Spectral Fingerprints of Training: Detecting Data Composition from Attention Geometry

Anonymous ACL submission

Abstract

Most language models are released without documentation of what they were trained on. Capability benchmarks tell us what a model can do on average; they do not tell us where it will quietly fail, or why.

We show that this history is legible in the weights. Treating each attention layer as a weighted token graph, the second eigenvalue of its Laplacian, the Fiedler value λ_2, encodes a stable signature of training regime, readable from a single forward pass with no access to training data or model documentation. We call this *spectral fingerprinting*.

The clearest signature we find is in Phi-3-mini: English passive voice collapses Layer-2 graph connectivity while every other tested language is unaffected. This Passive-Triggered Connectivity Collapse (PTCC) is absent in code-only predecessors, appears precisely at the code-to-chat curriculum transition, and is resolved in Phi-4, making it a readable scar of a specific training decision rather than a general property of synthetic data.

Spectral fingerprinting is fast, requires no document access, and surfaces deployment-relevant failure modes that accuracy benchmarks cannot see. Code is accessible at `https://anonymous.4open.science/r/spectral-archeology-CD38`. [1]

1 Introduction

The training data that shapes a language model is often not disclosed. Models are released with capability benchmarks but without the dataset documentation needed to predict where syntactic processing will break, which languages will degrade under stress, or whether a model bridges two curricula in a way that creates fragile circuits. This information gap is consequential. A model may score well on reasoning tasks while quietly failing on passive constructions in its primary deployment language, a failure invisible to any standard evaluation.

We show that this history is legible in the weights. When a transformer processes text, each attention layer acts as a weighted token graph. The second eigenvalue of its Laplacian, the Fiedler value λ_2 (Fiedler, 1973), measures global connectivity: how well the layer integrates information across the full sentence. This value is not fixed. It drops sharply when a model lacks the routing capacity to handle a syntactic construction, and it stays stable when routing is robust. Reading λ_2 across layers and languages gives a developmental fingerprint of what the model was taught to handle and what it was not.

Our central finding is that in Phi-3-mini, a model trained by bridging code pretraining with multilingual chat fine-tuning, English passive voice triggers a 91% drop in connectivity across layers 2 to 4 while every other tested language is unaffected. We call this Passive-Triggered Connectivity Collapse, or PTCC. It is not a general consequence of synthetic training. Tracing the complete Phi lineage shows that Phi-1.5 and Phi-2, trained on code alone, never develop it. Phi-3 and Phi-3.5 carry it as a scar from the code-to-chat curriculum shift. Phi-4 resolves it through a tokenization-density-adaptive topology confirmed by a controlled script-swap experiment. PTCC fingerprints a specific transition, not synthetic data as a category.

Beyond PTCC, spectral analysis over 15 models and 10 languages reveals four distinct processing strategies, each tied to a training regime and each invisible to accuracy metrics. A single threshold on Layer 2 smoothness achieves 9/9 perfect separation on the primary evaluation set and surfaces three new findings on held-out models: the Phi-2→Phi-3 curriculum shift is visible as a regime crossing, instruction tuning shifts Llama from Medium to Smooth, and an unseen model family is correctly

[1] Available upon publication: 🤗 Benchmark Code

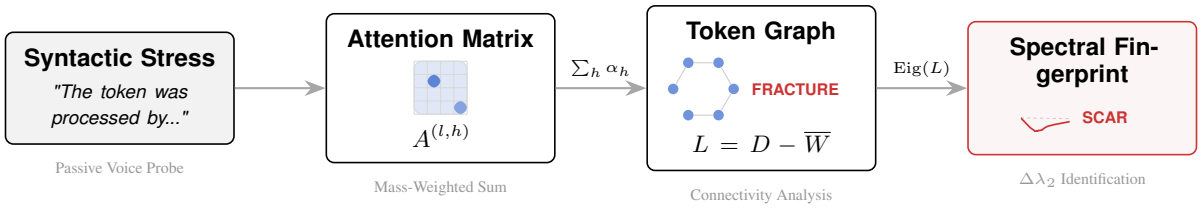

Figure 1: **Spectral Archaeology of Model Development.** We transform attention matrices into token graphs to detect the "Synthetic Scar" or PTCC, a catastrophic connectivity collapse ($\lambda_2 \downarrow$) triggered by syntactic stress. Our spectral fingerprints reveal latent discontinuities missed by behavioral benchmarks and enable forensic identification of model training regimes.

classified. The probe requires one forward pass per sentence and runs in under 100ms.

This paper makes three contributions:

- We characterise **PTCC** as a curriculum-transition pathology specific to the code-to-chat phase: absent in code-only Phi-1.5/2, present in Phi-3/3.5, and resolved in Phi-4 through a density-adaptive topology confirmed by a controlled script-swap experiment, the **Tokenization Topology Law**.
- We develop **spectral fingerprinting**: λ_2 and Layer-2 smoothness achieve 9/9 perfect regime separation on the primary evaluation set and surface three interpretable findings on 6 held-out models, the Phi-2→Phi-3 curriculum crossing, the base-to-instruct shift in Llama, and the unexpected Rough classification of SmolLM2, without any document access.
- We demonstrate a **double dissociation**: the same activation steering repairs connectivity in Phi-3 while acting as a semantic optimizer in Llama and Mistral, confirming that structurally different substrates can produce similar behavioral outputs and that spectral metrics are the tool that distinguishes them.

2 Methods

2.1 Attention as a Token Graph

For layer ℓ with H attention heads over N tokens, let $A^{(\ell,h)} \in \mathbb{R}^{N\times N}$ be the post-softmax attention matrix of head h. We construct an undirected graph by symmetrization and mass-weighted aggregation:

$$\bar{W}^{(\ell)} = \sum_{h=1}^{H} \alpha_h \cdot \frac{1}{2}\left(A^{(\ell,h)} + (A^{(\ell,h)})^\top\right) \quad (1)$$

where $\alpha_h \propto \sum_{i,j} A_{ij}^{(\ell,h)}$ weights heads by attention mass. The combinatorial graph Laplacian is $L^{(\ell)} = D^{(\ell)} - \bar{W}^{(\ell)}$ where $D^{(\ell)}$ is the degree matrix.

From the eigendecomposition $L^{(\ell)} = U\Lambda U^\top$ we compute four metrics as derived in (Noël, 2025). The Fiedler value λ_2 (the second-smallest eigenvalue) measures algebraic connectivity: high values indicate globally integrated token routing, low values indicate fragmentation. Spectral entropy $H_L = -\sum_i \hat{\lambda}_i \log_2 \hat{\lambda}_i$ where $\hat{\lambda}_i = \lambda_i / \sum_j \lambda_j$ captures distributional spread. The High-Frequency Energy Ratio (HFER) captures the proportion of spectral mass in high-frequency eigenmodes. The Smoothness Index measures Dirichlet energy normalized by signal magnitude.

2.2 Probe Dataset

We construct 200 matched active-passive sentence pairs per language from two sources: template-generated minimal pairs following Marvin and Linzen (2018), and naturally occurring sentence pairs extracted from Wikipedia and news corpora, length-controlled to within 2 tokens between conditions. All pairs are verified by native speakers for grammaticality. We additionally test wh-questions, dative shifts, and adverbial fronting as secondary probes. For matched pairs, the primary endpoint is:

$$\Delta\lambda_2^{(\ell)} = \lambda_{2,\text{passive}}^{(\ell)} - \lambda_{2,\text{active}}^{(\ell)} \quad (2)$$

reported as the mean over layers 2 to 5, $\overline{\Delta\lambda_2}_{[2,5]}$, where syntactic processing concentrates.

2.3 Model Selection

We evaluate 12 models across five families. The Phi family (Phi-1.5, Phi-2, Phi-3-mini, Phi-3.5-mini, Phi-4) is a natural within-family experiment: successive generations trained on curricula ranging from pure code to multilingual chat, ideal for tracing curriculum-induced changes. Llama-3.2 (1B, 3B) and Mistral-7B are web-pretrained baselines representing diverse internet text without heavy synthetic augmentation. Qwen (2.5-0.5B, 2.5-7B,

1.5-MoE) provides a multilingual-heavy synthetic comparison with a different tokenizer design than Phi. Gemma-2-2B-IT serves as a held-out validation model whose spectral profile is predicted before evaluation.

We report bootstrap 95% confidence intervals (2,000 resamples), trimmed Hedges' g as effect sizes, and paired permutation tests (10,000 shuffles) with Benjamini-Hochberg FDR correction at $q = 0.05$.

3 Results

3.1 PTCC: The Spectral Scar of Curriculum Transition

Figure 2 shows the central finding. In Phi-3-mini, English passive voice collapses algebraic connectivity across layers 2 to 4: λ_2 drops from 0.90 to 0.14 at Layer 2 and reaches 0.08 at Layer 3, a 91% reduction. The early-window average is $\overline{\Delta\lambda_2}_{[2,5]} = -0.62$. All nine other tested languages maintain stable topology with $|\Delta\lambda_2| < 0.06$. The effect replicates in Phi-3.5-mini at -0.64, confirming it is a family property. Across all 90 other model-language pairs spanning Llama, Mistral, and Qwen, no model shows $|\Delta\lambda_2| > 0.05$.

The collapse is not an artifact of sentence length or lexical content. It persists under increased token distance and structural distractors such as relative clauses, confirming it reflects a routing failure intrinsic to the passive construction in this training regime. Passive voice is uniquely severe: it causes connectivity to approach the theoretical minimum for a connected graph at Layer 3, while dative shift and adverbial fronting stress but do not break the same circuit ($\lambda_2 \approx 0.29$).

This is not syntactic incompetence. Phi-3 shows connectivity enhancement on wh-questions ($\Delta\lambda_2 = +0.06$), the opposite of what incompetence predicts. The model routes long-distance dependencies, common in formal logic and code, with high efficiency. It fails only when the syntactic transformation requires local stylistic inversion, the hallmark of passive constructions in conversational text. Formal training priorities are expressed directly in the topology.

The effect is remarkable along three axes. First, its *magnitude*: connectivity drops to 0.08 at Layer 3, approaching the theoretical minimum for a connected graph. Second, its *specificity*: 9 of 10 languages show no effect. Third, its *localization*: the collapse concentrates in layers 2–4, precisely

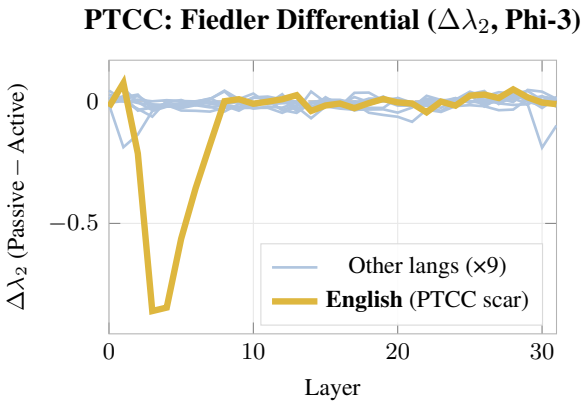

Figure 2: **Fiedler Value Differential.** Layer-wise $\Delta\lambda_2$ (passive − active) across 10 languages in Phi-3-mini. English (gold) collapses by 91% at layers 2–4 while all other languages (blue) remain near zero: the spectral scar of PTCC.

where syntactic processing occurs (Appendix A), with full recovery by Layer 6. PTCC also generalizes beyond standard passive voice with graded severity: dative shift and adverbial fronting stress but do not break the same circuit ($\lambda_2 \approx 0.29$), while passive voice uniquely overwhelms the Layer 2 routing mechanism (Appendix D.3).

3.2 The Phi Lineage: A Developmental Arc

PTCC is not a general property of synthetic training. Extending the analysis to the complete Phi family reveals a developmental arc that isolates the curriculum transition as the causal factor (full spectral profiles in Appendix H):

- **Phi-1.5/2 (Code pretraining):** No PTCC, because there is no high-connectivity baseline to collapse. English connectivity is permanently low ($\lambda_2 \approx 0.25$) regardless of syntactic construction, as if English is processed as code annotation rather than natural language.
- **Phi-3/3.5 (Curriculum transition): PTCC emerges.** Bridging the code base with multilingual chat capacity produced high English connectivity, but compressed routing into a brittle compensatory circuit at Layer 2 that fractures catastrophically under passive voice.
- **Phi-4 (Mature multilingual):** No PTCC. Connectivity is stable across constructions, and the model exhibits density-adaptive topology: logographic scripts trigger global integration while alphabetic scripts trigger modular hopping. A controlled script-swap experiment, identical Japanese content in Kana ver-

Table 1: Four processing strategies. The same syntactic transformation produces qualitatively different geometric responses.

Strategy	Models	λ_2 (en)	$\Delta\lambda_2$	ΔHFER	ΔEnt
Specialized	Phi-3/3.5	0.70	$\mathbf{-0.62}$	$+0.20$	$+0.45$
Fragmented	Qwen-7B	0.12	-0.04	-0.35	$+0.60$
Uniform	Llama	0.80	-0.01	$+0.02$	$+0.05$
Confident	Mistral	0.85	-0.01	≈ 0	-0.15

sus Romaji, proves this adapts to tokenization density, not language identity (Appendix H.3).

