Linguistic Gravity in Multilingual Reasoning: Quantifying Resource-Sensitive Representational Patterns in Multilingual LLMs

Anonymous ACL submission

Abstract

Multilingual large language models exhibit large reasoning gaps across languages: LLaMA-3-8B-Instruct solves 77.6% of MGSM problems in English but only 40.8% in Swahili. We propose the *Linguistic Gravity Well* hypothesis: lower-resource languages are associated with language-conditioned regions of representation space that make reasoning trajectories more likely to deviate from representations associated with correct solutions. Across 10 MGSM languages, we find three converging signatures. First, zero-shot accuracy strongly correlates with a Wikipedia-based resource proxy ($r = 0.840$, $p = 0.002$). Second, correct and incorrect trajectories become most separable around layers 13–14, where correctness is linearly decodable from hidden states (AUROC 0.82), and an SVD-estimated language subspace can be partially projected out from a contrastive correctness direction. Third, a probe-gated diagnostic intervention, Dynamic Contrastive Logic Steering (DCLS), yields resource-dependent effects: several higher-resource languages show positive point estimates, while lower-resource languages show limited gains or degradation. These results provide evidence that cross-lingual reasoning gaps are associated with layer-localized and resource-sensitive representational structure.

1 Introduction

When a multilingual large language model (LLM) is presented with identical grade-school math problems translated into different languages, a puzzling pattern emerges: the same model that correctly solves 77.6% of problems in English can only manage 40.8% in Swahili—a gap of nearly 37 percentage points. This performance chasm, reported across multilingual LLMs in prior work, is typically attributed to data scarcity during pretraining. But such an explanation raises a deeper question: *how, mechanistically, does training data volume translate into reasoning failure?*

A natural remedy is to "pivot" through English—prompting the model to reason in English regardless of the input language (Shi et al., 2023). Yet we find that English pivoting yields inconsistent results: it helps Swahili (+6.4pp) but actively *harms* French (-5.2pp), suggesting that the cross-lingual reasoning gap is not merely a surface-level translation problem but reflects something more fundamental about how multilingual representations interact with reasoning computations.

We propose the **Linguistic Gravity Well** hypothesis: lower-resource languages are associated with deeper language-conditioned regions in the model's hidden-state space, which may make reasoning trajectories more likely to deviate from representations associated with correct logic. The gravity metaphor captures the intuition that some language-conditioned trajectories require stronger representational support to remain near the model's effective reasoning manifold. We treat this as a mechanistic hypothesis that generates testable predictions, rather than as a claim that the model implements a literal physical potential.

This hypothesis yields five testable predictions, formalized in Section 3: a resource–accuracy gradient, layer-localized trajectory separation, linear decodability of correctness, partial decomposability of language and correctness directions, and resource-sensitive intervention effects.

We test all five predictions using LLaMA-3-8B-Instruct (Llama Team, 2024) on MGSM (Shi et al., 2023) across 10 languages spanning four families, from English (6.8M Wikipedia articles) to Swahili (78K). Our contributions are:

- **Resource-sensitive phenomenon.** We quantify a strong resource–accuracy gradient on MGSM using Wikipedia article count as a proxy for language resource availability ($r = 0.840$, $p =$

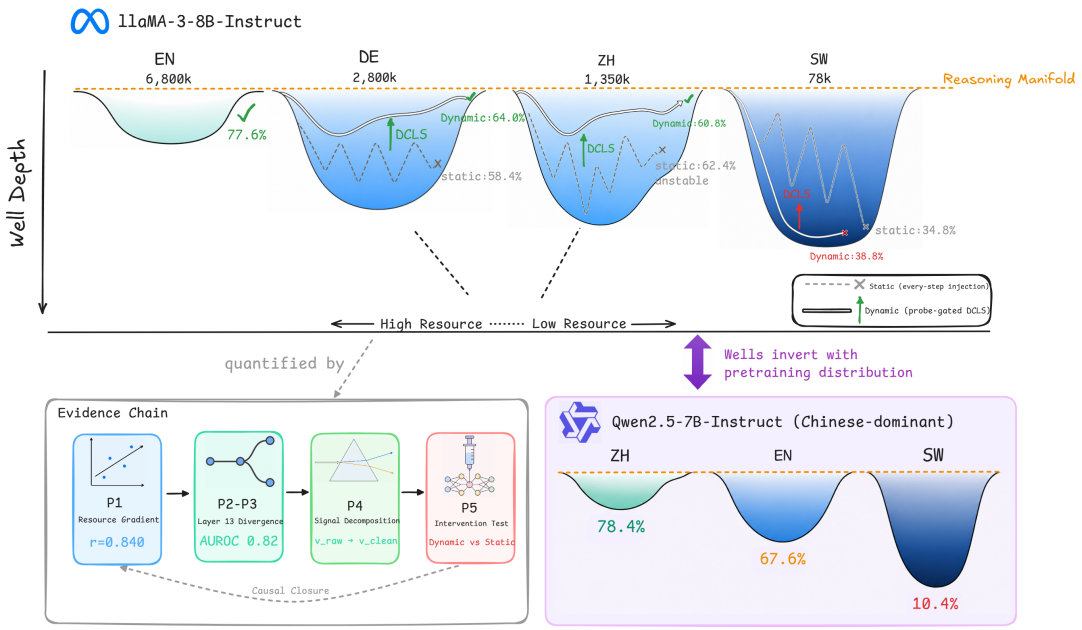


Figure 1: Overview of the Linguistic Gravity Well hypothesis. The figure links resource-associated performance gaps, layer-localized divergence, SVD debiasing, diagnostic DCLS intervention, and a cross-model check.

0.002).

- **Mechanistic localization.** We identify a mid-layer region around layers 13–14 where correct and incorrect trajectories are most separable and correctness is linearly decodable from hidden states (AUROC 0.82).
- **Diagnostic intervention and cross-model evidence.** We introduce DCLS as a probe-gated diagnostic intervention and show resource-dependent effects, with an additional Qwen2.5-7B-Instruct check.

2 Related Work

Cross-lingual reasoning and resource imbalance. Multilingual LLMs have advanced in cross-lingual transfer, yet reasoning performance remains uneven across languages. MGSM shows that chain-of-thought prompting can elicit multilingual mathematical reasoning while revealing large gaps between high- and low-resource languages (Shi et al., 2023). Related work likewise reports that language identity, script, typology, and resource availability shape model behavior (Wang et al., 2023a; Huang et al., 2023; Joshi et al., 2020). These patterns are often attributed to pretraining data imbalance: languages with more web text, Wikipedia, or instruction-tuning exposure receive stronger support. This framing leaves the internal reasoning mechanism open.

Still, imbalance alone does not explain how reasoning trajectories change across languages. We use Wikipedia article count as a resource proxy and ask whether the behavioral gradient has layer-localized representational signatures.

