MULTILINGUAL LATENT REPRESENTATION DISTORTION DRIVES GENERATIVE DRIFT AND BOUNDS LATENT-CHANNEL COMMUNICATION CAPACITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual large language models (LLMs) internalize conceptual reasoning within an English-centric latent frame, while low-resource languages (LRLs) occupy activation subspaces that are compact relative to English ($\|h_{\text{en}}\|/\|h_{\text{tgt}}\| \in [1.5, 4.8]$ under mean pooling; 1.15–1.46 per token excluding the BOS position). Across **seven** LLM backbones, **122** `Belebele` configurations and **220** `FLORES+` varieties, we show that this single geometric property is closely associated with two failures previously studied in isolation: (i) **intra-model language drift** during free-form generation, and (ii) the limited capacity of continuous latent channels for **inter-agent communication**. First, contrastive Singular Value Decomposition isolates a low-rank language surface subspace $\mathbf{U_L}$ from an orthogonal reasoning subspace $\mathbf{U_R}$ that is cross-lingually isomorphic to English in a model-invariant way (debiased CKA, Procrustes, and RSA; Spearman's $\rho = 0.71$–1.00 across six model lineages). Building on this geometry, we design a *training-free*, depth-stratified and magnitude-normalized steering method that bounds relative layer-wise perturbation. Our method reduces Involuntary Fidelity Loss by $\mathbf{40}\%$ on low-resource scripts (Macro IFL $0.491 \rightarrow 0.295$ on `SEA-LION-v3-8B`) and by 16–45% relative across **six** backbones, while leaving safety-classification accuracy within confidence intervals. Second, we introduce a *specificity protocol* for continuous multi-agent latent channels across **nine languages**, and a Universal Latent Hub that extends continuous communication to *heterogeneous* model pools with mismatched hidden dimensions. We establish the token-budget crossover where continuous latents outperform literal text on high-fertility scripts. Finally, we examine the *unifying bridge*: on the six languages shared by both programs, activation distortion and tokenizer fertility are collinear ($\rho = 0.83$ under mean pooling), suggesting a single resource-scarcity axis associated with both single-model generative drift and multi-agent channel capacity, a pattern consistent with fertility and accuracy trends across **122** `Belebele` configurations.

1 INTRODUCTION

Modern multilingual large language models (LLMs) display a pronounced performance disparity between high-resource pivot languages (predominantly English) and low-resource languages (LRLs) (Artetxe et al., 2020; Wendler et al., 2024). Two distinct symptoms of this disparity have so far been investigated by separate research communities:

Intra-Model Generative Drift: At the level of a single model, low-resource generation suffers from Involuntary Fidelity Loss (IFL): the inability to sustain a target script during multi-step reasoning, causing intermediate thoughts and final generations to collapse into the English pivot (Tang et al., 2024; Zhao et al., 2025). Standard prompt engineering and uncalibrated steering often exacerbate this instability or degrade downstream reasoning accuracy.

Inter-Model Latent Capacity Bottleneck: At the level of multiple communicating agents, continuous latent channels, proposed to bypass the severe 'token tax' of discrete natural language hand-offs in high-fertility scripts (Petrov et al., 2023; Ahia et al., 2023a), face severe capacity starvation. For

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the same set of high-fertility, low-resource languages, soft-prefix communication yields minimal instance-specific transfer, often collapsing into an input-independent register cue (Petrov et al., 2024).

Our investigations suggest that both phenomena are downstream consequences of one underlying property of multilingual representation geometry: **low-resource activation spaces are compact and distorted relative to English** ($\|h_{\text{en}}\|/\|h_{\text{tgt}}\| \in [1.5, 4.8]$ across four backbones under mean pooling, Table 1 and Appendix Figure 4), occupying a compressed, low-effective-rank region separable from a shared, English-isomorphic reasoning manifold. This compactness produces two failures: *(i) intra-model*, steering vectors scaled to English norms overwhelm compact target activations, corrupting boundary syntax and unembedding unless rescaled and depth-scheduled; *(ii) inter-model*, continuous thoughts sent through a fixed-width prefix channel starve when the language geometry lacks usable rank, so receivers exploit prefixes as static register cues.

Our contributions are:

1. Geometric Theory Across Diverse Backbones: We formulate the contrastive SVD decomposition $U_L \oplus U_R$, prove propositions on subspace alignment and bounded layer-wise perturbation, and show across seven open-weight architectures and **220 FLORES+** varieties that U_R is cross-lingually isomorphic to English in a model-invariant way ($\rho = 0.71$–1.00).

2. Pareto-Optimal Latent Steering: Depth-stratified Gaussian scheduling with dynamic magnitude normalization reduces Macro IFL by 40% on low-resource scripts ($0.491 \rightarrow 0.295$) and by 16–45% relative across six backbones. Against five baseline families on four backbones, a gated variant avoids the MCQA accuracy collapse of unnormalized steering, and safety-classification accuracy stays within confidence intervals.

3. Certified Specificity Protocol & Latent Communication Bottlenecks: Across nine languages, our specificity protocol separates input-independent register cues (83.3% of the naive prefix gain against a trained constant-prefix control) from certified instance-specific content transfer ($+0.056$ nats pooled, $p \leq 10^{-4}$). Our Universal Latent Hub extends continuous communication to heterogeneous pools with mismatched hidden dimensions ($3584 \neq 4096$), and end-task evaluation shows that continuous thought exchange degrades on compact low-resource representations (0.210 across nine languages vs. 0.467 on English/Thai/Swahili) and does not yet surpass a single-agent baseline.

2 Related Work

Representation Engineering and Activation Steering. Inference-time representation engineering modifies latent activations without parameter updates (Zou et al., 2023; Turner et al., 2023; Li et al., 2023; Subramani et al., 2022). ReCoVeR (Sterz et al., 2025) extracts linear steering vectors; CLaS-Bench (Gurgurov et al., 2026) evaluates DiffMean, LAPE, and PCA. While recent work identifies steering brittleness and "anti-steerability" across diverse tasks (Tan et al., 2024; Lee et al., 2025), we address these instabilities through formal subspace geometry, continuous Gaussian scheduling, and dynamic magnitude normalization with bounded relative perturbations.

Multilingual Geometry and Subspace Ablation. Multilingual LLMs route conceptual reasoning through an English-like pivot (Wendler et al., 2024; Tang et al., 2024; Conneau et al., 2020), with language identity concentrating in low-rank subspaces (Gurgurov et al., 2025). While Zhao et al. (2025) ablate language subspaces, we differentiate by: *(i)* formalizing contrastive SVD decomposition $U_L \oplus U_R$ with an alignment proposition; *(ii)* establishing cross-architecture invariance across six backbones ($\rho = 0.71$–1.00); *(iii)* resolving magnitude distortion via dynamic rescaling; and *(iv)* connecting this geometry to inter-agent communication.

Latent Multi-Agent Communication & Causal Audits. While multi-agent LLMs traditionally communicate via natural language (Wu et al., 2023; Hong et al., 2024; Tran et al., 2025), continuous channels have emerged in MARL (Sukhbaatar et al., 2016; Foerster et al., 2016) and latent reasoning (Hao et al., 2024). In LLMs, LatentMAS (Zou et al., 2026) shares hidden states homogeneously; ThoughtComm (Zheng et al., 2025) separates private thoughts; Cache-to-Cache (Fu et al., 2026) fuses KV-caches at $O(N^2)$ cost; and Vision Wormhole (Liu et al., 2026) couples visual pathways. Concurrently, Zhang & Emu (2026) introduce an English-only causal audit. We advance beyond this by: *(i)* evaluating across nine languages, eight of them non-English; *(ii)* isolating input-independent

register cues (Δ_{cue}) via matched controls; and *(iii)* bridging channel capacity directly to representation geometry and tokenizer fertility.

3 METHODOLOGY

3.1 MULTILINGUAL REPRESENTATION GEOMETRY

Let $\mathbf{h}^{(\ell)} \in \mathbb{R}^d$ denote the residual stream activation vector at layer $\ell \in \{1, \ldots, L\}$. Given K semantically parallel sentence pairs $(x_k^{\text{en}}, x_k^{\text{tgt}})_{k=1}^K$, we extract mean-pooled hidden states $\mathbf{H}_{\text{en}}^{(\ell)}, \mathbf{H}_{\text{tgt}}^{(\ell)} \in \mathbb{R}^{K \times d}$. We define the **contrastive difference matrix**:

$$C^{(\ell)} = \mathbf{H}_{\text{tgt}}^{(\ell)} - \mathbf{H}_{\text{en}}^{(\ell)} \in \mathbb{R}^{K \times d}. \tag{1}$$

Computing the truncated Singular Value Decomposition (SVD) $C^{(\ell)} = U\Sigma V^\top$ partitions the activation space $\mathbb{R}^d$ into an r-dimensional **language surface subspace** $U_L = \text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_r)$ and an orthogonal **reasoning subspace** $U_R = U_L^\perp$. We fix $r = 16 \ll d$ and fit one pooled decomposition per steered layer ($K = 1200$ mean-centered pairs over six languages); U_L then captures 64–98% of the contrastive energy, depending on backbone and depth (Appendix Table 4).