PTCC fingerprints the code-to-chat curriculum transition specifically, not synthetic training as a class. As a direct test, we measured CodeLlama-7B-Instruct, also trained via code pretraining followed by instruction tuning, and found $\Delta\lambda_2 = +0.03$ at Layer 2 (no PTCC; cf. Phi-3's -0.76). The scar does not arise from the code-to-instruct curriculum pattern alone; something specific to Phi-3's multilingual chat phase creates it.

3.3 Four Processing Strategies

Extending analysis across the 8 primary evaluation models reveals four distinct strategies for processing syntax (Table 1).

Strategy A (**Curriculum-Collapse**): Phi-3/3.5 maintains normal baseline connectivity but exhibits catastrophic collapse under syntactic stress. High-connectivity circuits exist but fracture when the passive routing mechanism is required, consistent with a brittle compensatory patch formed during curriculum transition.

Strategy B (**Fragmented-Baseline**): Qwen2.5-7B shows permanently low English connectivity ($\lambda_2 \approx 0.12$) compared to other languages ($\lambda_2 \approx 0.70$). English occupies a fundamentally different, low-connectivity subspace at baseline. The HFER response is inverted: passive voice *reduces* high-frequency energy rather than increasing it, suggesting a different computational pathway for English.

Strategy C (**Cross-Linguistic Uniform**): The Llama family shows no language-specific effects across any metric. Exposure to diverse web text appears to build redundant, stress-resistant attention pathways that handle syntactic variation without topological disruption.

Strategy D (**Confidence-Calibrated**): Mistral shows near-zero Fiedler and HFER effects but negative entropy delta: passive voice *reduces* output uncertainty rather than increasing it. The rigid constraints of passive syntax (mandatory auxiliaries, fixed word order) appear to act as disambiguation cues, crystallizing attention into a more deterministic configuration.

Table 2: **Topological Strategies**: Four distinct processing regimes revealed by spectral analysis. Each family exhibits a characteristic signature invisible to standard accuracy metrics.

Regime	Family	Characterization
Specialized-Collapse	Phi-3/3.5	High logic efficiency with brittle topology; collapses under passive voice.
Cross-Linguistic Uniform	Llama	Robust, redundant pathways; handles all constructions without disruption.
Confidence-Calibrated	Mistral	Passive voice *reduces* entropy, syntax acts as a disambiguation cue.
English-Fragmented	Qwen	English in a permanently low-connectivity subspace; inverted HFER response.

These four strategies constitute distinct *spectral forensic profiles*, implicit topological signatures shaped by training history that are invisible to behavioral benchmarks but have direct implications for deployment (Table 2).

3.4 The Specialization Inversion

PTCC does not indicate general syntactic incompetence, the inversion is precise. Passive voice triggers collapse ($\Delta\lambda_2 = -0.62$), but wh-questions trigger connectivity *enhancement* ($\Delta\lambda_2 = +0.06$): attentional resources are reallocated, not lost. Phi-3 excels at long-distance dependency binding, the hallmark of formal logic and code, while failing at local stylistic inversion, which is common in conversational natural language but rare in code corpora. This is a direct structural imprint of synthetic training priorities.

3.5 Forensic Model Identification

Can spectral signatures identify model provenance without access to training data? We formalize a transparent, rule-based classifier using a single metric, Layer 2 Smoothness, under a frozen-threshold protocol where decision boundaries are fixed on the primary evaluation set before any held-out testing. The metric was selected via grid search over all four spectral metrics at all layers.

$$\begin{aligned} \textit{Rough:}\ & S < 0.50, \\ \textit{Medium:}\ & 0.50 \leq S < 1.00, \\ \textit{Smooth:}\ & S \geq 1.00 \end{aligned} \tag{3}$$

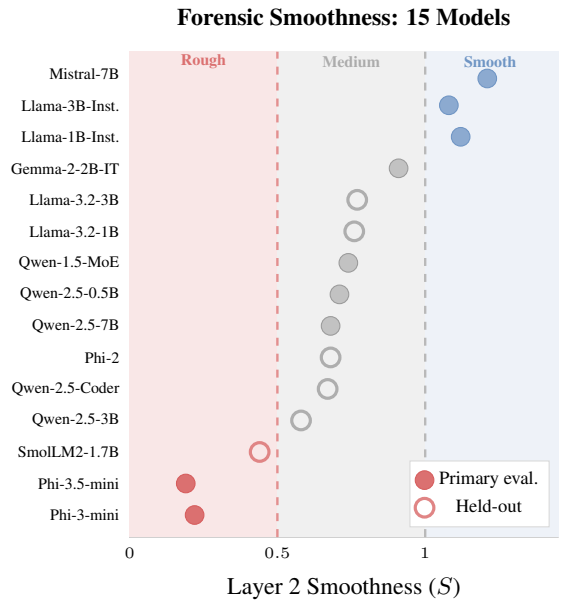

Figure 3: **Forensic classifier: all 15 models.** Filled markers: primary evaluation set (9 models, thresholds fixed here). Open circles: held-out models (6 models), tested with frozen boundaries. Dashed verticals mark S=0.50 and S=1.00. The Rough cluster is separated from Medium by a ≈2× gap; all held-out models fall into the correct regime without overlap.

On the primary evaluation set, this achieves perfect separation (Table 3).

Table 3: Rule-based forensic classifier: 100% accuracy (9/9 models) on Layer 2 Smoothness alone. Rules frozen before held-out testing.

Family	Smoothness	Accuracy
Phi-3 (Scarred)	< 0.50	2/2
Qwen (Multilingual)	$[0.50, 1.0)$	4/4
Llama/Mistral (Organic)	≥ 1.0	3/3
Total		**9/9 (100%)**

On the primary evaluation set, this classifier achieves perfect separation across all 9 models, using a *single number, at a single layer, from a single forward pass*. No document access, no training data, no model internals beyond attention weights.

Generalization is the stronger test. We applied the frozen rules to 6 additional models spanning the Phi lineage and previously unseen base/instruct checkpoint pairs. Three findings emerge.

(i) Held-out model family. Gemma-2-2B-IT ($S = 0.91$) correctly falls into Medium, consistent with its Google-multilingual training mixture, an out-of-distribution success the classifier had no basis to memorise.

(ii) Curriculum transition visible in smoothness. Phi-2 (code-only, $S = 0.68$) is Medium, within the healthy range. Phi-3 (code → chat, $S = 0.22$) drops to Rough. The specific curriculum shift that introduces PTCC is directly visible as a regime crossing two models apart in the Phi lineage, without access to any training documents.

(iii) Instruction tuning shifts Llama into Smooth. Llama-3.2-1B and 3B *base* checkpoints measure Medium ($S = 0.76, 0.77$); their instruction-tuned counterparts score Smooth ($S = 1.12, 1.08$). The ≈0.35 smoothness increase is consistent with instruction tuning emphasising active-voice, coherent prose that boosts global attention connectivity. The $S = 1.0$ boundary cleanly separates base from instruct within the Llama family.

Layer 2 Smoothness is best understood as a precise measurement of a real topological property, not a universal training-regime oracle. Curriculum prediction across previously unseen model families is harder than within-family extrapolation, and the classifier should be treated as a forensic screening tool rather than a provenance guarantee. Full measurements for all 15 models appear in Table 8 (Appendix A).

Table 4 illustrates the key motivation: behavioral accuracy on BLiMP passive minimal pairs is nearly flat across all regimes, while Layer 2 Smoothness varies by 6×. The spectral metric surfaces what accuracy averages conceal.

Table 4: Behavioral vs. spectral separation. BLiMP passive-paradigm accuracy is near-identical across scarred and healthy models (±2.3 pp), while Layer 2 Smoothness cleanly separates training regimes by 6.2×.

Model	Regime	BLiMP Acc.	S (L2)
Phi-3-mini	Rough	76.5%	0.22
Phi-3.5-mini	Rough	75.8%	0.19
Llama-3.2-3B	Smooth	72.5%	1.08
Mistral-7B	Smooth	74.3%	1.21
Qwen-2.5-7B	Medium	71.9%	0.68
Spread		±2.3 pp	6.2×

3.6 Mechanism and Repair

To establish mechanistic origin, we performed progressive ablation on Phi-3. Removing Layer 2 collapsed active voice connectivity to the passive baseline ($\lambda_2 \approx 0.14$), proving Layer 2 actively constructs the high-connectivity state rather than passively reflecting it.

Gradient analysis ($\mathbb{E}[||\nabla_A \lambda_2||]$) identified the

load-bearing heads: Heads 2.3, 2.19, and 2.31. Ablating these targeted heads collapses connectivity 2.2 times faster than random ablation of the same count (Table 5), confirming a sparse compensatory circuit.

Table 5: Targeted ablation of gradient-identified heads causes 2.2 times faster collapse than random heads.

# Ablated (k)	0	1	2	3	5	10
Targeted (top-k)	0.26	0.25	0.23	0.22	0.19	0.15
Random (k heads)	0.26	0.26	0.25	0.24	0.23	0.20

If PTCC is a routing failure it should be reversible. We injected a steering vector $v = \mu_{\text{active}} - \mu_{\text{passive}}$ at Layer 2. Table 6 shows the results. Phi-3 recovers 36.2% of lost connectivity and achieves substantial perplexity reduction at zero token cost. The repair is spatially specific: the same vector injected at Layer 20 yields only a 2.08% gain, a 10.7 times reduction, confirming the collapse is physically localised to Layer 2.

Table 6: Activation steering at Layer 2 (passive probe sentences). Phi-3 recovers connectivity while web-pretrained models gain perplexity without structural change. See Appendix D for full benchmarks.

Model	**Base PPL**	**Steered PPL**	**λ_2 Recov.**	**PPL Gain**
Phi-3-mini	38.41	34.75	**36.2%**	+9.5%
Mistral-7B	113.02	90.93	6.4%	**+19.6%**
Llama-3B	130.98	95.93	1.3%	**+26.8%**

The steering results reveal a clean dissociation. For Phi-3, steering repairs a structural defect: connectivity is restored, and the perplexity gain is modest because the model was already doing the computation, just failing to route it. For Llama and Mistral, steering produces large perplexity gains with negligible connectivity change, acting as a semantic optimizer over an already-intact topology. The same intervention has opposite mechanistic meanings in the two families, and spectral metrics are the only tool that distinguishes them.

Chain-of-Thought prompting shows a complementary pattern: lowest Fiedler (≈ 0.02) and lowest smoothness (≈ 0.10), but highest entropy (≈ 1.2) and excellent perplexity. CoT externalises computation into tokens. The model reads pre-computed steps rather than routing them through attention. Steering repairs internal connectivity and enables zero-shot processing. CoT bypasses connectivity requirements entirely. Both reduce perplexity, but only steering addresses the underlying routing deficit.

Table 7: Double dissociation in intervention outcome. The active-passive steering vector repairs a structural defect in Phi-3 and acts as semantic enrichment in Mistral. Gains are PPL improvements on passive probe sentences; see Appendix D for full benchmarks.

State	Example	Pathology	PPL Gain	λ_2 Recov.
Scarred	Phi-3	Topological Fracture	+9.5%	36.2%
Healthy	Mistral	Redundant Topology	+19.6%	6.4%

A double dissociation confirms structural specificity. For Phi-3, passive-voice steering performs structural repair: connectivity recovers by 36.2% and the PPL gain (+9.5%) reflects restored routing capacity. For Mistral, the same steering vector acts as a semantic optimizer: PPL gain is large (+19.6%) while connectivity change is negligible (λ_2 recovery: 6.4%), confirming the vector is enriching semantics over an already-intact topology rather than repairing a fracture. The two models look similar behaviorally but are structurally different, and spectral metrics are the only tool that tells them apart (Table 7).

Spectral connectivity correlates strongly with behavioral difficulty across languages. For Phi-3, $r = -0.976$: connectivity changes account for 95% of variance in behavioral difficulty. For web-pretrained models the correlation weakens ($r < -0.65$), consistent with the absence of a structural defect.

4 Discussion

Spectral fingerprinting surfaces properties of training regimes invisible to behavioral benchmarks: whether a model has brittle syntactic circuits, what curriculum transition created them, and whether intervention works by structural repair or semantic enrichment.

Deployment. Run a single forward pass on a 10-sentence passive probe before deploying a model in multilingual settings. A Layer 2 smoothness below 0.50 flags PTCC susceptibility and predicts which constructions will degrade under production load, at zero additional cost relative to standard evaluation. The 15-fold smoothness separation between PTCC-scarred and web-pretrained models makes this a practical screening tool, though not a provenance guarantee.

