Predicting How Transformers Attend, Part III: From Attention to Residual Computation

Separating Transport, Writing, and Commitment

continues Part I (distance axis / γ) and Part II (six-axis decomposition)

Carles Marín*

Independent Researcher

transformerkmarin@gmail.com

July 2026

Abstract

Parts I and II characterised the *distance* axis of attention: the power-law exponent γ and its six-axis decomposition. This paper opens the orthogonal *depth* axis and finds that a single question — "is the answer here yet?" — silently conflates four statements about the residual stream: that information is *present*, *linearly decodable*, *written*, and *used*. We show these are separate, independently measurable observables, and decompose residual computation into three: **transport** (the mean input–output Jacobian; the Jacobian lens, J-lens), **net injection** (direct logit attribution, DLA), and **commitment** (the depth at which the running decision locks). The DLA decomposition is *exact*: with the final normalisation frozen the output logit equals the sum of per-layer contributions (validated to 7.6×10^{-6}, norm-aware for LayerNorm and RMSNorm). Empirically we map J-lens readability across a ladder-complete Pythia suite (70M–2.8B) and cross-family anchors (Llama-2/3, Qwen2.5/3, GPT-2): the readability *knee* is a family property (Pythia 0.54–0.75 of depth, Llama 0.81–0.84) and the J-lens *advantage band* is strikingly heterogeneous (absent at d_{head}=256; up to 68× at layer 1 of Qwen2.5-3B). Three pre-registered hyperparameter predictors of the knee are *refuted* by discriminant designs. The paper's pivot is a structural theorem: the advantage band is not a "workspace" of content but a property of the transport *operator* itself ($J_\ell = \prod_{k \geq \ell}(I + w'_k)$), which forces a strong anticorrelation (Spearman -0.80 for net injection, pythia-410m) between Jacobian-lens advantage and net injection — sensitivity and injection are logically independent axes. Across families the *observables* generalise but their *chronology* does not: $\Delta_{\text{prep}} = \tau_{\text{write}} - \tau_{\text{decode}}$ classifies families cleanly (Type A ≥ 0.5 vs Type B ≤ 0.3, no overlap). Its relation to the Part I exponent γ we leave *inconclusive*: two γ measurement protocols disagree for the same models and the direction of any co-ordination flips between them. All experiments are pre-registered with frozen decision rules; refuted hypotheses are reported in the same register as surviving ones. Load-bearing identities are stated here and machine-verified in Lean 4/Mathlib in the companion formalization paper.

Artifacts: the only ladder-complete J-lens suite (70M–2.8B), all readability curves, the DLA traces, pre-registrations with dated decision rules, and the metric specification are released at `part3/experiments/workspace_h1/`.

Reader guide. §2 is the results table; each row links to its section. §3 fixes the lenses, metrics, and pre-registration protocol that judge everything downstream. §§4–6 build the theory (three observables, the exact DLA identity, and the structural bridge that tells us what the J-lens band *is*). §§7–11 are the data: the map, the refuted predictors, the cross-family chronology, prompt-frame loading, and training-time formation. §12 closes the loop to Part I. Refuted hypotheses are first-class citizens throughout — they are how the map earns its trust.

*Prepared with the assistance of an AI system (Claude, Anthropic).

1 Introduction

The conflation. A large fraction of interpretability practice reads a transformer one layer at a time and asks "is the answer here yet?". But "here" silently merges statements that are not equivalent: (i) the answer is *linearly decodable* from the residual (logit / Jacobian lens); (ii) the answer has been *written* to the residual by some layer (attribution); (iii) the network has *committed* to it (the running decision no longer changes); (iv) the network *uses* it downstream (causal). Different tools measure different members of this list — which is one reason apparently competing findings coexist in the literature. The linear-decodability/causal-use gap is the probing confound of Hewitt and Liang [7]; this paper adds the decodable/written/committed distinctions and measures all three.

An instrument, three days old. In July 2026 the Jacobian lens was introduced [1]: average the Jacobian of the final residual state with respect to layer ℓ over a corpus, and use J_ℓ as a linear readout of what layer ℓ carries *as the model itself would propagate it*. Applied to Claude models it revealed a mid-depth band (38–92% of layers) whose contents behaved like a global workspace [21]. The declared limitation was scope: every measurement was on Claude. An independent review [2] replicated the core phenomena on an open model (Qwen 3.6-27B), found a handful of prompts suffice to fit a usable lens, and framed the instrument as a *hypothesis generator* — one that "can miss concepts, produce noise, and should not be treated as a complete detector." What nobody had measured is how the structure the lens reveals varies across scale and family, and — the question this paper answers — *what the band physically is*.

Two axes. Part I [10] characterises the *distance* axis: attention decays as $A(d) \propto d^{-\gamma}$, with γ fixed in central scale by RoPE aliasing geometry and acting as an inverse temperature; Part II [11] decomposes γ_{obs} into six axes. This paper opens the orthogonal *depth* axis. The config (θ, T) induces two independent geometric scales — a ratio $u = T/(\theta\sqrt{2})$ governing attention decay ($\gamma = (1-u)/(1+u)$, a Möbius transform) and a logarithm $\alpha = \log_\theta(T/2\pi)$ governing layer structure (L_{crit}). We keep these axes explicitly distinct; the observables of *this* paper live on the second, and γ enters only in §12, where their relation is examined and found to depend on the γ measurement protocol.

Contributions.

C1. An operational decomposition of residual computation into three measurable observables — transport J, net injection (DLA), and commitment — with a demonstration that they are *not* equivalent (§4). ★

C2. The exact additive DLA identity under frozen normalisation, validated to 7.6×10^{-6} and norm-aware for LayerNorm and RMSNorm (§5). ✓

C3. A structural theorem — $J_\ell = \prod_{k \geq \ell}(I + w'_k)$ — that identifies the J-lens advantage band as a property of the transport *operator*, forcing the transport–injection anticorrelation (Spearman -0.80 net, pythia-410m). Sensitivity and injection are independent axes (§6). ★

C4. The first J-lens readability map across a controlled scale ladder and across families: knee as family property, advantage-band heterogeneity including the d_{head}=256 absence case (§7). ★

C5. Pre-registered refutation of three hyperparameter predictors of the knee, with discriminant designs (§8). ×

C6. Cross-scale and cross-family chronology: universal observables, family-specific timing; Δ_{prep} as a single-number classifier (§9). ✓

C7. An examination of whether the depth-axis chronology co-orders with the Part I exponent γ (§12); we find the direction is *protocol-dependent* and report it as inconclusive. ×

The paper follows the causal order of the argument: define the observables, make one exact, prove why they differ, then let the data show the separation is real and family-specific, and examine — inconclusively — its tie to γ.

