

An Analysis of Residual-Stream Geometry Across Transformer Depth

Sunit Bhattacharya, Ravi Kolli

ProRata AI

Correspondence: sunit@prorata.ai

Abstract

We propose a transition-centred geometric analysis of transformer residual streams. Relative displacement measures how *far* representations move between consecutive layers, and orthogonal Procrustes analysis separates each transition into a rigid rotation and a non-rigid residual. Across six instruction-tuned models, on code generation and cross-lingual translation, these measurements reveal reproducible depth regularities. Relative displacement is strongly layer-dependent; typically larger early and late, with a quieter middle third; and nearly invariant across conditions within each model. Rotation magnitude is nearly constant across depth, while Procrustes residual and angle concentration remain depth-modulated, with residual peaking at the final transition. During generation, non-English targets show larger final-layer displacement and residual than English targets. We present these as descriptive geometric regularities, not as measures of computational effort or causal explanations. The contribution is a measurement framework for residual-stream transitions and evidence that, in the settings studied here, depth curves are model-dependent and largely condition-stable.

1 Introduction

Transformers (Vaswani et al., 2017) process information by repeatedly updating a residual stream (Elhage et al., 2021; Olah et al., 2020). Most analyses of that stream ask what a layer represents or predicts: representational similarity (Kornblith et al., 2019), vocabulary projections such as the logit lens and tuned lens (nostalgebraist, 2020; Belrose et al., 2023), or circuit-level accounts of specific computations (Olah et al., 2020; Elhage et al., 2021; Bushnaq et al., 2024). We propose a complementary view: treat each layer transition as a geometric transformation of the token cloud in residual space. Under this view, the basic object of study is not a static layer activation, but the map from the cloud

at depth ℓ to the cloud at depth $\ell+1$. Relative displacement asks how far tokens move. Orthogonal Procrustes analysis (Schönemann, 1966) then separates that movement into an optimal rigid rotation and a non-rigid residual. Recent geometric work on transformers has studied kinematic acceleration (Fernando and Guitchounts, 2025), curvature along token paths (Damirchi et al., 2026), and the intrinsic dimension of per-layer clouds (Viswanathan et al., 2025). These approaches either snapshot a layer or follow individual tokens. Our contribution is to measure the collective inter-layer transformation of the full residual-stream cloud.

We apply this framework to six instruction-tuned models: Qwen3 0.6B – 8B, Gemma-2-2B-IT, and StableLM-2-1.6B-Chat, on code generation and translation between English and four Indo-European languages. In these settings, four regularities emerge:

1. Relative displacement is strongly structured by depth: updates are typically larger early and late, with a quieter middle third. Within each model, coding and translation conditions share essentially the same depth curve.
2. The magnitude of the global Procrustes rotation is nearly constant across depth (variation below a predeclared 3% bound), while Procrustes residual and angle concentration remain depth-modulated, with residual peaking at the final transition.
3. Conditions mainly rescale late-layer amplitude rather than rewrite the depth curve. During generation, non-English targets show larger final-layer displacement and residual than English targets.
4. These patterns are statistically robust under hierarchical bootstrap and layer-permutation tests with Holm correction.

Our contribution is measurement-first. We introduce a transition-centred geometric view of residual streams and document regularities that are stable across the models and conditions studied here. We do not claim that these metrics measure computational effort, nor do we offer a causal account of why the patterns arise. Near-constant rotation magnitude, in particular, may partly reflect high-dimensional Procrustes geometry; the more informative observation is that displacement and residual remain depth-structured even when rotational scale is flat. The middle-layer slowdown coincides with depths previously linked to more language-neutral processing (Wendler et al., 2024; Chang et al., 2022; Bhattacharya and Bojar, 2023), and the elevated non-English residual is consistent with English-pivot accounts (Wendler et al., 2024). We report these as associations. Taken together, the results support a compact descriptive claim: in coding and translation, residual-stream depth curves are model-dependent and largely condition-stable, while content mainly rescales late-layer amplitude.

2 Related Work

2.1 Multilingual Representations

Multilingual models exhibit structured geometry across depth: early and late layers often retain language-specific structure, while middle layers form a more shared, language-agnostic subspace (Chang et al., 2022; Bhattacharya and Bojar, 2023). Large language models also frequently process non-English inputs by routing through an English-centric latent space (Wendler et al., 2024). Within that shared space, some computational features appear largely independent of both the input language and the final output language (Schut et al., 2025; Kyriakou et al., 2026). These findings locate *where* language-specific structure appears; they do not measure how residual-stream geometry changes from one layer to the next.