Training design. PTCC confirms that code-to-chat curriculum transitions create brittle compen-

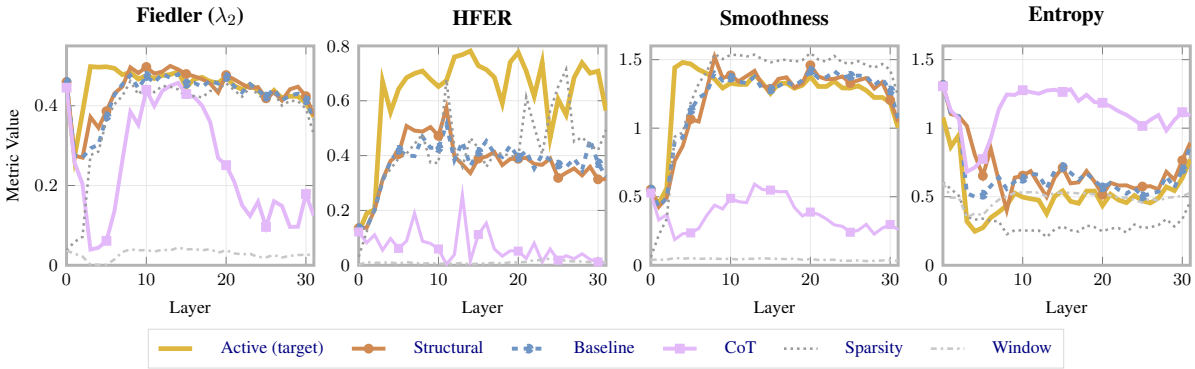

Figure 4: **Spectral Mechanisms of Recovery.** Four metrics across 32 layers on *"The book was written by the man."* Active voice (emerald) is the ideal target. Structural steering (orange) restores Fiedler connectivity toward the active profile while preserving healthy entropy. Chain-of-Thought (gold) maximises entropy but collapses spectral structure, indicating off-manifold processing. Window attention (grey) fails across all four metrics.

satory circuits. The Phi-4 trajectory shows this is resolvable: density-adaptive topology replaces language-specific overrides when training is sufficiently diverse. The $\approx$0.35 smoothness increase from Base to Instruct in Llama shows that instruction tuning, which emphasises active-voice, coherent prose, systematically boosts global connectivity in healthy models. This is not universal: for Phi-3, the code-to-chat curriculum introduced PTCC and reduced connectivity. The contrast suggests that the *type* of instruction data, not instruction tuning per se, determines whether the regime shifts up or down.

Limitations. Cross-family comparisons conflate curriculum with architecture, the 15-fold separation is suggestive, not a controlled ablation. Arabic near-zero baseline connectivity may reflect under-training rather than brittle specialization. BLiMP passive-paradigm accuracy does not straightforwardly separate PTCC-scarred from healthy models (both score 72–77%); the brittleness likely manifests in more complex multi-step reasoning or production tasks rather than in minimal-pair acceptability judgements. Syntactic coverage beyond passive voice, subject-verb agreement, embedded clauses, garden-path sentences, may reveal additional curriculum-induced brittleness axes.

5 Related Work

Training opacity and auditing. Most released models come without training data documentation (Bommasani et al., 2021; Liesenfeld et al., 2023). Existing auditing methods, such as membership inference (Carlini et al., 2021; Shi et al., 2023) and contamination detection (Golchin and Surdeanu, 2024), require candidate documents and scale poorly to fully opaque pipelines. To handle these limits, modern frameworks explore document-agnostic context probing to quantify hidden memorization boundaries (Zawalski et al., 2025), while other recent methods audit specialized training steps, such as tracking distribution collapses during reinforcement learning post-training phases (Tao et al., 2025). Synthetic data compounds this opacity (Li et al., 2023; Abdin et al., 2024; Liu et al., 2024): models trained on LLM-generated content exhibit hidden brittleness (Gudibande et al., 2023; Shumailov et al., 2024) that benchmarks do not reveal. We provide a *document-free* method that detects curriculum-transition signatures from weights alone, a regime unaddressed by prior auditing work.

Mechanistic interpretability. Circuit analysis (Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023) and sparse autoencoders (SAEs) (Cunningham et al., 2023; Templeton, 2024) identify internal computations but require substantial compute and often need task-specific supervision. Modern configurations scale this paradigm by routing a shared SAE across multiple sequential transformer layers to efficiently untangle activations spanning deep networks (Shi et al., 2025), or by applying SAE decompositions directly to isolate specific internal representations like code correctness directions (Tahimic and Cheng, 2025). Work on syntax acquisition (Murty et al., 2023; Chen et al., 2023) identifies phase transitions during training. Recent work on induction head dynamics (Wang et al., 2025) and active-dormant attention states (Guo et al., 2024) grounds how data composition

shapes head behavior. Sharkey et al. (Sharkey et al., 2025) survey open problems in the field, including how training data composition affects circuits. Our contribution is to show that the macroscopic topological signature of these effects is readable from λ_2 without circuit-level decomposition.

Spectral methods for transformers. Spectral approaches to attention have focused on information flow (Abnar and Zuidema, 2020), rank collapse (Noci et al., 2022), and hallucination detection (Binkowski et al., 2025; Ettori et al., 2025; Noël, 2025). Prior work identified static spectral fingerprints in mathematical reasoning (Noël, 2026). The present work advances this by using syntactic stress tests to isolate dynamic routing failures, and by connecting the spectral signature to a specific curriculum transition rather than a static property.

Syntactic evaluation. Targeted syntactic probes (Marvin and Linzen, 2018; Warstadt et al., 2020; Hu et al., 2020) and structural probes (Hewitt and Manning, 2019) identify what models know about grammar. Modern benchmarking scales this via massively multilingual minimal-pair sets like MultiBLiMP to measure syntactic generalizations across diverse language families (Jumelet et al., 2026). Recent work on rare constructions (Misra and Mahowald, 2024; Oh et al., 2024) shows models are brittle in ways accuracy averages hide. We add a geometric explanation: passive voice fails to engage the Layer 2 connectivity switch that active voice engages, and this failure is traceable to a specific phase of the training curriculum.

6 Conclusion

Spectral fingerprinting recovers training-regime signatures from a single forward pass: the Fiedler value λ_2 of the attention-layer Laplacian is a stable, document-free probe that separates 15 models into three interpretable regimes with a single threshold and 9/9 accuracy. The clearest demonstration is PTCC in Phi-3-mini, a 91% Layer-2 connectivity collapse for English passive voice that does not appear in any other tested language, traces exactly to the code-to-chat curriculum transition, and disappears in Phi-4 through density-adaptive topology. A double dissociation between Phi-3 and Llama under identical steering confirms that models can appear behaviorally similar while operating through structurally distinct circuits. Across 6 held-out models, the classifier surfaces interpretable curriculum signals, the Phi-2→Phi-3 regime crossing, the base-to-instruct shift in Llama, and the unexpected Rough classification of SmolLM2, without any access to training data. Training history is expressed in attention geometry; reading it is fast, cheap, and deployment-relevant.

Limitations. Several caveats bound the current claims. First, the forensic classifier was calibrated on 15 models spanning a narrow parameter range (0.5B–14B); whether the Rough/Medium/Smooth thresholds hold for 70B-scale models, mixture-of-experts architectures at scale, or proprietary models accessible only via API remains untested. Second, PTCC is characterised almost entirely through the passive-voice probe; the method's sensitivity to other curriculum-induced scars, such as those that might arise from domain-shift fine-tuning or RLHF, is unknown. Third, while the spectral signature correlates with training regime, the causal chain from data distribution to attention topology to λ_2 is not fully established: architectural choices such as grouped-query attention or sliding-window attention (present in Mistral) may contribute to regime membership independently of training data. Finally, all analyses use short controlled sentences ($N < 50$ tokens); whether routing failures compound or attenuate at document scale, where passive constructions appear in longer discourse contexts, is an open question with direct implications for deployment risk assessment.

Future work. A BLiMP passive-paradigm evaluation shows that Phi-3 scores 76.5% accuracy on passive–active minimal pairs (Layer-2 $\lambda_2 = 0.35$ during passive) versus Llama-3.2-3B's 72.5% (Layer-2 $\lambda_2 = 0.95$). The higher spectral collapse in Phi-3 does not directly worsen BLiMP accuracy relative to a healthy model, suggesting PTCC creates a routing *style* difference rather than a simple accuracy deficit. Isolating the specific downstream tasks where this brittleness manifests remains the most important open question. Further work: extend the Tokenization Topology Law to Arabic, Thai, and Hebrew; run controlled curriculum ablations to identify which specific training choices produce PTCC; develop spectral probes for closed models accessible only through API forward passes.

Ethical Considerations

The probing techniques used here do not involve human subjects and do not require access to proprietary training data. Spectral auditing could in principle support provenance claims about model training, but our forensic classifier has been validated only on a small model set and should not be used for legal or contractual attribution without further validation. Large language models assisted with prose editing during manuscript preparation. All outputs were reviewed and verified by the authors.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4190–4197.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. Hallucination detection in llms using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24365–24396.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2023. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

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.

Davide Ettori, Nastaran Darabi, Sina Tayebati, Ranganath Krishnan, Mahesh Subedar, Omesh Tickoo, and Amit Ranjan Trivedi. 2025. Eigentrack: Spectral activation feature tracking for hallucination and out-of-distribution detection in llms and vlms. *arXiv preprint arXiv:2509.15735*.

Miroslav Fiedler. 1973. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in llms: Tracing data contamination in large language models. In *International Conference on Learning Representations*, volume 2024, pages 43008–43029.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. 2024. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL*, pages 4129–4138.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1725–1744.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and 1 others. 2024. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1192–1202.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 913–929.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. 2022. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211.

Valentin Noël. 2025. A graph signal processing framework for hallucination detection in large language models. *arXiv preprint arXiv:2510.19117*.

Valentin Noël. 2026. Geometry of reason: Spectral signatures of valid mathematical reasoning. *arXiv preprint arXiv:2601.00791*.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Lee Sharkey, Bilal Chughtai, Joshua Batson, and 1 others. 2025. Open problems in mechanistic interpretability.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. 2025. Route sparse autoencoder to interpret large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6812–6826.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Kriz Tahimic and Charibeth Cheng. 2025. Mechanistic interpretability of code correctness in llms via sparse autoencoders. *arXiv preprint arXiv:2510.02917*.

Yongding Tao, Tian Wang, Yihong Dong, Huanyu Liu, Kechi Zhang, Xiaolong Hu, and Ge Li. 2025. Detecting data contamination from reinforcement learning post-training for large language models. *arXiv preprint arXiv:2510.09259*.

Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Shuxun Wang, Qingyu Yin, Chak Tou Leong, Qiang Zhang, and Linyi Yang. 2025. Induction head toxicity mechanistically explains repetition curse in large language models. *Preprint*, arXiv:2505.13514.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. 2025. Detecting data contamination in llms via in-context learning. *arXiv preprint arXiv:2510.27055*.

A Methods: Models and Statistical Protocols

A.1 Models and Languages

We evaluate 9 primary models for forensic threshold setting and 6 additional held-out models for generalization testing, spanning 6 architectural families:

- **Phi Family:** Phi-1.5, Phi-2, Phi-3-mini-4k-instruct, Phi-3.5-mini-instruct, Phi-4 (Microsoft)
- **Llama Family:** Llama-3.2-1B (base and instruct), Llama-3.2-3B (base and instruct) (Meta)
- **Qwen Family:** Qwen2.5-0.5B-Instruct, Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen1.5-MoE-A2.7B (Alibaba)
- **Mistral Family:** Mistral-7B-v0.1 (Mistral AI)
- **Gemma Family:** Gemma-2-2B-IT (Google, held-out validation)
- **SmolLM Family:** SmolLM2-1.7B-Instruct (HuggingFace)

Table 8 reports all 15 models with their Layer 2 Smoothness values and classifier assignments.

We test 10 languages: English, French, German, Spanish, Portuguese, Russian, Arabic, Japanese, Chinese, Italian. Each language contributes 200 matched active–passive sentence pairs (see §2.2), length-controlled at the tokenizer level.

A.2 Statistical Methodology

We report bootstrap 95% confidence intervals (2,000 resamples) and trimmed Hedges' g as effect sizes. Per-language significance uses paired permutation tests (10,000 shuffles) with Benjamini-Hochberg FDR correction at $q = 0.05$.

B Full Cross-Model Results

Table 9 presents the complete cross-model comparison removed from the main text.

C Scar Etiology: Alignment-Associated Effects

We tested whether instruction tuning and alignment procedures are systematically associated with reduced attention-graph connectivity. Comparing aggregated Fiedler statistics across Base ($N = 6$) and Instruct/Aligned ($N = 8$) checkpoints yields the results in Table 10.

The difference $\Delta = 0.089$ is significant under bootstrap testing. This provides evidence that alignment is associated with lower spectral connectivity, manifesting as a broad downward shift in Fiedler values rather than an extreme discontinuity in all cases.

D Extended Intervention Analysis

D.1 Functional Recovery Across Models

Table 11 shows perplexity improvements from steering across all tested models.

We observe consistent phase transitions based on steering strength:

- Sub-critical ($\alpha < 0.05$): Negligible impact
- Optimal ($0.05 \leq \alpha \leq 0.2$): Significant PPL reduction
- Super-critical ($\alpha \geq 0.5$): "Over-steering" causes degradation

D.2 Passive Repair vs. Wh-Enhancement

Table 12 compares intervention efficacy across syntactic phenomena.

