What a Transformer Retrieves and What It Computes: A Measured Theory of the Composition Core in Three Interpretations

J. Allan Scott
Independent Researcher
allan@jallan.scot

Abstract

We decompile small transformer language models into a flat *retrieval* store (n-gram, induction, grammar) plus a *composition* kernel. We then ask of every next-token decision whether the model is *retrieving* a symbolic rule, *selecting* within a rule-proposed set, or *computing* something new. Across two Qwen2.5-0.5B models the labour is roughly 25% retrieved, 60% selected, 15% composed. The same retrieval-dominated structure (retrieved + selected $\approx$ 85–92%) recurs across the Pythia ladder (70M–1B). The $\sim$15% composed is a small but irreducible core of genuinely computed tokens (the *irreducible computation*) that grows with scale. The mechanism is not selection by magnitude: the logit is a uniformly distributed $\sim$45-way additive sum (participation ratio PR ≈ 45, route-invariant). Retrieval and computation instead separate on two near-orthogonal, measurable axes: a power-diagram margin and a single-circuit readout multiplicity μ_t. Composed tokens are *emergent* (the argmax of the sum but of no summand) and *causally fragile*, with a diffuse, order-$1/\mathrm{PR}$ repair. We then give one theory in three interpretations, a single object at two temperatures: as a probabilistic *logic* ($T=1$) it is Bundy's incidence calculus with the Gram kernel $G_{vw} = \langle U_v, U_w \rangle$ as the carrier of non-truth-functionality; as a *geometry* ($T=0$) it is the tropical variety of the decision surface; and as a *computation* it is a semiring-weighted Datalog program whose two temperatures are the softmax and the greedy decode. The retrievable fragment exports to compact, verifiable Datalog; the irreducible computation is the dense-Gram region that no lookup table reproduces. Every quantitative claim is reproducible with the openly available `fieldrun` toolchain (Apache-2.0).

1. Introduction

A large language model, asked for the next token, sometimes does something a lookup table could do and sometimes does not. The received picture treats "`The capital of France is`" as the paradigm of the first case: the model has memorised a fact, and the only hard question is where it is stored [12, 19]. But the route a decision actually takes is more subtle than this intuition, and is not a property of the fact alone. Knowing "Paris" is one thing; *producing* it as the immediate next token under a constraint is another. On the models we study, even this canonical example is not a clean lookup: its routing is a mixture, and constraining the answer to a single word pushes the decision into the *computed* regime (Section 4). Many tokens, in short, are not in any table the model could be said to consult: they are computed on the fly, assembled from a distributed interaction of partial contributions. The central empirical question of this paper is how to tell retrieval and computation apart *operationally*, on a real model, token by token, and what the mechanism of each looks like.

We attack the question with a *decompiler*. The `fieldrun` toolchain splits a small transformer into two halves that together reproduce it: a flat *retrieval* store (n-gram successor tables, induction copy-rules, a closed-class grammar skeleton) that reconstructs roughly half of the model's predictions with no linear algebra at all, and a *composition* kernel that runs the attention+MLP forward pass as plain matrix multiplies over flat weight arrays [24]. Every weight tier is validated by exact top-1 agreement against a reference implementation, so the decomposition is faithful rather than approximate. Given the retrieval store's candidate set for a context, we classify the model's argmax token t into three routes: RETRIEVED (a single store idiom's top-1 is t, a pure symbolic lookup), SELECTED (t is in the candidate set but is no idiom's top-1; the set contains the answer, the choice within it is made elsewhere), and COMPOSED (t is in no rule's output, the *irreducible computation*, genuinely computed).

The headline measurement is stable and, we will argue, theoretically loaded. Across two Qwen2.5-0.5B models [25] the model's labour divides roughly 25% RETRIEVED, 60% SELECTED, 15% COMPOSED, and the same retrieval-dominated structure (RETRIEVED+SELECTED $\approx$ 85–92%) recurs on all four rungs of the Pythia ladder [3]. Composition is mostly *disambiguation within a retrieved set*, not generation from nothing; only the $\sim$15% irreducible computation is from scratch, and it grows monotonically with model scale against a fixed store. Three further facts sharpen the picture and resist the obvious explanations. First, the route split is *not* a magnitude distinction. The predicted logit is a uniformly distributed $\sim$45-way additive sum (participation ratio PR ≈ 45, route-invariant), so no circuit ever "decides" a token by dominating the sum. This holds even for a token that a single n-gram rule reproduces perfectly. Second, the split separates cleanly on two other, near-orthogonal axes that we can measure *exactly*: a decision *margin* that is the Euclidean distance from the resid-

ual stream to the nearest unembedding power-diagram facet, and a single-circuit *readout multiplicity* μ_t. Third, composed tokens are *emergent* (the argmax of the sum but of no summand, $\mu_t \approx 0$) and *causally fragile* under ablation, with a repair mechanism that is provably diffuse rather than localisable.

These measurements ask for a theory, and the theory we give is one object seen from three sides.[1] The thread that ties them is older than transformers: it is Alan Bundy's *incidence calculus* [5], invented in 1985 precisely to reason with probabilities that are *not truth-functional* by tracking the underlying *incidences* (the worlds in which a proposition holds) and recovering probability as a measure on them. We make three claims, each an interpretation of the same structure:

1. As a probabilistic **logic** at temperature $T = 1$ (the soft accumulation and the recovered output measure), the core is incidence calculus generalized in the two ways the substrate forces. Incidences become *signed measures in an inner-product space* rather than sets, and Boolean $\cap/\cup$ become a *weighted threshold*. The Gram kernel $G_{vw} = \langle U_v, U_w \rangle$ of the unembedding frame is the explicit operator carrying non-truth-functionality; for mutually exclusive outcomes the classical calculus forces this Gram diagonal (disjoint incidences), and that forced-disjoint case is recovered *exactly* as the diagonal-G limit.
2. As a **geometry** at temperature $T = 0$ (the hard argmax, its cells, margins and rank), the decision surface is the *tropical variety* of a max-logit polynomial whose monomials are the unembedding rows. Composition is the decision cells whose winning monomial appears only in the sum of sources, never in a single one; the irreducible computation is a *tropical-rank floor*: the gap between the model's tropical rank and that of any flat lookup table, a gap that linear (SVD) rank structurally cannot measure.
3. As a **computation** (the executable, statically checkable artifact), the core is a *semiring-weighted Datalog program* Π. Evaluated under the log-semiring it is sum-product and returns the softmax distribution; under the tropical semiring it is max-product and returns the greedy decode. The two temperatures are one program under two semirings, and *Maslov dequantization* is the homomorphism between them.

