Projective Incidence Calculus

A semiring-weighted incidence algebra for the transformer decode,
its frozen-model geometry, and its learnable frame

Draft — companion to `spec/PIC_SPEC.md`

Abstract

We give an algebraic account of the transformer next-token decision as a *semiring-weighted incidence calculus over a continuous inner-product geometry—Projective Incidence Calculus* (PIC). Propositions (tokens) are points U_v in a residual space $\mathcal{H} = \mathbb{R}^d$; sources (per-layer contributions, or learned rules) are vectors d_j; the incidence of a source on a proposition is the inner product $\langle d_j, U_v \rangle$; and the decision aggregates incidences through a commutative semiring selected by a temperature T. At $T = 1$ the semiring is the log-semiring and PIC *is* the forward pass (softmax over summed logits); as $T \to 0$ it is the tropical (max-plus) semiring and PIC is the geometry of margins, Laguerre power diagrams, and a head/tail decode certificate, the two ends joined by Maslov dequantization. PIC separates the *frame* (intrinsic geometry of $\{U_v\}$: Gram coherence, Welch floor, packing capacity) from the *decode* (margins, participation ratio, multiplicity, evaluated in the chosen semiring). This separation organizes a three-part program: analysis reads the decode off a frozen frame (a runtime that reconstructs the model's decode exactly), verification proves two-sided packing bounds on the structures (a machine-checked corpus), and learning holds the decode fixed while *retraining the frame*. We state the calculus, its semiring family, and the load-bearing theorems—each tagged as machine-proved, empirical, or open—and we cast the learning problem as frame optimization under explicit soundness invariants.

1 Introduction

A transformer computes a next-token distribution by projecting a residual vector onto an unembedding matrix and taking a softmax. Read literally this is a single linear map followed by a normalization; but the residual vector is itself an *additive* object—a sum of per-layer attention and MLP writes plus the embedding—so the decode logit of token v decomposes as a *sum of inner products*, one per contributing block. We take this decomposition as primitive and ask what algebra it lives in. The answer is a semiring-weighted incidence calculus: tokens are propositions carrying geometric directions, blocks are sources casting weighted votes, and the decision is the semiring aggregate of those votes.

The payoff of naming the algebra is that three otherwise-separate activities become facets of one object. *Analysis* of a frozen model is evaluation of the calculus in the log-semiring; the decomposition is exact, so a runtime can reconstruct the model's own decode argmax with no approximation. *Verification* is proving structural bounds—how many tokens a frame can decode, how many rules can write the logits, what interference overpacking forces—and these are theorems about the calculus's structures, machine-checked in a kernel. *Learning* is the recognition that the frame is a free parameter: one can hold the decode semantics fixed and deform the geometry so

that more structure becomes retrievable. PIC is the shared vocabulary that lets a definition in one activity be a hypothesis in another.

This paper defines PIC (§2–§4), separates frame from decode (§5), collects the proved theorems as a two-sided packing picture (§6), and casts frame learning as constrained optimization (§7). Throughout we tag each claim [PROVED, I-ORCA] (machine-checked), [EMPIRICAL] (measured), or [OPEN].

2 Core structures

Definition 2.1 (Frame and Gram kernel)**.** Let V be a finite set of *propositions* (tokens), $|V| = n_p$, and let $\mathcal{H} = \mathbb{R}^d$ be the residual space. A *PIC frame* is a map $U : V \to \mathcal{H}$, $v \mapsto U_v$. It induces the *Gram kernel* $G(v, w) = \langle U_v, U_w \rangle$ and, after row normalization $\bar{U}_v = U_v / \|U_v\|$, the *cosine Gram* $\rho(v, w) = \langle \bar{U}_v, \bar{U}_w \rangle$.

The Gram kernel supplies the continuous, non-truth-functional overlap between propositions: tokens are not independent atoms but directions with measurable coherence.