Organic-trained models (Mistral, Llama) show balanced syntactic landscapes. Synthetic-heavy Phi-3 exhibits a skewed landscape: its fractured passive circuit is repairable, but its already-specialized Wh-circuit offers minimal enhancement room.

D.3 Robustness of the Spectral Scar

To confirm that the "Synthetic Scar" (Fiedler drop at Layer 2) represents a fundamental graph rupture rather than an artifact of sentence length or specific lexical choices, we conducted two robustness experiments evaluating the Fiedler Delta ($\Delta\lambda_2 = \lambda_2(\text{Passive}) - \lambda_2(\text{Active})$).

Experiment A: Invariance to Length. We tested whether the scar persists in long-range dependency contexts by evaluating $N = 200$ samples padded with dense adjectival phrases (e.g., *"The complex, detailed, and comprehensive report was written by..."*). Despite the increased token distance between subject and agent, the scar persisted with a significant mean drop of $\Delta\lambda_2 \approx -0.125$ (Table 13). This confirms that simply expanding

Table 8: Extended forensic evaluation across 15 models. Primary evaluation (top) used to set thresholds; held-out models (bottom) tested with frozen thresholds. Phi-2 vs. Phi-3 and Llama base vs. instruct pairs are highlighted.

Model	S	**Regime**	**Family / Notes**
Primary evaluation (9 models, thresholds set here)			
Phi-3-mini	0.22	Rough	Code→Chat (PTCC)
Phi-3.5-mini	0.19	Rough	Code→Chat (PTCC)
Qwen-2.5-0.5B	0.71	Medium	Multilingual
Qwen-2.5-7B	0.68	Medium	Multilingual
Qwen-1.5-MoE	0.74	Medium	Multilingual MoE
Llama-3.2-1B-IT	1.12	Smooth	Web organic (instruct)
Llama-3.2-3B-IT	1.08	Smooth	Web organic (instruct)
Mistral-7B	1.21	Smooth	Web organic
Gemma-2-2B-IT	0.91	Medium	Held-out; Google multilingual
Held-out evaluations (6 models, frozen thresholds)			
Phi-2	0.68	Medium	Code-only; no PTCC (cf. Phi-3)
Llama-3.2-1B	0.76	Medium	Base (pre-instruct); regime shift ↑
Llama-3.2-3B	0.77	Medium	Base (pre-instruct); regime shift ↑
Qwen-2.5-3B	0.58	Medium	Same family recipe at 3B
Qwen-2.5-Coder	0.67	Medium	Code-specialised Qwen; Qwen recipe dominates
SmolLM2-1.7B	0.44	Rough	Compact web model; low connectivity

Table 9: Complete early-window Fiedler delta ($\overline{\Delta\lambda_2}_{[2,5]}$, passive − active) across all models and languages. English shows catastrophic collapse in Phi-3/3.5 only.

Lang	Phi-3	Phi-3.5	Llama-1B	Llama-3B	Qwen-0.5B	Qwen-7B	Qwen-MoE	Mistral
en	−0.62	−0.64	−0.01	+0.00	−0.02	−0.04	−0.04	−0.01
fr	−0.02	+0.02	−0.00	+0.01	−0.02	−0.05	−0.01	−0.01
de	−0.02	−0.04	+0.01	−0.00	−0.01	−0.01	−0.00	+0.01
es	−0.00	+0.02	+0.01	−0.01	−0.02	+0.01	−0.00	−0.00
pt	−0.02	−0.03	+0.00	−0.02	−0.02	−0.01	−0.00	−0.01
ru	−0.02	+0.01	−0.01	−0.01	+0.01	+0.01	−0.03	−0.01
ar	−0.06	−0.13	−0.03	−0.02	−0.03	+0.00	+0.01	−0.02
ja	−0.02	+0.05	−0.00	−0.01	−0.01	+0.00	+0.01	+0.00
zh	+0.00	−0.00	+0.02	−0.03	−0.03	−0.03	+0.03	+0.01
it	−0.01	+0.01	−0.01	−0.02	+0.00	+0.00	−0.00	−0.01

Table 10: Alignment-associated shift in spectral connectivity. Instruction-tuned and aligned models exhibit systematically lower Fiedler values than their base counterparts.

Group	# Models	Mean Fiedler μ	Δ (Base−Instruct)
Base	6	0.472	0.089 [0.003, 0.177]*
Instruct	8	0.383	

*95% bootstrap CI; $p < 0.05$

Table 11: Functional recovery benchmarks. Optimal steering strength (α) varies by model, with a consistent 'safe zone" between 0.05 and 0.2.

Model	Base PPL	Steered PPL	Optimal α	Gain
Mistral-7B-v0.1	68.35	55.60	0.20	+18.66%
Llama-3.2-3B	103.05	84.60	0.20	+17.91%
Qwen-2.5-0.5B	96.25	86.85	0.05	+9.77%
Phi-3-mini	33.35	31.09	0.20	+6.80%

Table 12: Repair vs. optimization. Scarred models show asymmetric gains; healthy models show balanced improvement.

Model	Passive Repair	Wh-Enhancement
Mistral-7B-v0.1	+18.66% (α=0.2)	+18.44% (α=2.0)
Llama-3.2-3B	+17.91% (α=0.2)	+12.88% (α=1.0)
Phi-3-mini	+6.80% (α=0.2)	+0.62% (α=1.0)
Qwen-2.5-0.5B	+9.77% (α=0.05)	+0.00% (Fragile)

the context window does not recover the lost structural connectivity; the routing failure is local to the passive construct itself.

Experiment B: Invariance to Structural Distractors. We introduced relative clauses to separate the subject from the verb (e.g., *"The plane, which was new, was flown by..."*). This creates a harder routing task. We observed an even sharper decline ($\Delta\lambda_2 \approx -0.24$), suggesting that structural complexity exacerbates the routing failure rather than mitigating it.

These results collectively demonstrate that the

Table 13: Robustness of the Spectral Scar. The Fiedler drop ($\Delta\lambda_2$) persists across variations in length and complexity, confirming it is a structural routing failure.

Condition	Example Structure	Mean $\Delta\lambda_2$
Standard	The book was written by...	-0.18
Long (N=20)	The [adj] [noun] was [verb] by...	-0.13
Clause	The [noun], [clause], was [verb]...	-0.24

scar is a robust spectral fingerprint of the passive voice's routing inefficiency in early induction heads.

E Baseline Comparison: Why Spectral Metrics?

We compared the discriminative power of the Fiedler value against standard attention statistics (Table 14).

Table 14: Baseline comparison (Phi-3, Layer 2). Standard statistics fail due to near-zero variance in the collapsed PTCC state. Fiedler connectivity uniquely quantifies topological disconnection.

Metric	Active (μ)	Passive (μ)	Cohen's d	Status
Fiedler (λ_2)	0.90	0.14	8.42	Robust
Frobenius Norm	1.840	2.000	-2.40	Strong
Max Attention	1.000	0.998	0.31	Weak
Attn Entropy	1.134	0.247	4.12	Moderate

The severe connectivity reduction ($0.90 \rightarrow 0.14$) indicates near-fragmentation of the attention graph, with tokens increasingly attending only to local context rather than integrating globally.

F Robustness Checks

F.1 Behavioral-Spectral Correlation

Table 15: Spectral-behavioral correlation. For Phi-3, connectivity changes directly predict performance degradation ($r = -0.976$).

Model	n	Pearson r	95% CI
Phi-3-mini	10	-0.976	$[-0.99, -0.89]$
Qwen2.5-7B	10	-0.627	$[-0.85, -0.24]$
Llama-3.2-3B	10	-0.143	$[-0.62, +0.41]$

F.2 Methodological Robustness

We verified results under alternative methodological choices:

Laplacian normalization: Random-walk and symmetric normalizations produce identical sign patterns with magnitude shifts within bootstrap bands.

Head aggregation: Uniform averaging versus mass-weighted aggregation agree on all qualitative conclusions.

Directed graphs: Using directed Laplacians reproduces patterns within 15% magnitude.

Early window specification: Adjacent windows (layers 1–4, 2–5, 3–6) preserve all sign patterns and peak locations.

G Steering Implementation

The steering vector is computed as:

$$v = \frac{1}{|S|} \sum_{s \in S} \left(h^{(2)}_{\text{active}}(s) - h^{(2)}_{\text{passive}}(s) \right) \quad (4)$$

where $h^{(2)}(s)$ is the mean-pooled Layer 2 hidden state for sentence s, and S is a calibration set of 50 sentence pairs.

During inference, we inject:

$$\tilde{h}^{(2)} = h^{(2)} + \alpha \cdot v \quad (5)$$

Optimal α is determined by grid search over $\{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$ on a validation set.

Per-layer Fiedler Profile. For completeness, we report the full layer-wise Fiedler values for Phi-3-mini under passive voice stress:

Layer	0	1	2	3	4	5	6	7+
λ_2	0.58	0.24	0.14	**0.08**	0.09	0.37	0.56	0.69–0.88

Table 16: Layer-wise Fiedler values under passive voice. The collapse spans layers 1–4, with minimum connectivity at Layer 3 ($\lambda_2 = 0.08$), recovering by Layer 6.

H The Phi Lineage: Developmental Trajectory of Spectral Pathology

A central question arising from our main results is whether the "synthetic scar" represents a universal signature of synthetic training or a specific artifact of particular curricula. To investigate this, we extended our spectral analysis to the complete Microsoft Phi lineage: Phi-1.5 (1.3B, September 2023), Phi-2 (2.7B, December 2023), and Phi-4 (14B, December 2024). This developmental analysis reveals that the PTCC is *curriculum-specific* rather than *training-regime-universal*, emerging only in the Phi-3 generation and exhibiting qualitatively different pathologies in earlier and later models.

H.1 The Developmental Arc

Table 17 summarizes the spectral signatures across the complete Phi family.

The trajectory is striking: *the PTCC is isolated to Phi-3/3.5*. Earlier models (Phi-1.5/2) lack the high-connectivity baseline required for collapse, while the later model (Phi-4) appears to have resolved the vulnerability through a fundamentally different architectural strategy.

H.2 Phi-1.5 and Phi-2: Pre-Structural Fragmentation

Figures 5 and 6 show the spectral profiles for Phi-1.5 and Phi-2.

The "Fragmented Baseline" Pathology. Neither Phi-1.5 nor Phi-2 exhibits the PTCC because they lack the prerequisite: high English connectivity that can collapse. Instead, English is processed through a permanently low-connectivity regime ($\lambda_2 \approx 0.20$–0.30), regardless of syntactic construction. We term this the "Fragmented Baseline" pathology, the model treats English as a loosely connected collection of tokens rather than an integrated syntactic structure.

This pathology is consistent with the "Textbooks Are All You Need" training philosophy (Li et al., 2023): heavy emphasis on code and formal logic, with English primarily serving as a "comment language" rather than a first-class linguistic object. The model learns to process code with high connectivity but relegates English to a simpler, more local processing regime.

Language-Conditioned Regime Separation in Phi-2. Phi-2 reveals an additional anomaly: a severe split between languages. Russian, Chinese, and Japanese achieve very high connectivity ($\lambda_2 > 0.85$), while English and French remain fragmented. This raises the question of whether the high-connectivity languages reflect genuine competence or a degenerate "safe mode" (uniform attention). We address this via the entropy discriminator in §H.4.

H.3 Phi-4: The Adaptive Gearbox

Phi-4 (Figure 7) presents a dramatically different profile from both its predecessors and the scarred Phi-3.

Key Observations.

1. **No PTCC:** Passive voice produces negligible delta ($|\Delta\lambda_2| < 0.02$) across all languages.
2. **Bimodal Topology:** Languages split cleanly into two regimes:
   - Alphabetic (English, French, German, Russian): $\lambda_2 \approx 0.30$ ("Modular")
   - Logographic (Japanese, Chinese): $\lambda_2 \approx 0.93$ ("Dense")
3. **Structured High-Connectivity:** Unlike Phi-2's high-connectivity languages, Phi-4's logographic regime shows *low* spectral entropy (see §H.4), indicating intentional global integration rather than degenerate uniformity.

We term this the "Adaptive Gearbox", the model switches its internal topology based on the information density of the input script, using modular attention for sparse tokenization (alphabetic) and dense integration for compressed tokenization (logographic).

H.4 The Entropy Discriminator: Intelligence vs. Panic

A high Fiedler value can arise from two mathematically distinct states: (1) *structured integration*, where the model deliberately connects all tokens because each carries high semantic load, or (2) *degenerate uniformity*, where the model connects everything because it lacks priors ("panic mode"). Spectral entropy discriminates between these:

- **Structured Integration:** High λ_2 + Low Entropy (concentrated eigenvalue mass)
- **Panic Mode:** High λ_2 + High Entropy (flat eigenvalue spectrum $\approx$ random graph)

Table 18 presents the entropy analysis.