The unification is not loose analogy. The same kernel G that hardens near-synonym competition in the logic is the facet geometry in the tropical picture and the provenance coupling in the Datalog picture; the same irreducible computation is the high participation-ratio residual and, we conjecture, the high tropical rank and the high-treewidth dense-G region; the same temperature parameter is the semiring choice. We present the measurements first (Sections 3–4), because they constrain every axiom of the theory, and then the three interpretations and their unification (Section 5), the reproduction recipe (Section 6), and the open frontier (Section 7). The tribute is plain throughout: this is Bundy's calculus, forty years on, meeting a substrate he could not have had.

[1] A companion guide, *Projective Incidence Calculus: A Guide* [22], develops this theory from first principles for readers new to incidence calculus or the transformer setup, with full derivations, worked numerical examples, visual intuition, a glossary, and machine-checkable (Isabelle/Isar) proof skeletons.

2. Background and Related Work

2.1 Mechanistic interpretability

Our measurement apparatus inherits the circuits program's central move: read computation off the model in the basis of its own weights. The residual stream is a linear sum of component writes, so a final-layer logit decomposes additively over heads and neurons; the *direct logit attribution* (DLA) of a component to token v is $\langle d_j, U_v \rangle$, the inner product of the component's write d_j with the unembedding row U_v [11]. *Superposition*, the storage of more features than dimensions as a near-orthogonal overcomplete code, predicts exactly the distributed, low-per-circuit-magnitude readout we measure [10], and sparse autoencoders attempt to recover that code as monosemantic features [4, 8]. *Induction heads*, the copy-from-earlier-match mechanism behind much in-context learning [20], are the one idiom that exports as a clean recursive rule (Section 5). Feed-forward layers as key–value memories [12] and the locate-and-edit line [19] are the retrieval end of the spectrum we are trying to delimit from the computed end. Where this work differs is in making the retrieve-versus-compute distinction a *measured, per-token classification* with an exact geometric and causal characterisation, rather than a case study of one circuit.

2.2 Incidence calculus

Incidence calculus [5, 6] was built on one load-bearing observation: probability is *not* truth-functional, since $P(A \wedge B)$ is not a function of $P(A)$ and $P(B)$, but *incidence* is. Bundy attaches to each proposition A a set $i(A) \subseteq I$ of "incidences" (possible worlds, or sampled situations) in which A holds, makes the connectives set operations ($i(A \wedge B) = i(A) \cap i(B)$), and recovers probability as the measure $P(A) = |i(A)|/|I|$. The joint that probability loses is carried explicitly by the set overlap. Our claim is that the transformer's composition core is this calculus carried to mutually exclusive next-token outcomes, with one change the data forces. Classical incidence calculus already carries a non-diagonal Gram, $G_{vw} = |i(v) \cap i(w)|$; but the incidences of *mutually exclusive* outcomes are forced *disjoint* ($G_{vw} = 0$ for $v \neq w$, a diagonal Gram). PIC removes that forced disjointness: propositions are directions in an inner-product space whose pairwise overlaps are a dense frame Gram, so competing outcomes (near-synonyms) carry nonzero correlation, and the intersection connective becomes a weighted threshold. The classical disjoint-outcome case reappears, exactly, when the Gram is diagonal.

2.3 Log-linear, tropical, and semiring views

The additive core is a log-linear / product-of-experts model [14, 15], and Markov logic networks [21] are its closest symbolic-

probabilistic cousin; what we add is the inner-product (frame-Gram) coupling and the explicit retrievable/computed split. The hard decision surface is tropical: an argmax over linear forms is the $(\max,+)$ semiring, and the ReLU/PWL-network-to-tropical lineage [26, 17] renders a feedforward network's decision map as a tropical rational function whose linear-region count is bounded by Newton polytope vertices (in a transformer attention is bilinear, so this holds only on the PWL skeleton; the residual-space decision we analyse is exactly tropical regardless). Tropical *rank* [9] supplies the irreducible-computation floor; power (Laguerre) diagrams [2] are the decision cells. The bridge between the soft ($T=1$) and hard ($T=0$) pictures is *Maslov dequantization*, the $T\to 0$ limit of the log-semiring into $(\max,+)$ [16]. The executable form lives in provenance / semiring Datalog [13, 1, 7]: sum-product over a semiring is database evaluation, and the temperature is the semiring parameter. The correlated-alternatives regime connects to discrete-choice econometrics (conditional logit / discrete choice) [18]. Finally, the phenomenon of "redundant distributed voting shading to emergent combination" has ingredients with precedent (distributed connectionist production systems [23], superposition, products of experts, and incidence calculus over a continuous learned space), but, as far as we are aware, the fusion is new.

3. The Decompiler and the Measurement

3.1 The two-tier decompile

`fieldrun` is a single static binary that runs a model from a flat *bundle* (a raw weight blob plus a small manifest) with no deep-learning framework at runtime. The build side decompiles a model into **Tier A**, a retrieval store (induction thresholds; quad/tri/bi/uni n-gram successor tables; an optional closed-class grammar skeleton), and **Tier B**, the attention+MLP composition kernel as plain matmuls. Faithfulness is enforced by a top-1 gate: Tier A reproduces the reference idiom predictions with zero per-position mismatches over 500 positions, and Tier B matches the numpy/torch reference exactly in `f32` (e.g. GPT-2 200/200, Qwen2.5 32/32). All probes below are *explain-only*: the decode path is untouched, so there is no faithfulness-gate risk and the numbers describe the real model, not a surrogate.

3.2 Objects

Fix a model with hidden width d, vocabulary V, and unembedding rows $\{U_v \in \mathbb{R}^d\}_{v\in V}$. For a position, let $r \in \mathbb{R}^d$ be the final residual stream; the model's logits are $L_v = \langle r, U_v\rangle + b_v$ and its prediction is $t = \arg\max_v L_v$. The residual is an additive sum of component writes $r = \sum_j d_j$ (embedding, each layer's attention block, each layer's MLP, down to individual heads and neurons), so each *source* j contributes $c_j^v = \langle d_j, U_v\rangle$ to every proposition and $L_v = \sum_j c_j^v$ exactly. The four measured quantities that organise everything are:

- the **participation ratio** $\mathrm{PR} = \left(\sum_j c_j^t\right)^2 / \sum_j (c_j^t)^2$, the effective number of sources carrying the winning logit;
- the **normalised margin** $\Delta/\|U_t - U_{v^*}\|$ with $\Delta = L_t - L_{v^*}$ and v^* the runner-up, i.e. the Euclidean distance from r to the nearest power-diagram facet (Section 5.2);
- the **readout multiplicity** $\mu_t = \#\{\text{top-12-by-DLA sources } j : \arg\max_v c_j^v = t\}$, how many sources would, in isolation, already pick t;
- the **differential incidence** (pivotality) $D_j = c_j^t - c_j^{v^*} = \langle d_j, U_t - U_{v^*}\rangle$, the amount by which ablating source j shifts the t-vs-v^* margin.

