Feature Geometry of Language Models Transfer Across Modalities to TimeSeries

Prateek Humane[1,2,*] Alexis Roger[3,2,*] Zhenghan Tai[4,*]
Gwen Legate[5,2] Andrei Mircea[1,2] Vasilii Feofanov[6] Irina Rish[1,2]

[1]Université de Montréal, [2]Mila - Quebec AI Institute, [3]McGill University,
[4]University of Toronto, [5]Concordia University, [6]42.com

Motivation and Aim

- LLMs transfer to time-series forecasting, but is this **reusable internal structure** or just *rapid relearning* under a familiar architecture?
- Time series and text share **no semantic surface** so successful transfer must be *geometric*.
- Central claim: autoregressive pretraining creates **reusable temporal feature geometry** (directions, sparse features, circuits) that finetuning **selects and specializes**.
- We give **four pieces of mechanistic evidence**: gradient alignment, effective-rank reshaping, hidden-state geometry, and shared crosscoder features + a causal Layer-1 circuit.

Forecasting *emerges* from auto-regressive pretraining, it is not built from scratch by finetuning.

1. Coherent Gradients From Step 1

- **Per-sample gradient alignment** $[1] = \frac{1}{N(N-1)} \sum_{i \neq j} \cos(\mathbf{g}_i, \mathbf{g}_j)$ over N=32 held-out.
- LangInit: **aligned from step 1**; loss descends immediately. RandInit: near-zero until a **phase transition** (∼step 64-300), where alignment and loss *co-emerge*.
- Holds in **every regime** (incl. IO-only, trunk frozen): the pretrained **residual stream itself** aligns the new objective.

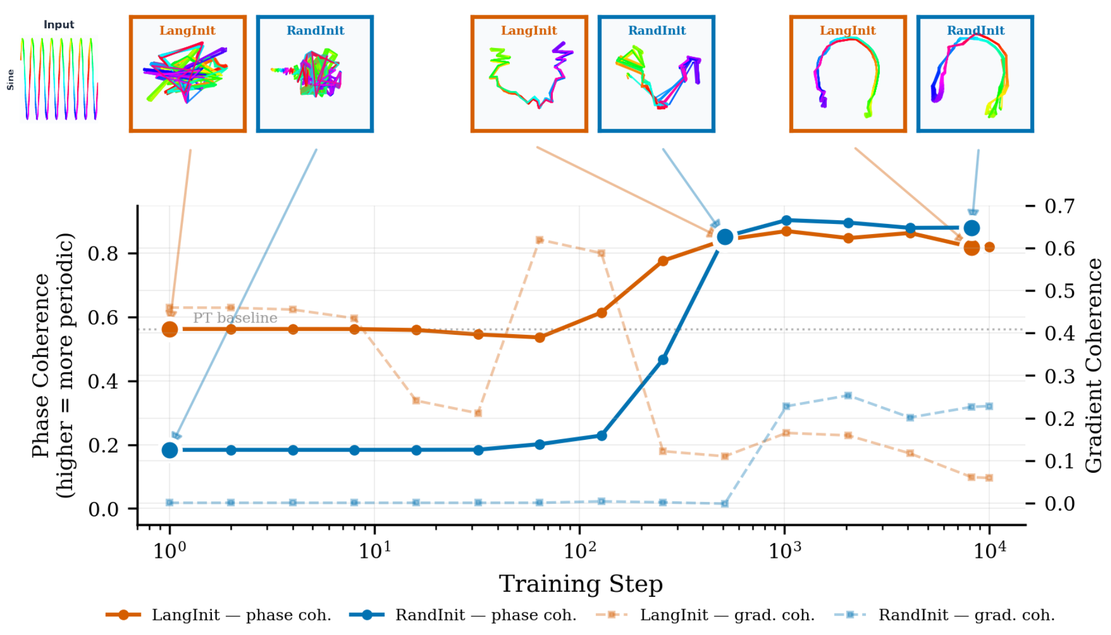

Figure 1: Phase & gradient coherence on a sine input. LangInit inherits phase-coherent states; RandInit phase-transitions around step 300.

2. Effective-Rank Reshaping

- Track erank(Σ)= $\exp(-\sum_i p_i \log p_i)$ on hidden-state covariance for every layer.
- RandInit starts **near-isotropic** (erank≈440); **collapses uniformly** to ≈10.
- LangInit **starts compressed** (erank≈50); changes *selectively* — **middle layers (8–17) expand** while periphery contracts.

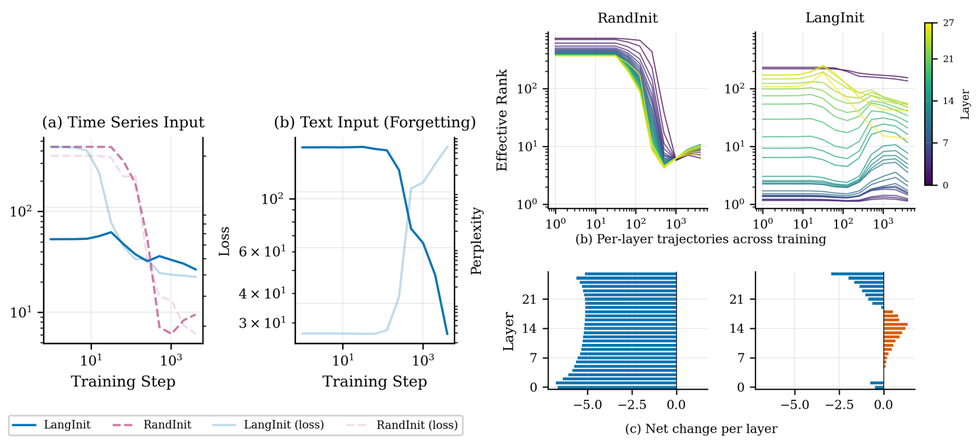

Figure 2: **Left:** mean erank over training. **Right:** per-layer $\log_2$ rank change.