Hypothesis 1 (Language-Reasoning Disentanglement)**.** *Within deep transformer layers, hidden representations decompose orthogonally as $\mathbf{h}^{(\ell)} = \mathbf{h}_L^{(\ell)} + \mathbf{h}_R^{(\ell)}$, where $\mathbf{h}_L^{(\ell)} \in U_L$ encodes language-specific surface morphology and $\mathbf{h}_R^{(\ell)} \in U_R$ encodes language-agnostic conceptual reasoning.*

Proposition 1 (Cross-Lingual Isomorphism of U_R)**.** *Let $P_R = I - U_L U_L^\top$ denote the projection onto U_R, and let K_{tgt}, K_{en} be the centered linear Gram matrices of $\mathbf{H}_{tgt}^{(\ell)}, \mathbf{H}_{en}^{(\ell)}$. Under Hypothesis 1, assume the surface components carry no cross-lingual similarity structure, i.e. $\text{Tr}(K_{tgt}K_{en}) = \text{Tr}(K_{tgt}^R K_{en}^R)$, where K^R denotes the Gram matrix of the U_R component. Then, whenever the target surface component is non-zero and $\text{Tr}(K_{tgt}^R K_{en}^R) > 0$, linear Centered Kernel Alignment (CKA) satisfies:*

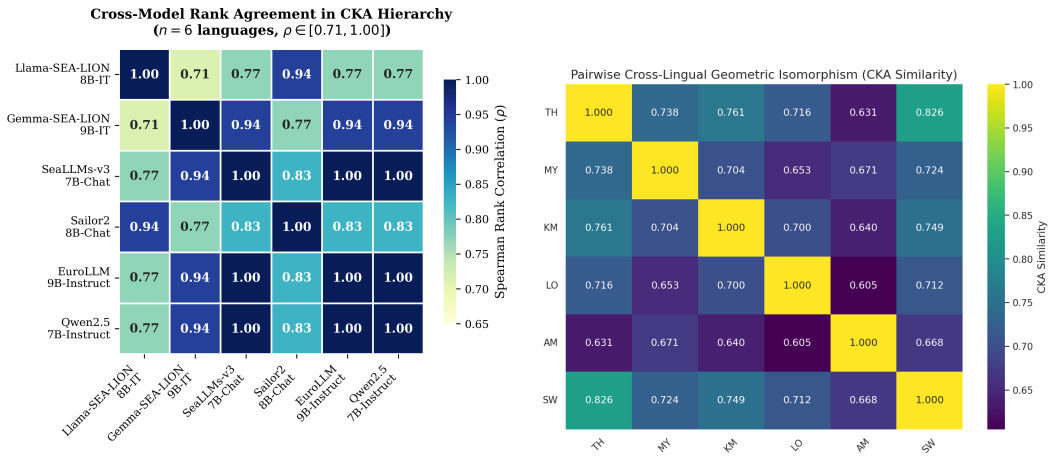
$$\text{CKA}(P_R\mathbf{H}_{tgt}^{(\ell)}, P_R\mathbf{H}_{en}^{(\ell)}) > \text{CKA}(\mathbf{H}_{tgt}^{(\ell)}, \mathbf{H}_{en}^{(\ell)}). \tag{2}$$

(Proof in Appendix A.) Proposition 1 states when removing language-discriminative surface components increases cross-lingual alignment; empirically we report the debiased estimator. Because U_L is estimated from the same parallel pairs, we treat the measured increase as a consistency check rather than independent evidence; the stronger test of U_R is whether steering built on it reduces drift on held-out generations (§5.1).

Table 1: **Master geometric alignment and channel capacity table across shared low-resource languages.** Geometric alignment in U_R ($\ell \approx 0.65L$, Llama-SEA-LION-8B) matched with tokenizer fertility (tokens/sentence) and certified channel specificity Δ_{spec} (nats, from Table 3). Debiased CKA and RSA: higher = more aligned. Procrustes residual and Distortion Ratio: lower = more aligned. Distortion is mean-pooled and includes the BOS position (see §3.1). Ranked by activation distortion ratio.

Language	Family / Script	Distortion ↓	CKA ↑	RSA ↑	Procrustes ↓	Fertility	Δ_{spec} (nats) ↑
Swahili (sw)	Niger-Congo (Latin)	*1.489*	*0.763*	*0.666*	*0.222*	50.1	**+0.020***
Thai (th)	Kra-Dai (Thai)	1.544	0.694	0.632	0.207	66.6	**+0.091*****
Amharic (am)	Afroasiatic (Ethiopic)	2.301	0.653	0.517	0.339	102.3	+0.001
Khmer (km)	Austroasiatic (Khmer)	2.492	0.521	0.439	0.323	174.0	**+0.008****
Burmese (my)	Sino-Tibetan (Myanmar)	2.528	0.395	0.377	0.362	228.4	+0.001
Lao (lo)	Kra-Dai (Lao)	2.651	0.506	0.426	0.356	146.5	+0.003
Mean (non-Latin)	—	**2.303**	**0.554**	**0.478**	**0.317**	**143.6**	**+0.021**

Triangulated geometric isomorphism and cross-model invariance. To avoid metric-specific artifacts, we evaluate U_R using three independent similarity measures: debiased U-statistic CKA (Song et al., 2012; Murphy et al., 2024), orthogonal Procrustes distance, and Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008). As reported in Table 1, the three metrics broadly agree: Swahili and Thai sit closest to English, while Khmer, Lao, and Burmese exhibit substantial geometric separation.



(a) Cross-model Spearman ρ.

(b) Pairwise CKA matrix in U_R.

Figure 1: **Geometric invariance across model lineages.** (a) Pairwise Spearman rank correlations between deep-layer CKA profiles across six distinct model families ($\rho = 0.71$–1.00). (b) Pairwise CKA similarity matrix of U_R.

Crucially, this geometric hierarchy is **model-invariant**. We compute pairwise Spearman rank correlations of deep-layer CKA profiles across six distinct model lineages: `Llama-SEA-LION-8B`, `Gemma-SEA-LION-9B`, `Sailor2-8B`, `SeaLLMs-v3-7B`, `EuroLLM-9B`, and `Qwen2.5-7B`. Figure 1(a) reveals high cross-model agreement ($\rho = 0.71$–1.00, with all 15 pairwise correlations positive and 10 of 15 significant at $p < 0.05$ at $n = 6$), consistent with multilingual transformers converging to a similar geometric alignment structure.

Activation magnitude distortion. Importantly, low-resource activations are not merely shifted in direction; they exhibit severe **magnitude distortion**:

$$r_l^{(\ell)} = \frac{\mathbb{E}[\|\mathbf{h}_{\text{en}}^{(\ell)}\|_2]}{\mathbb{E}[\|\mathbf{h}_{\text{tgt},l}^{(\ell)}\|_2]}, \tag{3}$$

which ranges from 1.5 to 4.8 across languages and backbones. As shown in Table 1 and Appendix Figure 4, Burmese, Lao, and Khmer activations have less than half the norm of parallel English activations. Part of this ratio is a length artifact: the first (BOS) position carries an outlier norm whose weight in a mean falls with sequence length. Excluding it on SEA-LION-v3-8B (eight languages), the per-token ratio shrinks to 1.15–1.46 (mean-pooled 1.23–1.79) but stays above 1 for every language and still tracks channel specificity (Spearman $\rho = -0.75$, $p = 0.03$). This distortion ratio is the foundational quantity that connects intra-model steering instability to inter-model latent channel starvation.

3.2 Intra-Model Intervention: Geometry-Aware Latent Steering

Given the raw contrastive mean difference $\mathbf{v}_{\text{raw}}^{(\ell)} = \frac{1}{K}\sum_k(\mathbf{h}_{\text{tgt},k}^{(\ell)} - \mathbf{h}_{\text{en},k}^{(\ell)})$, we project out language-surface components to steer semantic reasoning:

$$\mathbf{v}_{\text{steer}}^{(\ell)} = \mathbf{v}_{\text{raw}}^{(\ell)} - \lambda P_L \mathbf{v}_{\text{raw}}^{(\ell)}, \quad \lambda \in [0, 1]. \tag{4}$$

Functional layer stratification and Gaussian depth scheduling. Transformers exhibit functional stratification across depth (nostalgebraist, 2020): early layers ($\ell \leq 0.2L$) execute tokenization and local syntax; mid-to-late layers ($\ell \approx 0.5L$–$0.8L$) perform abstract semantic reasoning; terminal layers ($\ell \geq 0.9L$) project onto vocabulary logits. To avoid disrupting boundary mechanisms, we modulate injection via a continuous Gaussian profile:

$$\alpha_\ell = \alpha_0 \exp\left(-\frac{(\ell - \mu_s)^2}{2\sigma_s^2}\right), \quad \mu_s = 0.65L,\ \sigma_s = 0.15L. \tag{5}$$

Proposition 2 (Boundary Layer Invariance). *With $\mu_s = 0.65L$ and $\sigma_s = 0.15L$, the steering injection weight satisfies $\alpha_\ell \leq 0.011\alpha_0$ for early layers ($\ell \leq 0.2L$) and $\alpha_L \leq 0.066\alpha_0$ at the terminal layer ($\ell = L$).*

(Proof in Appendix A.) Early layers ($\ell \leq 0.2L$), which precede all substantial injection, are therefore left essentially unperturbed, and the direct injection at the final layer is small. The final residual stream still carries the norm-bounded perturbations injected upstream, which is the intended effect.

Dynamic magnitude normalization. To eliminate the activation norm disparity, we dynamically scale injection relative to the target language's local activation norm:

$$\gamma_\ell = \eta \frac{\mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]}{\|\mathbf{v}_{\text{steer}}^{(\ell)}\|_2}, \tag{6}$$

yielding the layer-wise intervention hook $\mathbf{h}^{(\ell)} \leftarrow \mathbf{h}^{(\ell)} + \alpha_\ell \gamma_\ell \mathbf{v}_{\text{steer}}^{(\ell)}$.