Prompting, pivoting, and translation-based remedies. Many practical approaches treat multilingual reasoning gaps as input or prompting problems. English-pivot prompting asks the model to reason in English for non-English problems; translate-then-reason pipelines first map inputs into a high-resource language and then apply standard reasoning prompts. Multilingual chain-of-thought and cross-lingual-thought prompting encourage explicit intermediate steps across languages (Kojima et al., 2022; Shi et al., 2023; Huang et al., 2023). These methods test whether poor performance mainly reflects surface translation difficulty or prompt misalignment. In our experiments, English-pivot prompting helps some languages but harms others, suggesting that the gap is not reducible to a single translation bottleneck. This motivates a complementary view in which language-conditioned hidden states interact with internal reasoning computations.

Multilingual representations and probing. Prior representation-learning work shows that multilingual models can align languages in shared spaces, but alignment remains incomplete and uneven across languages, scripts, and resource levels (Con-

neau et al., 2020; NLLB Team et al., 2022). Probing studies show that linguistic and task-relevant information can often be decoded from hidden states, while cautioning that decodability is not causality (Belinkov et al., 2017; Hewitt and Manning, 2019). We use layer-wise separation and correctness probes to identify where correct and incorrect reasoning trajectories become distinguishable, not to claim that a linear probe explains the computation. The diagnostic asks whether correctness-related information becomes recoverable in the same layer range across languages and varies with resource level.

Mechanistic interpretability and activation interventions. Mechanistic interpretability studies analyze behavior through hidden representations, causal tracing, circuit hypotheses, and activation interventions (Elhage et al., 2021; Olsson et al., 2022; Meng et al., 2022; Wang et al., 2023b; Stolfo et al., 2023). Representation engineering and activation steering show that learned residual-stream directions can change outputs at inference time (Zou et al., 2023; Turner et al., 2023; Li et al., 2023). Related intervention methods often target desired behavior, truthfulness, or reasoning-critical failures (Li et al., 2023; Dong et al., 2026; Pokharel et al., 2026). Our use of steering is diagnostic: Dynamic Contrastive Logic Steering (DCLS) applies the same correctness-associated perturbation under probe control to test whether languages respond differently to an intervention derived from English correct/incorrect trajectories. The object is not a reliable repair method, but language-dependent intervention sensitivity.

Positioning of this work. This paper connects multilingual reasoning evaluation, representational probing, and activation-level intervention. The contribution is not to show that one steering direction uniformly improves multilingual reasoning. Instead, we ask whether a resource-associated behavioral gap is mirrored by mid-layer trajectory separation, partial decomposability of language and correctness directions, and heterogeneous responses to a diagnostic intervention. This framing treats multilingual reasoning gaps as resource-sensitive representational phenomena while keeping the statistical and mechanistic claims scoped. It also separates diagnostic evidence from claims of a complete causal circuit. This distinction keeps the following intervention analysis diagnostic rather than prescriptive, and frames resource effects as testable representational signatures rather than post-hoc error labels.

3 The Linguistic Gravity Well

3.1 Informal Statement

We use the term *linguistic gravity well* as a mechanistic metaphor rather than as a claim that the model implements an explicit physical energy function. The hypothesis is intended to generate observable predictions about behavior, representation geometry, and intervention sensitivity.

Consider a multilingual LLM processing a reasoning task in language L. As the model generates tokens, its hidden-state trajectory $\mathbf{h}_{l,t}$ traces a path through representation space. For high-resource languages, this trajectory may remain closer to regions associated with valid reasoning operations. For lower-resource languages, we hypothesize that the trajectory is more likely to deviate from these regions due to language-conditioned representational structure.

Operationally, a deeper well refers to a language for which reasoning accuracy is lower, correct/incorrect trajectory separation is more diffuse or delayed, and correctness-oriented steering produces smaller gains or larger degradation. This is an operational definition based on observable signatures, not a direct measurement of a physical potential or attractor depth. We test these signatures through the predictions below.

3.2 Testable Predictions

The gravity well hypothesis generates five specific, falsifiable predictions:

P1 (Resource Gradient). Reasoning accuracy $A(L)$ should be positively associated with language resource availability $R(L)$ across languages, using Wikipedia article count as an observable proxy for resource level.

P2 (Spatial Localization). The divergence between correct and incorrect reasoning trajectories, measured by cosine separation between trajectory centroids, should be concentrated in a specific layer range rather than uniformly distributed across the network.

P3 (Linear Decodability). Near the divergence horizon, correctness should be linearly decodable from hidden states, indicating that correct and incorrect outcomes are associated with distinguishable representational directions.

P4 (Partial Signal Decomposability). A contrastive correctness-associated direction should be partially separable from an estimated language-identity subspace, suggesting that language-related

and correctness-related variation are not fully entangled.

P5 (Intervention Sensitivity). Given a fixed correctness-oriented steering procedure, languages associated with deeper representational wells should show smaller gains or larger degradation than languages associated with shallower wells.

4 Experimental Setup

4.1 Model and Data

We use **LLaMA-3-8B-Instruct** (Llama Team, 2024), a 32-layer Transformer with hidden dimension 4096, loaded in bfloat16 precision, as our primary model. For a cross-model check (§6), we replicate key components of the pipeline—including model-specific probe training, SVD debiasing, and DCLS evaluation—on **Qwen2.5-7B-Instruct** (Yang et al., 2024), a 28-layer Chinese-dominant model (hidden dimension 3584), to assess whether similar resource-associated patterns appear under a different architecture and pretraining distribution.

We evaluate on **MGSM** (Multilingual Grade School Math; Shi et al., 2023), which provides 250 elementary math problems in 10 languages. All experiments use greedy decoding (temperature 0) with a maximum of 1024 new tokens.

4.2 Language Selection

We evaluate 10 MGSM languages spanning high-, mid-, and low-resource settings: EN, DE, FR, ES, JA, ZH, TH, BN, TE, and SW. These languages cover multiple families and scripts, with Wikipedia article counts ranging from 6.8M for English to 78K for Swahili. The full language list is provided in Appendix A.2.

4.3 Methods Overview

Pipeline. Our pipeline has four stages. First, we evaluate zero-shot CoT and English-pivot CoT on all 10 languages and cache last-token hidden states from all 32 layers. Second, we diagnose trajectory divergence using layer-wise centroid separation and correctness probes. Third, we extract a contrastive correctness direction from English correct/incorrect trajectories and apply SVD-based projection to reduce language-associated components. Fourth, we evaluate DCLS, a probe-gated steering intervention at layer 13, and compare it with static injection.

Statistical testing. All per-language accuracies are reported with Wilson 95% confidence intervals (n=250). The resource–accuracy correlation is assessed with both Pearson and Spearman coefficients, supplemented by bootstrap 95% CIs (B=10,000) and leave-one-language-out robustness checks. For paired comparisons with per-sample predictions, such as English-pivot and Qwen DCLS, we report paired bootstrap CIs and McNemar's exact test. For LLaMA DCLS, where only aggregated accuracy counts are available, we report approximate bootstrap intervals and interpret deltas descriptively. To control for potential confounds, we report tokenizer fertility statistics and partial correlations after residualizing on average input token count, as well as invalid-answer rates across all languages. Probe AUROC estimates are accompanied by bootstrap CIs and random-label permutation baselines. Full statistical tables are in Appendix A.8.