2.2 Residual Streams, Readout, and Depth

Mechanistic interpretability treats the residual stream as the transformer’s main communication channel, updated additively by attention and MLP blocks (Elhage et al., 2021; Olah et al., 2020). Layer-wise readout methods such as the logit lens and tuned lens ask what each depth predicts about the next token (nostalgebraist, 2020; Belrose et al., 2023). A complementary line of work debates the functional role of depth itself: late layers have been

argued to mainly refine output probabilities (Csordás et al., 2026), while other results emphasise the need for additional computational steps rather than additional parameters (Saunshi et al., 2025). Related proposals allocate extra compute through looping or longer generation (Dehghani et al., 2018; Wei et al., 2022). Our metrics address a different question: not what a layer predicts, but how representations move between consecutive residual states.

2.3 Mechanistic Interpretability and Reasoning Dynamics

Mechanistic studies have reverse-engineered specific algorithmic operations in transformers, such as Fourier multiplication circuits during grokking (Nanda et al., 2023) and the flow of operand information from early attention into late MLP computations (Stolfo et al., 2023). More recent work adopts geometric and kinematic views of reasoning. Logical reasoning has been modelled as smooth geometric flows whose structure shapes trajectory velocity and curvature (Zhou et al., 2025). Fernando and Guitchounts (2025) treat the residual stream as a dynamical system with distinct phases of kinematic acceleration, while Damirchi et al. (2026) use discrete curvature to study model veracity, arguing that scalar token kinematics alone can be confounded by lexical effects. These approaches motivate transition-centred analysis, but they primarily track individual token paths rather than global cloud geometry across depth.

2.4 Geometric Analysis of Transformer Representations

Other work studies the geometry of token representations more directly. Viswanathan et al. (2025) characterise layer-wise token clouds using intrinsic dimension and cosine similarity, relating spatial structure to prediction loss; related efforts model cross-layer propagation with dynamical systems (Shang et al., 2026). Directional consistency has also been measured across the sequence axis: Hosseini et al. (2026) use cosine similarity of velocity vectors to quantify trajectory straightening during in-context learning. Much of this literature, however, treats each layer as a static snapshot. Kinematic approaches track transitions, but usually for isolated tokens rather than for the full token cloud. We instead use orthogonal Procrustes analysis (Schönemann, 1966) to decompose each layer transition into an optimal rigid rotation and a non-

rigid residual. This global, transition-centred view is what distinguishes our framework.

2.5 Synthesis

The multilingual literature maps where language-specific structure appears; mechanistic and read-out methods characterise what layers compute or predict; geometric and kinematic studies describe representation shape or token trajectories. Fewer works ask how residual-stream geometry is organised across layer transitions. Our contribution is to measure that organisation directly: relative displacement and Procrustes residual are strongly depth-structured, while rotation magnitude is nearly constant across depth. In short, prior work explains where language structure lives and what layers do; we quantify how geometric change is allocated through the stack.

3 Methods

3.1 Setup

For each example, we record residual-stream activations for the input prompt (*prefill*) and for autoregressively generated tokens (*generation*). If a model has L layers and hidden dimension d , token t is represented by $\mathbf{h}_t^\ell \in \mathbb{R}^d$ at depth ℓ . Prefill yields one cloud of all prompt tokens after a single forward pass; generation yields the hidden state of the newly produced token at each decoding step. Our analysis is transition-centered: instead of treating each layer as a static snapshot, we measure how the token cloud is transformed from depth ℓ to $\ell + 1$.

3.2 Geometric Metrics

We view each layer transition as a geometric transformation of the residual stream. The metrics below separate how far tokens move, how sharply their trajectory bends, how much of the cloud change is explained by a global rotation, and how that rotation is distributed across planes.

Relative displacement. We measure each token’s inter-layer movement relative to the norm of its source representation:

$$\delta_\ell(t) = \frac{\|\mathbf{h}_t^{\ell+1} - \mathbf{h}_t^\ell\|_2}{\|\mathbf{h}_t^\ell\|_2}. \quad (1)$$

This controls for growth in residual-stream norms with depth, which is common in Pre-LN models: a large absolute step may still be small relative to a

growing residual vector. Larger δ_ℓ means a larger relative update at that transition. Unless noted otherwise, we report the mean of $\delta_\ell(t)$ over tokens within a prompt, then average across prompts and conditions.