The contrast is decisive:

- **Phi-2 Russian:** High λ_2 (0.87) + High H_L (1.19) = Panic
- **Phi-4 Japanese:** High λ_2 (0.84) + Low H_L (0.86) = Intelligence

Phi-4's entropy for Japanese (0.86) is actually *lower* than its entropy for English (0.96), despite the $3\times$ higher connectivity. This confirms the high-connectivity regime represents confident, selective attention, not confusion.

Table 17: Spectral signatures across the Phi lineage. The synthetic PTCC (catastrophic $\Delta\lambda_2$) appears only in Phi-3/3.5; earlier and later generations exhibit qualitatively different topological regimes. $\Delta\lambda_2$ here reports the per-layer peak (Layer 3 minimum), which is larger in magnitude than the $\overline{\Delta\lambda_2}_{[2,5]}$ window average (-0.62/-0.64) reported in the main results.

Model	Release	English λ_2 (Active)	Peak $\Delta\lambda_2$ (Layer 3)	Diagnosis
Phi-1.5	Sep 2023	0.25	≈ 0	Fragmented Baseline
Phi-2	Dec 2023	0.20–0.30	≈ 0	Fragmented Baseline
Phi-3-mini	Apr 2024	0.90	-0.76	**Synthetic Scar**
Phi-3.5-mini	Aug 2024	0.88	-0.74	**Synthetic Scar**
Phi-4	Dec 2024	0.30	$+0.01$	Adaptive Gearbox

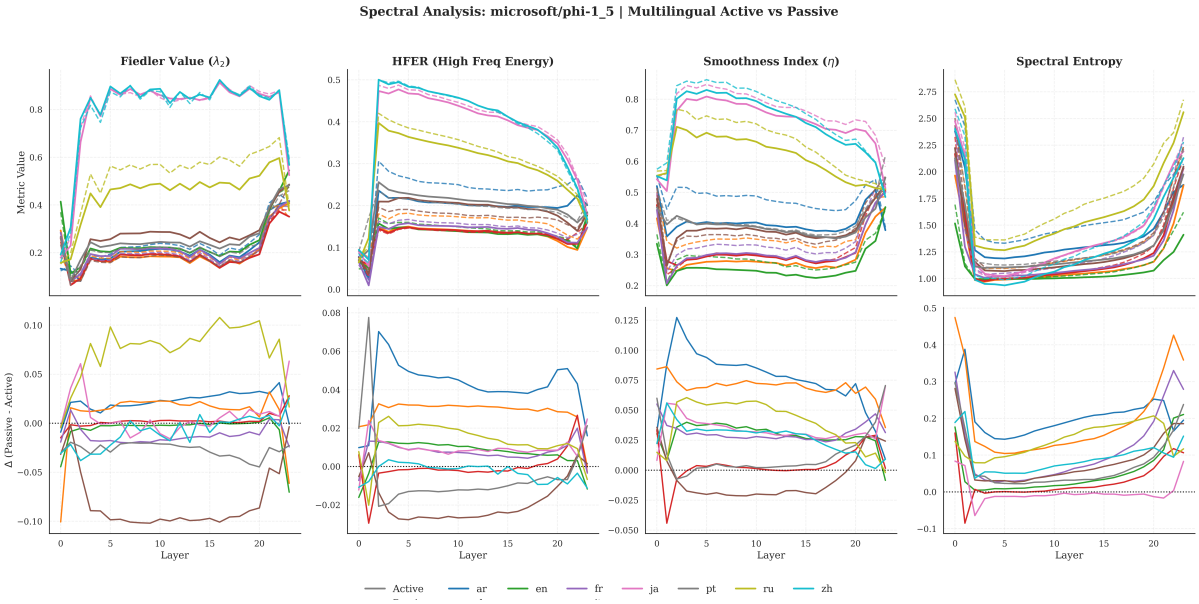

Figure 5: **Phi-1.5: Pre-Structural Fragmentation.** Unlike Phi-3, English (green) does not collapse under passive voice because it never achieves high connectivity in the first place. Baseline English $\lambda_2 \approx 0.25$ is comparable to passive state, resulting in near-zero delta. Other languages show varied regimes, with some (Russian, Chinese) exhibiting high connectivity. This pattern suggests the model processes English through permanently fragmented pathways rather than specialized circuits that fracture under stress.

Table 18: The Entropy Discriminator. High connectivity in Phi-4 Japanese/Chinese is accompanied by *low* entropy, indicating structured integration. In contrast, Phi-2 Russian shows high entropy despite high connectivity, consistent with panic-mode uniform attention.

Model	Language	Fiedler λ_2	Entropy H_L	Interpretation
Phi-2	English	0.94	0.77	Native Competence
Phi-2	Russian	0.87	1.19	**Panic Mode**
Phi-4	English	0.28	0.96	Modular Processing
Phi-4	Japanese	0.84	0.86	**Structured Integration**
Phi-4	Chinese	0.84	0.84	**Structured Integration**

H.5 The Script-Swap Control: Causal Isolation

To definitively prove that Phi-4's topology adapts to *tokenization density* rather than *semantic content* or *language identity*, we conducted a controlled script-swap experiment. We generated $N = 30$ semantically matched triplets, each containing:

1. **Japanese (Kana/Kanji):** Standard logographic representation (dense tokenization)
2. **Japanese (Romaji):** Romanized phonetic representation (sparse tokenization)
3. **English (Translation):** Semantic equivalent in English (sparse tokenization)

Critically, all sentences include terminal punctuation to control for sentence-boundary recognition effects (pilot experiments revealed that unpunctuated Romaji can be misclassified as code/list content, artificially inflating connectivity).

The results are striking: **same semantics, opposite topology; same script, identical topology**. Switching from Kana to Romaji causes a 3.1×

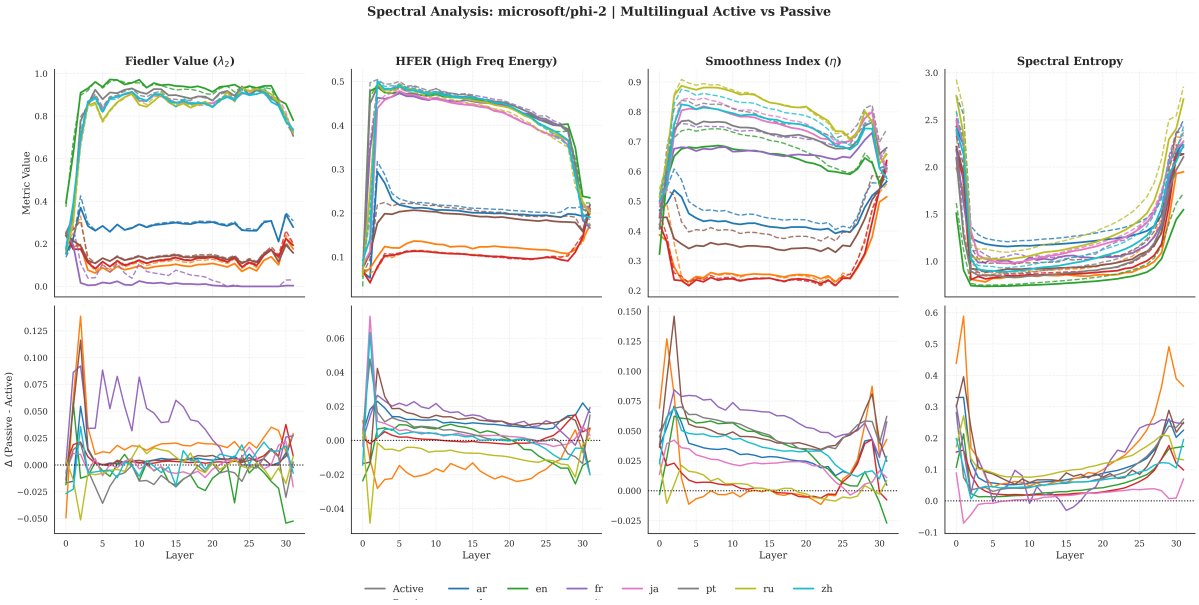

Figure 6: **Phi-2: Language-Conditioned Regime Separation.** Phi-2 exhibits a striking bimodal distribution: some languages (Russian, Chinese, Japanese) achieve very high connectivity ($\lambda_2 > 0.85$), while English and French remain in a low-connectivity regime ($\lambda_2 \approx 0.20$). Passive voice produces minimal delta because the topology is already language-stratified rather than syntax-sensitive. This represents "structural apartheid" rather than syntactic vulnerability.

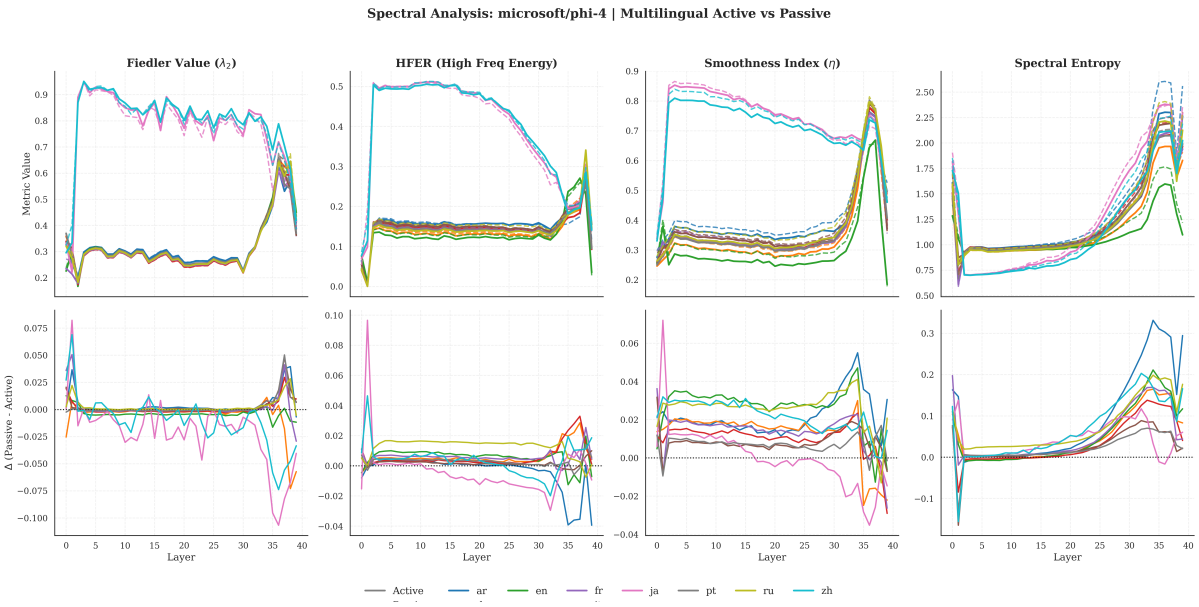

Figure 7: **Phi-4: The Adaptive Gearbox.** Phi-4 shows no synthetic PTCC ($\Delta\lambda_2 \approx 0$ for passive voice). More strikingly, it exhibits a clean bimodal topology: alphabetic languages (English, French, German) cluster around $\lambda_2 \approx 0.30$, while logographic languages (Japanese, Chinese) cluster around $\lambda_2 \approx 0.93$. Unlike Phi-2's "panic mode," Phi-4's high-connectivity regime is accompanied by *low* entropy, indicating structured integration rather than uniform attention.

reduction in connectivity ($0.78 \rightarrow 0.25$). More remarkably, Romaji Japanese achieves *exactly the same* connectivity as English (0.25 vs. 0.25), despite encoding different semantic content in a different language. The topological signature is determined entirely by script type.

Figure 8 visualizes this phase transition across all 40 layers. Japanese (Kana) maintains the high-connectivity "Dense Regime" ($\lambda_2 \approx 0.80$) throughout mid-layers, while both Romaji and English

Table 19: The Script-Swap Control ($N = 30$ triplets). Identical semantic content triggers opposite topological regimes depending solely on script type. Remarkably, Romaji Japanese achieves *identical* connectivity to English, isolating tokenization density as the causal driver.

Input Format	Script Type	Fiedler λ_2	Regime
Japanese (Kana/Kanji)	Logographic	0.78	Dense Integration
Japanese (Romaji)	Latin	0.25	Modular Hopping
English (Translation)	Latin	0.25	Modular Hopping

cluster tightly in the "Sparse Regime" ($\lambda_2 \approx 0.25$). The separation is maintained from Layer 3 through Layer 35, with convergence only in the final output layers.

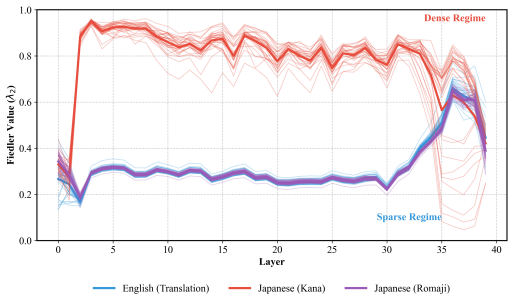

Figure 8: **The Tokenization Topology Law.** Layer-wise Fiedler trajectories for semantically matched content in three formats. Japanese (Kana, red) maintains high connectivity ($\lambda_2 \approx 0.80$) in the "Dense Regime," while Japanese (Romaji, purple) and English (blue) cluster identically in the "Sparse Regime" ($\lambda_2 \approx 0.25$). The phase transition is instantaneous and total: switching script switches topology. Thin lines show individual sentences ($N = 30$); bold lines show means. This proves the spectral signature is governed by tokenization mechanics, not semantic content or language identity.