Definition 2.2 (Sources, incidence, monomial, polynomial)**.** A *source set* S ($|S| = J$) assigns to each source j a contribution $c_j : V \to R$ valued in a semiring carrier R. The *residual reading* fixes a vector $d_j \in \mathcal{H}$ per source and sets the *incidence* $c_j(v) = \langle d_j, U_v \rangle$. The two semiring operations play different roles: $\otimes$ *combines sources within a token*, giving a *monomial*

$$L(v) = \Big(\bigotimes\nolimits_{j \in S} c_j(v) \Big) \otimes b_v = \sum_{j \in S} \langle d_j, U_v \rangle + b_v \qquad (T\text{-independent});$$

while $\oplus_T$ *combines tokens*, giving the decode *polynomial*

$$Z_T = \bigoplus\nolimits_{T,\, v \in V} L(v) = \bigoplus\nolimits_{T,\, v \in V} \bigotimes\nolimits_{j \in S} c_j(v) \otimes b_v.$$

Thus each token is a monomial and the decode is a $(\oplus_T, \otimes)$ polynomial whose monomials are the tokens and whose variables are the sources—the tropical-polynomial picture of a ReLU network. The decision reads off Z_T: at $T \to 0$, $Z_0 = \max_v L(v)$ and the decode is $\arg\max_v L(v)$; at $T = 1$, $Z_1 = \mathrm{logsumexp}_v L(v)$ and $p(v) = \exp(L(v) - Z_1)$. The monomials are shared across T; only $\oplus_T$ changes (Lemma 4.1).

In the transformer instance, the sources are the *DLA blocks*: an L-layer model has $n_b = 2L + 1$ blocks (one attention and one MLP/expert write per layer, plus the embedding), and d_j is block j's additive contribution to the residual at the decode position. The incidence $c_j(v) = \langle d_j, U_v \rangle$ is precisely the per-block contribution a decomposing runtime emits, and its source-sum reconstructs the model's decode logit.

Native notation. PIC borrows its carriers (semiring, Gram kernel, Welch bound, power diagrams) so as to inherit their theorems, but three things are native and get their own notation. The *incidence arrow* $j \triangleright v := \langle d_j, U_v \rangle$ names the one new primitive—provenance as an inner product in a learnable frame. The *coalition bracket* $[\![P]\!](v) := \bigotimes_{j \in P} (j \triangleright v)$ is the partial monomial of a coalition $P \subseteq S$ (no T: sources add). The *margin turnstile*

$$P \vdash_\gamma v \quad :\Longleftrightarrow \quad [\![P]\!](v) - \max_{w \neq v} [\![P]\!](w) \;\geq\; \gamma$$

("coalition P decides v with margin $\geq \gamma$"; write $P \vdash v$ for $\gamma = 0$) collapses the decision regimes to one line each: *retrieved* $\exists j.\ \{j\} \vdash_\gamma v$; *composed* $S \vdash v$ but $\forall j.\ \neg(\{j\} \vdash v)$; *irreducible* $S \vdash v$ and $\forall P \subsetneq S, P \neq \emptyset.\ \neg(P \vdash v)$.

3 The semiring family

PIC is parameterized by a temperature $T \geq 0$ selecting a commutative semiring $(R_T, \oplus_T, \otimes, 0, 1)$ with a fixed product $\otimes = +$ (incidences along a coalition add) and a T-dependent sum:

$$a \oplus_T b \;=\; T\,\log\!\big(e^{a/T} + e^{b/T}\big), \qquad a \oplus_T b \xrightarrow[T\to 0^+]{} \max(a, b).$$

T	semiring	$a \oplus_T b$	reading
1	log-semiring	$\log(e^a + e^b)$	probabilistic incidence → softmax
$\to 0^+$	tropical (max-plus)	$\max(a, b)$	argmax, margins, power diagrams
general	deformed	$T \log \sum e^{\cdot/T}$	temperature-controlled decoding

The two ends are joined by *Maslov dequantization*: the probabilistic calculus (the model in operation) and the geometric calculus (the proofs) are limits of one deformation, not different objects.

Definition 3.1 (The complete semiring)**.** For each $T \geq 0$, $R_T = (R, \oplus_T, \otimes, \mathbf{0}, \mathbf{1})$ has carrier $R = \mathbb{R} \cup \{-\infty\}$, sum $a \oplus_T b = T \log(e^{a/T} + e^{b/T})$ with $\oplus_0 := \max$, product $a \otimes b = a + b$, sum identity $\mathbf{0} = -\infty$, and product identity $\mathbf{1} = 0$. It satisfies the semiring axioms: $(R, \oplus_T)$ is a commutative monoid with identity $-\infty$; $(R, \otimes)$ is a commutative monoid with identity 0; $\otimes$ distributes over $\oplus_T$; and $\mathbf{0}$ annihilates ($-\infty \otimes a = -\infty$). It is commutative, and idempotent exactly at $T = 0$ ($a \oplus_0 a = a$ but $a \oplus_1 a = a + \log 2$)—the algebraic difference between geometry and probability.