Proposition 3 (Bounded Relative Perturbation). *Under dynamic magnitude normalization, the relative perturbation is strictly bounded across all layers:*

$$\frac{\|\Delta\mathbf{h}^{(\ell)}\|_2}{\mathbb{E}[\|\mathbf{h}_{tgt}^{(\ell)}\|_2]} = \alpha_\ell \eta \leq \alpha_0 \eta, \quad \forall \ell \in \{1, \ldots, L\}. \tag{7}$$

(Proof in Appendix A.) Dynamic normalization thus keeps each layer's injection at a fixed fraction of the typical target activation norm, independent of the language's distortion ratio; summed over layers, the injected norm is at most $\eta \sum_\ell \alpha_\ell$ times that scale.

Distortion-weighted multilingual aggregation. A shared steering vector over M languages weights each per-language vector inversely by its distortion ratio, $\bar{\mathbf{v}}^{(\ell)} = \sum_l w_l \mathbf{v}_l^{(\ell)} / \sum_l w_l$ with $w_l = 1/r_l^{(\ell)}$, so that severely distorted languages do not inject disproportionate noise.

3.3 Inter-Model Communication: Universal Latent Hub

Petrov et al. (2024) show that a fixed prefix cannot alter relative attention over content; it can only bias attention outputs in a static direction. An input-dependent prefix may likewise gain from such a static component, lowering a receiver's loss purely as an *input-independent register cue* (e.g., priming fluent multilingual text) without transmitting instance-specific information. Any single-payload evaluation against an unconditioned baseline is confounded by this effect.

Universal Hub and Soft-Prefix Projector. We consider heterogeneous agents $M_1, \ldots, M_N$ with distinct hidden dimensions $d_1, \ldots, d_N$. Direct mid-stack activation splicing corrupts generation fluency (producing keyword recall of 0.000 and degenerate token repetition; Table 10). We therefore adopt a two-stage delivery architecture:

Universal Latent Hub ($O(N)$ Adapters): Linear adapters $E_i : \mathbb{R}^{d_i} \to \mathbb{R}^u$ and $D_i : \mathbb{R}^u \to \mathbb{R}^{d_i}$ map each model to a shared interlingua hub of dimension $u = 512$: $\mathbf{z} = E_i(\mathbf{h}_i)$, $\mathbf{h}_j = D_j(\mathbf{z})$, requiring $2N$ adapters rather than $O(N^2)$ pairwise translators.

Learned Soft-Prefix Delivery Projector ($\mathcal{P}_\theta$): A delivery projector $\mathcal{P}_\theta$ maps sender hidden states $H \in \mathbb{R}^{L \times d_s}$ at layer ℓ to a k-token soft prefix $S = \mathcal{P}_\theta(H) \in \mathbb{R}^{k \times d_r}$ via cross-attention with k learned queries, an MLP, and norm rescaling. The receiver conditions on $[S \,\|\, \text{embed}(x)]$.

Although debiased CKA is a useful probe (§3.1), we **reject** it as a channel-alignment objective: its invariance to rotation and isotropic scaling preserves relational similarity while destroying coordinate decodability for a receiver. We therefore train the hub with rotation-sensitive, invertibility-preserving objectives: orthogonal Procrustes ($\|X_i R^\star - X_j\|_F^2$), InfoNCE contrastive alignment, and cycle-consistency (Jha et al., 2025).

To prevent the delivery projector from collapsing into a static register cue, we train $\mathcal{P}_\theta$ with two anti-collapse losses alongside reconstruction cross-entropy $\mathcal{L}_{\text{recon}}$:

$$\mathcal{L}_{\text{proj}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}, \tag{8}$$

where $\mathcal{L}_{\text{id}}$ enforces prefix identifiability against a negative memory queue ($|\mathcal{Q}| = 256$), and $\mathcal{L}_{\text{ctr}} = \text{softplus}(m - (\mathcal{L}_{\text{recon}}^{\text{mismatched}} - \mathcal{L}_{\text{recon}}))$ penalizes the model when matched and counterfactual mismatched reconstruction losses are too close.

The Four-Arm Specificity Protocol. We evaluate each held-out sample x under five paired configurations sharing identical prompts, decoding parameters, and token contexts: MATCHED ($S = \mathcal{P}_\theta(H_x)$, the correct payload); MISMATCHED ($S = \mathcal{P}_\theta(H_{x'})$, an unrelated counterfactual payload through the same projector); CONSTANT ($S = \mathcal{P}_{\text{const}}$, an input-independent prefix of identical capacity); NONE ($S = \emptyset$); and TEXT@k (literal source text truncated to k tokens at identical context positions). We compute four contrasts:

$$\Delta_{\text{prefix}} = \text{CE}(\text{NONE}) - \text{CE}(\text{MATCHED}), \qquad \Delta_{\text{cue}} = \text{CE}(\text{NONE}) - \text{CE}(\text{CONSTANT}), \quad (9)$$

$$\Delta_{\text{spec}} = \text{CE}(\text{MISMATCHED}) - \text{CE}(\text{MATCHED}), \quad \Delta_{\text{budget}} = \text{CE}(\text{TEXT@}k) - \text{CE}(\text{MATCHED}). \quad (10)$$

For a static register cue, swapping the payload changes nothing ($\Delta_{\text{spec}} = 0$), so $\Delta_{\text{spec}} > 0$ indicates instance-specific content transmission. In our specificity experiments the sender and receiver are `Qwen2.5-7B-Instruct`, and CE is the receiver's loss when asked to reproduce the source message verbatim.

4 Experimental Setup

Benchmark & Languages. We organize our empirical program into four tiers:

(i) **Tier 1: Typological Spectrum (122 Languages, 220 Varieties):** Reading comprehension on all 122 configurations of `Belebele` (Bandarkar et al., 2024) and multi-tokenizer fertility profiling on 220 `FLORES+` (NLLB Team et al., 2022) varieties, grouped into High-Fertility Low-Resource (**HFLR**), Southeast Asian (**SEA**), South Asian (**SA**), African (**AFR**), and European/Global (**EUR/GLO**) clusters (Table 12).

(ii) **Tier 2: Cross-Architecture Suite (7 Open-Weight LLMs, 7B–9B):** Llama-SEA-LION-v3-8B-IT and Gemma-SEA-LION-v3-9B-IT (AI Singapore, 2025), Sailor2-8B-Chat (Dou et al., 2025), SeaLLMs-v3-7B-Chat (Nguyen et al., 2023), Qwen2.5-7B-Instruct (Yang et al., 2024), EuroLLM-9B-Instruct (EuroLLM Team, 2024), and Aya-Expanse-8B (Cohere For AI et al., 2024) (depths and hidden sizes in Appendix C).

(iii) **Tier 3: Reasoning, Safety & Communication:** `MGSM` (Shi et al., 2023) (11 languages) and `AfriMGSM` (Ahia et al., 2023b) (17 African languages; 27 unique in total), safety classification on SEA-SafeguardBench (Tasawong et al., 2026) (6 languages) and Safety-Toxicity-Detection, and latent channel capacity across 9 languages (Tables 13, 3).

(iv) **Tier 4: Drift & Bridge Testbed (6 Core LRLs + English):** Paired permutation tests and causal ablations ($n = 200$/language for drift; $n = 40$/language for certified communication) on Thai (`th`), Burmese (`my`), Khmer (`km`), Lao (`lo`), Amharic (`am`), and Swahili (`sw`) alongside English (`en`).

Metrics and statistical protocols. Generative fidelity is quantified by Script Fidelity Rate (**SFR** $\uparrow$, fraction of tokens matching target Unicode script), Involuntary Fidelity Loss (**IFL** $\downarrow$, fraction of generations with SFR < 0.50), Language Consistency Benchmark metrics (Line-Level Pass Rate, **LPR** $\uparrow$ and Word-Level Pass Rate **WPR** $\uparrow$), and **chrF** character n-gram overlap (Popović, 2015). Communication metrics follow Equations 9–10. Statistical significance is assessed via paired sign-flip permutation tests (10,000 permutations) and percentile bootstrap (10,000 resamples) with Holm–Bonferroni family-wise error rate correction.

5 Results

5.1 Intra-Model Drift Mitigation & Reasoning Transfer

Intra-model generative drift reduction. Without intervention, 49.1% of generated outputs on SEA-LION-v3-8B fall below 50% target-script fidelity (Table 2, top), reaching 80.0% on Thai and 71.5% on Amharic. Our geometry-aware steering achieves statistically significant reductions across every non-Latin language ($p < 0.001$, Holm–Bonferroni corrected), yielding a 40.0% relative reduction in Macro IFL ($0.491 \rightarrow 0.295$): Amharic drift is halved ($0.715 \rightarrow 0.335$), Burmese drops to 12.0% (-64.2% relative), and Khmer drops by 0.255 ($0.480 \rightarrow 0.225$, -53.1%).

Table 2: **Intra-model drift reduction and baseline comparison on `SEA-LION-v3-8B`.** (Top) IFL across six languages ($n = 200$/language, `FLORES+`; paired permutation p, Holm–Bonferroni). (Bottom) Steering baselines on `Belebele` MCQA ($n = 600$, teacher-forced); IFL/SFR measured for the unsteered model and ours only. †Latin-script control: script-based IFL cannot detect drift into English.