2 Summary of Results

The data is the sap of this paper; the table is its cross-section.

	Result	Status
R1	Readability knee is a family property (Pythia 0.54–0.75; Llama 0.81–0.84)	★
R2	L_{crit} predicts the knee	× (discriminant)
R3	Universal constant knee fraction ≈ 0.69	× (full ladder)
R4	RoPE fraction $\log_\theta(T/2\pi)$ predicts the knee	inconclusive (frozen rule)
R5	Advantage-band heterogeneity (none / mid / early), up to 68×	★
R6	Three observables $\tau_{\text{decode}} < \tau_{\text{write}} < \tau_{\text{commit}}$ do not coincide (410m: 0.17/0.66/0.96)	★
R7	DLA decomposition exact (7.6×10^{-6}), norm-aware	✓
R8	Band = operator property: Spearman$(A_\ell, \mu_\ell) = -0.80$ net (-0.85 gross)	★
R9	Δ_{prep} classifies families (Type A ≥ 0.5 vs Type B ≤ 0.3, no overlap)	★
R10	$\gamma \leftrightarrow$ chronology: direction flips between γ protocols	× inconclusive
R11	Axiomatic frame worsens band loading (pythia-410m); does *not* replicate on Qwen2.5	★
R12	MLP-efficiency valley shallows with scale ($-0.66 \to -0.09$)	★ (refutes "deeper")

3 Background: Lenses, Metrics, and Pre-Registration

Models and data. We measure Pythia [8], GPT-2 [16], Llama-2/3 [17, 18], and Qwen2.5/3 [19, 20]; all use rotary position embeddings (RoPE) [13] except GPT-2 (absolute PE). Lenses are fit on The Pile [14] (`pile-10k`); the prompt-frame study (§10) uses SciQ [15]. Our fitted lenses are deliberately complementary to public collections such as Neuronpedia [22].

Two lenses on the residual stream. The logit lens reads layer ℓ's residual h_ℓ directly through the model's own unembedding, unembed(norm(h_ℓ)): what is *already output-aligned*. The J-lens first applies the mean Jacobian $J_\ell = \mathbb{E}[\partial h_L / \partial h_\ell]$, reading unembed(norm($J_\ell h_\ell$)): what the layer carries *as the network itself would transport it*. Every comparison below is J-lens *versus its own logit-lens control* — without the control, any lens shows mirages. We fit one lens per model on 200 pile-10k prompts of 128 tokens (seed 42; fp32 up to 1.4B, bf16 above, dtype recorded per lens), checkpointing every 10 prompts; evaluation uses 100 disjoint held-out sequences (seed 43).

Metrics (frozen before measurement). bestRR(ℓ): median over positions/sequences of the best reciprocal rank, *within the word-like vocabulary* (regex `^ ?[A-Za-z]{3,}$`), of the actual future tokens (k=16) under the lens readout at layer ℓ. *Knee*: first layer with logit-lens bestRR $\geq 0.5 \cdot \max_\ell$. *Advantage band*: longest contiguous run with $A_\ell \geq 2$, $A_\ell = \text{bestRR}_{\text{J-lens}}(\ell)/\text{bestRR}_{\text{logit}}(\ell)$.

Two design choices deserve justification. The *word-like mask*: the average-Jacobian estimator is dominated by corpus-frequency directions, so unrestricted rankings measure whitespace and punctuation rather than content; restricting both targets and ranking pool to word-like tokens

recovers a content metric. The *median*: per-position reciprocal ranks are heavy-tailed and means are hostage to single lucky readouts. Both were registered as a dated amendment (v1.2) after the first model and before any further model was evaluated.

Pre-registration protocol. Every experiment follows the same discipline: hypotheses, metrics, and numeric decision thresholds are frozen in a public register before measurement; amendments are dated; outcomes are reported against the frozen rules whichever way they fall. Where a result below reads "refuted," that is the pre-registered decision rule speaking, not a post-hoc judgement.

4 Three Observables of Residual Computation

The transformer does not compute states; it computes a flow, $h_{\ell+1} = h_\ell + \Delta h_\ell$, with $\Delta h_\ell = \mathrm{attn}_\ell + \mathrm{mlp}_\ell$. We attach one observable to each of three questions — *can it be read? was it written? has it locked?* — and show they answer at different depths.

Definition 4.1 (Transport)**.** The mean input–output Jacobian $\bar{J}_\ell = \mathbb{E}_x[\partial h_L / \partial h_\ell]$ (the fitted J-lens operator). Its off-identity norm $\|\bar{J}_\ell - I\|$ measures the *sensitivity* of downstream computation to the residual at ℓ. Onset depth τ_{decode}: first depth at which J-lens readability becomes non-trivial.

Definition 4.2 (Net injection)**.** $\mu_\ell = U(\Delta h_\ell)$, the DLA contribution of layer ℓ's write to a target direction, U the unembedding with frozen final normalisation. Decomposed by sign, $\mu_\ell = \mu_\ell^+ - \mu_\ell^-$. Centre of mass τ_{write}: the depth around which net injection concentrates.

Definition 4.3 (Commitment)**.** τ_{commit}: the first depth (as a fraction) at which the running argmax of the cumulative readout locks to the final answer.

A fourth, output-alignment marker — the logit-lens *knee* of §3 — sits between injection and commitment. For pythia-410m the four depths are ordered and distinct:

$$\underbrace{\tau_{\text{decode}} = 0.17}_{\text{transport onset}} < \underbrace{\text{band } 0.25\text{–}0.54}_{\text{transport advantage}} < \underbrace{\text{knee} = 0.71 \approx \tau_{\text{write}} = 0.66}_{\text{output formation}} < \underbrace{\tau_{\text{commit}} = 0.96}_{\text{lock}}.$$

The residual becomes J-lens-decodable long before it is written to the output direction, and is written well before the decision locks. These are not three readings of one event; they are a front sweeping across depth (Fig. 3).