The distributive law $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$ is the load-bearing one: it rewrites a sum-of-maxes (a ReLU stack) as a max-of-sums (a tropical polynomial). It and the two monoid laws are machine-checked at $T = 0$ [PROVED, I-ORCA] (`TropicalSemiring.thy`); for general T the distributive law is the log-domain identity $a + T \log(e^{b/T} + e^{c/T}) = T \log(e^{(a+b)/T} + e^{(a+c)/T})$.

4 PIC as a calculus

Syntax. Sources are weighted clauses; a *derivation* of v is a finite *coalition* $P \subseteq S$ whose incidences combine, under $\otimes$ within the coalition and $\oplus_T$ across alternatives, to support v.

Semantics. Semiring *provenance* in the style of Green–Karvounarakis–Tannen: the value of v is an element of R_T accumulated over derivations, with the Gram kernel supplying the continuous overlap a Boolean provenance lacks. The frozen-model reading is the log-semiring provenance; the geometric reading is the tropical provenance.

Decision rules (in the native turnstile of Definition 2.2).

- *Aggregate:* $L(v) = [\![S]\!](v) \otimes b_v$; decode $\arg\max_v L(v)$.
- *Retrieved:* $\exists j \in S.\ \{j\} \vdash_\gamma v$ (a singleton decides; high μ_t).
- *Composed:* $S \vdash v$ but $\forall j \in S.\ \neg(\{j\} \vdash v)$ (no singleton; $\mu_t = 0$).
- *Irreducible:* $S \vdash v$ and $\forall P \subsetneq S, P \neq \emptyset.\ \neg(P \vdash v)$.

Soundness anchors.

- *Log-semiring ⇒ transformer* [EMPIRICAL]. With inner-product incidences in the log-semiring, the calculus is next-token behaviour: a decomposing runtime reconstructs the model's decode argmax with measured reconstruction 1.00 across architectures (rotary, NeoX, mixture-of-experts, latent attention). PIC is the forward pass rebracketed, not an approximation.
- *Tropical ⇒ power diagram.* As $T \to 0$, $\arg\max_v \langle r, U_v\rangle + b_v$ is Laguerre (power-diagram) cell membership over sites $\{U_v\}$ with weights $\{b_v\}$; the capacity theorems of §6 are the geometry of that diagram.

The two anchors are consistent only because of a fact native to the T-family:

Proposition 4.1 (Decode temperature-invariance)**.** *For every $T > 0$, $\arg\max_v L(v)$ is the same token, equal to the tropical ($T \to 0$) decoder; only the normalized confidence $p_T(v) = \exp\big((L(v) - Z_T)/T\big)$ and the partition $Z_T = \bigoplus_{T,v} L(v)$ depend on T.*

Proof. The monomials $L(v) = [\![S]\!](v) \otimes b_v$ carry no T (sources combine by $\otimes = +$). T enters only the cross-token aggregator $\oplus_T$, which is strictly order-preserving in each argument with $\arg\max = (\oplus_0)$-selector; hence the arg max of the fixed vector $(L(v))_v$ is T-independent. $\oplus_T$ reweights how much the winner beats the field, never which token wins. □

This is the rigorous bridge between the anchors: the *proofs* run at $T \to 0$ (tropical margins, power-diagram capacity, head/tail) and the *model* runs at $T = 1$ (softmax), yet they decode the same token because they share the monomials. Every margin/decode theorem of §6 proved tropically therefore transfers verbatim to the operating model's argmax; the T-dependent part is exactly the confidence (entropy, margins), the locus of the forge tax.

5 Encode, frame, decode

PIC is a sandwich of three roles: the *encoder* writes the residual, the *frame* is the shared dictionary, and the *decoder* reads it out. The encode side has an explicit model (machine-checked, `PIC_Core.thy`).