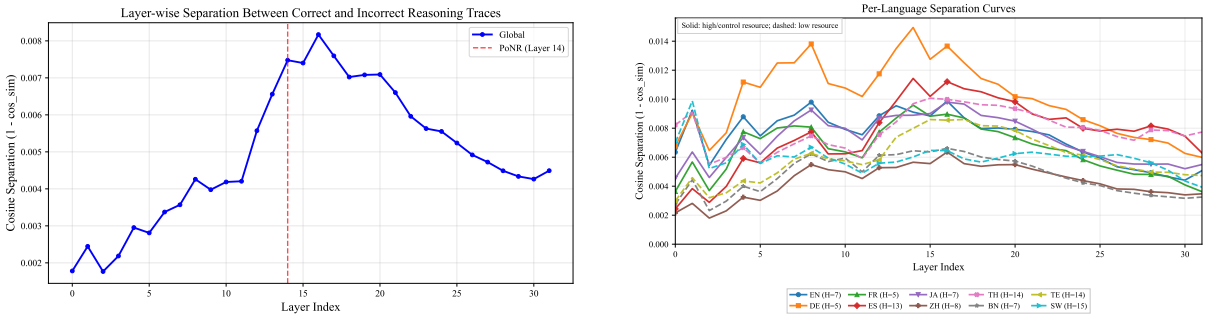
5 Evidence for the Gravity Well

We now present evidence for each of the five predictions derived from the linguistic gravity well hypothesis.

5.1 Behavioral Evidence: The Resource Gradient (P1)

Table 1 presents zero-shot CoT and English-pivot CoT accuracy across all 10 languages. English achieves 77.6%, while Swahili achieves 40.8%, yielding a 36.8-point gap. The Pearson correlation between log-scaled Wikipedia article counts and zero-shot accuracy is strong (r=0.840, p=0.002; bootstrap 95% CI $[0.57, 0.96]$), and the Spearman rank correlation is even stronger (ρ=0.915, p<0.001), supporting P1. Leave-one-language-out analysis confirms the gradient is not driven by any single language (all $r \geq 0.75$; Appendix A.8). A partial correlation controlling for tokenizer fertility retains the positive association (r_{partial}=0.803, p=0.005), indicating that the resource gradient is not solely explained by tokenization differences. Invalid-answer rates are negligible ($\leq 0.4\%$ across all languages), ruling out answer-extraction failure as a confound.

English-pivot prompting does not uniformly bridge the gap: it helps some languages but harms others. This suggests that the cross-lingual reasoning gap cannot be reduced to a simple translate-then-reason failure mode, motivating the represen-



(a) Global separation curve. The divergence horizon at layer 14 marks where correct and incorrect trajectories maximally diverge.

(b) Per-language curves. High-resource languages show sharper, earlier peaks; lower-resource languages show flatter or later peaks.

Figure 2: Layer-wise cosine separation between correct and incorrect reasoning trajectory centroids. Panel (a) shows the global separation curve, with the largest separation around layers 13–14. Panel (b) shows that the location and sharpness of this separation vary across languages, supporting the view that correct/incorrect decodability is modulated by language-conditioned representation structure.

Table 1: Zero-shot CoT and English-pivot CoT accuracy (%) on MGSM. Δ: pivot gain. English-pivot helps some languages (SW +6.4, DE +5.2) but hurts others (FR -5.2), indicating the gap is not merely translational.

Lang	Wiki (K)	Zero-shot	Eng-Pivot	Δ
EN	6,800	77.6	—	—
FR	2,550	69.2	64.0	-5.2
ES	1,950	68.8	68.8	0.0
DE	2,800	61.2	66.4	+5.2
TH	160	59.6	57.6	-2.0
ZH	1,350	58.4	61.6	+3.2
JA	1,420	55.2	58.0	+2.8
BN	145	53.2	54.4	+1.2
TE	83	50.0	50.8	+0.8
SW	78	40.8	47.2	+6.4

Table 2: Probe AUROC and centroid-separation scores at representative layers. The linear probe reaches its highest AUROC at layer 13, close to the best MLP result, indicating that correctness is linearly decodable near the divergence horizon. This supports P3 as a decodability claim, without implying that the linearly recoverable correctness signal is causal.

Layer	Linear	MLP	Sep. Score
0	0.712	0.715	0.0018
7	0.795	**0.830**	0.0036
13	**0.820**	0.810	0.0066
14	0.813	0.815	0.0075
16	0.785	0.806	**0.0082**
31	0.761	0.770	0.0045

tational analyses below.

5.2 Mechanistic Evidence: Locating the Well Boundary (P2 + P3)

Layer-wise separation. For each layer, we compute centroid separation between correct and incorrect hidden-state trajectories, $s_l = 1 - \cos(\bar{\mathbf{h}}_l^+, \bar{\mathbf{h}}_l^-)$. Figure 2(a) shows that separation rises sharply around layers 12–16, with the largest separation around layers 13–14. Figure 2(b) shows that this pattern varies by language: higher-resource languages tend to show sharper and earlier separation, while lower-resource languages show flatter or later separation. The resource–horizon correlation is negative ($r = -0.690$, $p = 0.027$), supporting P2. We therefore treat the boundary as a mid-layer region rather than a single exact layer, since neighboring layers carry related trajectory information.

Probe analysis. Table 2 reports probe results. The linear probe reaches AUROC 0.820 at layer 13, close to the best MLP result (0.830), indicating that correctness is linearly decodable near the divergence horizon. The small linear–MLP gap suggests that the signal is recoverable by a simple readout, not only by a high-capacity classifier. Bootstrap 95% CIs and random-label permutation baselines confirm that these AUROC values exceed chance (Appendix A.8). This supports P3 as a decodability claim, not as evidence that linear decodability itself is causal. In other words, the probe shows that correctness-related information is recoverable from hidden states, but it does not by itself identify the causal mechanism that produces correct reasoning.

5.3 Signal Decomposability (P4)

To test P4, we extract a contrastive correctness direction from English correct/incorrect trajectories:

$$v_{\text{raw}}^{(l)} = \frac{\bar{\mathbf{h}}_l^+ - \bar{\mathbf{h}}_l^-}{\|\bar{\mathbf{h}}_l^+ - \bar{\mathbf{h}}_l^-\|}.$$

Because this direction may contain English-specific components, we estimate a language subspace using 1,800 FLORES-200 (NLLB Team et al.,

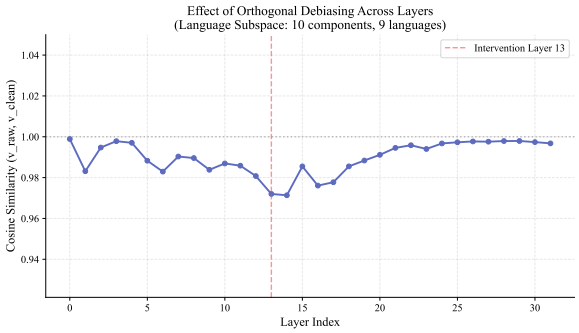


Figure 3: Cosine similarity between the raw contrastive direction v_{raw} and the SVD-debiased direction v_{clean} across layers. The decrease around layers 13–14 suggests that the estimated language subspace overlaps most with the raw contrastive direction near the divergence horizon. This provides evidence for partial, rather than complete, disentanglement of language-related and correctness-related components.