5 The Exact DLA Decomposition

Because the unembedding is affine once the final-layer normalisation statistics are frozen at the value taken by the full residual h_L, the output logit decomposes exactly:

$$\boxed{\mathrm{logit}(t) \;=\; \beta_t \;+\; \sum_\ell U(\Delta h_\ell) \cdot e_t} \tag{1}$$

with no approximation. We validate Eq. (1) to $\max|\mathrm{err}| = 7.6 \times 10^{-6}$ (fp32) and confirm it is *norm-aware*: LayerNorm (centre + std; Pythia/GPT-2) and RMSNorm (no centring; Qwen/Llama) require different U; the RMSNorm variant validates to 1.5×10^{-5} in fp32 (bf16 inflates the residual to 0.16, a numerical artefact, not a modelling error).

Lemma 5.1 (Frozen-normalisation affinity)**.** *Fix the final-norm gain/bias and, for RMSNorm, the scale $\sigma(h_L)$ at the full-residual value. Then $t \mapsto U(t)$ is affine, and by linearity of U over the additive residual $h_L = emb + \sum_\ell \Delta h_\ell$, Eq. (1) holds identically. (Statement here; mechanised proof in the formalization paper.)*

Erasure — writing is signed. Eq. (1) makes "writing" signed: a layer can add *and* remove evidence. The cumulative trajectory need not be monotone; on average $\sim 7/24$ layers write *against* the correct target per item in pythia-410m, with measurable drawdown. "The answer is written by layer ℓ" is therefore not well-posed without a sign convention — another way the naive reading conflates distinct events.

6 The Structural Bridge: What the Band Is

The map of §7 will show a J-lens *advantage band* — depths where the Jacobian lens reads future content that the logit lens cannot. It is tempting to call this band a "workspace" of content. The following identity says it is something more modest and more universal: a property of the transport *operator*, not of what the layer happens to hold.

Theorem 6.1 (Recurrence factorisation)**.** *For $h_{\ell+1} = h_\ell + w_\ell(h_\ell)$, the transport Jacobian factorises as $J_\ell = \prod_{k\geq\ell}(I + w'_k)$, hence $J_\ell - I = \sum_{k\geq\ell} w'_k + (\text{cross terms})$.*

Theorem 6.1 says the Jacobian at ℓ is a property of the *future* of ℓ, not of ℓ itself. A layer can therefore have large $\|J_\ell - I\|$ (its residual strongly drives later writes) yet near-zero net injection of its own — and vice versa. This makes the empirical anticorrelation

$$\text{Spearman}\big(\text{advantage } A_\ell,\ \text{net injection } \mu_\ell\big) = \mathbf{-0.80} \quad \text{(pythia-410m; } -0.85 \text{ for gross circulation)}$$

replicates on pythia-2.8b (-0.86) and Qwen3-1.7b (-0.79)

expected, not paradoxical. The band where the J-lens beats the logit lens marks *input hubs* whose content is generically transportable — denoised by the mean Jacobian — not *write sites* where evidence is injected. The methodological headline follows: a J-lens advantage is not evidence of a hidden workspace; it is the mean operator supplying generic transport the raw residual lacks, and the honest control is to match estimation regime, not to read the band as content. A controlled estimation-regime study bears this out: the mean-Jacobian readout saturates as an *estimator* at $N \approx 20$ prompts, and a small fixed target rank ($K^\star = 5$–20) already beats the per-prompt oracle — the J-lens advantage is denoising, not stored content (released as the $S(N)$/oracle sweeps).

Remark 6.2 (Load-bearing identities). Theorem 6.1, together with the Möbius form $\gamma = (1-u)/(1+u)$ used in §12, the tautology $d_{\text{horizon}} = T$, and $\beta = -2u/(1+u)^2$, are stated where they do work and machine-verified in Lean 4/Mathlib in the companion formalization paper (file `Taf/GammaMobius.lean`; 6 theorems, 0 `sorry`s). Proofs are deferred there so that Parts I–III state each result exactly once.

7 The Map: Knees and Bands Across the Ladder

Read through Theorem 6.1, the empirical map is no longer a catalogue of workspaces but a portrait of where each model's transport operator denoises content and where it commits to output. Table 1 is that portrait.

Model	Layers	Knee (frac.)	Advantage band	Max A_ℓ
pythia-70m	6	0.667	none	< 2
pythia-160m	12	0.750	0–0.25	L8: 5×
pythia-410m	24	0.708	0.25–0.54	L14: 3–6×
pythia-1b	16	0.688	none	< 2
pythia-1.4b	24	0.542	0–0.38	L30: 19×
pythia-2.8b	32	0.656	0–0.53	L52: **67**×
Llama-2-7B	32	0.813	— (logit only)	—
Llama-3-8B	32	0.844	— (logit only)	—
Qwen2.5-3B	36	0.917	**0–0.78**	L55: **68**×

Table 1: **The readability map.** Violet intensity in the last column scales with the maximum J-lens advantage; amber marks the late-knee Llama family (logit lens only); grey = no band. Bands migrate toward the input in larger / more heavily trained models.