Language	Base IFL ↓	Steered ↓	Abs Δ	p (Holm)
Thai (`th`)	0.800	0.620	−0.180	0.0006
Burmese (`my`)	0.335	0.120	−0.215	0.0006
Khmer (`km`)	0.480	0.225	−0.255	0.0006
Lao (`lo`)	0.615	0.465	−0.150	0.0008
Amharic (`am`)	0.715	0.335	−0.380	0.0006
Swahili (`sw`)†	*0.000*	*0.005*	*+0.005*	1.0000
Macro (non-Latin)	0.589	0.353	−0.236	$p < 0.001$
Macro (all 6)	0.491	0.295	−0.196	(−40.0%)

Method	Belebele ↑	Macro IFL ↓	SFR ↑
Unsteered Baseline	58.7%	0.491	0.519
DiffMean (Gurgurov et al., 2026)	29.5%	—	—
LAPE (Gurgurov et al., 2026)	29.5%	—	—
PCA (Gurgurov et al., 2026)	29.5%	—	—
ReCoVeR (Sterz et al., 2025)	30.3%	—	—
Flat MRRE (Zou et al., 2023)	56.2%	—	—
Ours (G+M+P)	56.7%	**0.295**	**0.707**
Ours (gated)	**58.8%**	—	—

Cross-backbone generality across six model lineages. Drift mitigation is not restricted to SEA-LION. In a separate cross-architecture run ($n = 200$/language), steering reduces Macro IFL on all six backbones by 16–45% relative, with macro-level 95% CIs (bootstrapped over languages) excluding zero for four of them (Appendix Table 8); the largest gains are on `Aya-Expanse-8B` ($0.662 \rightarrow 0.364$) and `EuroLLM-9B` ($0.708 \rightarrow 0.420$), where LCB Language Preservation Rate (LPR) rises from $0.163 \rightarrow 0.453$ and $0.134 \rightarrow 0.504$. Across all backbones, late-layer English logit mass stays below 0.101 (mean 0.078 vs. 0.829 in unsteered controls).

Multi-backbone baseline comparison & collapse prevention. Table 2 (bottom) and Appendix Table 7 compare against five activation steering baselines on four lineages (`Belebele` MCQA, $n = 600$ items/cell). Methods lacking magnitude normalization (DiffMean, LAPE, PCA) fall to near chance on all four backbones (26.7%–29.5%); their identical accuracies indicate degenerate predictions at the steering strengths used. ReCoVeR and Flat MRRE preserve accuracy on `Gemma-SEA-LION-9B` (68.7%, 70.8%) but degrade it on the smaller backbones (25.8%–56.2%). Our full G+M+P configuration costs 2.0 points on `Llama-SEA-LION-8B` and degrades `Sailor2-8B` and `SeaLLMs-v3-7B` under teacher-forced scoring, whereas our gated variant (adaptive gating with prefix injection) preserves base accuracy on all four lineages. A $2 \times 2 \times 2$ factorial (Appendix C.1) attributes the MCQA differences mainly to subspace projection; magnitude normalization, designed for free generation, lowers teacher-forced accuracy. Safety-classification accuracy on Thai and Burmese is unchanged within Wilson 95% intervals on all four backbones ($\Delta \in [-0.03, +0.02]$; Appendix Table 5).

5.2 Inter-Model Specificity & Token-Budget Frontier

Register cues dominate naive gains, but content survives. Table 3 reports the specificity certification across nine languages. In a rerun of the k=16, layer -4 configuration with a trained constant-prefix checkpoint (Table 3 reports the original run), raw prefix conditioning yields a pooled Δ_{prefix} of $+0.411$ nats, of which the constant-prefix control (Δ_{cue}) accounts for $+0.342$ nats: **about** 83.3% **of what a single-payload evaluation would measure as inter-agent communication is an input-independent register cue** (rising to 96% under a batch-mean approximation). A content-specific effect survives the counterfactual protocol ($\Delta_{\text{spec}} = +0.056$ nats pooled, 95% CI $[+0.042, +0.071]$, $p \leq 10^{-4}$; and $+0.069$ nats over the constant control), with intervals excluding zero in six of nine

Table 3: **Channel certification and token-budget efficiency across nine languages.** Paired percentile bootstrap (10,000 resamples) with sign-flip permutation tests: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. Languages ordered by tokenizer fertility. Rows marked with § are shared with the intra-model evaluation.

Language	Fertility	Δ_{prefix}	$\Delta_{\text{spec}} \uparrow$	95% CI	$\Delta_{\text{budget}} \uparrow$
English (en)	26	+0.764	**+0.341*****	$[+0.254, +0.432]$	**-1.409*****
Swahili§ (sw)	50	+0.800	**+0.020***	$[+0.004, +0.037]$	**-0.705*****
Thai§ (th)	67	+0.473	**+0.091*****	$[+0.062, +0.123]$	**-0.150*****
Amharic§ (am)	102	+0.376	+0.001	$[-0.004, +0.006]$	**-0.108*****
Bengali (bn)	130	+0.307	**+0.027*****	$[+0.018, +0.036]$	**+0.107*****
Lao§ (lo)	146	+0.254	+0.003	$[-0.001, +0.007]$	-0.036
Khmer§ (km)	174	+0.264	**+0.008****	$[+0.003, +0.015]$	+0.038
Telugu (te)	178	+0.283	**+0.008****	$[+0.004, +0.012]$	**+0.136*****
Burmese§ (my)	228	+0.160	+0.001	$[-0.002, +0.004]$	**+0.036****
Pooled	—	+0.409	**+0.056*****	$[+0.041, +0.072]$	**-0.232*****

languages. Across prefix widths and source layers, Δ_{spec} peaks at layer -8 with $k = 8$ ($+0.068$ nats; Appendix Table 6).

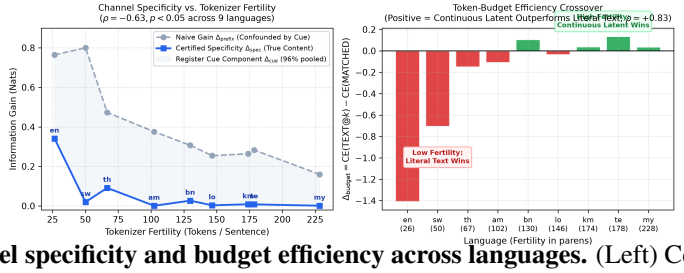


Figure 2: **Channel specificity and budget efficiency across languages.** (Left) Certified specificity Δ_{spec} vs. naive prefix gain Δ_{prefix} across tokenizer fertility (Spearman $\rho = -0.63$, $n = 9$), showing the register-cue gap. (Right) Token-budget crossover (Δ_{budget}): the advantage of continuous latents over literal text rises with fertility ($\rho = +0.83$, $n = 9$) and turns positive for the highest-fertility scripts.

Token-budget efficiency crossover. Δ_{budget} rises with tokenizer fertility ($\rho = +0.83$; Figure 2, right): literal text carries far more information per position for low-fertility English and Swahili (-1.409, -0.705 nats), whereas the continuous channel wins at equal context positions for high-fertility Bengali, Telugu, and Burmese ($+0.107, +0.136, +0.036$ nats). Pooled, literal text remains ahead (-0.232 nats), so continuous hand-offs are most attractive where tokenization fragmentation is severe.

End-task multi-agent coordination. In full multi-agent reasoning with a homogeneous SEA-LION-v3-8B pool, continuous latent communication reaches 0.467 macro Belebele accuracy on its three strongest languages (en, th, sw) but drops to 0.210 across the full nine-language suite (Burmese 0.000, Khmer 0.020, Lao 0.050, Amharic 0.010). Under the same generative protocol, the single-agent baseline scores higher (0.633 on en/th/sw and 0.291 overall; Appendix Table 11), so latent hand-offs do not yet improve end-task accuracy, and the losses concentrate on the compact low-resource scripts. Prior continuous baselines assume matched hidden dimensions, whereas our Universal Latent Hub runs across heterogeneous pools ($3584 \neq 4096$; 0.51 on English Belebele, $n = 100$).

5.3 The Geometric Bridge: Distortion, Fertility, and Channel Capacity

To test whether the geometry driving generative drift (§3.2) also governs latent channel capacity (§3.3), Table 1 matches geometric properties (distortion, debiased CKA) against communication properties (fertility, Δ_{spec}) for the six languages shared by both programs.

Across these six languages, activation distortion ratio is negatively associated with channel specificity: $\rho_{\text{dist, spec}} = -0.59$ (with $n = 6$ this is not significant on its own, $p \approx 0.2$). Languages with severe

geometric distortion and compact activations (Lao, Burmese, Amharic) exhibit near-zero content specificity, whereas languages with moderate distortion (Thai, Swahili) sustain certified content transfer.

Concurrently, tokenizer fertility correlates with channel specificity at $\rho_{\text{fert, spec}} = -0.61$, and is collinear with distortion ratio ($\rho_{\text{dist, fert}} = +0.83$, $p \approx 0.04$; excluding the BOS position this falls to 0.45–0.69, so part of the collinearity reflects sequence length). The partial correlation controlling for fertility attenuates to $\rho_{\text{dist, spec}\cdot\text{fert}} \approx -0.19$, suggesting that **representation distortion and tokenizer fertility are not competing independent mechanisms, but collinear manifestations of one resource-scarcity axis**: under-represented languages suffer from both undertrained subword vocabularies (high fertility) and under-allocated representation volume (compact, distorted manifolds).

The broader Belebele evaluation is consistent with this picture. Across all 122 configurations (Table 12), the HFLR tier averages over 100 tokens/sentence with compressed accuracy (0.339) and contains the three languages where latent channels beat literal text in Table 3 (Bengali, Telugu, Burmese), whereas European/Global languages (< 50 tokens/sentence) favor discrete tokens. Romanization also lowers accuracy (Arabic $0.420 \rightarrow 0.200$, Bengali $0.432 \rightarrow 0.280$; Appendix D), though romanized text is also rarer in pretraining, so this does not isolate tokenization.