2022) parallel sentence differences and take the top $k = 10$ SVD directions as U. This construction treats cross-lingual hidden-state differences on parallel sentences as an operational estimate of language-associated variation. The projection is therefore not intended to isolate a correctness-associated direction independent of language, but to reduce one measurable component of language identity from the contrastive correctness direction. Accordingly, we refer to the resulting vector as an SVD-debiased contrastive direction. We then project out this subspace:

$$v_{\text{clean}} = v_{\text{raw}} - U(U^{\top} v_{\text{raw}}).$$

Figure 3 shows that v_{raw} and v_{clean} differ most around layers 13–14, suggesting that the estimated language subspace overlaps most with the raw contrastive direction near the divergence horizon. Static injection of v_{clean} yields mixed effects, motivating the dynamic intervention in Section 5.4.

5.4 Intervention Resistance: The Depth Gradient (P5)

P5 provides an intervention-based diagnostic test: if language-conditioned representational structure differs by resource level, then the same steering procedure should have different effects across languages. DCLS applies a layer-13 forward hook that injects v_{clean} only when a smoothed probe score exceeds $\tau = 0.6$. The probe score is smoothed with an exponential moving average ($\alpha_{\text{ema}} = 0.3$), and the perturbation uses injection strength $\alpha = 1.0$. This makes DCLS selective, unlike static steering, which injects the same direction at every generation step.

Table 3: Representative intervention results on MGSM. ZS denotes zero-shot CoT, Static denotes every-step injection, and Dynamic denotes probe-gated DCLS. $\Delta_{\text{D}-\text{ZS}}$ reports Dynamic DCLS minus the zero-shot baseline. This compact table shows five representative languages; complete per-language results and approximate bootstrap intervals are reported in Appendix A.8, Table 18. No individual LLaMA DCLS delta reaches significance at α=0.05, so the pattern should be interpreted diagnostically rather than as evidence of a robust repair method.

Lang	ZS	Static	Dynamic	$\Delta_{\text{D-ZS}}$	Intv.%
EN	77.6	75.2	74.4	−3.2	36.6
DE	61.2	58.4	**64.0**	**+2.8**	54.0
ES	68.8	64.4	**70.4**	**+1.6**	44.2
ZH	58.4	62.4	**60.8**	**+2.4**	53.7
JA	55.2	57.6	**56.0**	**+0.8**	54.4

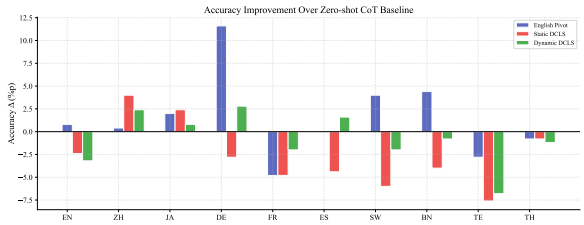


Figure 4: Accuracy change relative to the zero-shot CoT baseline. Dynamic DCLS exhibits language-dependent point estimates across the full language set, providing diagnostic evidence consistent with P5.

Table 3 reports a compact representative subset of the DCLS results, with complete per-language estimates and approximate bootstrap intervals reported in Appendix A.8, Table 18. In the representative subset, Dynamic DCLS yields positive point estimates for several higher- or mid-resource languages—DE (+2.8pp), ES (+1.6pp), ZH (+2.4pp), and JA (+0.8pp)—while EN provides a high-resource reference point with a different direction. Because MGSM contains only 250 examples per language, none of the individual LLaMA DCLS deltas reaches significance at α=0.05: approximate 95% bootstrap CIs cross zero (e.g., DE $[-5.7, 11.3]$; Appendix A.8). We therefore interpret per-language deltas descriptively; the main diagnostic signal is the qualitative, language-dependent variation in intervention sensitivity. This pattern is consistent with the intervention-resistance prediction P5, and supports using DCLS as a diagnostic intervention rather than as a reliable repair method.

The probe-triggered intervention rate provides an additional diagnostic signal: lower-resource languages tend to trigger intervention more frequently. However, this rate should not be interpreted as a

direct measurement of well depth, since it may also reflect probe calibration, tokenization, or generation fragility.

5.5 DCLS Ablation Evidence

Additional ablations support the diagnostic interpretation of DCLS. First, the injection-strength sweep shows that stronger perturbations do not monotonically raise accuracy: moderate values of α can yield positive point estimates for some languages, but large values degrade performance across languages (Appendix A.5, Figure 6). This suggests that the contrastive direction can disrupt generation coherence when applied too strongly, even if it carries correctness-associated information.

Second, threshold sweeps show language-specific calibration effects. Chinese performs best under a more aggressive triggering threshold (τ=0.3), whereas Swahili performs best under a more conservative threshold (τ=0.7; Appendix A.5, Table 8). Thus, the same probe score does not have identical operational meaning across languages. These ablations reinforce the view that DCLS measures language-dependent intervention sensitivity rather than providing a stable repair method.

6 Analysis and Discussion

English-pivot is model- and language-dependent. English-pivot prompting helps some languages but harms others, suggesting that cross-lingual reasoning gaps are not purely translation failures. Under our hypothesis, languages already close to the model's dominant reasoning substrate may not benefit from an English detour, while languages with weaker native-language reasoning support may benefit from partial routing through English.

Static steering is disruptive. Static DCLS often underperforms dynamic DCLS and the zero-shot baseline, suggesting that continuous perturbation can disrupt the model's native trajectory when the injected direction is only partially aligned with the language-conditioned representation. Dynamic steering reduces this disruption by intervening only when the probe detects likely drift. Additional ablations in Appendix A.5 show that large steering magnitudes degrade performance and that different languages prefer different thresholds, reinforcing our interpretation of DCLS as a diagnostic probe rather than a robust repair method.

Robustness and interpretation. Several auxiliary analyses support the main interpretation while clarifying its limits. Full-layer probe results show that the layer-13 peak is not an artifact of inspecting only selected layers, and the static steering comparison shows that SVD debiasing can reduce some language-associated components without making steering uniformly beneficial. Together, these results support a diagnostic, not repair-oriented, interpretation.

Typological and tokenization factors. The resource gradient is not the only factor shaping trajectory geometry. For example, DE and FR show relatively early divergence peaks despite being mid- to high-resource languages. This suggests that typological proximity, script, tokenizer overlap, and instruction-tuning exposure may interact with resource volume. We therefore interpret resource availability as a dominant but incomplete predictor of well geometry, rather than as a single causal explanation. Thus, the hypothesis is resource-sensitive, not resource-deterministic.