Definition 5.1 (Encoder)**.** A PIC encoder fixes a finite rule set K, fixed write directions $a : K \to \mathcal{H}$, and input-dependent gates $g : K \to (X \to \mathbb{R})$. On input x the residual is the gated sum $\mathrm{enc}(x) = \sum_{k\in K} g_k(x)\, a_k$, and the logit is $L(v; x) = \langle \mathrm{enc}(x), U_v\rangle + b_v = \sum_{k\in K} g_k(x)\langle a_k, U_v\rangle + b_v$.

This expresses any transformer block: a block's residual write is $W_{\text{out}} \cdot$ (input vector)—fixed output columns a_k gated by an input coefficient $g_k(x)$—for MLP and attention alike. The gate $g_k(x)$ is left *uninterpreted*: PIC formalizes the linear write/read algebra ($\langle a_k, U_v\rangle$ is rule k's incidence) and leaves the nonlinear gate as the un-modelled forge tax, so the encoder sharpens rather than enlarges the calculus. Three facts are proved [PROVED, I-ORCA]: the forward pass is one PIC term ($L(v; x) = \sum_k g_k(x)\langle a_k, U_v\rangle + b_v$); every encoded residual lies in $\mathrm{span}\{a_k\}$ of dimension $\le \min(|K|, d)$ (routing rank, now a property of the encoder— §6); and every fixed-input slice *is* a source model (Def. 2.2) with $d_k = g_k(x)a_k$. The remaining two roles split every other quantity:

Every other PIC quantity is either *frame-side* (a function of $\{U_v\}$ alone—intrinsic, label-free) or *decode-side* (a function of the aggregate L and a target, evaluated in the chosen semiring).

Decode-side. margin $m(t) = L(t) - \max_{v\neq t} L(v)$; participation ratio $\mathrm{PR} = \big(\sum_j |w_j|\big)^2 / \sum_j w_j^2$ over incidence magnitudes $w_j = |c_j(t)|$ (the effective number of contributing sources); multiplicity $\mu_t = \#\{j : \arg\max_v c_j(v) = t\}$ (unanimity; $\mu_t = 0$ is composed); ambiguity η^2 (between/total variance of candidate firing on the lowest-margin slice).

Remark 5.1 (Architecture-fair PR). When sources carry a common-mode component identical across candidates (e.g. a LayerNorm/embedding offset), it cancels in the argmax but inflates raw PR. The discriminative count centres each source across candidates first, $\mathrm{PR}\big((c_j(t) - \overline{c_j})_j\big)$; raw and centred agree for RMSNorm frames and diverge for LayerNorm. Use the centred form for cross-architecture comparison.

Frame-side. Gram coherence $\max_{v \neq w} |\rho(v, w)|$; frame potential $\mathrm{FP} = \mathrm{mean}_{v \neq w}\, \rho(v, w)^2$; Welch floor $W = (n_p - d)/\big(d(n_p - 1)\big)$ for $n_p > d$, else 0. Always $\mathrm{FP} \geq W$; the ratio FP/W reports frame quality, equal to 1 at an equiangular tight frame.

6 The two-sided packing theorems

The proved results split into a *frame side* (how many tokens a geometry can decode) and a *generator side* (how many rules can write the logits, and what interference overpacking forces). This split is not a coincidence: it is the *two arities* of the decode polynomial Z_T of Definition 2.2. A $(\oplus, \otimes)$ polynomial has a monomial-count and a variable-rank, and PIC bounds each—the frame side bounds the **monomials** (tokens that can be cleanly decoded, the $\oplus$-arity) at $(1 + 2\rho/\gamma)^d$, the generator side bounds the **variables** (independent sources, the $\otimes$-arity) at $\min(M, d)$, and Welch is the obstruction that appears exactly when the variable side is overpacked. "Two-sided packing" is thus a simultaneous bound on a polynomial's monomial-count and its variable-rank. The frame side is astronomically slack for realistic d; the binding side is the generator—superposition is forced on the writers, not on the readouts.

6.1 Frame side: decode capacity (monomial count)

Call v *γ-decodable* if some unit-ball residual r ($\|r\| \leq 1$) makes it win by margin γ: $\forall w \neq v,\ \langle r, U_w\rangle + b_w + \gamma \leq \langle r, U_v\rangle + b_v$.

