
Local-Geometry Signals of Capability Emergence During Portuguese Grammar Acquisition in a Small Language Model¹

Elvis Sikora
Independent

Carolina Oliveira
Mox SF

José Pedro Brito
Dovetail Research

With
Apart Research

Abstract

Capabilities can emerge abruptly during training yet leave no mark on a smoothly falling loss curve, while behavioral tests detect an ability only once it is already expressed. We test a third kind of signal, read from a model’s internal structure during training rather than from its behavior: the *local learning coefficient* (LLC), a singular-learning-theory measure of how much non-redundant internal structure a model is using, estimated by sampling rather than computed exactly. In a small, fully controlled setting we adapt TinyStories-8M to Portuguese by continued pretraining and measure both behavior and the LLC across 17 checkpoints. Modeling loss falls smoothly, but grammatical competence emerges abruptly: accuracy on a 538-item minimal-pairs benchmark moves from chance to 89.2%, with the preference margin flipping sign near 2.5M tokens. Over the same window the LLC rises steeply, from about +52 to +87, then plateaus, slightly leading behavior. Two content-matched controls—token-shuffled Portuguese and matched English—do not reproduce this rise-then-plateau, and the trajectory shape is robust to the SGLD localization. Hoogland et al. (2024) established the underlying phenomenon; our increment is the continued-pretraining setting with matched shuffled- and cross-language controls, showing the changepoint is specific to acquiring real linguistic structure. The natural follow-up is to test whether safety-like constraints learned in one language transfer to, or are eroded by, continued training in another.

1. Introduction

A model can acquire a new ability suddenly, partway through training, with no obvious sign in its training statistics. This makes such abilities hard to monitor. Two common ways of watching for them both have blind spots:

¹Research conducted at the Global South AI Safety Hackathon, June 2026.

- The **loss curve**—a running measure of how well the model predicts its training data—usually keeps falling smoothly, with no visible bump at the moment a new ability appears.
- **Behavioral evaluations**—testing the model by asking it questions—can only detect an ability after it exists, and only if someone thinks to test for it. Strictly reactive.

This work enlists a third kind of signal: one read from the model’s internal structure during training, independent of its behavior. Singular learning theory (SLT), a mathematical framework for analyzing models such as neural networks, offers a candidate—the **local learning coefficient** (LLC). The LLC estimates the local geometric degeneracy of the model around its current state: informally, how much genuine, non-redundant internal structure the model is currently using. The hypothesis is straightforward: if the model’s internal geometry reorganizes when an ability forms, the LLC should change at the same time.

The chosen setting is small and tightly controlled: teaching a model a second language by continued training. The model learning Portuguese is not itself the result but the test case; the actual contribution is the calibration between the LLC and behavior, together with control experiments that show the signal is specific.

Our main contributions are:

1. A demonstration, in a fully controlled second-language-acquisition setting, that an SLT-derived structural quantity (the LLC) exhibits a changepoint aligned with—and slightly leading—the abrupt onset of grammatical competence, while the loss curve stays smooth.
2. **Content-matched controls**: token-shuffled Portuguese and matched English, neither of which reproduces the rise-then-plateau, showing the changepoint is specific to acquiring real linguistic structure rather than a generic feature of continued training.
3. A **robustness check** showing the trajectory keeps its shape when the LLC estimator’s localization is changed ($100 \rightarrow 300$), so the changepoint is a trajectory feature rather than a localization artifact.

2. Related Work

Singular learning theory describes statistical models that are *singular*: many different internal settings can produce the same behavior, so the usual assumptions of classical statistics do not hold. Neural networks are such models [1]. The LLC measures how much of this redundancy is present near a given state and is estimated for neural networks using sampling based on Stochastic Gradient Langevin Dynamics (SGLD), a noisy gradient-based sampling procedure [2, 3].

Developmental interpretability tracks structural quantities such as the LLC through training and asks whether they shift when abilities or internal mechanisms form. The closest prior work follows the LLC across transformer training and in small theoretical settings [4, 5]. This project differs in three ways: it studies second-language acquisition by continued training, it pairs the LLC with a dense grammar benchmark, and it adds content-matched control conditions. In short, one would use this approach over a purely behavioral probe when the goal is to flag an ability *as it forms* rather than after it is expressed, and the controls supply information a behavioral test cannot—evidence that the structural signal is tied to real linguistic structure and not to generic continued training.

