4.4	**4.4**	**4.4**	4.4	2.8	2.4	6.0	4.0	4.8	4.1
Qwen3 4B	Base	**3.6**	7.6	**7.6**	**5.2**	**6.0**	**4.8**	**3.2**	**4.4**	**6.4**	4.4	**5.1**	**8.4**	8.4	10.8	**7.6**	8.8	8.4	2.8	7.2	**10.4**	**10.8**	8.1
	CPT	2.0	9.2	4.8	3.2	3.6	2.0	2.4	2.8	3.2	4.0	3.6	6.0	12.4	10.4	6.8	8.4	8.4	**3.2**	7.2	9.2	9.2	7.9
	+ CosReg	3.2	**10.4**	3.6	3.2	4.0	4.4	2.8	4.0	2.8	**5.6**	4.5	7.6	**15.2**	8.8	7.2	**9.6**	8.8	2.8	7.6	8.8	8.4	**8.4**
	+ I-STAR	2.0	9.2	4.4	3.6	5.2	4.0	2.4	3.2	2.4	**5.6**	4.2	**8.4**	13.2	**11.2**	6.4	8.0	**9.2**	2.8	**8.0**	8.0	9.2	8.1
Qwen3 8B	Base	7.6	6.4	**19.2**	**7.6**	6.8	**8.4**	**4.0**	**5.2**	6.8	**8.8**	6.8	10.4	8.8	15.2	10.8	9.2	9.2	3.6	9.2	12.4	13.2	9.6
	CPT	6.8	**12.4**	8.4	3.2	7.2	8.0	3.6	**5.2**	**9.6**	5.2	6.8	10.8	16.8	**16.8**	10.8	10.8	12.0	5.6	**10.8**	**15.6**	14.0	11.9
	+ CosReg	**8.4**	10.0	9.2	3.6	**9.2**	7.2	3.2	4.8	9.2	6.8	**6.9**	**11.2**	**23.2**	15.6	**12.4**	**13.2**	12.0	**6.0**	10.4	13.6	**14.8**	**13.0**
	+ I-STAR	6.0	10.0	8.0	4.4	6.8	6.0	3.6	4.4	8.8	5.6	6.2	10.8	**23.2**	**16.8**	10.8	10.8	**12.8**	4.8	**10.8**	14.8	12.8	12.4
Gemma 3 1B	Base	3.2	1.6	**2.4**	2.0	1.6	1.6	**1.6**	2.0	1.6	1.6	1.9	3.6	**2.8**	**3.2**	1.6	**3.2**	1.6	2.0	1.6	1.2	**2.8**	2.3
	CPT	**4.0**	**3.2**	1.6	1.6	2.4	2.0	1.2	2.0	**2.0**	2.4	**2.3**	3.2	1.6	2.8	2.8	2.0	2.8	2.8	**2.4**	**2.0**	1.6	2.4
	+ CosReg	2.8	2.4	1.6	2.0	**3.2**	1.6	0.8	1.6	**2.0**	**3.2**	2.2	**4.4**	2.4	2.0	**3.2**	**3.2**	**4.0**	2.0	1.6	1.2	**2.8**	**2.8**
	+ I-STAR	3.2	2.0	1.2	**2.4**	1.6	**2.8**	1.2	**2.4**	1.6	2.4	2.2	2.4	2.0	2.8	**3.2**	2.8	2.0	**3.2**	2.0	1.6	2.0	2.4
Gemma 3 4B	Base	5.2	7.6	**10.8**	6.8	4.4	5.2	**5.6**	2.8	5.6	3.2	5.2	9.2	15.2	19.2	10.0	14.4	12.8	5.6	6.0	10.4	4.0	9.7
	CPT	**10.0**	9.6	6.4	8.0	7.6	**7.6**	3.6	4.0	6.0	3.6	6.7	15.2	20.0	18.4	13.6	**18.4**	15.6	10.4	5.6	**14.8**	**8.8**	**13.6**
	+ CosReg	9.2	**10.0**	6.8	6.8	**9.2**	6.8	2.8	**5.2**	**7.2**	**5.6**	**7.0**	**16.4**	**21.2**	20.4	13.2	18.0	**16.4**	8.8	6.0	13.2	8.4	13.5
	+ I-STAR	9.6	9.6	7.2	**8.4**	6.8	6.8	3.6	4.4	6.4	4.4	6.7	15.6	18.4	**21.2**	**14.4**	18.0	14.4	**10.8**	**6.4**	14.4	8.0	13.4
Gemma 3 12B	Base	16.8	**23.6**	23.2	11.6	14.4	16.4	10.0	14.4	**18.8**	10.0	15.1	26.8	37.2	**46.4**	27.6	33.2	30.4	24.0	7.6	29.6	17.6	26.0
	CPT	23.2	22.4	21.6	9.6	16.4	16.0	**11.6**	**18.0**	14.0	16.4	16.4	**39.6**	41.2	40.8	31.2	**38.0**	37.6	**29.2**	28.4	36.4	32.8	34.9
	+ CosReg	**23.6**	21.2	**24.0**	**16.8**	**17.6**	15.6	10.4	**18.0**	16.0	**19.2**	**17.6**	**39.6**	**42.0**	42.8	30.4	34.4	38.0	28.8	**30.8**	**41.6**	32.8	**35.4**
	+ I-STAR	19.2	23.2	22.0	10.4	**17.6**	**19.6**	10.8	17.6	13.6	18.0	16.7	38.4	40.4	42.4	**31.6**	36.4	**38.4**	26.4	26.8	37.2	**33.2**	34.3

Table 6: AfriMGSM direct and chain-of-thought performance (%) across CPT languages. Best variants **boldfaced**.