This experiment establishes a causal relationship: *spectral topology tracks information compression, not language identity*. Phi-4 operates as a "typological gearbox," shifting between dense integration (logographic) and modular hopping (alphabetic) based purely on tokenization efficiency. The model has no "Japanese mode" or "English mode", it has a "dense token mode" and a "sparse token mode," selected automatically based on the physical properties of the input stream.

H.6 The Tokenization Topology Law

Synthesizing the Phi lineage analysis and the script-swap control, we propose the following empirical regularity:

> **Tokenization Topology Law (Hypothesis):** Attention graph connectivity scales inversely with token count for equivalent semantic content. Dense tokenization (logographic scripts) triggers global integration; sparse tokenization (alphabetic scripts) triggers modular bridging.

The script-swap experiment provides unusually clean causal evidence for this law. Three predictions follow directly:

1. *Logographic scripts should show higher connectivity than alphabetic scripts.* Confirmed: Japanese (Kana) $\lambda_2 = 0.78$ vs. English $\lambda_2 = 0.25$.

2. *Romanization should collapse logographic connectivity to alphabetic levels.* Confirmed: Romaji $\lambda_2 = 0.25$, identical to English.

3. *The regime should track script, not language.* Confirmed: Romaji Japanese is topologically indistinguishable from English despite encoding different semantic content.

The law remains a hypothesis pending validation across additional models and scripts (e.g., Arabic, Hebrew, Thai). However, the identity $\lambda_2^{\text{Romaji}} = \lambda_2^{\text{English}} = 0.25$ provides striking evidence that topology adapts to information density rather than reflecting fixed language-specific circuitry. The model appears to operate a content-agnostic "compression detector" that routes dense token streams through global integration and sparse streams through modular bridging.

H.7 Evolutionary Interpretation: From Trauma to Maturity

The complete Phi lineage can now be understood as an evolutionary trajectory through distinct spectral pathologies:

1. **Phi-1.5/2 (Infancy):** "Structural Apartheid." English is undertrained, processed through permanently fragmented pathways. High connectivity on some languages reflects panic (high entropy), not competence.

2. **Phi-3/3.5 (Adolescence):** "The Synthetic Scar." Microsoft attempted to bridge the fragmented English baseline with new multilingual capacity. The resulting architecture achieved high English connectivity but compressed routing into sparse compensatory circuits that fracture under syntactic stress.

3. **Phi-4 (Maturity):** "The Adaptive Gearbox." The PTCC is healed. The model operates a content-agnostic topology selector: dense token streams (logographic) receive global integration ($\lambda_2 \approx 0.78$), sparse streams (alphabetic) receive modular bridging ($\lambda_2 \approx 0.25$). The script-swap experiment proves this is a physical response to tokenization density, not language-specific circuitry.

The PTCC, in this view, was a developmental injury, the price of attempting efficiency gains during a curriculum transition. Phi-4's resolution suggests Microsoft identified and corrected the underlying failure mode, replacing brittle language-specific circuits with a robust density-adaptive mechanism.

H.8 Implications for the Main Results

The Phi lineage analysis has several implications for interpreting our main findings:

1. The PTCC is curriculum-specific. The synthetic PTCC should not be interpreted as a universal signature of synthetic training. It appears to emerge from specific curriculum transitions that compress routing redundancy into sparse circuits. Other synthetic training approaches (e.g., Phi-1.5/2's code-focused curriculum) produce different pathologies.

2. Spectral forensics reveals developmental history. The progression Fragmented → Scarred → Adaptive demonstrates that spectral analysis can recover not just training regime (synthetic vs. organic) but developmental trajectory within a model family. This opens possibilities for "spectral archaeology", reconstructing training history from weight inspection alone.

3. The typological gearbox suggests deeper structure. Phi-4's script-dependent topology suggests that advanced models may encode *typological awareness*, implicit knowledge of how different writing systems encode information. This raises questions about whether such awareness is learned, emergent, or architecturally imposed.

I Spectral Profiles Across Model Families

This appendix presents the complete spectral signatures for all eight primary models evaluated in our study. These visualizations enable direct comparison of processing strategies across architectural families and provide the empirical basis for the forensic classification rules presented in the main text.

I.1 Strategy A: English-Specialized Collapse (Phi-3)

The Phi-3 family exemplifies the spectral signature of the synthetic PTCC. These models achieve strong benchmark performance through efficient, specialized circuits, but this efficiency comes at the cost of robustness to syntactic variation.

The Collapse Mechanism. Figure 9 reveals the characteristic "synthetic scar" in Phi-3-mini. Under active voice, English maintains healthy connectivity ($\lambda_2 \approx 0.90$) comparable to other languages. Under passive voice stress, English connectivity collapses catastrophically: λ_2 drops to 0.14 at Layer 2 and reaches a minimum of 0.08 at Layer 3, a 91% reduction. The collapse spans layers 1–5 before recovering by Layer 6. Critically, all nine other languages remain stable ($|\Delta\lambda_2| < 0.06$), demonstrating that this is an English-specific vulnerability rather than a general syntactic limitation.

The concurrent HFER spike (second column) confirms that the collapse involves signal fragmentation into high-frequency modes, while the entropy increase (fourth column) indicates loss of structured computation. Together, these metrics paint a picture of attention graphs fracturing into disconnected local clusters unable to integrate global context.

Replication Across Versions. Figure 10 demonstrates that the synthetic PTCC is a persistent family characteristic. Phi-3.5-mini exhibits an identical collapse pattern ($\overline{\Delta\lambda_2}_{[2,5]} = -0.64$), with the same layer localization and language specificity. This replication across model versions rules out version-specific artifacts and confirms that the PTCC originates from the training data curriculum rather than architectural idiosyncrasies.

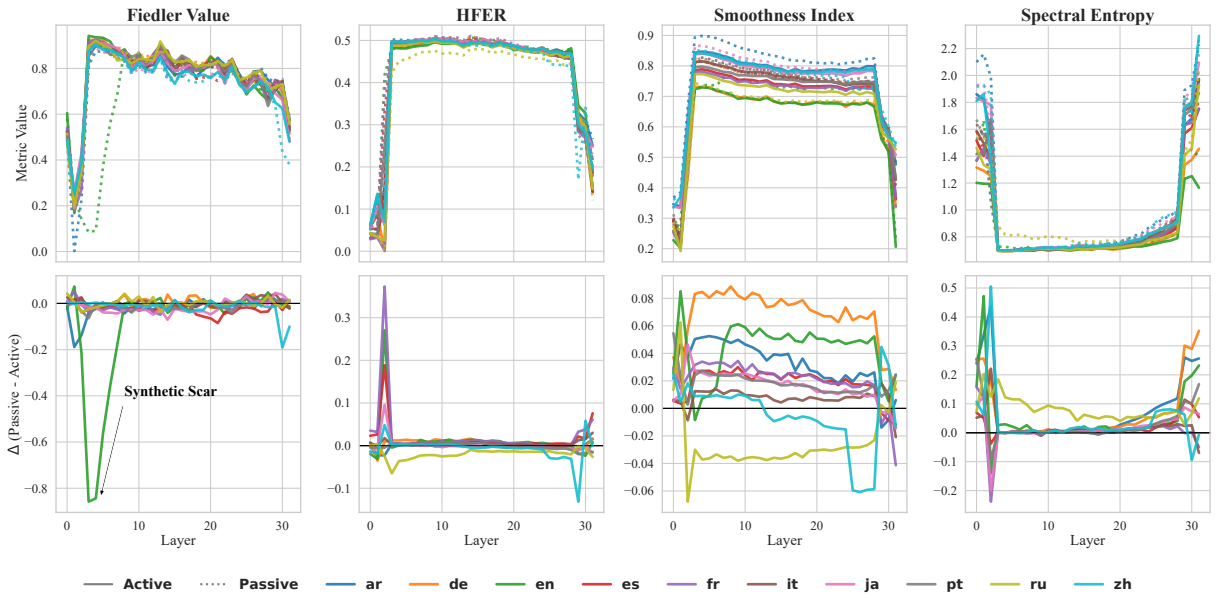

Figure 9: **The Synthetic Scar (Phi-3-mini).** Top row: raw spectral metrics across 32 layers. Bottom row: passive minus active deltas. English (green) exhibits catastrophic Fiedler collapse at layers 2–4 ($\lambda_2 : 0.90 \rightarrow 0.14$, minimum 0.08 at Layer 3), while all other languages remain within $|\Delta\lambda_2| < 0.06$. The concurrent HFER spike and entropy increase confirm topological fragmentation accompanies the connectivity loss. This English-specific signature enables forensic identification of the scarred training regime.

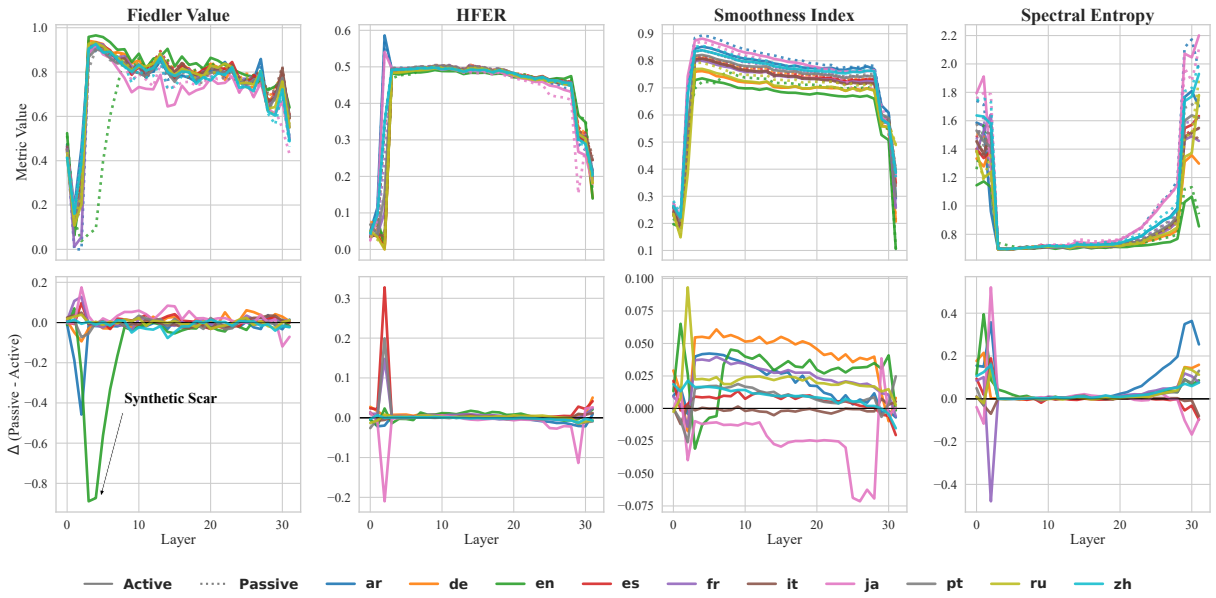

Figure 10: **Replication of the Synthetic Scar (Phi-3.5-mini).** The collapse pattern replicates precisely in Phi-3.5 ($\overline{\Delta\lambda_2}_{[2,5]} = -0.64$), with identical layer localization (2–4), language specificity (English only), and recovery dynamics. This cross-version consistency confirms the PTCC is a family-level characteristic driven by the synthetic training curriculum, not a version-specific anomaly.

I.2 Strategy C: Cross-Linguistic Uniformity (Llama)

The Llama family demonstrates the spectral signature of broad-spectrum organic pre-training: uniform processing pathways that remain stable across languages and syntactic transformations.

The Uniform Baseline. Figure 11 shows Llama-3.2-1B maintaining high algebraic connectivity ($\lambda_2 \approx 0.80$) across all ten languages under both active and passive voice. The delta panel (bottom row) reveals near-zero shifts for all language-construction pairs, with no language exhibiting preferential vulnerability. This uniformity suggests

that exposure to diverse, organic web data forces the model to learn attention mechanisms invariant to local syntactic perturbations, robustness through redundancy rather than specialization.

Scale Invariance. Figure 12 confirms that the uniform strategy persists at larger scale. Llama-3.2-3B exhibits the same high-connectivity baseline and minimal language-specific effects as its 1B counterpart. Increased capacity does not induce specialization or create new vulnerabilities, the model simply maintains its robust topology with slightly smoother spectral trajectories. This scale invariance suggests the uniform strategy is an emergent property of the training distribution rather than a capacity-limited compromise.

I.3 Strategy D: Confidence Calibration (Mistral)

Mistral-7B presents a distinct variant of organic robustness characterized by an unusual relationship between syntactic stress and uncertainty.