3.3 Models, holdouts, probes

The primary study uses **Qwen2.5-Coder-0.5B-Instruct** and **Qwen2.5-0.5B-Instruct** (shared vocabulary, hence a shared model-captured store), on a natural-text and a code holdout, ∼300–500 contexts at context window 64. The cross-architecture replication uses the **Pythia** ladder (70M/160M/410M/1B) under a corpus n-gram store held constant across architectures. The probes, each a `fieldrun` CLI mode, are: `--attribute` (the three-way route split), `--probe` (is selection a function of the firing state?), `--probe-dla` (PR, margin, μ_t), `--probe-facet` (exact nearest power-diagram facet over the full vocabulary), `--probe-ablate` (single- and multi-circuit causal ablation), `--probe-reconstruct` (per-block logit reconstruction), and `--probe-quant` (per-block quantisation sensitivity). Numbers are indicative ($n\sim 300$), not high-precision; the qualitative structure is what replicates.

4. Empirical Findings

4.1 The irreducible computation: a measured three-way split

Classifying each next-token decision by route gives, on natural text, roughly **25% RETRIEVED, 60% SELECTED, 15% COMPOSED**: the model's labour is retrieval-dominated, and most of what looks like "computation" is disambiguation *within* a retrieved candidate set. The genuinely-computed irreducible computation is the ∼15% (the COMPOSED route). This fraction is *store-relative*: a richer retrieval store shrinks it, so 15% is an upper bound for this store, not an absolute property of the model. The split is regime-dependent in the expected direction: at ∼540 store candidates, the candidate set covers the model's argmax ∼85% of the time on natural text but only ∼63% on code, because code is computed, not retrieved. Table 1 shows the split is robust across a second architecture family and a different store construction.

4.2 No selection in magnitude: a uniform 45-way sum

The obvious hypothesis, that retrieved tokens are decided by a dominant circuit and composed tokens are not, is false. The participation ratio is $\mathrm{PR} \approx 42$–49 on Qwen and is *route-invariant*: RETRIEVED, SELECTED, and COMPOSED tokens all have the same ∼45-way distributed sum, even for tokens a single n-gram rule reproduces perfectly. No source dominates the logit magnitude, *ever*. The mechanism is a uniformly distributed additive sum followed by an argmax, exactly the read-

Model (store)	RETRIEVED	SELECTED	COMPOSED
Qwen-0.5B×2 (captured)	∼25	∼60	∼15
Qwen-0.5B (corpus)	26.4	60.8	12.8
Pythia-70M (corpus)	35.6	56.6	7.8
Pythia-160M (corpus)	35.8	54.2	10.0
Pythia-410M (corpus)	36.8	51.6	11.6
Pythia-1B (corpus)	35.6	49.6	14.8

Table 1: Route split (% of next-token decisions). Retrieval dominates (RETRIEVED+SELECTED ≈ 85–92%; Pythia-1B is the low rung at 85.2%) across architectures and store constructions; the irreducible computation (COMPOSED) grows with scale against a fixed store. The Qwen rows use a model-captured store on natural text; the Pythia rows use a corpus n-gram store on a book-prose holdout (a higher-coverage regime), so the levels are not directly comparable across the two blocks.

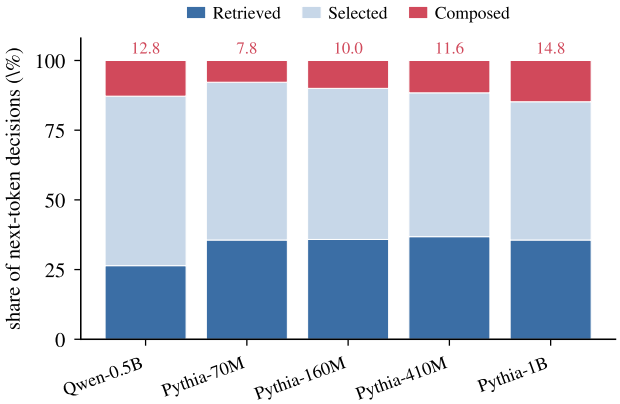

Figure 1: The same route split, visualised. Retrieval (RETRIEVED+SELECTED) dominates every model; the COMPOSED band (the irreducible computation, red, value annotated) grows monotonically up the Pythia ladder against a fixed store (7.8 → 14.8%), even as the geometry and margin structure stay scale-stable (Section 4).

out superposition predicts [10]. On the Pythia ladder, PR runs 13 → 27 → 63 → 35 (70M/160M/410M/1B): it does not track parameter count, but tracks head count (48/144/384/128) far more closely (Pythia-1B is shallower and wider). The "∼45-way sum" is thus an architecture-shape parameter (how many circuits exist to spread over), not a universal constant.

4.3 The two separating axes: margin and readout multiplicity

If magnitude does not separate the routes, two other quantities do, and they are nearly independent (per-position correlation $r \approx 0.15$, ∼2% shared variance). The first is **geometry**: the exact nearest power-diagram facet distance is monotone RETRIEVED ≫ SELECTED > COMPOSED (coder 2.23/1.34/1.03, instruct 2.78/1.45/1.22; the RETRIEVED ≫ {SELECTED, COMPOSED} gap on every Pythia rung), so retrieved tokens sit deep inside their decision cell and composed tokens sit near a facet. The robust part is the RETRIEVED ≫ {SELECTED, COMPOSED} gap; the fine SELECTED > COMPOSED ordering holds on Qwen and three of the four Pythia rungs but is within noise on the non-coder and reverses on Pythia-160M (n=13 composed). The second is **redundancy**: the readout multiplicity μ_t. Coverable tokens are *redundantly multiply-realised* (SELECTED mean $\mu_t = 1.45/1.06$), while ∼80% of COMPOSED tokens are *strictly emergent*: $\mu_t = 0$, the answer is the argmax of the ∼45-way sum but of *no* summand (COMPOSED μ_t=0 fraction 84%/76%; since μ_t counts only the top 12 circuits this is an *upper* bound on strict emergence). This is the operational definition of emergence we will formalise three ways in Section 5:

> *the argmax of a sum that is the argmax of no summand.*

The split survives a confidence control: within matched margin bins, the covered-versus-composed redundancy gap persists (coder 65/71% vs 17/16%), so it is not merely that the store covers the model's confident predictions.

A second, geometric *critical test* rules out the most natural story for composition. One might guess that a composed token is the residual r crossing the facet *out of the store's predicted cell*. It is not: the nearest facet is the bisector with the store's own prediction for only 14%/8% of COMPOSED tokens. For ∼85% of composed tokens the store's prediction is not even the nearest competitor, so composition is a *non-local* divergence from the rule, not a near-miss. The ∼14% near-miss subclass is one identifiable thing: closed-class and morphological coin-flips ($a \perp the$, $will \perp is$, $; \perp ,$, $\text{-}ful \perp \text{-}y$), the grammar regime where the store is strongest, not novel computation.

4.4 Causal ablation: composed is fragile, and the repair is diffuse

Turning the readout into an intervention, zeroing the top-DLA circuit in the forward pass and asking whether the prediction flips, converts these correlations into causal structure. Five results stand out.