Curvature. Treating depth as a trajectory, let $\mathbf{v}_\ell(t) = \mathbf{h}_t^{\ell+1} - \mathbf{h}_t^\ell$ be the update at transition ℓ . Curvature compares consecutive updates:

$$\kappa_\ell(t) = \frac{\|(\mathbf{h}_t^{\ell+2} - \mathbf{h}_t^{\ell+1}) - (\mathbf{h}_t^{\ell+1} - \mathbf{h}_t^\ell)\|_2}{\|\mathbf{h}_t^{\ell+1} - \mathbf{h}_t^\ell\|_2^2}. \quad (2)$$

Large values indicate that the update changes strongly relative to its magnitude, whether by turning direction or by changing step size. We use curvature only as a descriptive complement to displacement.

Global rotation. At each transition, the tokens at depths ℓ and $\ell + 1$ form two point clouds in $\mathbb{R}^d$. We center both clouds and use orthogonal Procrustes alignment (Schönemann, 1966) to find the rotation R_ℓ^* that best maps the source cloud onto the target cloud. We summarize the size of this rigid component by

$$\mathcal{R}_\ell = \|R_\ell^* - I\|_F. \quad (3)$$

A larger value means a larger global reorientation of the token cloud. Because R_ℓ^* is orthogonal, $\mathcal{R}_\ell$ depends only on how far the fitted rotation lies from the identity, not on residual-stream scale.

Procrustes residual. A single global rotation does not explain every token’s movement. After applying R_ℓ^* , we compute the Euclidean distance between each aligned source token and its centered target, then average across tokens within each prompt and transition. The residual is the part of the cloud change that cannot be absorbed by one shared rotation: anisotropic stretching, local rearrangements, and other non-rigid effects. We treat it as a measure of non-rigid mismatch, not as a direct measure of computational effort.

Rotation-angle concentration. A high-dimensional rotation may be spread across many planes or dominated by a few. Let $D_\ell = R_\ell^* - I$ and $q = \lfloor d/2 \rfloor$. We define

$$\mathcal{C}_\ell = \frac{\|D_\ell\|_F^2}{q\|D_\ell\|_2^2}. \quad (4)$$

This is an efficient proxy derived from $R_\ell^* - I$, not a full eigendecomposition of the rotation. Larger

$\mathcal{C}_\ell$ means rotational change is distributed across more planes; smaller values mean fewer planes dominate. Together with $\mathcal{R}_\ell$ and the Procrustes residual, it lets us ask whether constant rotation *size* co-occurs with structured rotation *geometry*.

3.3 Statistical Validation

We analyze models and phases separately using the previously computed prompt-level metrics; activations are not recomputed. For each prompt, token-level values are averaged at every transition; prompts are then averaged within condition, with conditions weighted equally. This yields one mean depth curve per model, phase, and metric.

Depth modulation. For relative displacement, Procrustes residual, and angle concentration, we score depth structure as the standard deviation of the mean curve divided by its absolute mean. A large score means some transitions are consistently larger than others. We compare the observed score with 2,000 null scores obtained by shuffling transition labels within each prompt, which preserves observed values but removes shared depth order. Holm-adjusted $p < 0.05$ indicates significant depth modulation.

Rotation constancy. For rotation magnitude, we ask whether the mean value across depth is essentially constant. We measure this with relative peak-to-trough variation:

$$V = \frac{\max_\ell \bar{\mathcal{R}}_\ell - \min_\ell \bar{\mathcal{R}}_\ell}{|\text{mean}_\ell(\bar{\mathcal{R}}_\ell)|}. \quad (5)$$

Before looking at the data, we defined rotation as practically constant when $V < 0.03$ (less than 3% variation across depth).

We estimate uncertainty with 2,000 hierarchical-bootstrap samples, resampling conditions and then prompts within conditions. A model passes the constancy test only if both of the following hold after Holm correction: (i) the one-sided 95% upper bound on V is below 0.03, and (ii) the bootstrap probability that $V \geq 0.03$ yields an adjusted p -value below 0.05. The same bootstrap procedure provides confidence intervals for the other reported effects. Holm correction is applied separately within each metric family across model-phase comparisons.

Dominant transitions and middle-layer slowdown. For relative displacement, Procrustes residual, and angle concentration, we identify the

transition with the largest mean value across depth and normalize its index to $[0, 1]$, where 0 is the first transition and 1 is the last. This puts models with different depths on a common scale. Bootstrap samples yield a 95% interval for this normalized peak depth.