Role of alignment. Our analysis focuses on pretraining resource imbalance, but instruction tuning and preference optimization may also reshape language-conditioned representations. SFT and RLHF-style alignment can change how models follow multilingual instructions, while preference optimization methods such as DPO (Rafailov et al., 2023) may amplify or suppress reasoning trajectories differently across languages. Understanding whether alignment stages deepen, flatten, or relocate linguistic wells is an important direction for future mechanistic work.

6.1 Cross-Model Check

We replicate key components of the pipeline on Qwen2.5-7B-Instruct, which exhibits a different language-performance profile. Unlike LLaMA, Qwen achieves higher zero-shot accuracy on Chinese than English (78.4% vs. 67.6%). DCLS further shows an asymmetric effect: an English-derived contrastive direction raises English accuracy from 67.6% to 84.4% but decreases Chinese from 78.4% to 54.0%. English-pivot results show the same model dependence, with pivoting often harmful on Qwen. This suggests that the extracted direction is not universal, but depends on how each model represents the target language. Table 4 summarizes the comparison; full Qwen results are in Appendix A.7. Together, these findings suggest that balanced multilingual pretraining and

Table 4: Cross-model comparison between LLaMA-3-8B-Instruct and Qwen2.5-7B-Instruct. Qwen shows a different language-performance profile, and DCLS indicates that an English-derived contrastive direction is not universally beneficial across models or languages.

Lang	LLaMA-3-8B		Qwen2.5-7B	
	Zero	DCLS	Zero	DCLS
EN	77.6	74.4	67.6	**84.4**
ZH	58.4	60.8	**78.4**	54.0
DE	61.2	64.0	68.0	68.0
SW	40.8	38.8	10.4	15.2
TE	50.0	43.2	29.6	28.4

language-aware calibration may be more reliable long-term directions than inference-time steering alone.

7 Limitations

Dataset and statistical scope. Our experiments use MGSM, so the evidence is limited to arithmetic word-problem reasoning and may not transfer unchanged to commonsense, causal, logical, or instruction-following tasks. Each language has only 250 problems, making small accuracy deltas difficult to interpret. We report Wilson confidence intervals, paired bootstrap intervals, and McNemar's exact test where paired per-sample data are available (Appendix A.8); these reduce uncertainty around the main estimates, but do not make individual DCLS deltas significant at α=0.05 for most languages.

Resource proxies and surface confounds. Wikipedia article count is a transparent proxy for language resource availability, not a measurement of the true language distribution in LLaMA-3 or Qwen2.5 pretraining data. It can understate or overstate actual web, instruction-tuning, or benchmark exposure. We partially address surface confounds by reporting tokenizer fertility, a partial resource–accuracy correlation after residualizing on average input token count (r_{partial}=0.803, p=0.005), language metadata controls, and invalid-answer rates ($\leq$0.4%). These controls make tokenization and answer extraction less plausible as sole explanations, but do not replace corpus-level pretraining statistics.

Mechanistic interpretation and intervention scope. The mechanistic analysis is primarily on LLaMA-3-8B-Instruct. Qwen2.5-7B provides a partial cross-model check through behavioral results, model-specific probes, SVD debiasing, and DCLS evaluation, but not a full replication of the layer-wise separation and divergence-horizon pipeline. Probe results establish decodability, not causality. DCLS should therefore be read as a diagnostic intervention revealing language-dependent sensitivity, not as a stable repair method; its effects can depend on probe calibration, generation fragility, tokenization, and direction alignment.

8 Conclusion

We presented the Linguistic Gravity Well hypothesis as a resource-sensitive representational account of cross-lingual reasoning gaps. Across behavioral evaluation, layer-wise separation, probe decodability, SVD-based projection, and diagnostic steering, we find converging evidence that multilingual reasoning gaps are associated with layer-localized and resource-sensitive internal structure. This suggests that cross-lingual reasoning failures are not merely surface-level translation errors and motivates future causal analyses of multilingual reasoning.

Ethics Statement

This work analyzes multilingual reasoning behavior in publicly available benchmark tasks and does not involve human-subject experiments, private user data, or newly collected personal information. The main ethical risk is interpretive: resource-linked performance differences should not be read as claims about the intrinsic value, complexity, or reasoning capacity of languages or speaker communities. We use Wikipedia article count only as a transparent proxy for language resource availability, not as a direct estimate of model-specific pretraining exposure or cultural importance. DCLS is likewise presented as a diagnostic intervention for studying language-dependent sensitivity, not as a deployable method for reliably repairing multilingual model behavior.

Reproducibility Statement

We report the datasets, language set, prompts, model families, layer choices, probe settings, intervention hyperparameters, and statistical tests needed to reproduce the main analyses. Appendix A lists the languages and hyperparameters, and Appendix A.8 provides the main robustness and control tables, including Wilson confidence intervals, correlation analyses, paired or approximate bootstrap intervals, tokenizer-fertility controls, invalid-answer rates, and probe baselines. Reproduction requires access to the corresponding model checkpoints and sufficient compute for

hidden-state extraction and generation; exact runtime may vary with hardware, caching, and inference backend.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Fangan Dong, Zuming Yan, Xuri Ge, Zhiwei Xu, Mengqi Zhang, Xuanang Chen, Ben He, Xin Xin, Zhumin Chen, and Ying Zhou. 2026. Identifying and transferring reasoning-critical neurons: Improving llm inference reliability via activation steering. *arXiv preprint arXiv:2601.19847*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, and Nicholas Joseph. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of EMNLP*.

Pratik Joshi, Sebastin Santy, Amar Buber, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Llama Team. 2024. The Llama 3 herd of models. *Leibniz-Zentrum für Informatik (Schloss Dagstuhl)*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Rhitabrat Pokharel, Ameeta Agrawal, and Tanay Nagar. 2026. Cross-lingual activation steering for multilingual language models. *arXiv preprint arXiv:2601.16390*.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *Proceedings of the International Conference on Learning Representations*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of EMNLP*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David S. Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via ChatGPT. *arXiv preprint arXiv:2302.14229*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: a circuit for indirect object

identification in GPT-2 small. In *Proceedings of ICLR*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Additional Experimental Details

A.1 Additional Behavioral Results

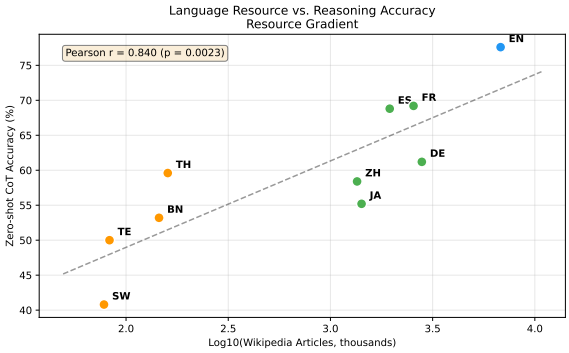


Figure 5: Resource gradient: log-scaled Wikipedia article count versus zero-shot CoT accuracy. Each point represents one language. The positive correlation supports a resource-associated view of the cross-lingual reasoning gap, while relying on Wikipedia size only as an indirect proxy for language resource availability.