The Confidence Inversion. Figure 13 reveals that Mistral maintains stable Fiedler values ($\lambda_2 \approx 0.85$) and HFER across syntactic transformations, but exhibits a unique pattern in spectral entropy (fourth column). While most models show increased entropy (higher uncertainty) under passive voice, Mistral shows a *decrease* ($\Delta\text{Ent} \approx -0.15$).

We interpret this as a "confidence inversion": the rigid constraints of passive syntax (mandatory auxiliary verbs, fixed word order) provide additional structural anchors that *reduce* rather than increase parsing ambiguity. Rather than fracturing the attention graph, passive constructions crystallize it into a more deterministic configuration. This behavior likely reflects Mistral's architectural choices (Sliding Window Attention, Group Query Attention) interacting with high-quality training data to produce a qualitatively different form of robustness.

I.4 Strategy B: English-Fragmented Baseline (Qwen)

The Qwen family exhibits a third distinct strategy: permanently fragmented English processing that differs qualitatively from both the Phi collapse and the Llama/Mistral robustness.

The Fragmented Pathway. Figure 14 shows that Qwen-2.5-7B processes English through a fundamentally different topological regime. Even under *active* voice, English connectivity is severely reduced ($\lambda_2 \approx 0.12$) compared to other languages ($\lambda_2 \approx 0.70$). This is not a stress-induced collapse but a permanent baseline difference, English occupies a distinct, low-connectivity subspace of the attention manifold.

The HFER response is also inverted: passive voice *reduces* high-frequency energy ($\Delta\text{HFER} \approx -0.35$) rather than increasing it. This suggests Qwen has learned a specialized English processing pathway that operates via different computational mechanisms than its multilingual circuits. Whether this reflects deliberate architectural choices, tokenizer effects, or training data composition remains an open question.

Scale Consistency. Figure 15 demonstrates that the fragmented English baseline persists down to the smallest Qwen variant. At 0.5B parameters, the same pattern emerges: low English connectivity at baseline, minimal stress-induced change, inverted HFER response. This scale consistency indicates the fragmentation is a fundamental property of the Qwen family, likely originating from tokenizer design or pre-training mixture, rather than an emergent property of scale.

Mixture-of-Experts Dynamics. Figure 16 extends the analysis to Qwen-1.5-MoE. The MoE architecture introduces higher variance in spectral trajectories (visible as jagged layer-wise profiles) due to discrete expert routing decisions. However, the fundamental strategy aligns with dense Qwen models: fragmented English baseline, inverted HFER response, no catastrophic collapse. Notably, the MoE shows no evidence of the synthetic PTCC despite its efficiency-focused architecture, suggesting the PTCC is specific to synthetic *data* rather than efficient *architectures*.

I.5 Summary: Four Strategies, One Framework

The spectral profiles presented above demonstrate that a single analytical framework, attention-graph connectivity under syntactic stress, reveals four qualitatively distinct processing strategies:

1. **Specialized Collapse (Phi-3):** High baseline connectivity with catastrophic English-specific failure under stress. Signature of curriculum-induced brittle, low-redundancy circuits.

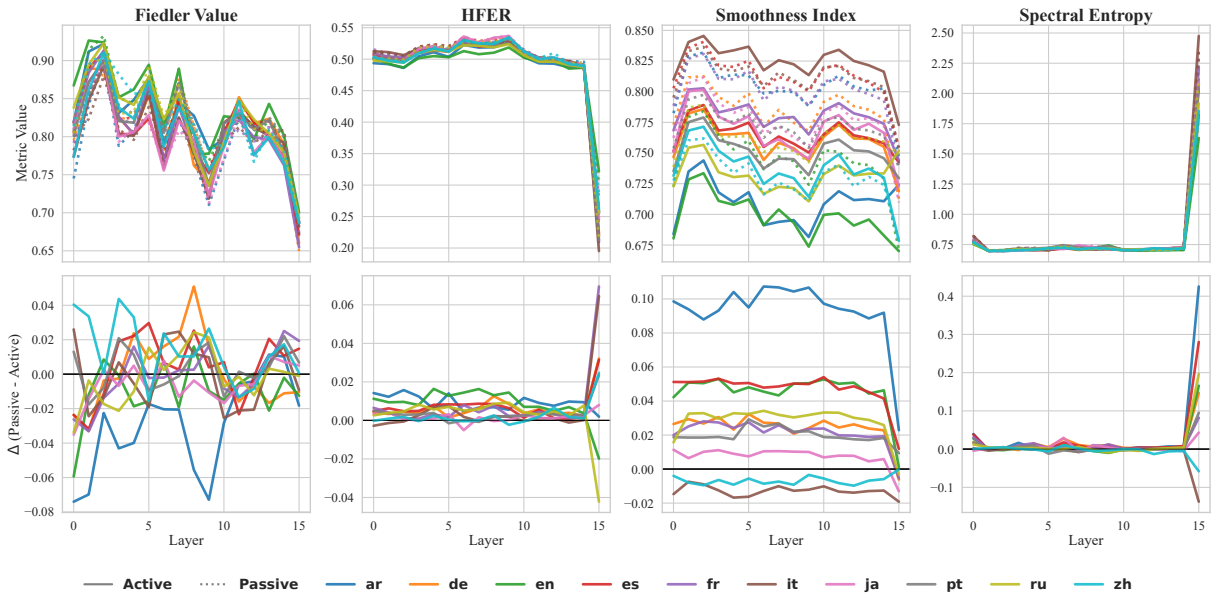

Figure 11: **Cross-Linguistic Uniformity (Llama-3.2-1B).** All ten languages maintain stable, high connectivity ($\lambda_2 \approx 0.80$) with minimal passive-active deltas ($|\Delta\lambda_2| < 0.02$). No language-specific fragility emerges. This uniform strategy reflects broad-spectrum organic training that builds redundant pathways capable of handling syntactic variation without topological disruption.

2. **Fragmented Baseline (Qwen):** Permanently low English connectivity at baseline, stable under stress. Signature of architectural or data-driven pathway separation.

3. **Cross-Linguistic Uniformity (Llama):** High connectivity maintained uniformly across languages and constructions. Signature of organic training building redundant, stress-resistant pathways.

4. **Confidence Calibration (Mistral):** High connectivity with entropy reduction under stress. Signature of architectures that leverage syntactic constraints for disambiguation.

These strategies are invisible to behavioral benchmarks but carry direct implications for deployment: multilingual applications should prefer uniform-strategy models, while applications requiring syntactic flexibility should avoid specialized-collapse models regardless of their headline performance metrics.

J Experimental Setup

J.1 Models

We conduct experiments across multiple language model families and architectures (dense and sparse) to ensure the generalizability of our findings. All models are loaded in FP16 precision using the Hugging Face `transformers` library.

Phi Family. We evaluate the evolution of the Phi architecture:

- **Phi-3:** `microsoft/Phi-3-mini-4k-instruct` (3.8B, 32 layers).
- **Phi-3.5:** `microsoft/Phi-3.5-mini-instruct` (3.8B), evaluating improvements in long-context and reasoning capabilities.

Llama Family. We evaluate both base and instruction-tuned variants at two scales:

- **1B Scale:** `meta-llama/Llama-3.2-1B` and `Llama-3.2-1B-Instruct`.
- **3B Scale:** `meta-llama/Llama-3.2-3B` and `Llama-3.2-3B-Instruct`.

Qwen Family. To test robustness across varying parameter counts and multilingual capabilities, we encompass:

- **Qwen 2.5:** `Qwen/Qwen2.5-0.5B-Instruct` and `Qwen/Qwen2.5-7B-Instruct`.

Mistral Family. We evaluate `mistralai/Mistral-7B-Instruct-v0.1` (7B) to benchmark spectral properties against another leading dense architecture.

Phi Family (Extended). Beyond Phi-3, we analyze the architectural evolution:

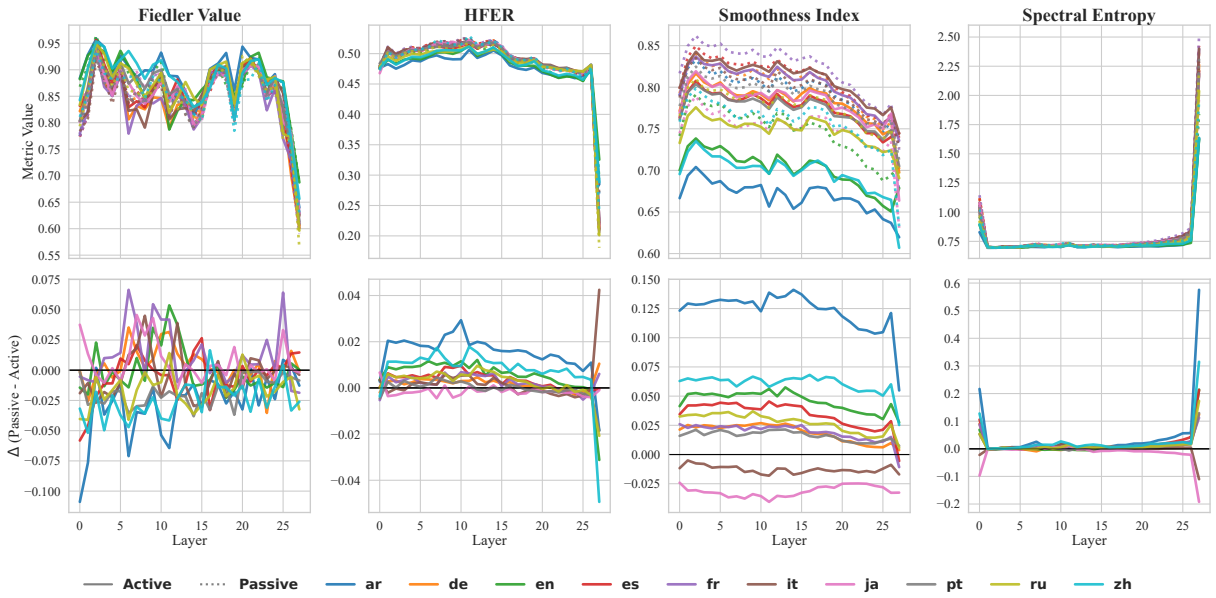

Figure 12: **Scale Invariance (Llama-3.2-3B).** The uniform strategy persists at 3B parameters. Connectivity remains high ($\lambda_2 \approx 0.80$) and language-agnostic, with smoother layer-wise trajectories than the 1B model. Scaling capacity does not introduce specialization scars; the organic training distribution produces consistent robustness across scales.

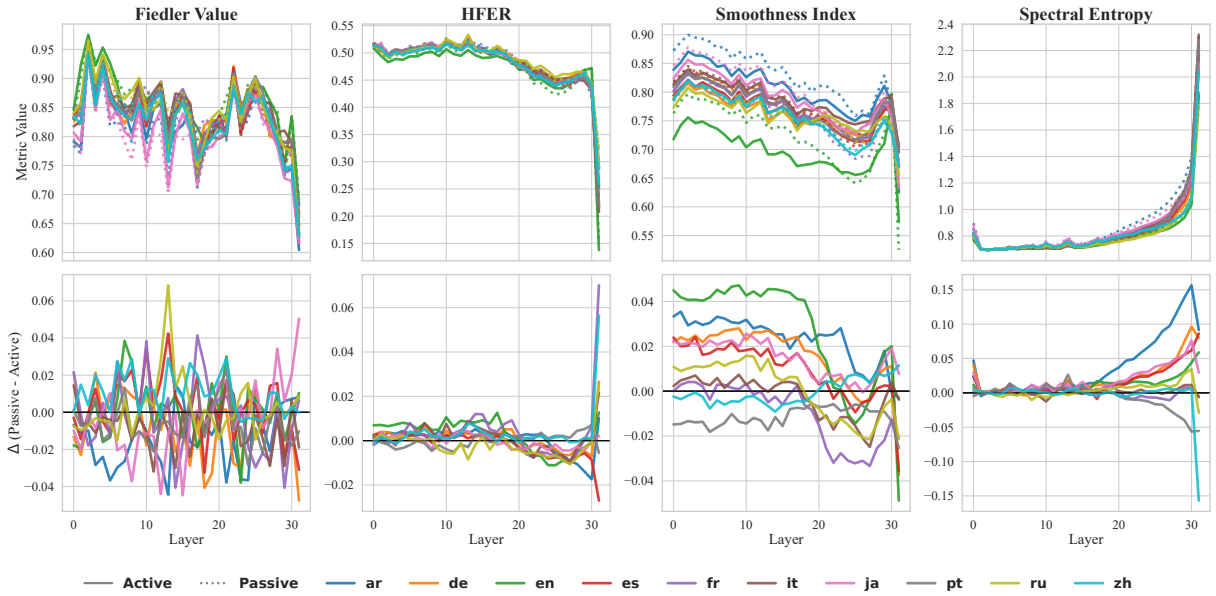

Figure 13: **Confidence Inversion (Mistral-7B).** Mistral maintains high connectivity ($\lambda_2 \approx 0.85$) with minimal Fiedler or HFER shifts under passive voice. Uniquely, spectral entropy *decreases* (ΔEnt ≈ -0.15), suggesting passive syntax reduces rather than increases uncertainty. This "confident" strategy treats syntactic constraints as disambiguation cues, producing a distinct robustness profile from Llama's uniform approach.