For relative displacement, we additionally test whether values are lower in the middle third of depth than in the early and late thirds:

$$S = \frac{\frac{1}{2}(\mu_{\text{early}} + \mu_{\text{late}}) - \mu_{\text{middle}}}{|\mu|}, \quad (6)$$

where thirds are defined on normalized depth and μ is the mean across all transitions. The numerator compares the outer thirds with the middle; dividing by $|\mu|$ makes the score scale-free. Positive S therefore indicates a middle-layer slowdown. Coding and translation are tested separately in prefill and generation with the same bootstrap and permutation procedures. Holm correction is applied across model-task-phase comparisons. Slowdown is supported when the adjusted p -value is below 0.05 and the 95% bootstrap interval lies above zero.

4 Experiments

4.1 Models

Six instruction-tuned models across three families: Qwen3 (0.6B, 1.7B, 4B, 8B), Gemma-2-2B-IT, and StableLM-2-1.6B-Chat. All use Pre-LN architectures. This gives us variation in both family and parameter count¹.

4.2 Tasks

Code generation. Problems from LiveCodeBench (Jain et al., 2025), stratified by difficulty (easy, medium, hard). Models generate Python solutions with greedy decoding up to 512 tokens. Monolingual (English in, English out).

Translation. 100 sentence pairs per direction from Europarl v7 (Koehn, 2005): French, German, Czech, and Swedish paired with English (eight directions). Sentences under ten words filtered out. Greedy decoding up to 128 tokens. All four languages are Indo-European; we note this as a typological limitation.

4.3 Activation Capture

Forward hooks record the residual stream $H_\ell \in \mathbb{R}^{T \times d}$ after each layer’s attention and MLP. For

¹Implementation details [here](#)

input: all T prompt tokens from a single prefill pass. For generation: the hidden state of the last generated token at each autoregressive step. All runs use PyTorch MPS in float16 on Apple M-series hardware. Layer indices are normalised to $[0, 1]$ for cross-model comparison.

5 Results

5.1 Displacement

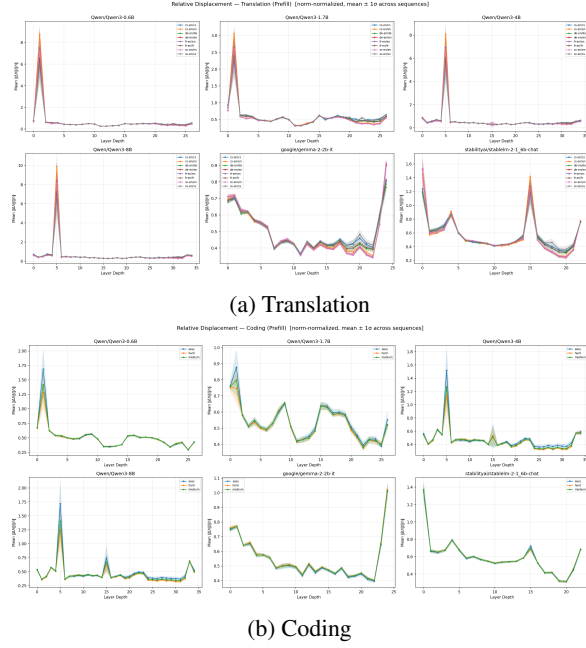


Figure 1: Relative displacement during prefill. Within each model, the depth curve is nearly identical across conditions: the eight language directions (a) overlap, and coding difficulties (b) share the same shape, with only peak height varying.

Relative displacement follows a model-specific depth curve that is largely condition-invariant. In prefill, the eight translation directions collapse onto one curve per model (Figure 1a), and coding difficulties reproduce the same shape (Figure 1b). Qwen models are dominated by a sharp early peak; Gemma decays more smoothly before a final rise; StableLM shows multiple early peaks and a middle valley. Within each model, the *shape* of change through depth is therefore more stable across conditions than across models.

This structure is statistically robust. Depth modulation of relative displacement is significant for all six models in both prefill and generation (Holm-adjusted $p < 0.05$; Figure 2). Prefill modulation is large (median 97.8%; model range 25.2%–166.4%). Depth structure persists in generation,

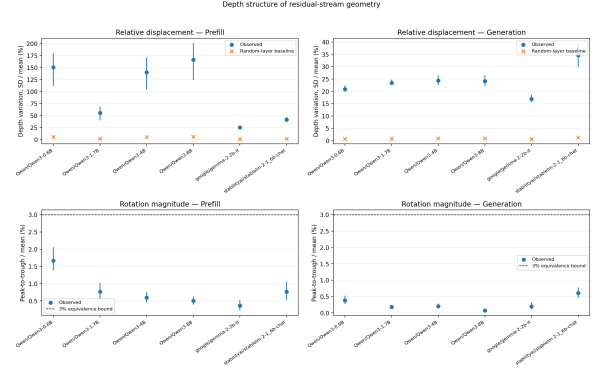


Figure 2: Depth structure of relative displacement (top) and rotation magnitude (bottom). Displacement is strongly modulated by depth; rotation magnitude stays below the predeclared 3% constancy bound in every model and phase.