(i) Route-ordered fragility, margin-governed. flip@k=1 is RETRIEVED 22/26% < SELECTED 40/48% < COMPOSED 54/61%: knock out the single top circuit and a composed token flips ∼2.4× as often as a retrieved one. But this tracks the margin, and a margin-matched split shows the margin is the governor: flips collapse across margin terciles for both $\mu_t \geq 2$ (76 → 41 → 17%) and μ_t=0 (69 → 27 → 7%), leaving only a small anti-protective gap that the $t \to$ control accounts for.

(ii) Redundancy is non-compensatory. Holding the ablated circuit fixed to a confirmed t-supporter and splitting on whether a backup remains (μ_t=1 vs $\mu_t \geq 2$), backups confer no robust protection (pooled flip: low-Δ 90%/80%, high-Δ

4%/21%). Removing one supporter is not caught by the others: apparent agreement among many weak readers is *not* fault tolerance. The redundant supporters, individually $< 10\%$ of the logit at PR$\approx$40, provide essentially no cushion.

(iii) The causal variable is pivotality D_j versus margin Δ, not μ_t. The *linear flip identity* flip $\iff \Delta < D_j$ holds as a near-perfect *necessary* condition: $\text{sign}(D_j - \Delta)$ mispredicts a non-flip only 3/300 (coder) and 17/300 (instruct) times. A logistic control flip $\sim \Delta + D_j + \mathbf{1}[\mu_t \geq 2]$ gives standardised weights $\Delta = -4.21/-3.11$, $D_j = +2.82/+1.16$, and μ_t essentially zero ($-0.60/+0.06$, opposite signs across models); dropping μ_t costs $+0.0035/+0.000$ mean log-loss. So μ_t is a *proxy* for (Δ, D_j) position, not an independent cause.

(iv) Non-truth-functionality is a kernel. With $\rho = \cos(U_t, U_{v^*})$ the runner-up coherence, pivotality and flip rate both fall as competitors become near-synonyms (coder $|D_j|$ $1.47 \to 0.86$, flip $53 \to 18\%$): common-mode evidence cancels in $D_j = \langle d_j, U_t - U_{v^*} \rangle$. The competition geometry is read directly off the frame Gram, the structural fact Section 5.1 builds on.

(v) The repair is diffuse, not a lever. When the linear identity predicts a flip, an indirect downstream recomposition *rescues* the original token about half the time, and this cushion scales with the margin (rescue $14 \to 39 \to 50 \to 61\%$ across Δ). But the rescue is not localisable: 18–34% of rescues survive ablating *every* single downstream block, and the per-head un-rescue rate is order-1/PR: measured 4.0–4.1% against $1/\text{PR} = 2.5$–2.8%, a $\sim 1.5\times$ constant. There is no surgical repair target; the repair is distributed for the same reason the readout is (PR$\approx$45).

4.5 Per-block reconstruction: exact, block-sparse, circuit-dense

Decomposing the predicting logit into its 49 per-block residual writes (embedding + attention + MLP per layer) and checking $\sum_{\text{blocks}} = L_t$ gives *floating-point exact* reconstruction (mean error $5.9 \times 10^{-6}/7.2 \times 10^{-6}$). Residual-stream additivity holds, so the static decomposition reconstructs every logit exactly, the empirical face of the soundness theorem of Section 5.3. The decision is *block-sparse* (effective supporting blocks $\approx$ 8–10 of 49; top block to flip $\sigma \approx 1.1$–1.6) but *circuit-dense within a block*: a block's write is itself a dense sum over its $\sim$14 heads and $\sim$4.9k neurons, and the circuit-level PR ≈ 45 (Section 4) lives below block granularity. The readable program is compact at block granularity and bottoms out there for composed tokens; below the block it is the dense computed remainder.

4.6 Scaling and a research-to-speed bridge

On the Pythia ladder the robust core (the three-way split, route-invariant PR, the RETRIEVED $\gg$ {SELECTED, COMPOSED} margin geometry, the refuted critical test, and margin-governed fragility) recurs at every scale. Two findings are scale-dependent. The irreducible computation *grows* with scale ($7.8 \to 14.8\%$ up the ladder against the same store; Figure 1): bigger models compute more. And μ_t-redundancy is something models *grow into*: at 70M/160M, $\mu_t \approx 0$ everywhere, while the full transition appears between 160M and 410M. The power-diagram geometry (margin) is scale-stable from 70M; the readout multiplicity is scale-emergent then stable, consistent with μ_t being a readout property, not the causal variable. Finally, quantisation sensitivity tracks the split: single-block int4 flip rates are RETRIEVED $3.6/2.4\% \ll$ COMPOSED $5.9/18.8\%$, so the retrievable tier survives aggressive quantisation while the irreducible computation is where quantisation error concentrates. This is the research result with a deployment consequence (quantise the retrievable tier hard, protect the computed core).

5. One Theory in Three Interpretations

The measurements above constrain a theory tightly. It must be *additive* (the logit is an exact sum, Section 4.5); *cardinality-inert* (the count μ_t has no causal weight, only D_j and Δ do, Section 4.4); carry a *non-truth-functional coupling* that hardens with $\rho = \cos(U_t, U_{v^*})$; have a *weighted-threshold* decision beyond Horn clauses (μ_t=0 for most composed tokens); and possess a second, *diffuse, non-localisable repair layer* bounded by 1/PR. We now give a structure with all five properties, and observe that it is a single object viewed as logic, geometry, or computation, according to a temperature/semiring parameter.

5.1 Interpretation I, Logic (T=1): Projective Incidence Calculus

Objects. Sources S (circuits) are vectors d_j in an inner-product space $\mathscr{H}$; propositions V (tokens) are directions $U_v \in \mathscr{H}$. The *projective pairing* $c_j^v = \langle d_j, U_v \rangle$ is the direct logit attribution; aggregated evidence is the residual $r = \sum_j d_j$ with logits $L_v = \langle r, U_v \rangle = \sum_j c_j^v$. Two objects carry the theory. The *Gram kernel* $G_{vw} = \langle U_v, U_w \rangle$ is the proposition frame's inner-product structure, and the *differential incidence* $D_j^{I;v} = \langle d_j, U_t - U_v \rangle$ is the atomic causal quantity. Where Bundy's incidence of a proposition is a *set* $i(v) \subseteq I$, ours is the linear functional $\langle \cdot, U_v \rangle$, and the overlap of two propositions is not a set intersection but the kernel value G_{vw}. This is the central move: propositions share an inner-product space, so their incidences overlap *intrinsically*, and the structure that probability is not truth-functional over is carried by G.