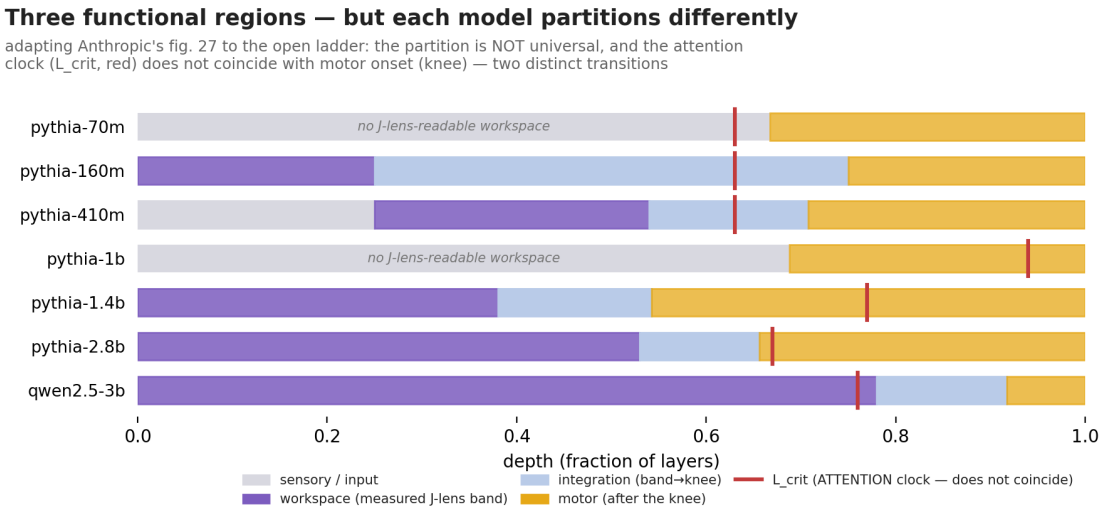


Figure 1: **Functional regions are not universal.** Adapting the three-region reading (sensory / workspace / motor) to the open ladder: each model partitions depth differently, two models expose no J-lens-readable band at all, and the attention-regime clock (L_{crit}, red ticks — a *distinct* transition, §8) does not coincide with the knee.

Three regularities organise Table 1. First, *band position migrates toward the input* as models grow and train longer: the mid-band of pythia-410m becomes the early bands of 1.4b and 2.8b, and Qwen2.5-3B (18T tokens) reads content at layer 1 that its logit lens cannot see at all (68×). Second, the *absence case*: pythia-1b — the one ladder member with d_{head}=256 — shows no band anywhere despite unremarkable knee behaviour; independently, our $\dim_{\text{eff}}$ conservation of W_O (Part II) also fails exactly at d_{head}=256, making wide heads a two-anomaly architectural point. Third, the practical warning for the growing user base of public lenses: *whether a J-lens reads anything at a given depth is model-specific.* A lens silent on one architecture is not evidence of absent content, and a band map like Table 1 should precede any single-model interpretive claim.

8 Refuted Predictors of the Knee

Could the knee's position have been predicted from architecture alone? Three candidates, each frozen with a discriminant design before measurement.

(R2) $L_{\text{crit}} = N \log_\theta(T/2\pi) \cdot (d_{\text{head}}/64)^{1/\sqrt{12}}$. This formula demonstrably governs a *different* depth transition — the attention-regime change, validated causally in Part I. The hypothesis was

that it also fixes the readability knee. The discriminant was chosen in advance: pythia-1b, whose wide heads push the prediction to 0.94 of depth, far from every alternative. Verdict: observed 0.688, within the null band. *Refuted* (Fig. 2). Corollary: the ladder has **two distinct depth transitions** — attention regime (L_{crit}, causal) and output commitment (the knee) — which the refutation cleanly decouples. This is the depth-axis analogue of the distance/depth scale separation of §1.

(R3) A universal constant fraction. The first three models measured (0.667/0.708/0.688) tempted a "≈ 0.69 constant." The full ladder killed it: 0.542 (1.4b) through 0.750 (160m), and the Llama family at 0.81–0.84. *Refuted.*

(R4) The RoPE fraction $\log_\theta(T/2\pi)$ without head correction. Tested with a θ-discriminant pair (Llama-3 vs Llama-2 control). The frozen control-validity rule fired: the control landed at 0.813, outside the pre-registered window $[0.64, 0.75]$, revealing the knee as family-dependent before the discriminant could speak. *Inconclusive by rule.*

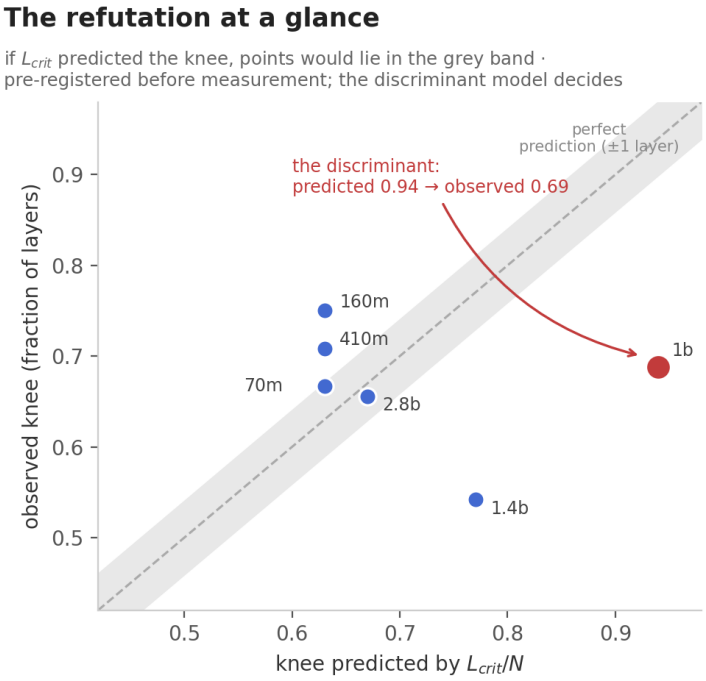


Figure 2: **The refutation at a glance.** Observed knee vs L_{crit}/N prediction; grey band is perfect prediction ± 1 layer. The discriminant model (pythia-1b, red) was chosen *before* measurement precisely because its prediction (0.94) departs most from the null; it landed at 0.69.

The technical moral: within what we measured, *training recipe dominates architecture* for output-commitment depth. Knowing $(\theta, T, d_{\text{head}}, N)$ does not tell you where a family's knee sits; family identity largely does (R1).

9 Cross-Family Chronology: Type A and Type B

The map's "knee is a family property" (R1) is one face of a deeper regularity: the *whole chronology* of the three observables is family-specific while the observables themselves are universal.

The classifier. $\Delta_{\text{prep}} = \tau_{\text{write}} - \tau_{\text{decode}}$ — the interval over which content is linearly decodable before it is written — separates families with no overlap (Table 2).