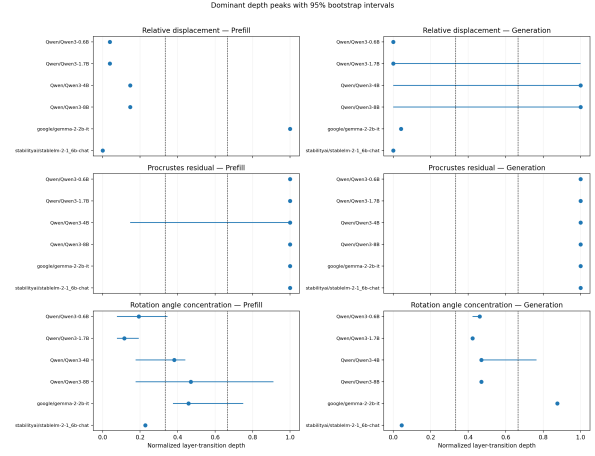


Figure 3: Normalized depth of the maximizing transition, with 95% bootstrap intervals. Displacement peaks early for most models in prefill; Procrustes residual peaks at the final transition in every model and phase.

though with smaller relative variation across layers (median 23.8%; 16.9%–34.6%).

Dominant peaks during prefill fall in the early third for five of six models (median normalized depth 0.09); Gemma is the exception, peaking at the final transition (Figure 3). During generation, peak location is less stable for several Qwen models, with bootstrap intervals that span much of the depth axis.

A second regular feature is a middle-layer slow-down (Figure 4). Relative displacement is lower in the middle depth third than in the early and late thirds for translation in all six models (prefill median 45.0%; generation 19.3%). Coding shows the same sign but a smaller effect (medians $\approx 10\%$); five of six models pass the Holm-corrected test, with Qwen3-1.7B the sole exception in both cod-

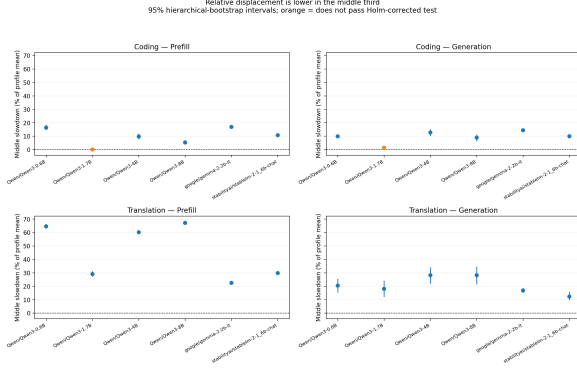


Figure 4: Middle-layer slowdown for relative displacement. Positive values indicate lower displacement in the middle third. Orange markers fail the Holm-corrected test.

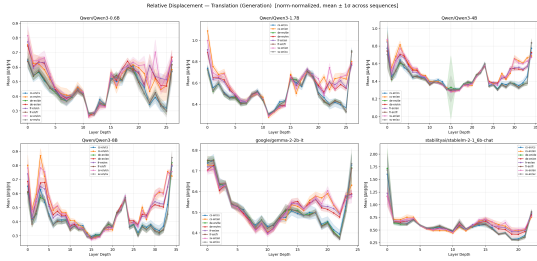


Figure 5: Relative displacement during translation generation. Early and middle layers remain nearly condition-invariant; final layers separate by target language, with larger updates for non-English targets.

ing phases. The early-middle-late rhythm is thus shared across the coding and translation conditions studied here, while its strength depends on condition and phase.

Generation makes the late-layer condition effect visible (Figure 5). Early and middle layers still share one curve across directions, but final-layer displacement splits by target language: non-English targets show larger updates than English targets. Coding generation preserves the same qualitative depth rhythm, with only modest difficulty-related magnitude differences. As shown next, the same English/non-English separation appears in Procrustes residual.