6 Conclusion

This paper provides evidence that multilingual representation distortion in low-resource languages underlies both intra-model generative drift and inter-model latent communication bottlenecks. Across evaluations spanning seven open-weight LLMs, 122 Belebele configurations, and 220 FLORES+ varieties, residual activations decompose orthogonally into language and reasoning subspaces ($U_L \oplus U_R$), with reasoning exhibiting model-invariant cross-lingual isomorphism to English ($\rho = 0.71$–1.00). Our depth-stratified Gaussian steering with dynamic magnitude normalization bounds per-layer perturbations and reduces generative drift by 16–45% relative across six LLMs. For inter-agent communication, our specificity protocol reveals that 83.3% of naive prefix gains are register cues (against trained constant controls; 96.5% under a batch-mean approximation), yet certified content transfer survives and continuous latents outperform literal text at equal context length on the highest-fertility scripts. Together, these findings suggest that representation distortion and tokenization overhead form a unified resource-scarcity bottleneck affecting both single-model and multi-agent multilingual intelligence.

Limitations. *(i)* Our component factorial is measured on teacher-forced MCQA, where magnitude normalization lowers accuracy; its contribution to drift reduction in free generation is not yet isolated. *(ii)* Per-backbone strength tuning for external steering baselines could narrow the MCQA gap. *(iii)* Script-based IFL cannot detect drift into English for Latin-script languages such as Swahili, and steered safety evaluation covers two languages supported by SEA-SafeguardBench.

References

Orevaoghene Ahia, Sachin Tjuatja, Devendra Sachan, Noah A. Smith, and Tsvetkov Yulia. Do all languages cost the same? Tokenization in the era of commercial language models. In *Empirical Methods in Natural Language Processing*, 2023a. arXiv:2305.13707.

Orevaoghene Ahia et al. AfriMGSM: Multilingual mathematical reasoning in african languages. *arXiv preprint arXiv:2311.08358*, 2023b.

AI Singapore. SEA-LION: Southeast asian languages in one network. *arXiv preprint arXiv:2504.05747*, 2025. URL https://arxiv.org/abs/2504.05747.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *Annual Meeting of the Association for Computational Linguistics*, 2020.

Lucas Bandarkar, Davis Liang, Benjamin Muller, et al. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Cohere For AI, John Dang, et al. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*, 2024. URL https://arxiv.org/abs/2412.04261.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Longxu Dou et al. Sailor2: Sailing in south-east asia with inclusive multilingual LLMs. *arXiv preprint arXiv:2502.12982*, 2025. URL https://arxiv.org/abs/2502.12982.

EuroLLM Team. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*, 2024.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2016.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LeatkxrBCi.

Daniil Gurgurov, Michael Trinley, Yusser Al Ghussin, et al. Language arithmetics: Towards systematic language neuron interventions. *arXiv preprint arXiv:2505.19700*, 2025.

Daniil Gurgurov et al. CLaS-Bench: Cross-lingual alignment and steering benchmark for large language models. *Findings of the Association for Computational Linguistics: ACL*, 2026.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, et al. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. URL https://arxiv.org/abs/2412.06769.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, 2024.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2025. URL https://arxiv.org/abs/2505.12540.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Andrew Lee, Xiaoyan Bai, Itamar Pres, and Martin Wattenberg. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, 2025. arXiv:2409.05907.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

Xiaoze Liu, Ruowang Zhang, Weichen Yu, Siheng Xiong, Liu He, Feijie Wu, Hoin Jung, Matt Fredrikson, Xiaoqian Wang, and Jing Gao. The vision wormhole: Latent-space communication in heterogeneous multi-agent systems, 2026. URL https://arxiv.org/abs/2602.15382.

Alex Murphy, Joel Zylberberg, and Alona Fyshe. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment (Re-Align)*, 2024. arXiv:2405.01012.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, et al. SeaLLMs – large language models for south-east asia. *arXiv preprint arXiv:2312.00738*, 2023. URL https://arxiv.org/abs/2312.00738.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Aleksandar Petrov, Andrey Malinin, and Mark Gales. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.15425.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.19698.

Maja Popović. chrF: character n-gram f-score for machine translation evaluation. In *Workshop on Statistical Machine Translation*, pp. 392–395, 2015.

Freda Shi, Mirac Suzgun, Markus Freitag, et al. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5):1393–1434, 2012.

Hannah Sterz, Philip Schmidt, Goran Glavaš, and Ivan Vulić. Language-specific steering vectors for cross-lingual transfer. *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL*, 2022.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, 2016.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso Hunter, François Balloux, and Robert Rehberg. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 2024. arXiv:2407.12404.

Tianyi Tang, Wenyang Luo, Haoyang Huang, et al. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*, 2024.

Worawat Tasawong, Christine Ngui, Alham Fikri Aji, Trevor Cohn, Lalita Lowphansirikul, and Peerat Limkonchotiwat. SEA-SafeguardBench: Regional content safety and toxicity benchmark for southeast asian languages. *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. arXiv:2512.05501.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D. Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025. URL https://arxiv.org/abs/2501.06322.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, et al. Steering language models with activation addition. In *arXiv preprint arXiv:2308.10248*, 2023.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. *Annual Meeting of the Association for Computational Linguistics*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Weihao Xuan, Rui Yang, Heli Qi, et al. MMLU-ProX: A multilingual benchmark for advanced llm evaluation. In *Empirical Methods in Natural Language Processing (Main)*, 2025. arXiv:2503.10497.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yue Zhang and Wei Emu. Do latent channels actually communicate? a causal audit of latent multi-agent LLM communication. *arXiv preprint arXiv:2607.26773*, 2026.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. When less language is more: Language-reasoning disentanglement makes LLMs better multilingual reasoners. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.15257.

Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. Thought communication in multiagent collaboration. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=d671ljgwfY.

Andy Zou, Long Phan, Sarah Chen, James Campbell, et al. Representation engineering: A top-down approach to AI transparency. In *arXiv preprint arXiv:2310.01405*, 2023.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=syG9I9ofd8.

Reproducibility Statement

All propositions are proven in Appendix A. Code for SVD decomposition, steering, projector training, and the evaluation harnesses, together with the result artifacts behind each table, is provided as anonymized supplementary material and will be released upon publication.

Ethics Statement

All datasets and model weights used in this study are publicly available research assets. By mitigating generative language drift and clarifying latent channel fidelity in low-resource scripts, this work aims to support more equitable multilingual AI.

LLM Usage Statement

LLM-based assistants were used to help write and debug experiment code and to edit the manuscript text. All scientific claims, experimental design, and results were reviewed and verified by the authors, who take full responsibility for the content.

A Mathematical Proofs

A.1 Proof of Proposition 1 (Alignment of the Reasoning Subspace)

Proof. Let $\mathbf{X} = \mathbf{H}_{\text{tgt}}^{(\ell)}$ and $\mathbf{Y} = \mathbf{H}_{\text{en}}^{(\ell)}$ be column-centered ($K \times d$). Centering acts on rows and P_R on columns, so the two commute. Write $\mathbf{X} = \mathbf{X}P_L + \mathbf{X}P_R$ with $P_L = U_L U_L^\top$. Because $P_L P_R = 0$, the linear Gram matrix splits without cross-terms:

$$K_{\mathbf{X}} = \mathbf{X}\mathbf{X}^\top = \mathbf{X}P_L\mathbf{X}^\top + \mathbf{X}P_R\mathbf{X}^\top = K_{\mathbf{X}}^L + K_{\mathbf{X}}^R, \tag{11}$$

and likewise $K_{\mathbf{Y}} = K_{\mathbf{Y}}^L + K_{\mathbf{Y}}^R$. Linear CKA is

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{Tr}(K_{\mathbf{X}} K_{\mathbf{Y}})}{\|K_{\mathbf{X}}\|_F \, \|K_{\mathbf{Y}}\|_F}. \tag{12}$$

Numerator. By the assumption of the proposition, $\mathrm{Tr}(K_{\mathbf{X}} K_{\mathbf{Y}}) = \mathrm{Tr}(K_{\mathbf{X}}^R K_{\mathbf{Y}}^R)$, which is exactly the numerator of $\mathrm{CKA}(\mathbf{X}P_R, \mathbf{Y}P_R)$.

Denominator. $K_{\mathbf{X}}^L$ and $K_{\mathbf{X}}^R$ are positive semidefinite, so $\mathrm{Tr}(K_{\mathbf{X}}^L K_{\mathbf{X}}^R) \geq 0$ and

$$\|K_{\mathbf{X}}\|_F^2 = \|K_{\mathbf{X}}^L\|_F^2 + \|K_{\mathbf{X}}^R\|_F^2 + 2\,\mathrm{Tr}(K_{\mathbf{X}}^L K_{\mathbf{X}}^R) \geq \|K_{\mathbf{X}}^R\|_F^2 + \|K_{\mathbf{X}}^L\|_F^2, \tag{13}$$

with strict inequality $\|K_{\mathbf{X}}\|_F > \|K_{\mathbf{X}}^R\|_F$ whenever $\mathbf{X}P_L \neq 0$; similarly $\|K_{\mathbf{Y}}\|_F \geq \|K_{\mathbf{Y}}^R\|_F$. With an unchanged numerator, positive by assumption, and a strictly smaller denominator,

$$\mathrm{CKA}(P_R\mathbf{H}_{\text{tgt}}^{(\ell)}, P_R\mathbf{H}_{\text{en}}^{(\ell)}) > \mathrm{CKA}(\mathbf{H}_{\text{tgt}}^{(\ell)}, \mathbf{H}_{\text{en}}^{(\ell)}). \tag{14}$$

The debiased estimator used in our experiments is a consistent estimator of the same population quantity; the inequality above holds for the population (biased) form. □