Theorem 6.1 (Margin separation [PROVED, I-ORCA])**.** *If v and w are both γ-decodable then $\gamma \leq \|U_v - U_w\|$.*

Proof sketch. Add the two witness inequalities; the biases cancel, and Cauchy–Schwarz with $\|r_v - r_w\| \leq 2$ gives the bound. (Machine-checked in `DecodeCapacity.thy`.) □

Corollary 6.2 (Head capacity [PROVED, I-ORCA])**.** *Any set S of γ-decodable tokens is a γ-code: $\|U_v - U_w\| \geq \gamma$ for distinct $v, w \in S$. Hence $|S| \leq (1 + 2\rho/\gamma)^d$ with $\rho = \max_v \|U_v\|$.*

This is the cell-capacity half; the bound is astronomically slack for realistic d, so this side rarely binds.

6.2 Generator side: routing rank and Welch interference (variable rank)

M rules write logit adjustments along fixed readout vectors $\{a_1, \ldots, a_M\}$ through input-dependent gates $h_k(z)$.

Theorem 6.3 (Routing superposition [PROVED, I-ORCA])**.** $\sum_{i \in I} h_i\, a_i \in \mathrm{span}(a_I)$ *and* $\dim \mathrm{span}(a_I) \leq |I|$. *Every rule adjustment lives in a subspace of dimension $\leq \min(M, d)$; superposition is forced when the number of routing decisions exceeds M.*

Theorem 6.4 (Routing Welch [PROVED, I-ORCA]). *For n unit-norm routing features in M coordinates, $\sum_{i\neq j}\langle f_i, f_j\rangle^2 \geq n(n-M)/M$, strictly positive exactly when $n > M$. Cross-talk-free routing requires $M \geq n$.*

Remark 6.5 (The missing link [OPEN]). The step "routing interference ⇒ margin degradation" is *not* a kernel theorem. Trained models pack features at ≈ the Welch floor with healthy margins; the coherence-to-margin implication is measured and mild, not proved.

6.3 The decode certificate and its robustness

With $\text{decode}(L, S) = \max_{v\in S} L(v)$:

Theorem 6.6 (Head/tail certificate [PROVED, I-ORCA]). $\text{decode}(L, H\cup T) = \max(\text{decode}(L, H), \text{decode}(L, T))$, *and if* $\text{decode}(L, T) \leq \text{decode}(L, H)$ *then the full decode equals the head decode and its argmax lies in H. If instead the head does not dominate, the decision lies in the tail—the explicit, uncertified residue.*

Theorem 6.7 (Margin certificate, threshold tight [PROVED, I-ORCA]). *If $|L'(v) - L(v)| \leq \delta$ for all v and $m(t) > 2\delta$, then t remains the strict argmax under L'. The constant 2δ is tight (a $m(t) = 2\delta$ token can be tied by a δ-perturbation), and the certificate is silent for $m(t) \leq 2\delta$—exactly the small-margin residue.*

Theorem 6.6 is the algebra behind the empirical observation that a per-position sparse head (a handful of late-layer blocks) reproduces the decode *when* it dominates the tail; harder tokens push mass into the residue.

6.4 Irreducibility

For a composition matrix c : sources × outcomes $\to \mathbb{R}$, S *decides* t if $\sum_{j\in S} c(j, v) < \sum_{j\in S} c(j, t)$ for all $v \neq t$; μ_0 holds if no singleton decides t; S is *irreducible* for t if it decides t and no proper nonempty subset does.

Theorem 6.8 ($\mu_0 \not\Rightarrow$ irreducible [PROVED, I-ORCA]). *There is an explicit instance with μ_0 (no single source decides t) in which a proper pair already decides t; and a separate explicit fully-irreducible triple. Hence "no singleton decides" does not imply "all sources needed".*

This is the precise statement of $\mu_t = 0 \not\Rightarrow$ irreducible: low multiplicity is necessary but not sufficient for genuine irreducibility. Whether a composed token is irreducible under *every* admissible frame—the real necessity question—is [OPEN].