5.2 Rotation and Non-Rigid Structure

Rotation magnitude is nearly constant across depth (Figure 6; Table 1). Peak-to-trough variation is below 3% for every model and phase (prefill median 0.68%; generation 0.20%), and all twelve comparisons pass the practical-constancy test (Figure 2, bottom). Absolute magnitudes grow with hidden dimension d , consistent with the high-dimensional

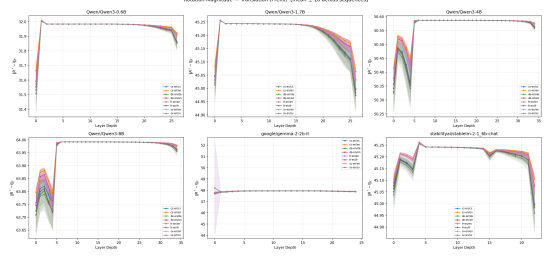


Figure 6: Rotation magnitude during translation prefill. Within each model the depth curve is nearly flat and overlaps across language directions; absolute scale grows with model size.

Table 1: Depth-structure tests (medians over models). Modulation = SD/mean; rotation = peak-to-trough variation. All comparisons pass after Holm correction (max $p_{\text{Holm}} = 0.006$).

Observable	Prefill		Generation	
	Effect	Pass	Effect	Pass
Rel. displacement	97.8%	6/6	23.8%	6/6
Rotation magnitude	0.68%	6/6	0.20%	6/6
Procrustes residual	101.5%	6/6	114.7%	6/6
Angle concentration	13.4%	6/6	9.1%	6/6

scale $\|R - I\|_F \approx \sqrt{2d}$. We therefore do not interpret flatness alone as evidence of a learned computational invariant; in high dimensions, Procrustes alignments can yield stable magnitudes for geometric reasons. The depth curve itself, however, remains nearly flat across conditions.

Near-constant magnitude does not imply unstructured geometry. Procrustes residual is strongly depth-modulated in all models (prefill median 101.5%; generation 114.7%; Figure 7). Its maximizing transition is the final layer for every model in both phases (normalized depth 1.0). During translation generation, residual curves stay low through most of the stack and then rise sharply near the output (Figure 8), again separating by target language: non-English targets leave a larger non-rigid mismatch than English targets. As in Methods, we treat residual as geometric mismatch after global alignment, not as computational effort.

Angle concentration is also depth-structured, though more weakly (Table 1). Prefill peaks fall early to mid-depth (median 0.30); generation peaks are more mid-depth (median 0.47). The raw curves show a broad mid-depth plateau with sharper changes near the beginning and end of the stack (Figure 9). Together these results separate two claims: the *size* of the global rotation is nearly depth-invariant, while the *distribution* of that ro-

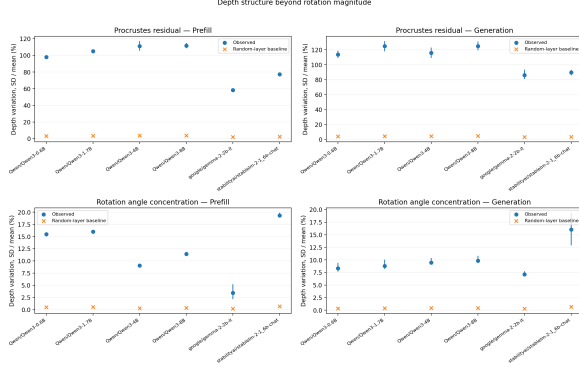


Figure 7: Depth structure beyond rotation magnitude. Procrustes residual and angle concentration are both significantly modulated by depth, even though rotation magnitude itself is nearly constant.

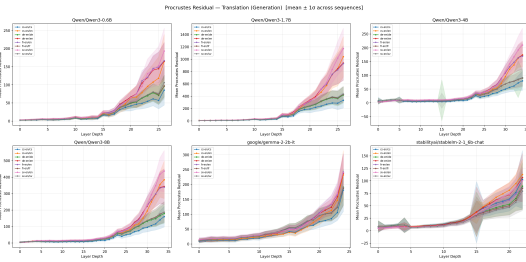


Figure 8: Procrustes residual during translation generation. Residual stays low through early and middle layers, then rises sharply near the output and separates by target language.

tation across planes and the residual mismatch after alignment both vary systematically with depth. The informative signal is this dissociation, not rotational flatness alone.