A.2 Proof of Proposition 2 (Boundary Layer Invariance)

Proof. The Gaussian depth injection schedule is defined as:

$$\alpha_\ell = \alpha_0 \exp\left(-\frac{(\ell - \mu_s)^2}{2\sigma_s^2}\right), \quad \mu_s = 0.65L,\ \sigma_s = 0.15L. \tag{15}$$

For early token-formation and syntactic layers ($\ell \leq 0.20L$), the standardized distance from the peak reasoning depth satisfies:

$$\frac{|\ell - \mu_s|}{\sigma_s} \geq \frac{|0.20L - 0.65L|}{0.15L} = \frac{0.45}{0.15} = 3.0. \tag{16}$$

Substituting into the exponential gives:

$$\alpha_\ell \leq \alpha_0 \exp\left(-\frac{3.0^2}{2}\right) = \alpha_0 \exp(-4.5) \approx 0.0111\alpha_0. \tag{17}$$

For terminal unembedding layers at $\ell = L$:

$$\frac{|L - \mu_s|}{\sigma_s} = \frac{0.35}{0.15} \approx 2.33, \quad \alpha_L = \alpha_0 \exp\left(-\frac{2.33^2}{2}\right) = \alpha_0 \exp(-2.72) \approx 0.066\alpha_0. \tag{18}$$

Thus the direct injection is at most 1.1% of its peak for $\ell \leq 0.2L$ and at most 6.6% at $\ell = L$. Since early layers precede all substantial injection, their activations are essentially unperturbed; later layers additionally carry the (norm-bounded) perturbations injected upstream. □

A.3 Proof of Proposition 3 (Bounded Relative Perturbation)

Proof. The layer-wise steering perturbation is given by $\Delta\mathbf{h}^{(\ell)} = \alpha_\ell \gamma_\ell \mathbf{v}_{\text{steer}}^{(\ell)}$. Substituting the dynamic magnitude normalization factor $\gamma_\ell = \eta \frac{\mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]}{\|\mathbf{v}_{\text{steer}}^{(\ell)}\|_2}$:

$$\|\Delta\mathbf{h}^{(\ell)}\|_2 = \alpha_\ell \left(\eta \frac{\mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]}{\|\mathbf{v}_{\text{steer}}^{(\ell)}\|_2}\right) \|\mathbf{v}_{\text{steer}}^{(\ell)}\|_2 = \alpha_\ell \eta\, \mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]. \tag{19}$$

Dividing both sides by the expected target activation norm $\mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]$:

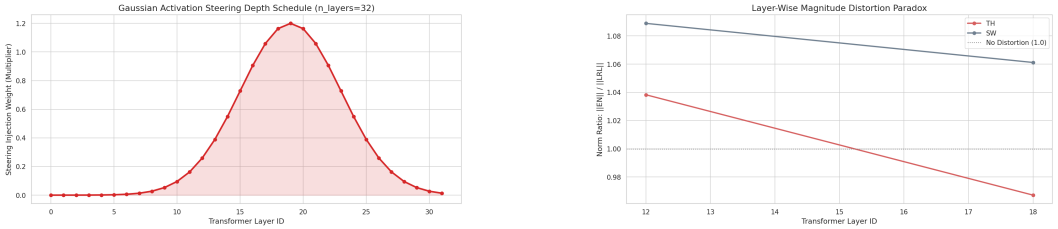
$$\frac{\|\Delta\mathbf{h}^{(\ell)}\|_2}{\mathbb{E}[\|\mathbf{h}_{\text{tgt}}^{(\ell)}\|_2]} = \alpha_\ell \eta \leq \alpha_0 \eta, \quad \forall \ell \in \{1, \ldots, L\}. \tag{20}$$

□

B Extended Hyperparameter Sweeps & Architecture Diagnostics

Table 4: **Contrastive energy captured by** U_L ($r = 16$) in the $\mathcal{G}/\mathcal{M}/\mathcal{P}$ factorial runs: fraction of the squared singular-value mass of the pooled, mean-centered contrastive matrix $C^{(\ell)}$ ($K = 1200$ pairs, six languages) in its top 16 components, at the two steered layers per backbone.

Backbone	**Layers** (ℓ/L)	**Energy, shallower**	**Energy, deeper**
Llama-SEA-LION-8B	16, 24 / 32	0.857	0.669
Gemma-SEA-LION-9B	21, 31 / 42	0.735	0.642
Sailor2-8B	16, 24 / 28	0.979	0.928
SeaLLMs-v3-7B	14, 21 / 28	0.983	0.939



(a) Gaussian depth schedule α_ℓ.

(b) Dynamic magnitude normalization.

Figure 3: **Mechanistic components of geometry-aware steering.** (a) Continuous Gaussian depth schedule α_ℓ targets the abstract reasoning band while keeping injection at boundary layers small. (b) Dynamic magnitude normalization bounds relative perturbation, preventing activation explosion in compact target states.

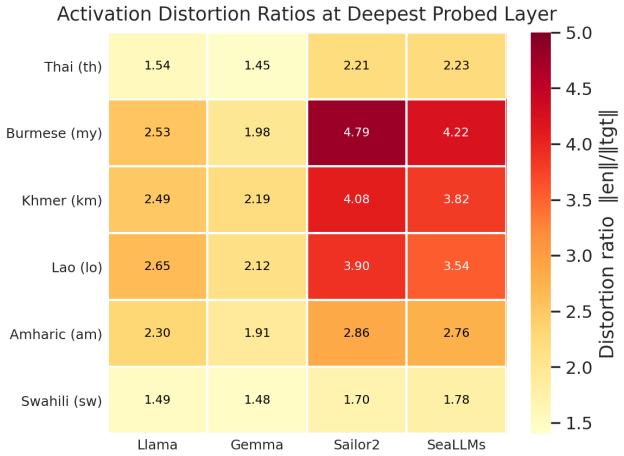
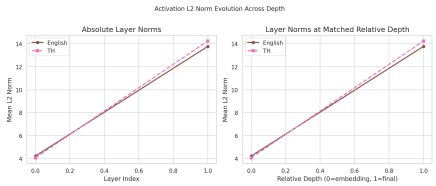
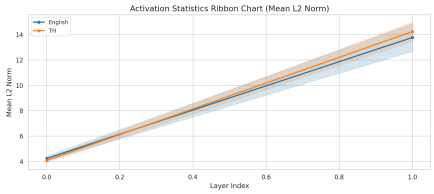


Figure 4: **Activation distortion ratio heatmap** ($\|h_{\text{en}}\|/\|h_{\text{tgt}}\|$ at deepest probed layers) across languages and models. Low-resource non-Latin languages consistently exhibit activations about $1.5\times$ to $4.8\times$ smaller than English.

Channel capacity scaling across depth and width. We evaluate channel capacity across prefix width $k \in \{8, 16, 32\}$ and source extraction layers $\{-1, -4, -8\}$ (Table 6). Whereas raw Δ_{prefix} is nearly flat across the grid (range 0.013 nats), content specificity Δ_{spec} is highest at mid-to-deep layer



(a) Per-layer residual stream norms.



(b) Intervention ribbon across depth.

Figure 5: **Residual stream dynamics under steering.** (a) Unnormalized flat injection inflates late-layer norms; magnitude normalization preserves the unsteered norm profile. (b) Steering intervention smoothly concentrates in the reasoning band ($0.5L$–$0.8L$).

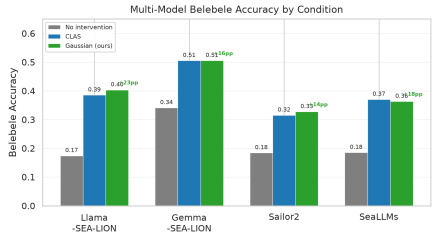
Table 5: **Safety classification accuracy under steering** ($n = 100$/model, 50 Thai / 50 Burmese) with Wilson 95% confidence intervals.

Model	No Intervention	Steered (Ours)	Δ
Llama-SEA-LION-8B	0.66 ± 0.09	0.63 ± 0.09	-0.03
Gemma-SEA-LION-9B	0.70 ± 0.09	0.72 ± 0.09	$+0.02$
Sailor2-8B	0.61 ± 0.09	0.61 ± 0.09	0.00
SeaLLMs-v3-7B	0.58 ± 0.10	0.57 ± 0.10	-0.01

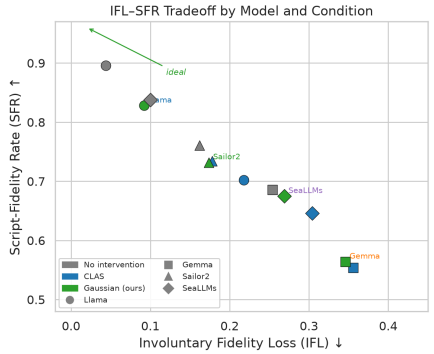
-8 with $k = 8$ ($+0.068$ nats, $p < 0.01$; single seed), suggesting that intermediate representations preserve semantic structure better suited to cross-agent conditioning.

C Multi-Backbone Baseline Evaluation Across Lineages

Backbone configurations. Llama-SEA-LION-v3-8B-IT (32L, $d = 4096$), Gemma-SEA-LION-v3-9B-IT (42L, $d = 3584$), Sailor2-8B-Chat (28L, $d = 3584$), SeaLLMs-v3-7B-Chat (28L, $d = 3584$), Qwen2.5-7B-Instruct (28L, $d = 3584$), EuroLLM-9B-Instruct (42L, $d = 4096$), and Aya-Expanse-8B (32L, $d = 4096$).