Family	Model	γ_{obs}	τ_{decode}	Δ_{prep}
Pythia	pythia-410m	0.94	0.17	+0.49
Qwen2.5	qwen2.5-3b	0.94	0.09	+0.75
Qwen3	qwen3-1.7b	0.79	0.63	+0.28
Qwen3	qwen3-4b	0.81	0.69	+0.22
GPT-2	gpt2-small	1.02†	0.82	−0.12

Table 2: **Family chronology.** Blue = Type A (early-decode, $\Delta_{\text{prep}} \geq 0.5$); orange = Type B (late-decode, $\Delta_{\text{prep}} \leq 0.3$). †GPT-2 uses absolute positional encoding, not RoPE — excluded from the γ comparison; a special case (flat residual, high identity baseline). The γ_{obs} column is the *depth-batch* fit; the Part I atlas gives materially different values for the same models (qwen2.5-3b 0.772, pythia-410m 1.022), which is why the γ–chronology direction is inconclusive (§12).

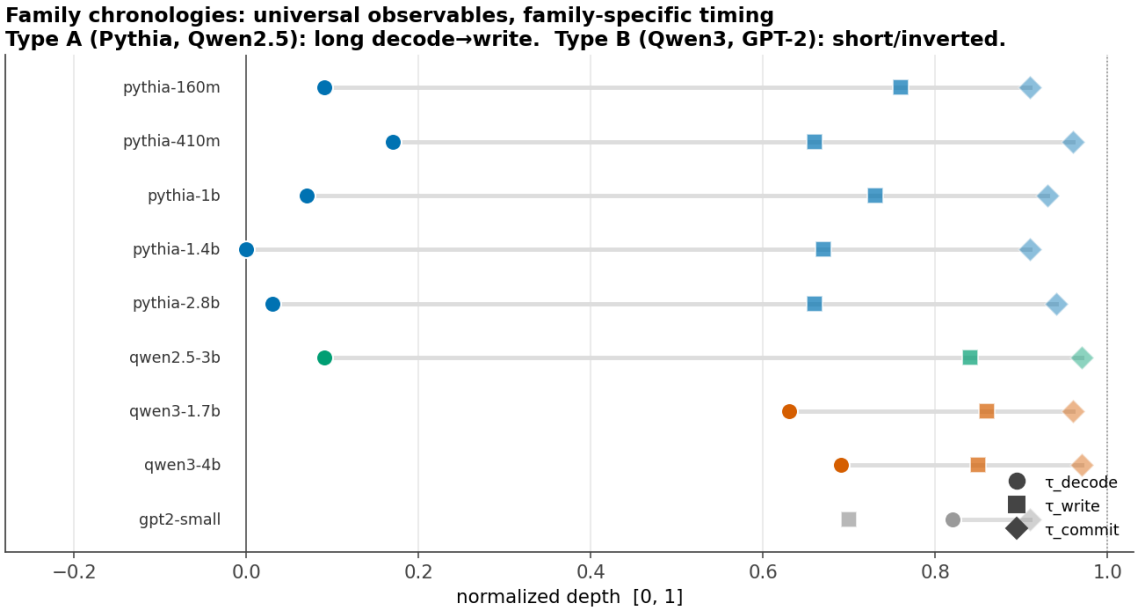


Figure 3: **Family chronologies: universal observables, family-specific timing.** Each row is one model; markers are τ_{decode} (transport onset), τ_{write} (injection centre of mass), τ_{commit} (commitment) on normalised depth. Type A models (Pythia, Qwen2.5) show a long decode→write interval; Type B (Qwen3, GPT-2) a short or inverted one.

τ_{commit} is architectural, not a law. Commitment always falls in the last 1–3 layers because the final layer dominates the write to the unembedding (in pythia-160m it carries 93% of the injection). We therefore do *not* treat τ_{commit} as a deep invariant. ×

Scale (five Pythia). The *structure* is scale-invariant — a three-phase MLP efficiency profile with a mid-depth suppression valley, attention as an efficient positive accumulator, and $\tau_{\text{decode}} < \tau_{\text{write}} < \tau_{\text{commit}}$ throughout — while the *intensity* is scale-dependent: the MLP valley shallows monotonically with N ($-0.66 \to -0.09$), *refuting* a pre-registered "deeper with scale" prediction (R12).

Instrument vs mechanism. A natural alternative — that Type B lacks early decodability because the mean Jacobian approximates the operator badly — is refuted: the operator improves over the identity by a comparable amount in all families (+0.18 to +0.32), and the per-prompt dispersion $\text{Var}_p(J)$ is comparable (Pythia 1.43 vs Qwen3 1.29). The Type A/B difference is real — in *when* the answer becomes linearly decodable, and (via the norm-aware DLA of Qwen3, τ_{write} : $0.66 \to 0.91$) in *where* evidence is injected — not an artefact of the lens.

10 Prompt Frames and Workspace Loading

H13: axiomatic frames in a base model. Does prompt *structure* change what the band carries? We wrapped 150 SciQ questions in two frames — bare (`Question: ... Answer:`) and "axiomatic" (the same content preceded by four explicit rules) — and measured, at the final prompt position, the J-lens rank of the gold answer's first word-like token within pythia-410m's band (L8–13), with the late-layer logit lens as an output-shift control. All rules frozen in advance.

> The rule frame makes band loading *worse*: median paired $\Delta \log_{10}$ rank $= -0.077$, Wilcoxon $p = 4.5 \times 10^{-9}$, band-specific (specificity -0.199 against the late-logit control), $n = 150$.

For a base model, a rule frame is not guidance; it is a competitor for limited transport capacity (Fig. 4). Two follow-ups bound the interpretation: a length-matched neutral filler control (does token count dilute?), and the base/instruct pair below.