5.3 Curvature

Curvature is used only descriptively and is not included in our statistical tests. Language directions again overlap within each model (Figure 10). Prefill splits by model family: Qwen and StableLM show high-variance early spikes, whereas Gemma decays more smoothly. In generation, several models approach an early-high, late-low decay; Qwen3-4B and StableLM remain spike-dominated. Curvature is therefore a secondary complement to displacement, with more model-to-model variability.

Summary. Treating residual streams as geometric transformations reveals nearly constant rotational scale together with strongly depth-dependent displacement and residual: early repositioning, a quieter middle third, and rising late-layer non-rigid mismatch. Across the coding and translation conditions studied here, depth curves are model-

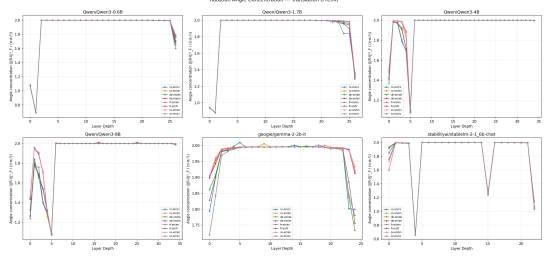


Figure 9: Rotation-angle concentration during translation prefill. Most models form a broad mid-depth plateau, with sharper changes in early and final transitions; language directions largely overlap.

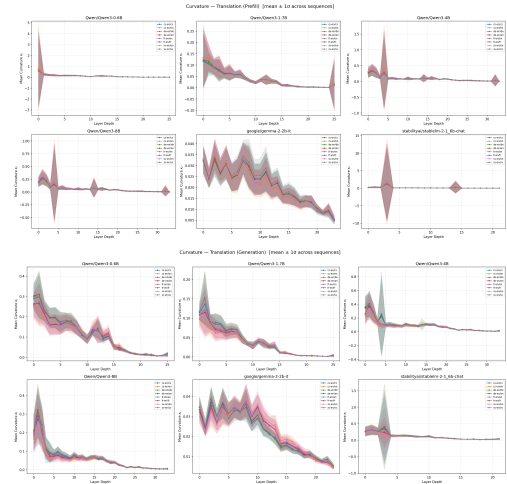


Figure 10: Curvature for translation (descriptive). Prefill is spike-dominated for Qwen/StableLM and smoother for Gemma; generation often shows an early-high, late-low decay.

dependent and largely condition-stable; conditions mainly rescale late-layer amplitude rather than rewrite the depth schedule.

6 Discussion

Scope of claims. This paper offers a measurement framework and descriptive regularities, not a mechanistic explanation. The metrics quantify the geometry of residual-stream transitions; they are not validated as measures of computational effort, reasoning difficulty, or circuit activity. Our evidence shows that, for code generation and translation, depth curves are highly consistent across conditions within each model and differ across models. We therefore describe the schedule as *model-dependent* and *condition-stable*, rather than as architecture-caused: the experiments compare prompts within fixed models and do not manipulate architecture. Likewise, near-constant rotation magnitude should be interpreted cautiously. In high

dimensions, orthogonal Procrustes alignments often yield $\|R - I\|_F$ values near the concentration scale $\sqrt{2d}$, so flatness may partly be a geometric baseline rather than a learned computational invariant. What is informative is the dissociation: even when rotational scale is nearly constant, relative displacement, Procrustes residual, and angle concentration remain strongly depth-structured.

What the measurements show. Under the transformation view, residual-stream change is not uniform across depth. Relative displacement is typically larger early and late, with a quieter middle third; Procrustes residual rises sharply near the output; and rotation magnitude stays nearly flat. Within each model, coding difficulties and translation directions largely share the same depth curve. The clearest condition effect appears late in generation: non-English targets show larger displacement and Procrustes residual than English targets. This is consistent with English-centric accounts of multilingual processing (Wendler et al., 2024; Conneau et al., 2020; Ahuja et al., 2023), but it remains an observational association restricted to the language pairs studied here. These measurements complement residual-stream and readout work. Mechanistic studies treat the residual stream as a communication channel (Elhage et al., 2021; Olah et al., 2020); logit-lens methods ask what each depth predicts (nostalgebraist, 2020; Belrose et al., 2023). Our question is different: how representations move between layers. The resulting early-middle-late pattern is compatible with late-layer refinement of token probabilities (Csordás et al., 2026) and with arguments that some problems need additional compute steps (Saunshi et al., 2025), but compatibility is not explanation.