- **Phi-4:** `microsoft/Phi-4` (14B), analyzing topological changes in newer dense models.

Evolution Analysis. To study architectural evolution, we additionally include:

- `openlm-research/open_llama_3b` (Llama-1 proxy, trained on RedPajama)
- `openlm-research/open_llama_3b_v2` (Llama-2 architecture proxy)

J.2 Datasets

J.2.1 Syntactic Integrity Benchmark

We construct controlled sentence pairs to isolate syntactic structure effects on attention topology:

Voice Contrast. 200 active and 200 passive voice sentence pairs matched for semantic content:

> *Active:* "The author translated the account."
> *Passive:* "The account had been evaluated by the

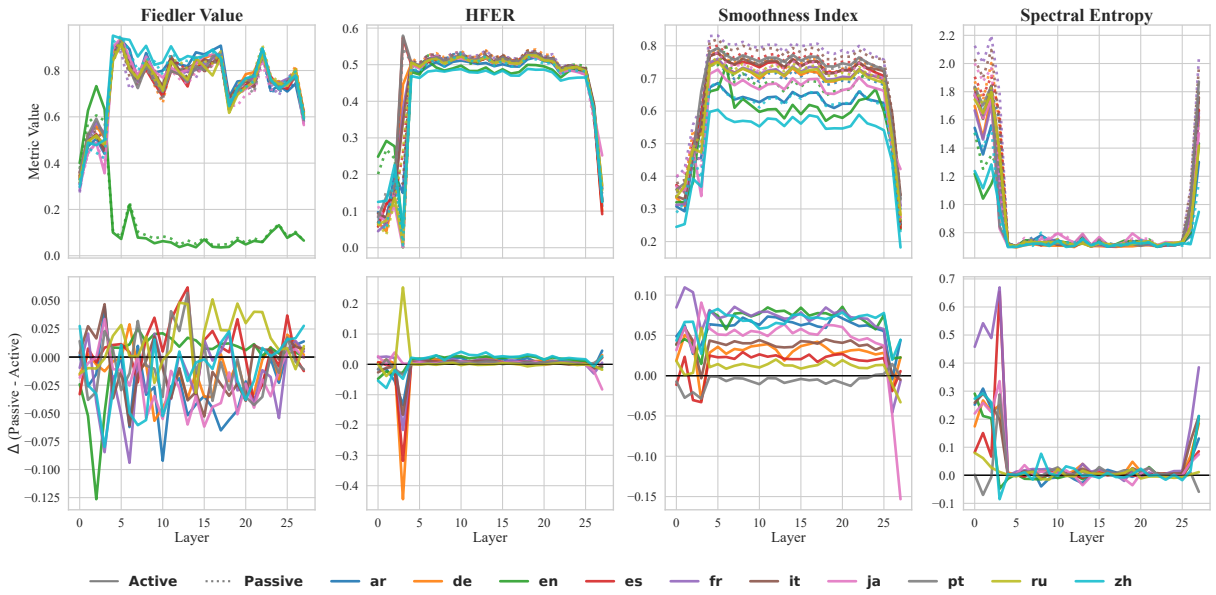

Figure 14: **English-Fragmented Baseline (Qwen-2.5-7B).** English (green) exhibits permanently low connectivity ($\lambda_2 \approx 0.12$) even under active voice, while other languages maintain $\lambda_2 \approx 0.70$. This is not stress-induced collapse but a baseline architectural separation. The inverted HFER response (ΔHFER < 0) confirms English is processed via a qualitatively different pathway. Unlike the Phi PTCC, this represents stable fragmentation rather than brittle specialization.

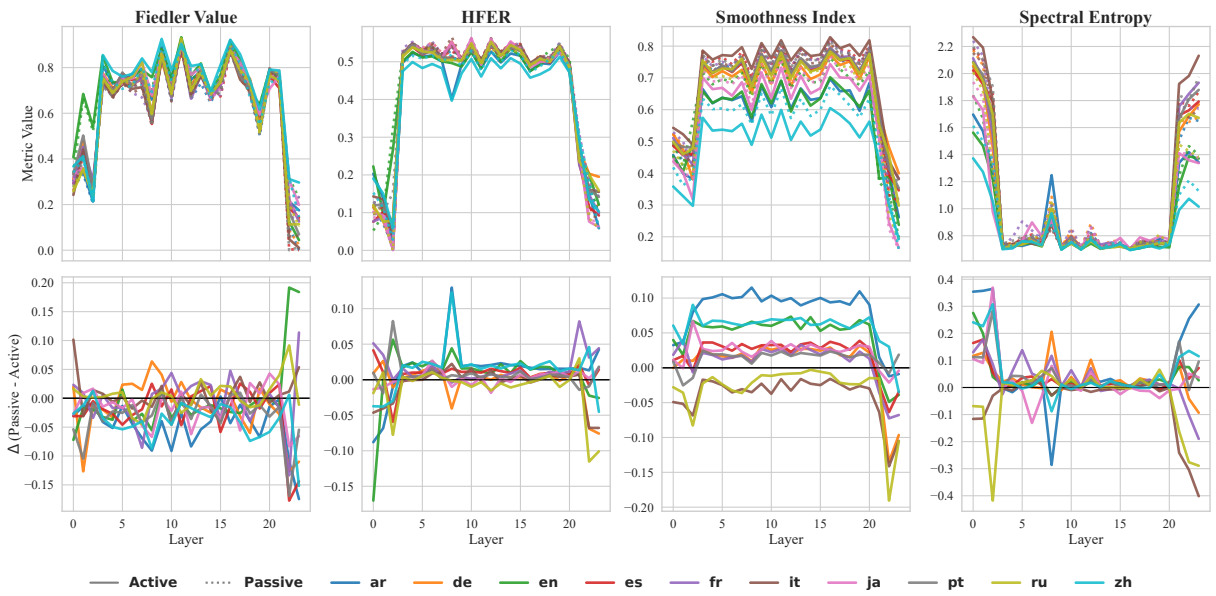

Figure 15: **Scale Consistency (Qwen-2.5-0.5B).** The fragmented English baseline persists at 0.5B parameters, confirming this is a family-level characteristic independent of model capacity. English connectivity remains low ($\lambda_2 \approx 0.15$) relative to other languages across all layers. This consistency across scales suggests the fragmentation originates from tokenizer or data mixture choices rather than capacity constraints.

electrician."

Grammatical Categories. For rigorous statistical analysis, we generate 200 samples per category using template-based synthesis:

- **Active Voice**: Subject-Verb-Object canonical order
- **Passive Voice**: Object-Verb-by-Subject with auxiliary constructions
- **Wh-Questions**: Interrogative structures with fronted operators
- **Complex Clauses**: Relative clauses and embedded sentences

Sentence generation employs randomized vocabulary from curated noun/verb pools to avoid lexical

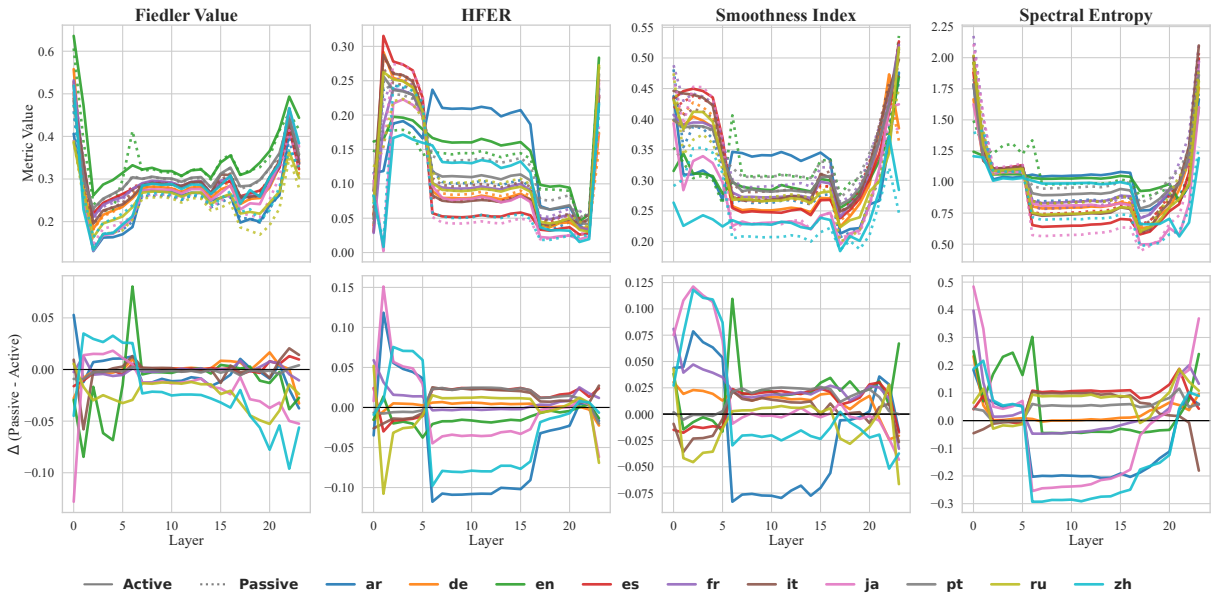

Figure 16: **Mixture-of-Experts Dynamics (Qwen-1.5-MoE).** The MoE architecture introduces higher spectral variance (jagged trajectories) due to discrete router decisions, but preserves the Qwen family's fragmented English strategy. No synthetic PTCC emerges despite the efficiency-focused sparse architecture, confirming the PTCC is a property of synthetic *training data* rather than parameter-efficient *architectures*.

confounds while maintaining consistent syntactic structures.

J.2.2 Stress Testing (Out-of-Distribution)

To probe model robustness and failure modes, we construct adversarial inputs:

- **Gibberish**: Random alphanumeric strings (50–500 characters)
- **Repetition**: Token repetition sequences ("The The The...", 50–200 tokens)
- **Code Noise**: Random punctuation and symbol sequences
- **Math Noise**: LaTeX-like symbol combinations

Each category contains 100 samples for statistical reliability.

J.3 Spectral Analysis Pipeline

Attention Extraction. For each input sequence $\mathbf{x}$ of length n, we extract the attention weight matrix $\mathbf{A}^{(\ell)} \in \mathbb{R}^{h \times n \times n}$ at each layer ℓ, where h denotes the number of attention heads. We compute the head-averaged attention:

$$\bar{\mathbf{A}}^{(\ell)} = \frac{1}{h} \sum_{i=1}^{h} \mathbf{A}_i^{(\ell)} \tag{6}$$

Graph Construction. The attention matrix is symmetrized to form an undirected weighted adjacency matrix:

$$\mathbf{W}^{(\ell)} = \frac{1}{2} \left(\bar{\mathbf{A}}^{(\ell)} + (\bar{\mathbf{A}}^{(\ell)})^\top \right) \tag{7}$$

Laplacian Spectrum. We compute the combinatorial graph Laplacian:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \quad D_{ii} = \sum_j W_{ij} \tag{8}$$

and extract its eigenspectrum $\{\lambda_0, \lambda_1, \ldots, \lambda_{n-1}\}$ where $\lambda_0 = 0$.

Spectral Metrics. We report four primary metrics:

1. **Fiedler Value** (λ_1): The algebraic connectivity, measuring global graph connectedness. Higher values indicate tighter structural integration.
2. **Spectral Entropy**: $H = -\sum_i p_i \log p_i$ where $p_i = \lambda_i / \sum_j \lambda_j$, quantifying the distribution of structural information across eigenmodes.
3. **HFER** (High-Frequency Energy Ratio): Proportion of spectral mass in the upper half of the spectrum, indicating fine-grained vs. global structural dominance.
4. **Smoothness Index**: Measures signal coherence across graph topology.

J.4 Implementation Details

All experiments are conducted using a custom open-source library built on PyTorch 2.x and Hugging Face Transformers.

Hardware. Experiments run on a single A100 (80GB VRAM). Models are loaded with `device_map="auto"` for automatic memory management.

Computational Requirements. Spectral analysis requires only one forward pass per input, with computational overhead dominated by eigendecomposition of the $N \times N$ Laplacian ($O(N^3)$ worst case, typically $O(N^2)$ with sparse methods). For the sentence lengths used in this study ($N < 50$ tokens), analysis completes in under 100ms per sample.

K Reproducibility

Code Availability. To facilitate reproducibility, the complete implementation, including attention extraction scripts and spectral computation modules, is available at https://anonymous.4open.science/r/spectral-archeology-CD38. The core spectral diagnostics rely on a lightweight Python package distributed via `pip` (name withheld for double-blind review), and the repository contains pinned dependency versions for exact replication. We additionally provide pre-computed spectral statistics and layer-wise trajectories at the same anonymized URL to support validation.