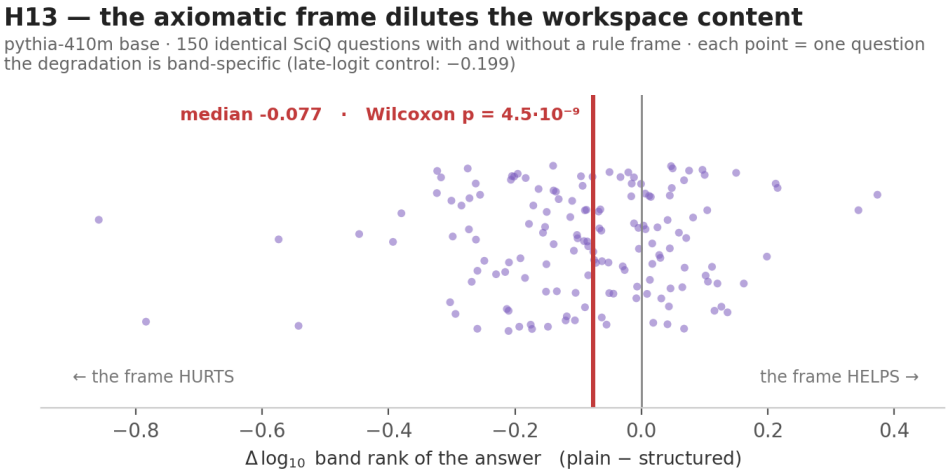


Figure 4: **H13.** Per-item paired difference in band-readable rank of the gold answer (plain − structured), $n = 150$ SciQ items on pythia-410m base. Negative = the axiomatic frame *worsens* loading. Median -0.077, Wilcoxon $p = 4.5 \times 10^{-9}$; band-specific.

H13b: does instruction tuning invert the sign? [pre-registered: Qwen2.5-3B base/instruct, same items, band frozen from the base model] The frozen rules asked whether the base replicates Pythia's negative sign ($\Delta_{\text{band}} \leq 0$) and whether the instruct model flips it ($\Delta > +0.10$ with specificity > 0). **Both fail.** On the Qwen2.5-3B *base* the effect does not replicate — it *reverses* ($\Delta_{\text{band}} = +0.072$, $p = 2.0 \times 10^{-6}$) — so H13's "an axiomatic frame worsens band loading" is **specific to Pythia, not universal**. On the *instruct* model the sign-flip also fails ($\Delta = +0.034 < +0.10$, specificity $-0.155 < 0$, $p = 7.6 \times 10^{-5}$): the clean instruction-tuning story is **null**. A single base/instruct pair is n=1; a second pair is needed before even the non-replication generalises. ×

11 Training-Time Embryology of the Band

When does the band form? We fit J-lens/logit curves on nine training checkpoints of pythia-410m (step 128 to step 64000), with the priority order frozen in advance: (Q1) the formation step; (Q2) whether it coincides with the induction-head window [9] and the Hagedorn transient of Part I; (Q3) knee migration; (Q4) an early-and-strong advantage in the young model.

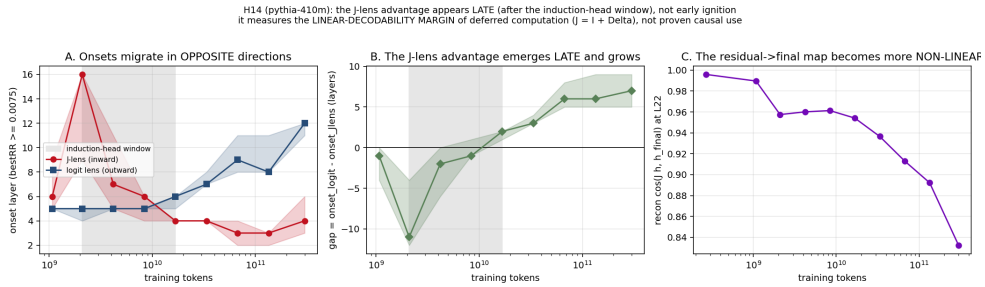


Figure 5: **H14: band embryology.** J-lens advantage by layer across nine pythia-410m checkpoints. *Caveat, pre-registered:* A_ℓ is a ratio against the logit-lens baseline, which itself matures during training; the temporal reading is valid only after the baseline is audited and held fixed (see below).

Confound (declared before interpretation). $A_\ell = \text{bestRR}_{\text{J-lens}}/\text{bestRR}_{\text{logit}}$ is a quotient against a baseline that *itself* matures during training. A rising band can reflect a numerator gain or a denominator collapse. Until the logit-lens baseline is audited and a fixed reference is used, H14 reports the raw advantage curves only, not a causal "the band forms at step s" claim.

This is the depth-axis instance of the Part I discipline: an apparent temporal signal must survive the moving-baseline control before it is read as mechanism.

12 Co-ordination with the Attention Exponent γ

We now close the loop to Part I. The distance-axis exponent γ and the depth-axis chronology are measured on entirely different quantities, yet they co-order.

An apparent co-ordination — and why we do not claim it. Measured with the depth-batch γ protocol (4-bit; the fit used for the chronology models), the four RoPE points (pythia-410m γ=0.94, qwen2.5-3b 0.94, qwen3-1.7b 0.79, qwen3-4b 0.81; all $R^2 \geq 0.97$) rank as $\text{Spearman}(\gamma, \Delta_{\text{prep}}) = +0.8$ and $\text{Spearman}(\gamma, \tau_{\text{decode}}) = -0.8$ (Fig. 6), suggesting $\gamma \uparrow \Leftrightarrow \tau_{\text{decode}} \downarrow$. But the *same models* carry a very different γ under the Part I attention-decay atlas protocol — qwen2.5-3b is $\gamma_{\text{obs}} = 0.772$ (not 0.94) and pythia-410m is 1.022 (not 0.94), a disagreement of up to 0.25 — and under *that* protocol qwen2.5-3b is a *low*-γ model with the *earliest* τ_{decode}, which *reverses* the trend. The direction of the co-ordination is therefore an artefact of the γ protocol; we release both sets of values and **claim no trend** (R10 is inconclusive).