A.2 Language List

Table 5: Languages used in our study, ordered by Wikipedia article count as a proxy for language resource availability.

Lang	Code	Family	Script	Wiki (K)
English	EN	Germanic	Latin	6,800
German	DE	Germanic	Latin	2,800
French	FR	Romance	Latin	2,550
Spanish	ES	Romance	Latin	1,950
Japanese	JA	Japonic	Mixed	1,420
Chinese	ZH	Sinitic	Hanzi	1,350
Thai	TH	Kra-Dai	Thai	160
Bengali	BN	Indo-Aryan	Bangla	145
Telugu	TE	Dravidian	Telugu	83
Swahili	SW	Bantu	Latin	78

The language set is designed to span several resource levels, scripts, and families while keeping per-language evaluation size fixed. Wikipedia article count is used only as a transparent ordering proxy: it provides a reproducible resource signal, but it should not be read as a direct estimate of model-specific pretraining exposure. The script and family annotations are therefore descriptive controls rather than causal explanations for the observed accuracy gradient.

A.3 Hyperparameters

Table 6 summarizes all hyperparameters used across the four experimental stages.

All decoding and feature-extraction settings are fixed across languages, using greedy decoding, final-token hidden states, and calibrated intervention parameters for the full evaluations.

Table 6: Complete hyperparameter listing.

Parameter	Value	Stage
Model	LLaMA-3-8B-Instruct	All
Precision	bfloat16	All
Temperature	0.0 (greedy)	All
Max new tokens	1024	All
Hidden state position	Last token	1–4
PCA dimensions	2	2
Separation metric	Cosine separation	2
Divergence horizon window	3 layers	2
Linear probe regularization	$C = 0.1$, L2	2
Linear probe class weights	Balanced	2
MLP probe architecture	4096→256→1	2
MLP probe training	100 epochs, Adam	2
Parallel pairs per language	200 (FLORES-200)	3
SVD components k	10	3
Intervention layer	13	4
Injection strength α	1.0	4
Drift threshold τ	0.6	4
EMA smoothing α_{ema}	0.3	4

A.4 Full Probe AUROC Across All Layers

Table 7 reports linear and MLP probe AUROC for all 32 layers.

Table 7: Full 32-layer probe AUROC. Bold: best per probe type.

L	Linear	MLP	L	Linear	MLP
0	.712	.715	16	.785	.806
1	.718	.720	17	.780	.800
2	.720	.722	18	.773	.795
3	.735	.740	19	.770	.790
4	.795	.800	20	.753	.780
5	.790	.798	21	.750	.775
6	.792	.810	22	.748	.770
7	.795	**.830**	23	.745	.765
8	.800	.820	24	.741	.760
9	.804	.815	25	.740	.758
10	.808	.812	26	.738	.755
11	.810	.810	27	.737	.752
12	.815	.808	28	.739	.750
13	**.820**	.810	29	.742	.755
14	.813	.815	30	.750	.760
15	.800	.810	31	.761	.770

The layer sweep shows a broad mid-layer region with strong linear decodability. Linear AUROC peaks at layer 13, while the MLP probe peaks earlier at layer 7; both remain well above chance across the middle layers. This supports using the mid-layer region for diagnostic intervention while treating the exact layer choice as a calibrated design decision rather than a universal optimum.

A.5 DCLS Ablations

Table 8 and Figure 6 report the DCLS threshold and injection-strength ablations summarized in Section 5.5.

A.6 Static Steering: v_{raw} vs. v_{clean}

Table 9 compares static injection of the raw vs. SVD-debiased contrastive direction across all 10

Table 8: Threshold ablation. ZH prefers aggressive intervention (τ=0.3), while SW prefers a more conservative threshold (τ=0.7), indicating language-specific calibration sensitivity.

	Accuracy (%)			Intervention Rate (%)		
τ	EN	ZH	SW	EN	ZH	SW
0.3	77.6	**62.8**	38.0	73.6	84.7	87.2
0.4	77.2	60.0	36.4	62.1	76.2	79.5
0.5	75.6	61.6	33.2	49.2	65.3	69.8
0.6	74.4	60.8	38.8	36.6	53.7	58.2
0.7	76.8	62.0	**39.6**	24.6	39.4	44.3
0.8	76.0	61.6	38.8	14.9	27.2	31.4

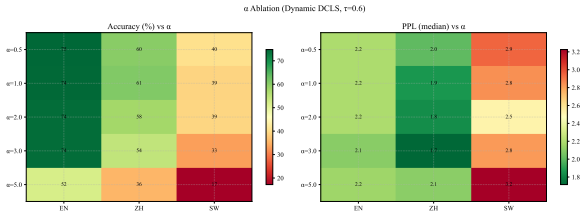


Figure 6: Injection-strength ablation for Dynamic DCLS. Increasing the steering magnitude does not monotonically raise accuracy; large values of α degrade performance across languages. This supports the view that DCLS should be interpreted as a diagnostic probe of representational sensitivity rather than as a robust repair method.

languages.

Table 9: Static steering comparison. $\mathbf{v}_{\text{clean}}$ outperforms $\mathbf{v}_{\text{raw}}$ for ZH and JA, suggesting that debiasing can reduce some language-associated components. However, both static methods hurt most languages overall, motivating dynamic intervention.

Lang	Baseline	$\mathbf{v}_{\text{raw}}$	$\mathbf{v}_{\text{clean}}$	$\Delta_{\text{clean}-\text{raw}}$
EN	77.6	74.4	75.2	+0.8
ZH	58.4	60.8	**62.4**	**+1.6**
JA	55.2	57.2	**57.6**	+0.4
DE	61.2	61.6	58.4	−3.2
FR	69.2	63.2	64.4	+1.2
ES	68.8	62.8	64.4	+1.6
SW	40.8	37.2	34.8	−2.4
BN	53.2	52.4	49.2	−3.2
TE	50.0	45.2	42.4	−2.8
TH	59.6	60.8	58.8	−2.0

A.7 Cross-Model Check Details

As a partial cross-model check of the linguistic gravity well hypothesis, we replicated the behavioral experiments on Qwen2.5-7B-Instruct, a 28-layer Chinese-dominant model. Tables 10 and 11 present the complete Qwen results across all 10 languages.

The cross-model comparison reveals three key patterns: (1) a positive resource–accuracy association appears in both models (Qwen r=0.659, LLaMA r=0.840); (2) the language-performance profile changes with model-specific training choices, with Chinese becoming Qwen's strongest language among those evaluated; and (3) English-pivot's effectiveness depends on the model, not universal properties of English as a reasoning language.