3. Methods

Model, training, and conditions

The starting model is **TinyStories-8M**, a small model already trained on simple English. It is adapted to Portuguese by continued training on Portuguese Wikipedia, up to 100M tokens, saving 17 checkpoints spaced across training. This model was chosen because it is small enough to measure the LLC across the whole run while still being capable of genuinely learning a language. The recipe was fixed in advance: AdamW, learning rate $1e-4$, 3% warmup, cosine decay, gradient clipping 1.0, weight decay 0.01, batch size 64, sequence length 128.

Three conditions were run with identical settings, differing only in training data (Table 1).

Table 1: The three training conditions. All share model, optimizer, and budget; only the data differs.

Condition		Training data		Purpose
Structured Portuguese	Por-	Real Wikipedia, 128-token chunks	Portuguese split into	Main condition: can the LLC curve reveal the model learning grammar?
Token-shuffled Portuguese	Por-	The same chunks, word order randomly scrambled within each chunk		Control: same Portuguese vocabulary exposure, but grammar is impossible to learn, so any LLC rise cannot be “learning grammar.”
Matched English		English Wikipedia, chunking	same	Control: what does the LLC do during ordinary continued training in a language the model already knows?

Measuring behavior and the LLC

At each checkpoint, three quantities were measured:

- **Prediction quality** in Portuguese and English, reported as bits per byte (BPB); lower is better. The English score tracks how much original English ability the model retains.
- **Grammatical competence**, using a fixed 538-item benchmark of *minimal pairs*—sentence pairs identical except for one grammatical feature, one correct and one not, covering ten kinds of grammatical agreement. The model passes an item if it assigns higher probability to the correct sentence. **Accuracy** is the fraction passed; **margin** is the average probability gap between correct and incorrect (a positive margin means the model systematically prefers correct grammar).
- **The LLC**, estimated with the `devinterp` SGLD method using one fixed sampler configuration per run: three independent chains, SGLD learning rate $1e-5$, `n_beta` = 10, `localization` = 100, batch size 64, 200 burn-in steps plus 100 draws. Each condition was scored against text from its *own* training distribution, because the LLC is only meaningful relative to a loss the model was actually trained to reduce.

4. Results

Smooth loss, abrupt grammar

The model genuinely learns Portuguese: its Portuguese BPB falls steadily from 6.69 to 2.54. But grammatical competence appears abruptly. Accuracy stays near chance through 1.5M tokens; the margin flips from negative to positive between 1.5M and 2.5M tokens; and accuracy reaches 89.2% by 27M tokens (Figure 1, Table 2). The loss curve, by contrast, gives no such signal—it just slides down.

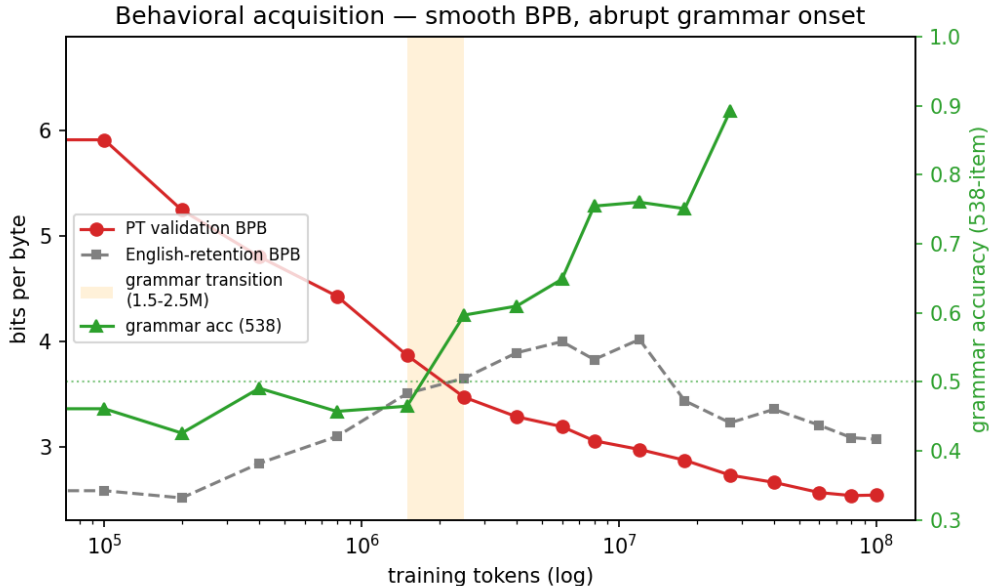


Figure 1: Behavioral acquisition. PT BPB falls smoothly, while grammar accuracy is flat near chance and then rises sharply through the 1.5M–2.5M token window.