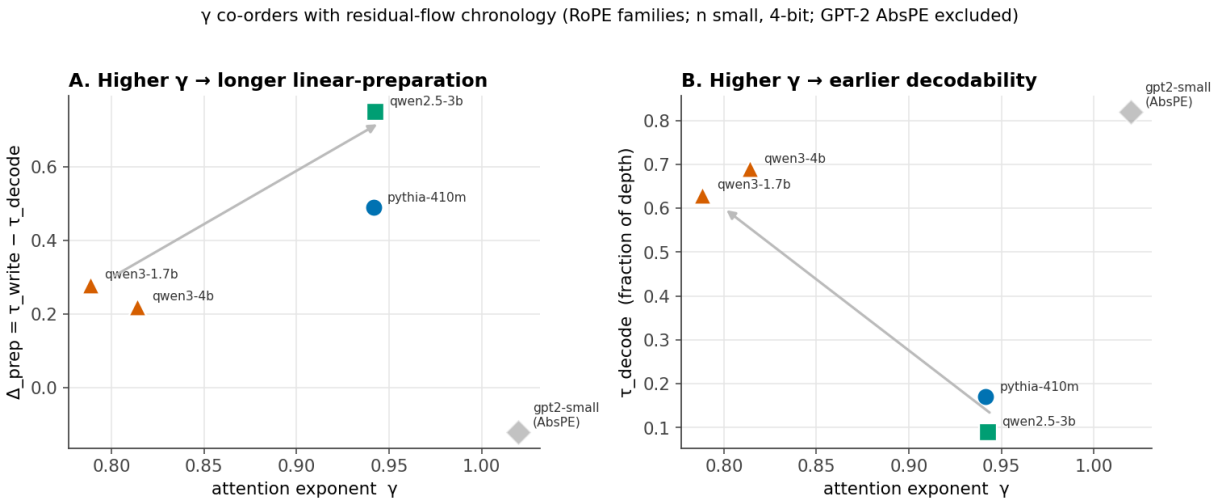


Figure 6: **Apparent γ–chronology co-ordination under the depth-batch protocol** (RoPE families; n small, 4-bit; GPT-2 AbsPE excluded). *A:* higher $\gamma \to$ longer linear-preparation Δ_{prep}. *B:* higher $\gamma \to$ earlier decodability τ_{decode}. The direction *reverses* under the Part I atlas γ (§12); we claim no trend.

A displacement bridge (exploratory). Joining this chronology with the Part I attention-decay atlas suggests the band's *position* tracks not the raw exponent γ but its trained displacement $\delta = \gamma_{\text{obs}} - \gamma_{\text{geom}}$ from the RoPE-geometric baseline: $\delta < 0$ (training delocalised attention beyond geometry) co-occurs with early bands (4/4), $\delta \approx 0$ with absence, $\delta \gg 0$ with the single mid-band model, while band *strength* tracks scale instead (rank $+0.79$ vs $+0.18$ for position; $+0.29$ vs $+0.75$ for strength; $n = 7$). The discriminating instance: two models with nearly identical raw γ (0.748 vs 0.772) sit at opposite phenomenological extremes, and δ separates them ($+0.001$ vs -0.183). [pre-registered: blind out-of-sample test on the gemma-3 / qwen3 harvest]

R10 is inconclusive, not a result. Two γ protocols (the depth-batch 4-bit fit and the Part I attention-decay atlas) disagree by up to 0.25 for the *same* models, and the sign of the $\gamma \leftrightarrow \tau_{\text{decode}}$ co-ordination flips between them. We release both and assert no direction. The displacement bridge above (δ) is offered in the same spirit — suggestive at $n = 7$, not a claim; if anything it shows raw γ does *not* discriminate the chronology, while its trained displacement δ might.

13 Discussion

A band-to-knee regularity (registered, robust to definition). In every strong-band model the band *ends* a fixed distance before the knee: gap = knee − band-end $\in [0.11, 0.21]$, stable across $1.5\times/2\times/3\times$ band definitions (410m 0.167, 1.4b 0.167, 2.8b 0.125, Qwen2.5 0.139 at the registered $2\times$). Read through the observables: generic transport hands off to output formation at a fixed relative distance before commitment. [pre-registered: prediction P2 for the incoming out-of-sample models]

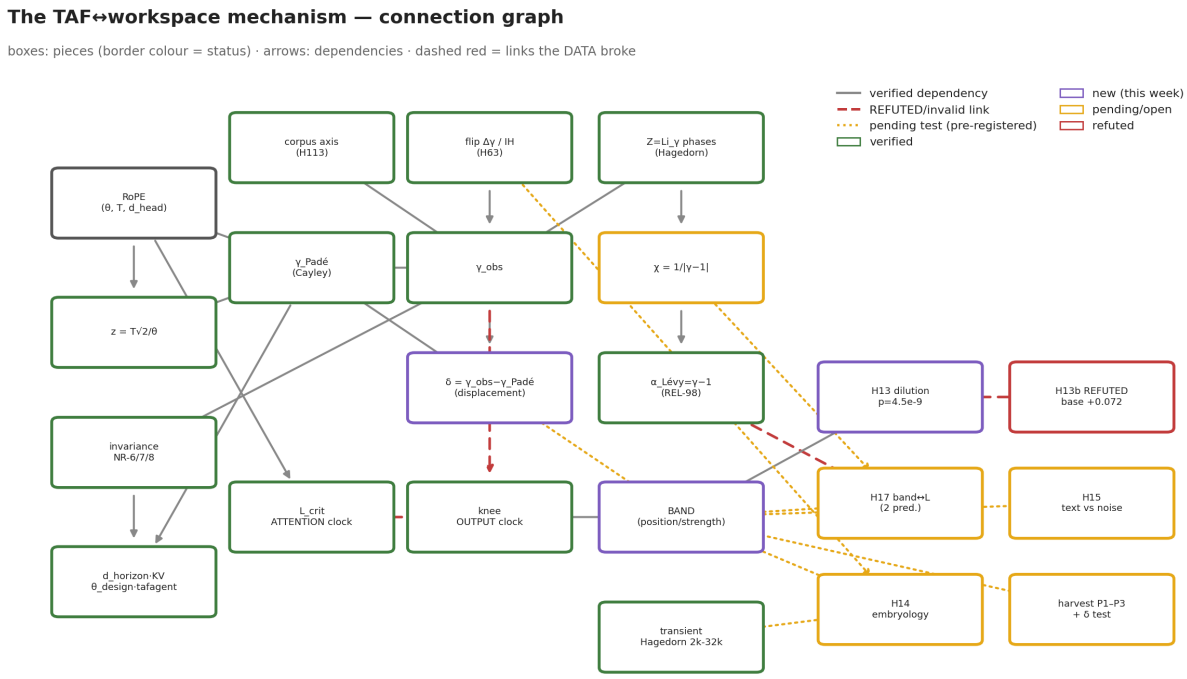


Figure 7: **The mechanism as a graph.** Nodes are the quantities and hypotheses of this paper and Part I (border: green = verified, violet = new, amber = pre-registered pending); arrows are dependencies. Red dashed edges are links *refuted by data*: L_{crit} does not predict the knee (§8), and the Lévy bridge is invalid for $\gamma < 1$ (Part II). The band is the most connected new node.