Connectives. Combination is signed linear accumulation in log-space, $L_v = \sum_j c_j^v$, equivalently a *product of experts*, each source multiplying a proposition's mass by $\exp(c_j^v)$ [14]. The decision is a weighted threshold: $\arg\max_v L_v$ is the Laguerre power diagram of $\{U_v\}$ with weights $\|U_v\|^2$, and the normalised margin is the facet distance (Section 5.2). The inferential analogue of incidence-calculus bounding is a *coalition bound*: from a known subset $S' \subseteq S$, the decision margin is bounded by the partial sum $\sum_{j \in S'} D_j$ plus a residual bound; the measured coalition additivity ($\text{sign}(\sum_{j \in S'} D_j - \Delta)$ predicts

joint-ablation flips at ∼75–83%) is its empirical face.

Theorems. The desiderata are recovered as consequences, not posited; proof sketches and the hypotheses each result rests on are collected in Appendix A.

Theorem 1 (Cardinality-inertness)**.** *Under the projective pairing, the decision depends only on* $\{D_j, \Delta\}$ *and is invariant to the count* μ_t.

This recovers the measured fact that μ_t carries no causal weight given (Δ, D_j).

Theorem 2 (Non-truth-functionality budget)**.** *Competition hardness between t and v is a monotone function of* $\rho_{tv} = G_{tv}/\sqrt{G_{tt}G_{vv}}$*; as* $\rho \to 1$*, the differential incidence* $D_j = \langle d_j, U_t - U_v \rangle$ *becomes common-mode and collapses. PIC reduces to the classical disjoint-outcome case (mutually exclusive incidences, diagonal overlap)* exactly *when G is diagonal.*

This recovers the measured ρ-boundary and identifies the forced-disjoint (mutually exclusive) case as the diagonal-G limit.

Definition 1 (Reducible, irreducible)**.** A subset $S' \subseteq S$ *decides* t if t is the strict argmax of the isolated sum $\sum_{j \in S'} c_j$. A composed token is *reducible* if some proper non-empty $S' \subsetneq S$ decides it, and *irreducible* if S decides t but no proper non-empty subset does. The readout condition $\mu_t{=}0$ (no *singleton* decides t) is strictly weaker than irreducibility.

Theorem 3 (Weighted-threshold expressivity)**.** *(a) Composed tokens with* $\mu_t{=}0$ *exist: t is the strict argmax of* $\sum_j c_j$ *but of no single source. (b)* $\mu_t{=}0$ *does not imply irreducibility; some* $\mu_t{=}0$ *tokens are decided by a proper sub-coalition. (c) Irreducible composed tokens nonetheless exist, and an irreducible t has no sufficient sub-conjunction over S: it is realised by a weighted threshold* $\sum_i w_i x_i > \theta$ *but by no Horn / ∩-∪ formula over those sources.*

Theorem 4 (Recovered probability)**.** *Maintain a measure over propositions with uniform base* M_0*; let each source reweight every proposition's mass by* $\exp(c_j^v)$*. Then* $m(v) = M_0 \exp(\sum_j c_j^v) = M_0 \exp(L_v)$ *and* $m(v)/\sum_w m(w) = \exp(L_v)/Z$*, exactly the model's softmax. The Gibbs measure is recovered as a PIC incidence frequency, exactly and parameter-free.*

Theorem 4 is the direct descendant of Bundy's "probability as proportion of worlds." Crucially, G need not be injected as correlated noise (which would trade off exact recovery): it is *already* structural as the Gram of the proposition frame, so it structures the static logits (Theorem 4) and the competition geometry (Theorem 2) at once, with no variance trade-off.

The second layer. PIC has a monotone additive core (the map $r \mapsto (L_v)$ and the threshold) and a non-monotone fixpoint closure. Model the forward pass as an iterated operator T on the evidence vector; the additive core is its linearisation at the operating point, and the diffuse repair (the irreducible computation) is the higher-order closure $T^\infty - (\text{linearisation})$. It appears only under intervention (the off-diagonal Jacobian) and is diffuse in the high-PR regime, where the contributions are near-equitable:

Theorem 5 (Diffuseness)**.** *Any causal property realised as* $E = \sum_m e_m$ *with equitable* $e_m \sim E/\mathrm{PR}$ *has single-source influence* $O(1/\mathrm{PR})$*; hence no bounded-size PIC formula localises it, and* $P(\text{single-module intervention alters } E) = O(1/\mathrm{PR})$.

This recovers the measured per-head un-rescue rate (Section 4.4(v)), order-of-magnitude confirmed at 4.0–4.1% versus $1/\mathrm{PR} = 2.5$–2.8%. The $\sim 1.5\times$ excess over $1/\mathrm{PR}$ is consistent with a mild within-layer correlation of repair (a layer's heads share rescue), the one place the equitable-contribution approximation is loosest. The clean one-line statement of the architecture: *the recovered probability lives in the static decomposition (additive, exact); the irreducible computation lives in the intervention response (non-additive, diffuse); the same model exhibits both as different layers.*

5.2 Interpretation II, Geometry ($T{=}0$): the Tropical Decision Surface

The decision is $\arg\max_v(\langle r, U_v \rangle + b_v)$, so the *max-logit*

$$M(r) = \bigoplus_v \left(b_v \otimes x^{U_v}\right) = \max_v \left(\langle r, U_v \rangle + b_v\right)$$

is a *tropical polynomial* in r over the $(\max, +)$ semiring ($a \oplus b = \max(a,b)$, $a \otimes b = a + b$): its monomial exponents are the unembedding rows U_v, its Newton polytope $\mathrm{conv}\{U_v\}$ is the set of tokens that can ever win (with an output bias this is the upper hull of the lifted points (U_v, b_v); the models here have none, $b_v = 0$, so it is exact), and its tropical hypersurface (where the max is attained by ≥ 2 monomials) is the decision boundary [26, 17]. Two of our measurements are this geometry, exactly:

Proposition 1 (Cells are a power diagram)**.** *The linear regions of M are the Laguerre power diagram of* $\{U_v\}$ *with weights from* $(b_v, \|U_v\|^2)$*; the cell containing r is the predicted token.*

Proposition 2 (Margin is tropical distance)**.** *The normalised margin* $(L_t - L_{v^*})/\|U_t - U_{v^*}\|$ *is the exact Euclidean distance from r to the nearest facet of the tropical hypersurface.*

These are not conjectures: `--probe-facet` computes the exact nearest facet over all 151,936 tokens, and the measured monotone ordering RETRIEVED ≫ SELECTED > COMPOSED (Section 4) is the facet-distance signature. Emergence translates sharply: decomposing $M(r) = \max_v \sum_j c_j^v$, a position is RETRIEVED when some single source's monomial already attains the max at r (its isolated argmax is the winner, $\mu_t \geq 1$) and COMPOSED when none does: the winning monomial appears only in the sum of sources, a *mixed* term in the Minkowski-sum subdivision of the sources' Newton polytopes ($\mu_t = 0$). This is the tropical reading of "argmax of a sum that is the argmax of no summand."