Table 2: Behavioral trajectory for structured Portuguese, seed A. Grammar was scored to 27M; BPB was scored to 100M.

Tokens	PT BPB	EN BPB	Grammar acc.	Margin
0	6.686	2.714	0.446	−0.277
1.5M	3.869	3.506	0.465	−0.394
2.5M	3.471	3.651	0.597	+0.461
8M	3.057	3.829	0.755	+1.540
27M	2.731	3.228	0.892	+2.403
100M	2.541	3.071	—	—

The LLC rises and aligns with the transition

The LLC is positive at every checkpoint, and the three sampling chains agree closely. It rises from about +52 at 400k tokens to about +87 at its plateau, with the steepest movement between 400k and 2.5M tokens. That rise brackets the grammar transition and slightly precedes the point where

grammatical behavior becomes visible (Figure 2, Table 3).

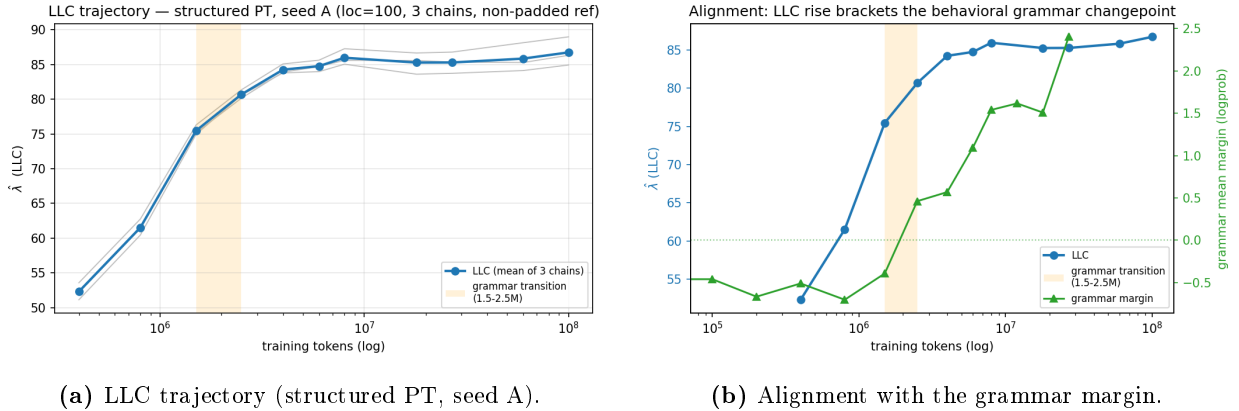


Figure 2: LLC trajectory and alignment. The LLC rises steeply, then plateaus; the rise brackets the grammar-margin sign change in the 1.5M–2.5M window.

Table 3: Primary LLC trajectory for structured Portuguese, seed A, at localization 100.

Tokens	400k	800k	1.5M	2.5M	8M	100M
LLC	+52.3	+61.5	+75.5	+80.7	+86.0	+86.7

Controls and robustness

Neither control reproduces the rise-then-plateau pattern (Figure 3):

- **Scrambled Portuguese** rises weakly to about +60, then *declines* to about +28. With no grammar available to learn, the model settles into simple word-frequency statistics and its internal complexity falls.
- **Matched English** rises smoothly from about +51 to +70—no steep transition and no high plateau, consistent with ordinary adaptation in a language the model already knows.

Changing the SGLD localization from 100 to 300 shifts the structured-Portuguese LLC trajectory about four units downward but preserves its shape: the same steep rise through the transition window and the same plateau. This makes the changepoint a trajectory feature rather than a localization artifact.

5. Discussion and Limitations

Across an identical model, optimizer, and training budget, the three conditions produced clearly different LLC trajectories. Only the condition with real grammatical structure to learn showed a steep LLC rise aligned with the moment grammatical behavior appeared, while the loss curve stayed smooth throughout. We therefore describe the result as a *changepoint in an SLT-derived local geometric estimate aligned with a behavioral transition*. We do not claim a formal SLT phase transition or grokking.