Table 10: Qwen2.5-7B zero-shot CoT results. Chinese achieves the highest accuracy (78.4%), surpassing English (67.6%), suggesting that the language-performance profile depends strongly on model-specific data and training choices. Thai's anomalously high performance may reflect tokenizer coverage, instruction-tuning exposure, or other model-specific data effects. Swahili's extremely low accuracy (10.4%, vs. LLaMA's 40.8%) is consistent with weaker model support for Swahili, though the source may include tokenization, instruction tuning, corpus composition, or benchmark-specific effects.

Language	Correct	Total	Accuracy (%)
ZH	196	250	**78.4**
TH	188	250	75.2
DE	170	250	68.0
FR	170	250	68.0
EN	169	250	67.6
JA	165	250	66.0
ES	165	250	66.0
BN	148	250	59.2
TE	74	250	29.6
SW	26	250	10.4

Table 11: Qwen2.5-7B English-pivot results. Unlike LLaMA, where pivoting helps low-resource languages, English-pivot is actively harmful for most languages on Qwen, especially Chinese (−13.2pp) and Thai (−12.8pp). One possible interpretation is that English is not Qwen's dominant reasoning substrate, so English-pivot prompting may introduce a representational mismatch for high-performing non-English languages. Only Swahili (+3.2pp), which has an extremely low baseline accuracy (10.4% baseline), shows marginal benefit.

Language	Zero-shot	Eng-Pivot	Δ
EN	67.6	67.6	0.0
ZH	78.4	65.2	**−13.2**
JA	66.0	59.6	−6.4
DE	68.0	64.0	−4.0
FR	68.0	63.2	−4.8
ES	66.0	62.8	−3.2
TH	75.2	62.4	**−12.8**
BN	59.2	54.8	−4.4
TE	29.6	29.2	−0.4
SW	10.4	13.6	+3.2

A.7.1 Qwen-Specific DCLS Pipeline

To ensure methodological rigor, the DCLS pipeline for Qwen was built entirely from Qwen's own representations:

1. Probe training: A logistic regression probe was trained on Qwen's hidden states at all 28 layers using 250 MGSM problems. Best layer: 17 (AUROC 0.851).

2. Contrastive vector extraction: $\mathbf{v}_{\text{raw}}$ was com-

puted as the mean difference between Qwen's layer-17 hidden states for English correct vs. incorrect samples.

3. SVD debiasing: Using 200 FLORES-200 parallel sentence pairs across 9 target languages (1,800 pairs total), we extracted EN–target hidden state differences at layer 17, applied SVD to obtain the top-10 components as a language subspace $U \in \mathbb{R}^{3584\times10}$, and projected: $\mathbf{v}_{\text{clean}} = \mathbf{v}_{\text{raw}} - U(U^\top \mathbf{v}_{\text{raw}})$, followed by renormalization. Resulting $\cos(\mathbf{v}_{\text{raw}}, \mathbf{v}_{\text{clean}}) = 0.970$.

4. Threshold calibration: τ sweep on 50 samples $\times$ 3 languages (EN, ZH, SW) over $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Optimal $\tau = 0.3$ (maximizing mean accuracy).

5. Evaluation: Full DCLS with α=1.0, τ=0.3, EMA smoothing α_{ema}=0.3 on 5 languages $\times$ 250 problems.

No components from the LLaMA pipeline were reused; all probe weights, SVD-debiased contrastive directions, and language subspace matrices were derived from Qwen's own hidden representations.

A.8 Statistical Robustness and Control Analyses

This section reports the full statistical tables referenced in the main text.

Probe robustness. Table 12 reports bootstrap confidence intervals for selected linear-probe AUROC values, using 5-fold out-of-fold scores from cached LLaMA hidden states. Random-label permutation baselines remain near chance.

Table 12: Probe AUROC with bootstrap 95% CI and random-label permutation baseline for selected layers. Reported AUROC is the held-out validation score from Phase 2; CV AUROC is computed from 5-fold out-of-fold linear-probe scores on cached hidden states. The random-label AUROC remains near chance ($\sim$0.50).

L	Rep.	CV	95% CI	Rand.
7	.795	.771	[.752, .790]	.501±.012
10	.785	.762	[.744, .781]	.501±.012
13	.820	.787	[.768, .804]	.500±.012
14	.813	.784	[.766, .803]	.500±.012
17	.760	.777	[.758, .795]	.500±.012
20	.785	.761	[.742, .780]	.500±.012
25	.750	.758	[.739, .778]	.500±.012
31	.753	.756	[.736, .776]	.500±.012

Resource–accuracy correlation. Table 13 reports Pearson and Spearman correlations between log-scaled Wikipedia article counts and zero-shot CoT accuracy, with bootstrap 95% CIs (B=10,000).

Table 13: Resource–accuracy correlation with bootstrap 95% CI (B=10,000, n=10 languages).

Model	Pearson	p	Spearman	p
LLaMA	.840 [.57, .96]	.002	.915 [.60, 1.00]	$<$.001
Qwen	.663 [−.45, .95]	.037	.530 [−.26, .98]	.115

Leave-one-language-out robustness. Table 14 confirms that the resource gradient is not driven by any single language.

Table 14: LLaMA-3-8B leave-one-language-out resource–accuracy correlation. The gradient remains positive regardless of which language is excluded (all $r \geq 0.75$).

Excluded	Pearson r	p	Spearman ρ	p
EN	0.786	0.012	0.883	0.002
DE	0.862	0.003	0.933	0.000
FR	0.826	0.006	0.917	0.001
ES	0.833	0.005	0.917	0.001
JA	0.880	0.002	0.933	0.000
ZH	0.854	0.003	0.933	0.000
TH	0.886	0.002	0.933	0.000
BN	0.836	0.005	0.883	0.002
TE	0.825	0.006	0.883	0.002
SW	0.794	0.011	0.883	0.002

Tokenizer fertility. Table 15 reports tokenizer fertility (tokens per character) for both models. Table 16 shows that the resource–accuracy partial correlation remains positive after residualizing on average input token count.

Table 15: Tokenizer fertility across languages. M abbreviates model; higher Tok/Char indicates less efficient tokenization.

M	Lang	Chars	Tok.	Tok/Char	Char/Tok
LLaMA	EN	241	60	0.249	4.0
	DE	271	81	0.298	3.4
	FR	266	80	0.300	3.3
	ES	257	77	0.299	3.4
	JA	112	84	0.744	1.3
	ZH	85	72	0.847	1.2
	TH	214	108	0.507	2.0
	BN	241	283	1.173	0.8
	TE	266	448	1.684	0.6
	SW	266	104	0.391	2.6
Qwen	EN	241	61	0.254	3.9
	DE	271	82	0.302	3.3
	FR	266	80	0.303	3.3
	ES	257	78	0.303	3.3
	JA	112	80	0.717	1.4
	ZH	85	66	0.771	1.3
	TH	214	130	0.607	1.6
	BN	241	258	1.069	0.9
	TE	266	384	1.444	0.7
	SW	266	106	0.400	2.5

Invalid answer rates. Invalid-answer rates are negligible across all languages and models: all Qwen rates are 0.0%, and LLaMA has only one invalid Swahili answer (0.4%) with all other languages at 0.0%. This rules out answer-extraction failure as a confound while avoiding a low-information table

Table 16: Resource–accuracy partial correlation after residualizing on average input token count. For LLaMA, the partial correlation remains strong (r_{partial}=0.803, p=0.005), indicating that tokenization differences do not solely explain the resource gradient.