The irreducible computation as a tropical-rank floor. A flat retrieval table ("context key → stored next-token logits") is a tropical map whose monomials are exactly its stored keys, one tropical term per row, its tropical (Barvinok) rank bounded by the table size. Composition is precisely the decision regions requiring monomials *not* in the table: sums of stored keys that create new mixed cells. Write $\rho_{\text{trop}}(\text{core})$ for the tropical (Barvinok) rank of the core's decision map (the fewest tropical rank-one terms reproducing its cells) and $\rho_{\text{trop}}(\text{KB})$ for the best flat table at matched coverage [9].

Conjecture 1 (Tropical-rank floor)**.** The irreducible computation is lower-bounded by the rank gap $\rho_{\text{trop}}(\text{core}) - \rho_{\text{trop}}(\text{KB})$: its decision cells are exactly those requiring composed (mixed) monomials that no lookup table reproduces.

The measured COMPOSED fraction is the empirical shadow of this gap. This also indicates *why* a linear (SVD) rank misranks the core: its hardness is the number of *tropical* monomials (decision cells), which a Frobenius rank does not measure. We therefore predict that the core's tropical and linear ranks diverge, and that a data-aware low-rank update beats plain SVD at matched rank, a falsifiable consequence.

5.3 Interpretation III, Computation: Semiring-Weighted Datalog

The core's next-token computation is semiring accumulation over a finite relational domain, evaluated bottom-up: contributions sum along the residual stream, then one competitive aggregation picks the token. That is exactly Datalog evaluated under a *provenance semiring* [13, 1]. We export the model as a semiring-weighted Datalog program Π: propositions V with directions U_v and Gram G; sources S with pairings c_j^v; the retrievable fragment as compact *stratified* clauses (an induction head is a recursive clause `next(T):-match_prefix(P),follows(P,T)`; an n-gram is a weighted fact; grammar is a unary constraint); the computed fragment as a flagged dense aggregate; and a provenance evaluator with a semiring/temperature parameter. Reading Π as a Functional Aggregate Query, the decision is the aggregation $\bigoplus_v \bigotimes_j (\text{factor})$.

Datalog rather than Prolog is a semantics decision: provenance semirings are a Datalog construct (Prolog has no semiring parameter, and cut / negation-as-failure / clause order break the semiring laws); forward residual accumulation matches Datalog's bottom-up least-fixpoint; aggregation is native; and the program is decidable and statically analysable (PTIME data complexity, a unique least fixpoint): a mathematical object one can bound and verify, rather than a procedure. The retrievable rules are finite-domain over tokens and need none of Prolog's extra power; the irreducible computation is dense arithmetic Prolog could not compact either.

Theorem 6 (Two-temperature soundness)**.** *Read Π as a semiring FAQ, with sources combined by $\otimes$ along the residual derivation (so the accumulated value of v is $\bigotimes_j c_j^v = \sum_j c_j^v = L_v$) and competing propositions combined by $\oplus$. Then*

- *under the* log-semiring *($\oplus = $ log-sum-exp, $\otimes = +$), $[\![\Pi]\!]$ is the sum-product: the aggregate is $\log \sum_v \exp(L_v) = \log Z$ and the per-proposition share is $\exp(L_v)/Z = P(v)$, the model's softmax distribution (Interpretation I, T=1);*
- *under the* tropical semiring *($\oplus = \max$, $\otimes = +$), $[\![\Pi]\!]$ is the max-product: the aggregate is $\max_v L_v$ with witness $\arg\max_v L_v$, the model's greedy decode (Interpretation II, T=0).*

This is the standard sum-product / max-product duality instantiated on Π, plus the product-of-experts recovery of Theorem 4; the log ↔ max homomorphism is Maslov dequantization, so the export is correct at both temperatures by one identity. It is confirmed numerically by the exact per-block reconstruction (Section 4.5): $\sum_{\text{blocks}} = L_t$ to floating point means the static export is faithful. The emitter is implemented: `fieldrun export --logic` writes a runnable, Soufflé-compatible semiring-Datalog program for one decode (candidate facts, the Tier-A retrievable clauses, per-block `contrib` facts with the dense remainder folded into a `"rest"` block, and the $(\max, +)$ decode), self-checking $\sum$`contrib` $= L_t$ and that the decode equals the model's token; a built-in evaluator runs it under either semiring (`eval --semiring max` returns the token, `--semiring log` the distribution): *one program, two semirings, two temperatures*.

The provenance gap. Theorem 6 evaluates Π to the right value for *any* G: the $\sum_{\text{blocks}} = L_t$ identity holds as measured, dense G and all (Section 4.5). What dense G breaks is not the evaluation but the *provenance* semantics, the per-source explanations and ablation counterfactuals, since $G_{vw} = \langle U_v, U_w \rangle$ couples every pair. Hence:

Conjecture 2 (Provenance gap)**.** Faithful *provenance and intervention* queries on Π require provenance *valued in the frame geometry* (carrying the U_v directions, or the operator G), not scalars in a commutative semiring over $\mathbb{R}$. Scalar semiring Datalog is provenance-exact only on diagonal G (the disjoint / independent case); for dense G, scalar provenance cannot carry the cross-outcome correlation the frame Gram encodes, the joint that mutually exclusive incidences force to zero.

This is the same wall as the irreducible computation seen from the computation side: a dense G coupling is a high-treewidth factor graph, and sum-product is exponential in treewidth, so the correlated evaluation is at once intractable and non-compact. Solving the provenance gap would solve Interpretation I's non-truth-functionality theorem; that one-to-one correspondence is the strongest evidence the three interpretations are one theory.

5.4 The unification

The three interpretations are one object under a single degree of freedom, the temperature, which is the semiring choice. As $T \to 0$, $T \log \sum_v \exp(L_v/T) \to \max_v L_v$ and $\text{softmax}(L/T) \to \arg\max$, by Maslov dequantization [16]. The power diagram

is $\lim_{T\to 0}$ of the softmax cells; Interpretation I's coherence kernel $\rho_{tv} = \cos(U_t, U_{v^*})$ becomes the tropical facet angle; Interpretation I's smoothed competition is the $T > 0$ viscosity regularisation of the tropical variety. The same irreducible computation is the high-PR residual (Interpretation I) and, we conjecture, the high tropical rank (Interpretation II) and the high-treewidth dense-G region (Interpretation III): *one wall, we expect, seen through three measures* (participation ratio, tropical rank, treewidth). These coincide on the retrievable / diagonal-G fragment and move together in the data; whether they remain equivalent on the dense fragment, or diverge there, is open. And the same kernel G is the non-truth-functionality operator, the facet geometry, and the provenance coupling. The retrievable fragment is, in all three, the additive / low-rank / low-treewidth part that exports to a compact verifiable program; the computed fragment is the diffuse / high-rank / dense-G remainder that admits no compact symbolic form at the retrievable (low-rank, low-treewidth) granularity. This is the program's logic-programming thesis made precise: *the model is a semiring-weighted Datalog program, decoding is provenance evaluation, and its two temperatures are the softmax measure and the greedy argmax over one program.*