What the next data adjudicates. The compression view (more training ⇒ more linearly transportable content) and the channel view (wide heads ⇒ overlapping writes cancel the average Jacobian) make *opposite* predictions for heavily-trained wide-head models; the pre-registered P1 test on the gemma-3 / qwen3 harvest decides.

Conceptual result. There is no universal chronology, but there are universal observables. State variables are common; trajectories are not — as in dynamical systems. Interpretability gains a set of observables that *distinguish* how families distribute computation across depth, and a structural reason (Theorem 6.1) not to mistake a transport advantage for a stored workspace.

14 Limitations

(i) The J-lens is correlational — a hypothesis generator; nothing here is causal evidence, and we make no swap/ablation claims. The causal step (intervening on Δh_ℓ, not h_ℓ) is future work. (ii) Readouts are single-token, biased against multi-token concepts. (iii) The knee is threshold-sensitive: the family separation of R1 is clean at the pre-registered $0.5 \cdot \max$ but degrades at 0.3/0.7 (Appendix A), so R1 should be read as "at the registered definition"; the band fares better (its band-to-knee regularity survives 1.5–3×). (iv) Several map cells rest on 1–2 models (notably the d_{head}=256 absence case). (v) The γ co-ordination is $n = 4$ and 4-bit; §12 is a hypothesis. (vi) GPT-2 (AbsPE) is a special case outside the RoPE γ story. (vii) All measurements come from one platform (single 16 GB GPU); dtype and environment are recorded per lens, and we have previously measured large cross-platform discrepancies in adjacent quantities.

15 Released Artifacts and Reproducibility

We release: the fitted lenses (the only *ladder-complete* Pythia suite, 70m–2.8b, plus the Qwen pair, complementary to public collections); all bestRR curves and DLA traces as CSV/JSON; corpus manifests with hashes and seeds; an implementation-ready metric specification (definition, tensor origin, normalisation, extraction, expected values — including the known negatives, so nobody re-derives dead predictors); every pre-registration document with dated decision rules and outcomes; and all fitting/evaluation scripts. Repository: `part3/experiments/workspace_h1/` (running log: `DINAMICA_ESCRITURAS.md`). Load-bearing identities are machine-verified in the companion formalization paper.

A Knee Threshold Sensitivity

Recomputing the knee at $0.3/0.5/0.7 \cdot \max$ on the same curves:

Model	thr 0.3	thr 0.5	thr 0.7
pythia-70m	0.500	0.667	0.667
pythia-160m	0.667	0.750	0.833
pythia-410m	0.667	0.708	0.750
pythia-1b	0.688	0.688	0.812
pythia-1.4b	0.500	0.542	0.583
pythia-2.8b	0.625	0.656	0.938
Qwen2.5-3B	0.889	0.917	0.944
Llama-2-7B	0.688	0.812	0.906
Llama-3-8B	0.750	0.844	0.938

The Pythia–Llama separation holds at 0.5 (gap +0.062) and breaks at 0.3 and 0.7; pythia-2.8b's long plateau makes its knee especially definition-fragile ($0.656 \to 0.938$). We found this by attacking our own claim before a reviewer could; a full-curve statistic (knee v2) is queued.

References

[1] Anthropic Interpretability Team. *Verbalizable Representations Form a Global Workspace in Language Models.* Transformer Circuits Thread, 6 July 2026. https://transformer-circuits.pub/2026/workspace/index.html. Companion code: https://github.com/anthropics/jacobian-lens.

[2] N. Nanda. *A Review of Anthropic's Global Workspace Paper.* LessWrong, 6 July 2026. https://www.lesswrong.com/posts/zFJ3ZdQwrTWE9jT5S/a-review-of-anthropic-s-global-workspace-paper. Replicates the J-lens on Qwen 3.6-27B.

[3] N. Belrose et al. *Eliciting Latent Predictions from Transformers with the Tuned Lens.* arXiv:2303.08112, 2023.

[4] V. Lad, W. Gurnee, M. Tegmark. *The Remarkable Robustness of LLMs: Stages of Inference?* arXiv:2406.19384, 2024.

[5] nostalgebraist. *Interpreting GPT: The Logit Lens.* LessWrong, 2020.

[6] N. Elhage et al. *A Mathematical Framework for Transformer Circuits.* Transformer Circuits Thread, 2021. (Direct logit attribution.)

[7] J. Hewitt, P. Liang. *Designing and Interpreting Probes with Control Tasks.* EMNLP, 2019.

[8] S. Biderman et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. arXiv:2304.01373, 2023.

[9] C. Olsson et al. *In-context Learning and Induction Heads*. Transformer Circuits Thread, 2022.

[10] C. Marín. *Predicting How Transformers Attend: Analytic Power-Law Theory, Phase Transitions, and Practical Compression Tools*. Zenodo, 2026. DOI:10.5281/zenodo.20314038 (Part I).

[11] C. Marín. *Predicting How Transformers Attend, Part II: A Six-Axis Decomposition with the Learned Imprint* $\nu = -1/(2\pi)$. Zenodo, 2026. DOI:10.5281/zenodo.19960573.

[12] C. Marín. *A Machine-Checked Backbone for the Thermodynamic Attention Framework* (Lean 4/Mathlib). 2026 (companion formalization paper).

[13] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, Y. Liu. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. arXiv:2104.09864, 2021.

[14] L. Gao et al. *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv:2101.00027, 2020.

[15] J. Welbl, N. F. Liu, M. Gardner. *Crowdsourcing Multiple Choice Science Questions* (SciQ). arXiv:1707.06209, 2017.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. *Language Models are Unsupervised Multitask Learners* (GPT-2). OpenAI, 2019.

[17] H. Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv:2307.09288, 2023.

[18] A. Grattafiori et al. *The Llama 3 Herd of Models*. arXiv:2407.21783, 2024.

[19] Qwen Team. *Qwen2.5 Technical Report*. arXiv:2412.15115, 2024.

[20] Qwen Team. *Qwen3 Technical Report*. arXiv:2505.09388, 2025.

[21] B. J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

[22] J. Lin, J. Bloom et al. *Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks*. https://neuronpedia.org, 2024.