7 Frame learning as constrained optimization

Fix the decode semantics (log-semiring L, softmax) and optimize the frame U—equivalently, deform the Gram kernel—so that more propositions become retrievable:

$$\mathcal{L}(U) = \underbrace{\text{NLL}(L; t)}_{\text{decode, data}} + \lambda_m \underbrace{\text{ReLU}\big(\gamma^\star - m(t)\big)}_{\text{decode, margin}} + \lambda_f \underbrace{\text{FP}(U)}_{\text{frame, structure}} .$$

The decode-side terms anchor the aggregate to the targets and push the worst-competitor margin past $\gamma^\star$; the frame-side term decorrelates propositions toward the Welch floor. A Generate–Gate–Refine loop wraps the gradient step: *generate* candidate sources (random or frame-aligned, or real DLA blocks loaded through the analysis seam), *gate* them by the PIC diagnostics (η^2, activation variance, firing decorrelation), *refine* by a step on $\mathcal{L}$.

Soundness invariants. The loop should preserve: (S1) *decode soundness*—the refined frame still realizes a valid log-semiring decoder on held data; (S2) *margin monotonicity*—the certified margin mass $\#\{t : m(t) > 2\delta\}$ of Theorem 6.7 does not decrease; (S3) *frame admissibility*—$\mathrm{FP}(U) \geq W$ is structural, so a report must target $\mathrm{FP}/W \to 1$ and never claim to beat the floor. These are natural targets for learning-side soundness theorems.

What is established [empirical]. On synthetic benchmarks no theory-guided knob (frame-regularization, allocation/targeting, propose-score-select, coherence-regularization) beats plain SGD on $\mathcal{L}$; PIL training *is* gradient descent on the frame, cheap only as a shallow head on frozen sources. The win that exists is *retraining the frame/subspace*—the movable side of the boundary—rather than fitting an encoder into a frozen dictionary, consistent with the verification-side result that frozen compression hits a $\Theta(d)$ floor a rank-bottlenecked retrained update clears.

The logic-programming reading. Structurally PIC is a semiring-weighted logic program: the turnstile $P \vdash_\gamma v$ is a weighted rule (body coalition derives head proposition with margin γ), the answer is the least fixpoint of the immediate-consequence operator $T_P(I) = \{v : \exists P \subseteq I,\ P \vdash v\}$ evaluated in R_T, and the query is the decode. The semiring chooses the logic—$T = 0$ is shortest-path/Viterbi (min-cost) logic, $T = 1$ is probabilistic (ProbLog-style) logic, Boolean is classical Datalog—with PIC's one native move that the fact weights are *derived from geometry*, $j \triangleright v = \langle d_j, U_v \rangle$, rather than given. The demand-closure theorem (§6, $\mathrm{lfp}(\mathrm{restrict}\, T\, D) = \mathrm{lfp}\, T \cap D$ [PROVED, I-ORCA]) is exactly the soundness of the magic-sets/demand optimization for such a language, and the margin certificate bounds how far a clause weight may drift before an answer flips. The frame/decode separation is its type discipline: frame-side clauses (what overlaps what) plus decode-side queries (margins, multiplicity).

8 An empirical parameter the proofs do not pin down

The packing exponent that binds in practice is an effective rank $\tau^\star$, not the ambient d. A natural conjecture, $\tau^\star \approx \min(e^H, d)$ (effective output support sets the rank), is *refuted*, robustly, on a fourteen-model sweep to 72B (Pythia 14m–12b, Qwen 0.5B–72B; reconstruction 1.00 throughout). The architecture-fair effective decode-block count r_{eff} *anti*-correlates with e^H (Pearson -0.54) and associates with *capacity* (hidden Spearman $+0.63$, n_b/depth $+0.58$); the ratio r_{eff}/e^H climbs monotonically $0.07 \to 1.74$ across the ladder—larger, more confident models use *more* effective blocks, not fewer.