6. Reproducibility

Every quantitative claim traces to an explain-only probe in the openly available `fieldrun` toolchain (Apache-2.0; source and bundles public, with the exact holdout and store manifests behind Table 1 included), which runs a model from a flat bundle in pure Rust with no deep-learning framework. After converting a checkpoint (`fieldrun convert --model <hf-id> --arch rope`) and building a store, the probes reproduce the paper:

```
# three-way route split (Sec. 4.1)
fieldrun --bundle <m> --ids holdout.json \
  --store store.json --attribute
# PR, normalized margin, mu_t (Sec. 4.2-4.3)
fieldrun ... --probe-dla --n-eval 500
# exact nearest power-diagram facet (Sec. 5.2)
fieldrun ... --probe-facet
# causal ablation; mu_t-falsifier; 1/PR lemma
fieldrun ... --probe-ablate --n-eval 300 \
  --head-sweep
# per-block reconstruction: sum(blocks)==logit
fieldrun ... --probe-reconstruct --n-eval 300
# quantization sensitivity by route (Sec. 4.6)
fieldrun ... --probe-quant --bits 4 --n-eval 80
# emit a runnable semiring-Datalog program and
# evaluate it under either semiring (Sec. 5.3)
fieldrun ... export --logic --ctx 32 \
  --candidates 24 --out decode.dl
fieldrun eval decode.dl --semiring max   # decode
fieldrun eval decode.dl --semiring log   # softmax
```

The cross-architecture replication (Table 1) uses the same probes on Pythia bundles (`--arch neox`) under a corpus n-gram store, with the faithfulness gate validated top-1 against a pure-numpy reference. All probes leave the decode path untouched.

7. Open Problems

The theory is anchored where it can be measured and conjectural exactly where the measurements stop. The sharpest open questions are: **(O1)** the soundness and completeness of weighted incidence resolution, the coalition bound $\sum_{j\in S'} D_j$ as a sound inference rule; **(O2)** whether the support number $\sigma(t)$ (the smallest sufficient *circuit* set; at block granularity it mildly reverses, so the circuit level is where the question is well posed) scales as the participation ratio, and, relatedly, a general criterion separating *irreducible* composed tokens (no proper sub-coalition decides t) from merely $\mu_t{=}0$ ones, which already differ on small machine-checked witnesses; **(O3)** a derivation of the measured $\sim 1.5\times$ constant in $P(\text{single-module repair}) = \kappa/\text{PR}$ from the fixpoint closure, accounting for the within-layer correlation of repair; **(O4, the central one)** a non-scalar, geometry-valued provenance semiring that carries G faithfully and reduces to the scalar log/tropical semiring on diagonal G (solving it closes the provenance gap and Interpretation I's non-truth-functionality theorem at once); **(O5)** the treewidth of the core's factor graph as a third quantitative irreducible-computation measure, related to the participation ratio and the tropical rank, with whether the three coincide (one wall) or diverge on the dense fragment itself open; and **(O6)** cross-scale invariance of G's spectrum and the retrievable/computed split as a program-wide thesis. A PIC *syntax* (compiling formulae directly from DLA traces, with composed positions flagged as "no compact formula") would make the retrievable fragment a statically verifiable extracted program and the irreducible computation a named, bounded region.

8. Discussion and Conclusion

A transformer, asked for a token, spends most of its labour retrieving and selecting, and a small, scale-growing remainder computing something new. We have made that distinction measurable on real models, given it an exact geometric and causal characterisation, and shown that the same structure is a probabilistic logic, a tropical geometry, and an executable Datalog program according to a single degree of freedom, the temperature. The retrievable fragment is additive, low-rank, low-treewidth, and compactly verifiable; the irreducible computation is the diffuse, high-rank, dense-Gram region that no lookup table reproduces and no single module localises. The three interpretations are not analogies bolted together: they share one kernel G, one wall measured three ways, and one homomorphism (Maslov dequantization) between their temperatures.

The lineage is the point. Incidence calculus was built in 1985 on the insight that probability is not truth-functional but incidence is, and that one recovers probability by measuring incidences [5]. Forty years on, a transformer's composition core

turns out to be exactly that calculus with the forced disjointness of mutually exclusive outcomes removed: its propositions live in an inner-product space whose Gram kernel is the explicit carrier of the non-truth-functional joint, its connective is a weighted threshold rather than Boolean intersection, and its output probabilities are recovered, exactly, as incidence frequencies. The classical disjoint-outcome case sits inside ours as the diagonal-G limit. We have only added a substrate, a thermometer, and a fixpoint engine; the central idea is Bundy's.

8.1 Limitations

Three limitations bound the claims. **Scale.** The evidence is from sub-billion models: two Qwen2.5-0.5B variants and the Pythia ladder up to 1B. Whether the same three-way structure and the participation-ratio mechanism hold at GPT scale is open. **Sample and context length.** The route statistics are $n \approx 300$ contexts at a 64-token window and are *indicative*, not high-precision; how the split behaves at long context is untested. **Cross-architecture comparability.** The Qwen rows use a model-captured store and the Pythia rows a corpus n-gram store on a book-prose holdout, so the two blocks are not directly comparable. The cross-architecture claim is therefore that the *structure* recurs (route-invariant PR, monotone COMPOSED growth), not that absolute levels match.

Acknowledgements

This work is, before anything else, a tip of the hat to Alan Bundy's incidence calculus. The author also gratefully acknowledges the extensive use of AI systems as research tools: Grok and Claude (Opus and Fable) were used for brainstorming and for working through the mathematical derivations (among them the product-of-experts recovery and the correlated-alternatives framing), and Claude additionally for the design, implementation, and execution of the `fieldrun --probe-*` programs from which the empirical constraints come. All derivations and results were checked by the author, who takes sole responsibility for the paper's claims and for any errors. J. Allan Scott carried out this work independently and in a personal capacity; it was not sponsored by, and does not represent the views of, his employer, Vista Higher Learning (VHL).

A. Proof Sketches

We sketch the argument and state the hypotheses each result rests on; the additive (static) results are exact, and the diffuseness bound is given under an explicit equitability hypothesis. Propositions 1–2 are immediate: the decision $\arg\max_v(\langle r, U_v\rangle + b_v)$ is by definition the Laguerre power diagram of $\{U_v\}$ with weights $(b_v, \|U_v\|^2)$, and rearranging $\langle r, U_t - U_v\rangle = \Delta$ and dividing by $\|U_t - U_v\|$ gives the signed Euclidean distance from r to the t–v bisector, so the normalised margin is exactly that facet distance. Full proofs, additional worked examples, and a transcription of each result into a machine-checkable proof skeleton (Isabelle/Isar) appear in the companion guide [22]; a passing static check covers the skeleton and method coverage but is not, on its own, a kernel-checked proof.