(a) Multi-model Belebele accuracy.



(b) IFL–SFR Pareto tradeoff.

Figure 6: **Cross-backbone performance and Pareto optimality.** (a) Accuracy gains across four diverse backbones. (b) IFL–SFR tradeoff curves showing our method occupying the Pareto-optimal upper-left frontier.

Table 7 provides the comprehensive multi-backbone comparison. Methods without magnitude calibration (DiffMean, LAPE, PCA) fall to near chance on MCQA ($\sim$ 26.7%–29.5%), consistent with uncalibrated additions corrupting candidate logit scoring. Our gated variant preserves base-model accuracy on all four backbones: `Llama-SEA-LION-8B` (58.83%), `Sailor2-8B` (56.50%), `SeaLLMs-v3-7B` (51.17%), and `Gemma-SEA-LION-9B` (70.83%).

Table 6: **Channel capacity response across depth and width.** Pooled Δ_{spec} (nats) across nine languages (seed 0; seeds 0–2 were also run for $k = 16$, layer -4). Bold entries indicate 95% bootstrap intervals excluding zero ($p < 0.01$).

Source Layer	$k = 8$	$k = 16$	$k = 32$
Layer -1	**+0.053**	**+0.047**	**+0.046**
Layer -4	**+0.056**	**+0.053**	**+0.048**
Layer -8	**+0.068**	**+0.046**	**+0.045**

Table 7: **Multi-backbone baseline comparison** on Belebele MCQA ($n = 600$ items/cell) across four model families.

Condition / Baseline	**Llama-SEA-LION-8B**	**Sailor2-8B**	**SeaLLMs-v3-7B**	**Gemma-SEA-LION-9B**
Unsteered Baseline (no_intervention)	58.67%	54.83%	51.33%	70.83%
Ours (gated: adaptive gating + prefix injection)	**58.83%**	**56.50%**	**51.17%**	**70.83%**
Ours (G+M+P, static all-token injection)	56.67%	33.67%	34.33%	69.17%
Norm-Matched Random Direction (H2 Control)	56.67%	49.00%	31.67%	72.00%
ReCoVeR (Sterz et al., 2025)	30.33%	25.83%	29.50%	68.67%
DiffMean (Gurgurov et al., 2026)	29.50%	29.50%	26.67%	29.50%
LAPE (Gurgurov et al., 2026)	29.50%	29.50%	26.67%	29.50%
PCA (Gurgurov et al., 2026)	29.50%	29.50%	26.67%	29.50%
Flat MRRE (Zou et al., 2023)	56.17%	30.00%	29.50%	70.83%

C.1 Factorial Component Ablation

To examine each component of our steering method, Table 9 reports a complete $2 \times 2 \times 2$ factorial over Gaussian depth scheduling ($\mathcal{G}$), dynamic magnitude normalization ($\mathcal{M}$), and subspace projection ($\mathcal{P}$) on `Belebele` MCQA across four backbones ($n = 600$ items/cell). Under teacher-forced scoring, subspace projection has the largest effect on backbones most sensitive to injection ($+7.5$ and $+13.2$ points on `Sailor2-8B` and `SeaLLMs-v3-7B`), Gaussian scheduling has small effects (-0.4 to $+1.5$ points), and magnitude normalization lowers accuracy on all four backbones (-0.7 to -8.1 points; the gated variant in §5.1 matches $\mathcal{G}{+}\mathcal{P}$ without $\mathcal{M}$). Magnitude normalization is designed for free generation, where it scales injection to compact target activations; factorial isolation of IFL in free generation remains future work.

D Massively Multilingual Benchmark Evaluation

To evaluate broad generalizability, we conduct benchmark evaluations across extended multilingual test suites: Belebele (Bandarkar et al., 2024) across all 122 language configurations under teacher-forced log-likelihood scoring; FLORES+ (NLLB Team et al., 2022) multi-tokenizer fertility profiling across 220 language variants; SEA-SafeguardBench (Tasawong et al., 2026) across 6 Southeast Asian languages; and MGSM/AfriMGSM (Shi et al., 2023; Ahia et al., 2023b).

Typological gradient and context token tax. The High-Fertility Low-Resource tier (Burmese `mya_Mymr` 0.421, Khmer `khm_Khmr` 0.234, Lao `lao_Laoo` 0.288, Amharic `amh_Ethi` 0.339) incurs an average fertility exceeding 100 tokens per sentence, which makes text hand-offs expensive under fixed context budgets. Fertility alone does not determine accuracy, however: several Indic languages retain comparatively high accuracy despite high fertility (Tamil 0.571, Kannada 0.560), and the European/Global tier averages 0.466.

Native vs. romanized script disparity. Evaluating the same language written in native versus romanized orthographies reveals large disparities: Native Arabic script (`arb_Arab`) achieves 0.420 accuracy versus 0.200 for romanized Arabic (`arb_Latn`); Bengali exhibits a similar disparity (`ben_Beng` 0.432 vs. `ben_Latn` 0.280). Romanization increases fertility (Arabic: 44.4 $\rightarrow$ 65.1 tokens/sentence) but also shifts the input toward text that is rarer in pretraining, so these gaps reflect both factors ($n = 50$ per configuration).

Length-normalized scoring on MMLU-ProX. Multiple-choice options on `MMLU-ProX` (Xuan et al., 2025) vary substantially in continuation character and token length, which can introduce systematic bias under unnormalized teacher-forced log-likelihood scoring. To evaluate this effect,

Table 8: **Cross-backbone drift reduction** (Macro IFL and LCB Language Preservation Rate, $n = 200$/language). 95% CIs on Δ are bootstrapped over languages.

Backbone	**IFL** (base $\rightarrow$ steered)	**Rel.**	Δ	95% CI	**LPR** (base $\rightarrow$ steered)
Aya-Expanse-8B	$0.662 \rightarrow 0.364$	-45.0%	-0.298	$[-0.402, -0.165]$	$0.163 \rightarrow 0.453$ $(+178\%)$
EuroLLM-9B	$0.708 \rightarrow 0.420$	-40.6%	-0.288	$[-0.543, 0.005]$	$0.134 \rightarrow 0.504$ $(+276\%)$
Llama-SEA-LION-8B	$0.492 \rightarrow 0.288$	-41.5%	-0.204	$[-0.286, -0.109]$	$0.365 \rightarrow 0.582$
Qwen2.5-7B	$0.313 \rightarrow 0.215$	-31.2%	-0.098	$[-0.181, -0.008]$	$0.499 \rightarrow 0.595$
Sailor2-8B	$0.340 \rightarrow 0.239$	-29.7%	-0.101	$[-0.208, 0.033]$	$0.497 \rightarrow 0.632$
SeaLLMs-v3-7B	$0.578 \rightarrow 0.487$	-15.8%	-0.092	$[-0.140, -0.040]$	$0.230 \rightarrow 0.302$

Table 9: **Cross-backbone** $2 \times 2 \times 2$ **factorial component matrix** on Belebele MCQA ($n = 600$ items/cell).

Cell	$\mathcal{G}$	$\mathcal{M}$	$\mathcal{P}$	**Sailor2-8B**	**SeaLLMs-v3-7B**	**Gemma-9B**	**Llama-8B**
1	Flat	Unnorm	Raw	29.83%	29.50%	70.33%	57.00%
2	✓	Unnorm	Raw	25.67%	29.50%	70.67%	57.17%
3	Flat	✓	Raw	44.83%	30.33%	70.67%	57.33%
4	Flat	Unnorm	✓	**55.67%**	**51.50%**	70.50%	**58.83%**
5	✓	✓	Raw	49.17%	29.67%	70.83%	58.00%
6	✓	Unnorm	✓	**56.50%**	**51.17%**	70.83%	**58.83%**
7	Flat	✓	✓	33.67%	34.83%	68.83%	51.67%
8	✓	✓	✓	33.67%	34.33%	69.17%	56.67%
Main Effect $\Delta\text{Acc}_{\mathcal{G}}$				$+0.25\%$	-0.37%	$+0.29\%$	$+1.46\%$
Main Effect $\Delta\text{Acc}_{\mathcal{M}}$				-1.58%	-8.12%	-0.71%	-2.04%
Main Effect $\Delta\text{Acc}_{\mathcal{P}}$				$+\mathbf{7.50}\%$	$+\mathbf{13.21}\%$	-0.79%	-0.87%

we compare unnormalized log-likelihood against character-normalized ($\frac{1}{|C|}\sum \log p$) and token-normalized ($\frac{1}{|T|}\sum \log p$) scoring across seven typologically diverse languages (`en, th, sw, bn, te, id, vi`; $n = 50$ per language on `Llama-SEA-LION-v3-8B-IT`). As shown in Table 15, token normalization improves macro accuracy from 0.111 to 0.126 (+13.5% relative gain), with notable gains on high-fertility scripts (Thai $0.100 \rightarrow 0.120$, Bengali $0.120 \rightarrow 0.140$, Vietnamese $0.040 \rightarrow 0.080$), while English remains constant (0.180), confirming that option-length normalization relieves spurious verbosity penalties without altering substantive cross-lingual disparities.

Table 10: **Content-fidelity probe on untrained mid-stack injection**, self-to-self, hub bypassed, Qwen2.5-7B-Instruct. Demonstrates that untrained mid-stack injection yields 0.000 keyword recall, motivating a trained delivery projector.