Finetuning *reshapes* pretrained geometry selectively; random init must *construct* a temporal coordinate system.

Key References

[1] Andrei Mircea, Supriyo Chakraborty, Nima Chitsazan, Irina Rish, and Ekaterina Lobacheva. Training dynamics underlying language model scaling laws: Loss deceleration and zero-sum learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28154–28188, Vienna, Austria, July 2025. Association for Computational Linguistics.

Structured Forecasting Setup

- **Backbone:** Qwen3-0.6B as a **probabilistic forecaster** via next-token prediction over discretized values.
- **Tokenization:** per-series z-scale on context window C=512; V=1024 uniform bins in $[-5, 5]$ mapped to the first vocab slots; output logits used to build CDF; CDF inversion gives non-crossing forecasts in one pass.
- **Loss:** weighted quantile loss over $\mathcal{Q}$=$\{0.1, \ldots, 0.9\}$.
- **Data: GiftEval**; CRPS / MASE / MSE on held-out windows.
- **Initialisations**: LangInit where the training starts from the language pretrained model Vs RandInit where weights are randomly initialised.
- **Training regimes**: full finetune (**Full FT**); input/output layers only (**IO Only**); LoRa on attention matricies (**LoRA Attn**); LoRa and full I/O layers (**LoRA Attn+IO**).

3. Reuse vs Reinvention

- Probe hidden states on **synthetic waveforms** (sine, square, sawtooth, two-freq, trend).
- RandInit: **clean low-D arcs**, 62–92% variance in 2 PCs; a simple solution *constructed* from scratch.
- LangInit: **richer layer-specific loops**, 34–98% 2-PC variance; temporal structure realized through *inherited heterogeneous geometry*.
- **Same relational content, different coordinates:** CKA 0.86–0.99 at layer 13; t-SNE shows near-identical local neighborhoods.

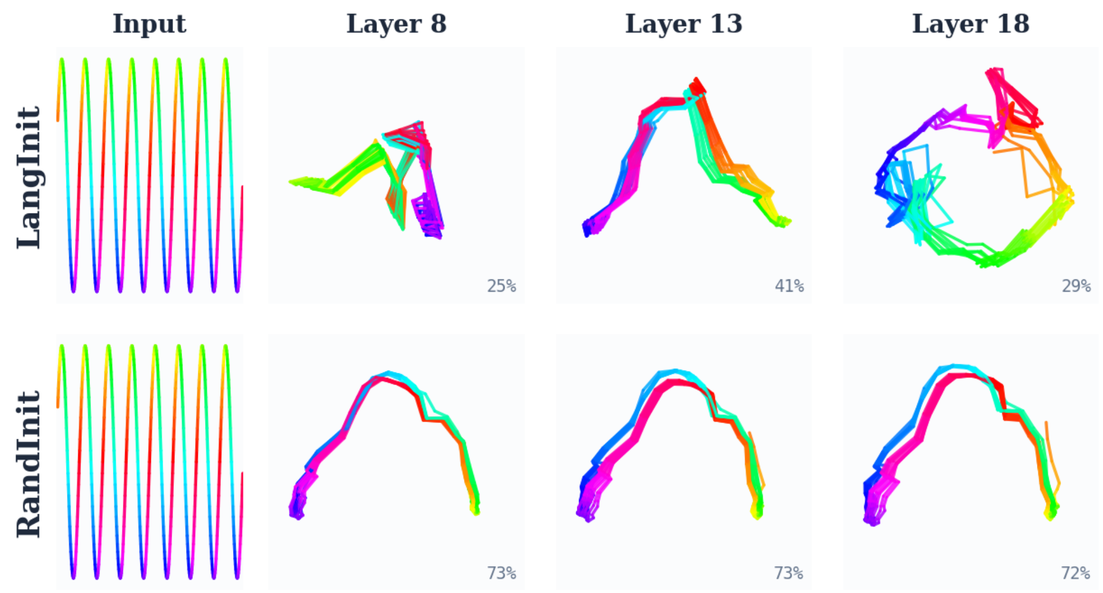

Figure 3: Sine-wave PCA across layers. RandInit (bottom) learns clean low-D arcs at every depth; LangInit (top) keeps layer-specific loops inherited from pretraining.

The pretrained model is not just *faster*, it brings an inherited geometry that finetuning *reshapes*, not rebuilds.

4. Shared Features & Causal Layer-1 Circuit

- One sparse crosscoder per layer; shared $\mathbb{R}^{1024} \to \mathbb{R}^{4096}$ encoder, Top-K=64, per-domain decoders mapping **LLM activations on text** and **LangInit activations on TS** into a common latent.
- Cross-domain features **concentrate in middle layers (7–10)**, with structural motifs realized in *both* domains:

Layer	Feat.	Time-series motif	Text motif
10	1712	magnitude jumps / plateaus	measurements & unit conversions
9	2469	volatile regime changes	tropical cyclone narratives
8	2567	missing / NaN windows	`<unk>` & incomplete refs
7	3888	isolated spikes	timestamped naval events

- **Causal ablation** over 476 components in IO-only finetuned model identifies a **super-additive Layer-1 pair**: head L1H4 ↔ MLP_{L1}.
- Ablation **selectively increases loss** on **periodic** TS ($\Delta\mathcal{L}$=13.34 vs 2.59 on controls; Cohen's d=0.61).
- *Same* circuit, on text: preferentially degrades **structurally repetitive WikiText** passages (Spearman ρ=0.161, p=4.9×10^{-13}).

A transferred *periodicity / repetition primitive* is *causally* shared between text and time series.