But there is *no clean functional form*: it is **architecture-dependent** (Figure 1A). Within Pythia, $r_{\text{eff}} \approx (0.4\text{–}0.5)\, n_b$ and grows with depth ($8 \to 30$ as $n_b : 13 \to 73$); within Qwen, r_{eff} is *flat* ($\sim$ 11–17) while the block *fraction* shrinks $0.27 \to 0.10$ from 0.5B to 72B. (The once-tempting "≈12 scale-invariant blocks" reading was a Qwen raw-PR artifact—Qwen's raw PR is flat 7.7–13.8 to 72B, while Pythia's grows to 37.) A genuine but modest *within-model* effect survives (higher-entropy tokens use somewhat more blocks; Spearman $\approx$ 0.2–0.47 in capable models)—a second-order grain, not the cross-model driver. The same architecture split shows up in a *wholly independent, causal* probe—the convergence depth $k^\star/L$ (the smallest layer prefix whose recompute already reproduces the decode), which declines with scale in Pythia but not Qwen (Figure 1B)—so it is a property of the models, not an artifact of the attribution measure. So $\tau^\star$ is capacity-associated rather than entropy-bound, with *no universal law* across architectures; its functional form is [OPEN]. We keep it as the sharpest example of the discipline the tags enforce: a plausible law, measured to 72B, and

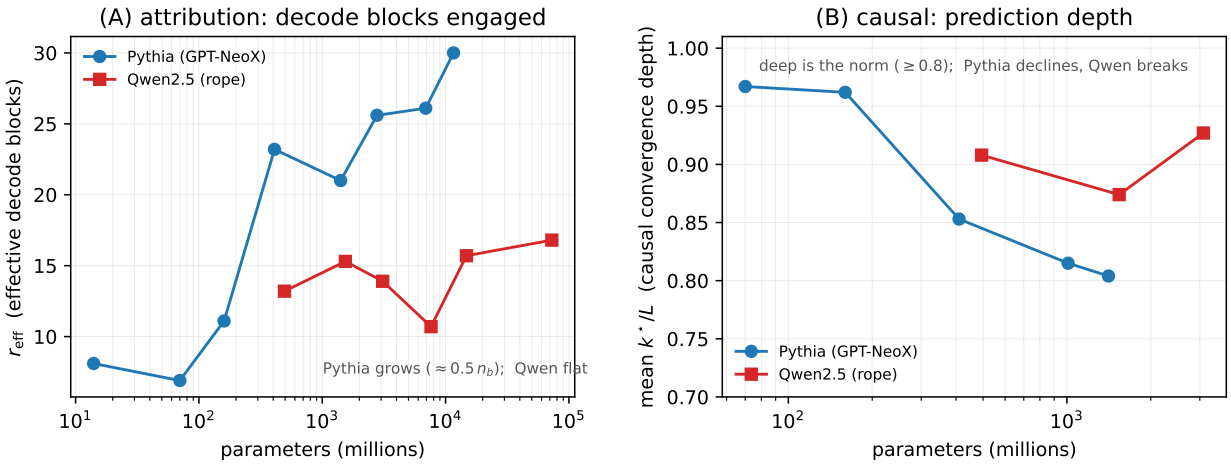

Figure 1: Two independent probes of how much of the model the next-token decode engages, both splitting Pythia from Qwen. **(A)** *Attribution*: the effective number of decode blocks r_{eff} across the $\tau^\star$ sweep (14m–72B, reconstruction 1.00): within Pythia it grows ($\approx 0.5\, n_b$), within Qwen it is flat (∼11–17) as the block fraction shrinks to 0.10. **(B)** *Causal*: the convergence depth $k^\star/L$ (`--converge-depth`) — the smallest layer prefix whose recompute reproduces the decode. Deep computation is the norm ($k^\star/L \geq 0.8$); Pythia's fraction declines with scale while Qwen's does not. That two unrelated measures — one attributional, one causal — agree on the architecture split is the load-bearing point: there is no universal scaling law for decode engagement.

reported by what the data support—including the architecture split—rather than by what would have been tidy.

Provenance of claims

[PROVED, I-ORCA] statements are machine-checked in the i-orca Isabelle corpus (`tropical/{TropicalSemiring,He superposition/{Welch,RoutingWelch,Superposition}.thy`, `provable_opt/ProvableOpt_Common.thy`, `fieldrun/separation/Separation.thy`). [EMPIRICAL] statements are measured by the fieldrun runtime and the pil experiments. [OPEN] statements are flagged hypotheses. See `spec/PIC_SPEC.md` for the operational definitions and file-level pointers.