Hypotheses for future work. Several non-exclusive hypotheses are consistent with the measurements. Early displacement peaks may reflect remapping from embeddings into a working residual space. Middle-layer slowdown may indicate gradual refinement with smaller relative updates. Late residual growth may reflect preparation for unembedding, where one global rotation is insufficient. The late English/non-English gap may reflect uneven output-space familiarity rather than a different depth schedule. Testing these hypotheses will require interventions that lie outside the present study.

Implications. Within the settings studied here, two practical points follow. First, methods that allocate extra compute through longer generation or layer looping (Dehghani et al., 2018; Wei et al., 2022) can be checked against whether they preserve or disrupt this measured depth pattern. Second, multilingual evaluation may benefit from layer-resolved geometric observables: shared depth curves can coexist with larger late-layer residual for non-English generation. Neither point assumes that residual equals computational cost.

7 Conclusion

We introduced a geometric view of residual streams: each layer transition is a transformation of the token cloud, measured by relative displacement and decomposed by orthogonal Procrustes analysis into rigid rotation and non-rigid residual. Applied to six models on code generation and translation, this view reveals reproducible depth regularities—early and late updates, a quieter middle third, nearly constant rotation magnitude, and rising late-layer residual—that are largely shared across conditions within each model. Content mainly rescales late-layer amplitude in these settings, with non-English generation showing larger final-layer displacement and residual than English generation. We present these findings as descriptive geometric measurements, not as explanations of computation. The contribution is a framework for quantifying residual-stream transitions and evidence that, for the tasks studied here, depth curves are model-dependent and largely condition-stable.

Limitations

We evaluate only code generation and translation, with Indo-European languages paired with English; broader tasks and language families remain untested.

The metrics describe geometric change and are not validated as measures of computational effort. Near-constant rotation magnitude may partly follow from high-dimensional Procrustes geometry, so our claim emphasises the dissociation with structured displacement and residual rather than flatness alone. Curvature is descriptive only.

All models are instruction-tuned Pre-LN decoders under greedy decoding, and because we do not manipulate architecture, model-dependent schedules should not be read as causal architectural effects.

References

- Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. Mega: Multilingual evaluation of generative ai. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.
- Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.
- Sunit Bhattacharya and Ondřej Bojar. 2023. [Unveiling multilinguality in transformer models: Exploring language specificity in feed-forward networks](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 120–126, Singapore. Association for Computational Linguistics.
- Lucius Bushnaq, Stefan Heimersheim, Nicholas Goldowsky-Dill, Dan Braun, Jake Mendel, Kaarel Hänni, Avery Griffin, Jörn Stöhler, Magdalena Wache, and Marius Hobbhahn. 2024. The local interaction basis: Identifying computationally-relevant and sparsely interacting features in neural networks. *arXiv preprint arXiv:2405.10928*.
- Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. [The geometry of multilingual language model representations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6022–6034.
- Róbert Csordás, Christopher D Manning, and Chris Potts. 2026. Do language models use their depth efficiently? *Advances in Neural Information Processing Systems*, 38:160313–160362.
- Hamed Damirchi, De la Jara, Ignacio Meza, Ehsan Abbasnejad, Afshar Shamsi, Zhen Zhang, and Javen Shi. 2026. Truth as a trajectory: What internal representations reveal about large language model reasoning. *arXiv preprint arXiv:2603.01326*.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.
- Jesseba Fernando and Grigori Guitchounts. 2025. Transformer dynamics: A neuroscientific approach to interpretability of large language models. *arXiv preprint arXiv:2502.12131*.
- Eghbal A Hosseini, Yuxuan Li, Yasaman Bahri, Declan Campbell, and Andrew Kyle Lampinen. 2026. Context structure reshapes the representational geometry of language models. *arXiv preprint arXiv:2601.22364*.
- Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pages 58791–58831.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMIR.
- Athina Kyriakou, Dennis Ulmer, and Ivan Titov. 2026. Shared doubt: Zero-shot cross-lingual confidence estimation for language models. *arXiv preprint arXiv:2605.31220*.
- Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.
- nostalgebraist. 2020. [interpreting GPT: the logit lens](#). *LessWrong*. <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. An overview of early vision in inceptionv1. *Distill*, 5(4):e00024–002.
- Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. 2025. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pages 14855–14881.
- Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

- Jiaqi Shang, Gabriel Kreiman, and Haim Sompolsky. 2026. Unraveling the geometry of visual relational reasoning. *Scientific Reports*.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. 2025. The geometry of tokens in internal representations of large language models. *arXiv preprint arXiv:2501.10573*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.
- Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R Zhang. 2025. The geometry of reasoning: Flowing logics in representation space. *arXiv preprint arXiv:2510.09782*.