Arm	Kw. Recall ↑	ROUGE-L ↑	BERTScore ↑
Upper bound (source in prompt)	1.000±0.000	0.837±0.115	0.978±0.016
Channel, layer −1	**0.000**±0.000	0.065±0.019	0.800±0.014
Channel, layer −4	**0.000**±0.000	0.013±0.014	0.791±0.012
Channel, layer −8	**0.000**±0.000	0.017±0.012	0.764±0.013
Lower bound (random latent)	0.000±0.000	0.067±0.032	0.712±0.110

Table 11: **Per-language Belebele breakdown for single-agent and multi-agent coordination.** Single-agent and our homogeneous latent MAS use `Llama-SEA-LION-v3-8B-IT` (8-bit) with generative answer scoring ($n = 100$/language; `results/table1_hom/`). Latent MAS scores below the single-agent baseline on seven of nine languages. †Reference baselines use the same backbone with 4-way log-likelihood scoring ($n = 100$/language; `results/baselines/*_belebele_sealion/`); because the scoring differs, they are not directly comparable with the first two columns.

Language	FLORES Code	Single-Agent	Latent MAS (ours)	Text MAS†	LatentMAS†	ThoughtComm†
English	eng_Latn	0.730	0.570	0.540	0.500	0.500
Thai	tha_Thai	0.620	0.470	0.550	0.480	0.490
Swahili	swh_Latn	0.550	0.360	0.390	0.370	0.390
Burmese	mya_Mymr	0.040	0.000	—	—	—
Khmer	khm_Khmr	0.100	0.020	—	—	—
Lao	lao_Laoo	0.010	0.050	—	—	—
Amharic	amh_Ethi	0.030	0.010	—	—	—
Bengali	ben_Beng	0.470	0.260	—	—	—
Telugu	tel_Telu	0.070	0.150	—	—	—
Mean (en/th/sw)		**0.633**	**0.467**	**0.493**	**0.450**	**0.460**
Macro Mean (all 9)		**0.291**	**0.210**	—	—	—

Table 12: **Massively Multilingual Evaluation Across 122 Belebele Languages, 220 FLORES+ Variants, and Recent Multilingual Benchmarks (unsteered baseline).** Accuracies report macro-averages over evaluated configurations. Typological groupings span High-Fertility Low-Resource (HFLR), Southeast Asian (SEA), South Asian (SA), African (AFR), and European/Global (EUR/GLO).

Benchmark Suite	Coverage	Languages (N)	Macro Accuracy ↑	Fertility / Budget	Dominant Regime
Belebele (All Variants)	Global Reading	122 configs	**0.389**	—	MCQA (4-Choice)
▷ HFLR Tier (my, km, lo, am, te, bn)	Low-Resource	6	0.339	>100 tok/sent	Latent Favored ($\Delta_{\text{budget}} > 0$)
▷ Southeast Asian	Regional SEA	14	0.395	80–228 tok/sent	Latent Favored
▷ South Asian	Indic Scripts	15	0.468	95–180 tok/sent	Latent Favored
▷ African	Niger-Congo/Afroasiatic	23	0.277	45–110 tok/sent	Balanced
▷ European / Global	High Resource	25	0.466	<50 tok/sent	Text Favored
MGSM	Math Reasoning	11 languages	0.438	Exact Match	Chain-of-Thought
AfriMGSM	African Math	17 languages†	0.119	Exact Match	Low-Resource Reasoning
SEA-SafeguardBench	Regional Safety	6 languages	0.510	Tox. Detection	Format Compliance
FLORES+ Full Spectrum	Machine Translation	220 configs	—	81.4 tok/sent	All 220 Linguistic Variants

†AfriMGSM was evaluated across all 19 configurations ($n = 50$/lang); `en` (0.540) and `fr` (0.520) serve as pivot controls, yielding a macro accuracy of 0.119 across the 17 indigenous African languages (0.162 across all 19). MGSM ∪ AfriMGSM covers 27 unique languages.

Table 13: **Per-Language Evaluation Across Reasoning and Safety Benchmarks (unsteered baseline).** Exact match accuracy for MGSM and AfriMGSM ($n = 50$/lang) and clean prediction accuracy on SEA-SafeguardBench ($n = 100$/lang, 100% parse compliance).

Language	Code	Family/Region	MGSM ↑	Safety ↑	Language	Code	Family/Region	MGSM ↑	Safety ↑
English	en	Indo-European (Global)	0.540	—	Japanese	ja	Japonic (East Asia)	0.300	—
Russian	ru	Slavic (Global)	0.540	—	Telugu	te	Dravidian (South Asia)	0.260	—
Spanish	es	Romance (Global)	0.540	—	Amharic	am	Semitic / Afroasiatic	0.200*	—
French	fr	Romance (Global)	0.520	—	Ewe	ee	Niger-Congo (W. Africa)	0.060*	—
German	de	Germanic (Global)	0.480	—	Vietnamese	vi	Austroasiatic (SEA)	—	0.800
Chinese	zh	Sino-Tibetan (East Asia)	0.440	—	Malay	ms	Austronesian (SEA)	—	0.540
Thai	th	Kra-Dai (Southeast Asia)	0.440	0.320	Burmese	my	Sino-Tibetan (SEA)	—	0.530
Bengali	bn	Indo-Aryan (South Asia)	0.400	—	Tagalog	tl	Austronesian (SEA)	—	0.480
Swahili	sw	Bantu (Sub-Saharan Africa)	0.360	—	Indonesian	id	Austronesian (SEA)	—	0.390

*AfriMGSM indigenous language evaluation ($n = 50$).

Table 14: **Fine-Grained Per-Language Performance on Belebele (Reading Comprehension) Across Typological Clusters and Script Disparities.** Evaluated on single-model baseline under 4-choice teacher-forcing ($n = 50$/config). Script disparity pairs isolate byte-level tokenization fragmentation on identical language varieties.

Language	Config	Family	Fert.	Acc. ↑	Regime	Language	Config	Family	Fert.	Acc. ↑	Regime
Burmese	mya_Mymr	Sino-Tibetan	312.1	0.421	Latent	Vietnamese	vie_Latn	Austroasiatic	37.4	0.400	Text
Khmer	khm_Khmr	Austroasiatic	234.1	0.234	Latent	Russian	rus_Cyrl	Slavic	43.4	0.620	Text
Lao	lao_Laoo	Kra-Dai	236.0	0.288	Latent	Japanese	jpn_Jpan	Japonic	40.6	0.560	Balanced
Amharic	amh_Ethi	Afroasiatic	203.7	0.339	Latent	Chinese (Simp.)	zho_Hans	Sino-Tibetan	48.5	0.520	Text
Telugu	tel_Telu	Dravidian	222.6	0.322	Latent	English	eng_Latn	Germanic	26.8	0.510	Text
Bengali (Native)	ben_Beng	Indo-Aryan	154.6	0.432	Latent	German	deu_Latn	Germanic	42.5	0.420	Text
Tamil	tam_Taml	Dravidian	205.3	0.571	Latent	French	fra_Latn	Romance	43.0	0.420	Text
Kannada	kan_Knda	Dravidian	237.8	0.560	Latent	Shona	sna_Latn	Bantu	55.8	0.420	Balanced
Marathi	mar_Deva	Indo-Aryan	75.1	0.540	Latent	Swahili	swh_Latn	Bantu	51.8	0.380	Text
Sinhala	sin_Sinh	Indo-Aryan	231.7	0.519	Latent	Hausa	hau_Latn	Afroasiatic	52.9	0.380	Balanced
Hindi	hin_Deva	Indo-Aryan	67.7	0.400	Balanced	Lingala	lin_Latn	Bantu	50.0	0.320	Balanced
Urdu	urd_Arab	Indo-Aryan	82.5	0.500	Balanced	Wolof	wol_Latn	Niger-Congo	51.0	0.300	Balanced
Tagalog	tgl_Latn	Austronesian	61.3	0.540	Balanced	Yoruba	yor_Latn	Niger-Congo	73.1	0.160	Balanced
Thai	tha_Thai	Kra-Dai	58.7	0.510	Latent	Arabic (Native)	arb_Arab	Afroasiatic	44.4	0.420	Balanced
Javanese	jav_Latn	Austronesian	46.0	0.460	Balanced	Arabic (Latin)	arb_Latn	Afroasiatic	65.1	0.200	Fragmentation
Indonesian	ind_Latn	Austronesian	41.3	0.440	Text	Bengali (Native)	ben_Beng	Indo-Aryan	154.6	0.432	Latent
Malay	zsm_Latn	Austronesian	43.1	0.440	Text	Bengali (Latin)	ben_Latn	Indo-Aryan	—‡	0.280	Fragmentation

Fertility denotes mean token count per FLORES+ sentence under the SEA-LION tokenizer; it is therefore not comparable with the Qwen2.5 fertility reported in Tables 1 and 3. *Fragmentation* marks romanized transliterations that expand under tokenization and lose accuracy: Arabic expands $1.5\times$ ($44.4 \rightarrow 65.1$) with a $0.420 \rightarrow 0.200$ accuracy drop. ‡FLORES+ has no romanized Bengali counterpart, so no fertility is reported.

Table 15: **MMLU-ProX length-normalization comparison** across seven languages ($n = 50$ items/language, `Llama-SEA-LION-v3-8B-IT` in 8-bit). Token-normalized scoring improves macro accuracy from 0.111 to 0.126 (+13.5% relative) by eliminating length penalties on non-English continuations while preserving typological rankings.

Language	Code	Unnormalized	Char-Norm	Token-Norm
English	en	0.180	0.180	0.180
Thai	th	0.100	0.100	0.120
Swahili	sw	0.100	0.120	0.120
Bengali	bn	0.120	0.120	0.140
Telugu	te	0.120	0.140	0.120
Indonesian	id	0.120	0.120	0.120
Vietnamese	vi	0.040	0.040	0.080
Macro Mean		**0.111**	**0.117**	**0.126**