Model	r_{resource}	r_{tokens}	r_{partial}	p_{partial}
LLaMA-3-8B	0.840	−0.463	0.803	0.005
Qwen2.5-7B	0.663	−0.474	0.526	0.118

dominated by zeros.

Language-level controls. Table 17 aggregates per-language metadata for reference. With only n=10 languages, script and family controls are descriptive rather than inferential.

The control table is also intended to make the evaluation set auditable. It records the surface properties most likely to affect interpretation before any claim about resource-sensitive behavior is made. Because several attributes are correlated in a small multilingual sample, we avoid fitting a larger regression model and instead report the partial tokenizer control separately.

Table 17: Language-level controls for LLaMA-3-8B zero-shot CoT. Because n=10, script and family controls are descriptive.

Lang	Family	Script	Wiki K	Acc.	Tok.
EN	Germanic	Latin	6,800	77.6	59.8
DE	Germanic	Latin	2,800	61.2	80.7
FR	Romance	Latin	2,550	69.2	79.7
ES	Romance	Latin	1,950	68.8	76.8
JA	Japonic	Mixed	1,420	55.2	83.5
ZH	Sinitic	Hanzi	1,350	58.4	72.1
TH	Kra-Dai	Thai	160	59.6	108.3
BN	Indo-Aryan	Bangla	145	53.2	283.1
TE	Dravidian	Telugu	83	50.0	447.6
SW	Bantu	Latin	78	40.8	103.9

These controls are included to make the language-level comparisons transparent rather than to support high-dimensional causal claims. In particular, script, family, Wikipedia size, and average token count are reported together so that readers can inspect whether the resource gradient coincides with obvious surface properties. The main statistical control used in the paper is the tokenizer partial correlation, while the remaining metadata should be read as descriptive context. This presentation keeps the robustness checks auditable without implying that the small language sample supports a full multivariate model.

A.9 Additional Notes on Scope

Model scope. The full mechanistic pipeline for P2–P4 is conducted primarily on LLaMA-3-8B-Instruct, covering layer-wise separation, probe training, and SVD-based signal decomposition.

Table 18: LLaMA-3-8B significance tests. Pivot denotes English-pivot CoT vs. zero-shot; Dyn. denotes Dynamic DCLS vs. zero-shot. DCLS CIs are approximate bootstraps from aggregated counts, so p is not reported.

L	Cmp.	B	I	Δ	CI	p
DE	Pivot	61.2	66.4	+5.2	[−1.2, 11.6]	.319
FR	Pivot	69.2	64.0	−5.2	[−11.6, 1.2]	.341
ES	Pivot	68.8	68.8	0.0	[−6.4, 6.4]	.992
JA	Pivot	55.2	58.0	+2.8	[−4.4, 9.6]	.791
ZH	Pivot	58.4	61.6	+3.2	[−3.2, 9.6]	.690
TH	Pivot	59.6	57.6	−2.0	[−8.4, 4.4]	.890
BN	Pivot	53.2	54.4	+1.2	[−5.2, 7.2]	.968
TE	Pivot	50.0	50.8	+0.8	[−5.6, 7.2]	.993
SW	Pivot	40.8	47.2	+6.4	[−0.4, 13.2]	.228
EN	Dyn.	77.6	74.4	−3.2	[−10.7, 4.3]	—
DE	Dyn.	61.2	64.0	+2.8	[−5.7, 11.3]	—
ES	Dyn.	68.8	70.4	+1.6	[−6.5, 9.7]	—
FR	Dyn.	69.2	67.2	−2.0	[−10.2, 6.2]	—
ZH	Dyn.	58.4	60.8	+2.4	[−6.2, 11.0]	—
JA	Dyn.	55.2	56.0	+0.8	[−7.9, 9.5]	—
TH	Dyn.	59.6	58.4	−1.2	[−9.8, 7.4]	—
BN	Dyn.	53.2	52.4	−0.8	[−9.6, 8.0]	—
TE	Dyn.	50.0	43.2	−6.8	[−15.5, 1.9]	—
SW	Dyn.	40.8	38.8	−2.0	[−10.6, 6.6]	—

Table 19: Qwen2.5-7B paired significance tests. Pivot denotes English-pivot CoT vs. zero-shot; DCLS denotes Qwen-specific DCLS vs. zero-shot. $^{**}p<0.01$ under McNemar's exact test.

L	Cmp.	B	I	Δ	CI	p
ZH	Pivot	78.4	65.2	−13.2	[−19.2, −7.2]	.000**
TH	Pivot	75.2	62.4	−12.8	[−19.2, −6.4]	.001**
JA	Pivot	66.0	59.6	−6.4	[−12.8, 0.0]	.191
FR	Pivot	68.0	63.2	−4.8	[−10.0, 0.8]	.283
BN	Pivot	59.2	54.8	−4.4	[−10.8, 2.0]	.484
DE	Pivot	68.0	64.0	−4.0	[−9.6, 1.6]	.445
ES	Pivot	66.0	62.8	−3.2	[−8.8, 2.4]	.600
SW	Pivot	10.4	13.6	+3.2	[−1.6, 8.0]	.525
EN	DCLS	67.6	84.4	+16.8	[11.6, 22.0]	.000**
ZH	DCLS	78.4	54.0	−24.4	[−31.2, −17.6]	.000**
DE	DCLS	68.0	68.0	0.0	[−5.2, 5.2]	.989
TE	DCLS	29.6	28.4	−1.2	[−7.2, 4.8]	.967
SW	DCLS	10.4	15.2	+4.8	[0.0, 10.0]	.237

Qwen2.5-7B is used as a partial cross-model check with model-specific probes, SVD-debiased directions, and DCLS evaluation, so we interpret it as evidence of model dependence rather than as a complete replication.

Task scope. MGSM provides a controlled multilingual arithmetic reasoning setting, but it does not cover all reasoning modes. Extending the analysis to commonsense, causal, symbolic, and instruction-following tasks would test whether the same resource-sensitive representational patterns hold beyond arithmetic word problems.

Proxy and intervention scope. Wikipedia article count is used as a transparent proxy for language resource availability, not as a direct estimate of model-specific pretraining exposure. Similarly, DCLS should be read as a diagnostic intervention: probe error, threshold calibration, and language-

specific direction alignment can all affect whether an intervention helps or harms a particular language.

Compute budget. All experiments were conducted on a single NVIDIA A100-80GB GPU with bfloat16 model weights. The LLaMA pipeline required approximately 28 wall-clock hours across all four stages, and the Qwen2.5-7B cross-model check required approximately 22 additional hours under the same hardware and precision settings. These totals include hidden-state extraction, probe fitting, SVD debiasing, threshold calibration, and final behavioral evaluation, but exclude one-off plotting and manuscript compilation.