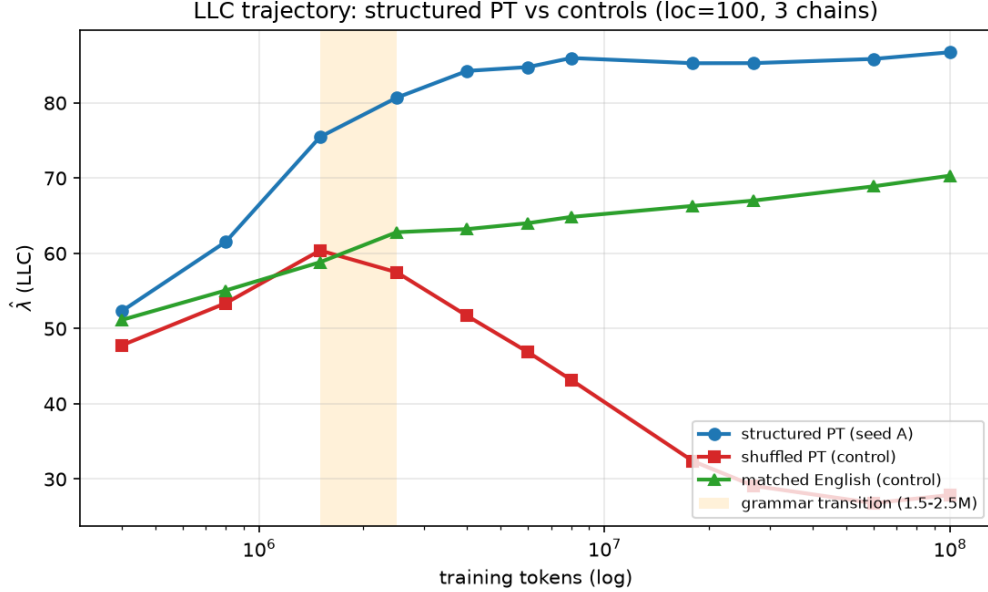


Figure 3: Three-condition contrast. Only structured Portuguese has both the steep grammar-window rise and the high plateau.

For AI safety, the implication is **developmental monitoring**: structural, training-time measurements might eventually complement behavioral tests by flagging an ability as it forms, before it is fully expressed in behavior. A practical note for research outside well-funded labs: this demonstration used an 8M-parameter model on a single small GPU, for roughly \$15 of compute.

Limitations

- The main result rests on a single Portuguese run; a second, independently-seeded run would help establish robustness.
- The model is tiny compared with frontier systems, so whether any of this transfers to high-stakes monitoring is open.
- LLC estimation is sampling-based; a deployable monitor would need cheaper, sturdier estimators.
- The grammar benchmark is templated, and the robustness sweep has only two settings; interpretation assumes both are adequate stand-ins for the competence and geometry being studied.

Dual-use. The method is primarily defensive: it measures capability formation rather than creating capability.

Future Work

A second-seed replication; formal statistical tests of the LLC–behavior alignment; and testing safety-like constraints across languages in both directions—whether a constraint learned in language A transfers to language B, and whether continued pretraining on B erodes a constraint already installed in A.

6. Conclusion

In a small, controlled language-acquisition setting, a structural quantity from singular learning theory shows a changepoint aligned with the onset of grammatical competence. Two content-matched controls establish that the signal is specific to real linguistic structure. Pending replication, this is a clean proof of concept that the LLC can align with behavior during second-language acquisition.

Code and Data

Code repository: <https://github.com/elvis-sik/slt-portuguese-acquisition>. Run artifacts and data live under `results/02_final_training/...`, `results/03_llc_campaign`, and `reports/submission/figures`: manifests, frozen configurations, sampler settings, source summaries, figures, and phase reports. The Portuguese and English corpora are derived from Wikipedia-like text pipelines described in the repository; content hashes are recorded in the manifests.

References

- [1] Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press.
- [2] Lau, E., Furman, Z., Wang, G., Murfet, D., & Wei, S. (2023). *The Local Learning Coefficient: A Singularity-Aware Complexity Measure*. arXiv:2308.12108.
- [3] Furman, Z., & Lau, E. (2024). *Estimating the Local Learning Coefficient at Scale*. arXiv:2402.03698.
- [4] Hoogland, J., Wang, G., Farrugia-Roberts, M., Carroll, L., Wei, S., & Murfet, D. (2024). *Loss Landscape Degeneracy Drives Stagewise Development in Transformers*. arXiv:2402.02364.
- [5] Chen, Z., Lau, E., Mendel, J., Wei, S., & Murfet, D. (2023). *Dynamical versus Bayesian Phase Transitions in a Toy Model of Superposition*. arXiv:2310.06301.

LLM Usage Statement

We used LLM assistance for planning, drafting code and text, and formatting. All empirical claims, figures, and reported measurements were verified against repository artifacts and run outputs before inclusion.