Theorem 1 (Cardinality-inertness). The prediction is $t = \arg\max_v L_v$ with $L_v = \sum_j c_j^v$. The argmax is a function of the totals L_v, hence of the pairwise differences Δ and $\{D_j\}$; the count $\mu_t = \#\{j : \arg\max_v c_j^v = t\}$ is a statistic of the per-source argmaxes, which do not enter L_v. Conditioned on $\{D_j, \Delta\}$ the decision is therefore independent of μ_t. (The empirical face is the Section 4.4 logistic control, in which μ_t adds ≈ 0 log-loss given (Δ, D_j).)

Theorem 2 (Non-truth-functionality budget). Write $D_j = \langle d_j, U_t - U_v\rangle$ and split $U_t - U_v$ into a common-mode part (along $U_t + U_v$) and a differential part. For unit directions $\|U_t - U_v\|^2 = 2(1-\rho_{tv})$, so as $\rho_{tv} \to 1$ the differential part vanishes and every $D_j \to 0$ (common-mode cancellation); competition between near-collinear propositions collapses the differential incidence, and competition hardness rises in ρ_{tv} in the form measured in Section 4.4. When G is diagonal the cross terms vanish and the accumulation reduces to independent per-proposition evidence, i.e. the classical disjoint-outcome case. The limiting collapse is exact; the monotone form is the empirically observed one.

Theorem 3 (Weighted-threshold expressivity). Take the Horn / $\cap$-$\cup$ fragment over S to be the conclusions derivable from a *sufficient sub-conjunction*: a proper non-empty $S' \subseteq S$ whose isolated sum already selects t. Parts (a) and (c) are existence claims, each witnessed by a small, machine-checked source system: a two-source system in which t is the strict argmax of the pair but of neither source and no proper subset decides t (irreducible); and a three-source system in which each source defends a distinct competitor, so only the full triple clears every defence and every source is necessary (irreducible; the formal face of the fragility of Section 4.4). Part (b) is the separation $\mu_t{=}0 \neq$ irreducible: a three-source witness has $\mu_t{=}0$ yet a two-source sub-coalition already decides t. For an irreducible t no sub-conjunction suffices, so it is expressible by a weighted threshold but by no Horn / $\cap$-$\cup$ formula over S. These witnesses are kernel-checked; a *general* criterion for which composed tokens are irreducible is open (Open Problem O2).

Theorem 4 (Recovered probability). Maintain mass M_0 uniform over propositions and let each source j multiply v's mass by $\exp(c_j^v)$. After all sources $m(v) = M_0 \exp(\sum_j c_j^v) = M_0 \exp(L_v)$; normalising, $m(v)/\sum_w m(w) = \exp(L_v)/Z = \mathrm{softmax}(L)_v$. The recovery is exact and parameter-free, and is equivalent to the iid-Gumbel argmax over $\{L_v\}$ (Gumbel-max).

Theorem 5 (Diffuseness). *Hypothesis: equitable contributions.* Suppose a causal quantity decomposes as $E = \sum_{m=1}^{\mathrm{PR}} e_m$

with $e_m \approx E/\mathrm{PR}$. Removing one module changes E by $\approx E/\mathrm{PR}$, so single-source influence is $O(1/\mathrm{PR})$; a bounded (k-source) formula captures only $\approx kE/\mathrm{PR}$, so no bounded-size formula localises E as $\mathrm{PR} \to \infty$, and the probability that a single-module intervention changes the argmax is $O(1/\mathrm{PR})$. The measured per-head un-rescue rate (4.0–4.1% vs $1/\mathrm{PR} =$ 2.5–2.8%) confirms the order; the $\sim 1.5\times$ constant is the departure from perfect equitability (mild within-layer correlation) that the hypothesis idealises away.

Theorem 6 (Two-temperature soundness). Read Π as a semiring FAQ with $\otimes$ along the residual derivation, so proposition v accumulates $\bigotimes_j c_j^v = \sum_j c_j^v = L_v$, and competitors combined by $\oplus$. Under the log-semiring ($\oplus =$ log-sum-exp, $\otimes = +$) the aggregate is $\log \sum_v \exp L_v = \log Z$ and the per-proposition share is $\exp(L_v)/Z = P(v)$ (sum-product; the recovery is Theorem 4). Under the tropical semiring ($\oplus = \max$, $\otimes = +$) the aggregate is $\max_v L_v$ with witness $\arg\max_v L_v$ (max-product). The two are related by Maslov dequantization, $T \log \sum_v \exp(L_v/T) \to \max_v L_v$ as $T \to 0$, so one identity gives both temperatures.

References

[1] Mahmoud Abo Khamis, Hung Q. Ngo, and Atri Rudra. FAQ: Questions asked frequently. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 13–28, 2016.

[2] Franz Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[3] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, pages 2397–2430, 2023.

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

[5] Alan Bundy. Incidence calculus: A mechanism for probabilistic reasoning. *Journal of Automated Reasoning*, 1(3):263–283, 1985.

[6] Alan Bundy. Incidence calculus. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, pages 663–668. Wiley, 2nd edition, 1992.

[7] Stefano Ceri, Georg Gottlob, and Letizia Tanca. *Logic Programming and Databases*. Springer, 1990.

[8] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[9] Mike Develin, Francisco Santos, and Bernd Sturmfels. On the rank of a tropical matrix. In *Combinatorial and Computational Geometry*, volume 52 of *MSRI Publications*, pages 213–242. Cambridge University Press, 2005.

[10] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/toy_model/index.html`.

[11] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. `https://transformer-circuits.pub/2021/framework/index.html`.

[12] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *Proceedings of EMNLP*, 2021.

[13] Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *ACM Symposium on Principles of Database Systems (PODS)*, pages 31–40, 2007.

[14] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[15] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

[16] Grigory L. Litvinov. The maslov dequantization, idempotent and tropical mathematics: A brief introduction. *Journal of Mathematical Sciences*, 140(3):426–444, 2007.

[17] Diane Maclagan and Bernd Sturmfels. *Introduction to Tropical Geometry*, volume 161 of *Graduate Studies in Mathematics*. American Mathematical Society, 2015.

[18] Daniel McFadden. *Conditional Logit Analysis of Qualitative Choice Behavior*. Frontiers in Econometrics. Academic Press, 1974.

[19] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[21] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1–2):107–136, 2006.

[22] J. Allan Scott. Projective incidence calculus: A guide to the composition core of an autoregressive transformer as a probabilistic logic, a tropical geometry, and an executable program, 2026. Explanatory companion to this paper; full derivations, worked examples, glossary, and machine-checkable (Isabelle/Isar) proof skeletons. https://github.com/jascal/fieldrun.

[23] David S. Touretzky and Geoffrey E. Hinton. A distributed connectionist production system. *Cognitive Science*, 12(3):423–466, 1988.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[25] An Yang, Baosong Yang, Binyuan Hui, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[26] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 5824–5832, 2018.