Projective Incidence Calculus

A Guide to the Composition Core of an Autoregressive Transformer as a Probabilistic Logic, a Tropical Geometry, and an Executable Program

An explanatory companion to
"What a Transformer Retrieves and What It Computes" (J. Allan Scott)

June 14, 2026

Abstract

An autoregressive transformer language model, asked for the next token, sometimes performs a lookup and sometimes computes something genuinely new. *Projective Incidence Calculus* (PIC) is a theory that makes that distinction precise and, more ambitiously, identifies the model's irreducible "computing" part with a forty-year-old idea from symbolic AI: Alan Bundy's *incidence calculus*. This guide builds the theory from the ground up for a reader who knows linear algebra and a little probability but is not assumed to know either transformers or incidence calculus. We first develop just enough transformer mathematics (the residual stream, logits, the softmax, direct logit attribution) and then work through classical incidence calculus with full numerical examples. We then explain, conceptually, visually, and mathematically, what PIC is: a single object viewed at two "temperatures" as a probabilistic logic, a piece of tropical geometry, and an executable Datalog program. We walk through the key proofs that hold the theory together, ground every abstraction in measurements from a real model (Qwen2.5 and the Pythia ladder, via the `fieldrun` toolchain), and close with the open questions the theory leaves on the table.

Contents

1 Notation and conventions

This section collects the notation used throughout, so that a reader who has not seen some of it can refer back. Nothing here is needed before it appears in context; skim it now and return as required. Each symbol is also re-explained at first use, and the most important ones are restated in the Glossary (Appendix B).

Standing conventions. Lowercase italic letters such as r, d_j, U_v denote *vectors* in a real space $\mathbb{R}^d$ (we do not bold them); the same letters subscripted are still vectors unless they appear as a single coordinate. The index j ranges over *sources* (the model's components: the embedding, each attention head, each MLP), and v, w range over *tokens* in the vocabulary V. Capital L_v, c_j^v, G_{vw} are real *scalars*. We write $:=$ for "is defined as."

The one symbol to fix first: $\langle\cdot,\cdot\rangle$. The angle brackets $\langle u, v\rangle$ denote the **inner product** (the generalised dot product) of two vectors: a single number measuring how much they point the same way. For ordinary real vectors it is

$$\langle u, v\rangle \;=\; \sum_i u_i v_i \;=\; u_1v_1 + u_2v_2 + \cdots \;=\; u^\top v \;=\; \|u\|\,\|v\|\cos\theta,$$

so it is large and positive when u, v are aligned, zero when orthogonal, and negative when opposed.

A **dot used as an argument** marks an *empty slot*: $\langle\cdot, U_v\rangle$ is not a number but a *function*, the rule "feed me any vector x and I will return the number $\langle x, U_v\rangle$." Writing the dot is just a way to name that function without inventing a letter for its input; the same object could be written $x \mapsto \langle x, U_v\rangle$. (Such a "fix one slot, leave the other open" inner product is called a *linear functional*: it eats a vector and outputs a scalar, linearly.) So $L_v = \langle r, U_v\rangle + b_v$ is what you get by feeding the specific residual vector r into the slot of $\langle\cdot, U_v\rangle$ and adding the bias.

(Watch one reuse: in Theorem 6 the notation $P\langle E\rangle$ means "the probability of event E," where the brackets merely delimit a described event rather than pair two vectors.)

Symbol	Meaning (section of first use)
Spaces and vectors	
$\mathbb{R}^d$	the d-dimensional real representation space (§3)
$\langle u, v\rangle$	inner product / generalised dot product of u and v (§3)
$\|u\|$	Euclidean length (norm) of u, $\|u\| = \sqrt{\langle u, u\rangle}$ (§3)
r	the residual-stream vector at the final position (§3)
d_j	the write of source j into the residual stream, $r = \sum_j d_j$ (§3)
U_v	the unembedding direction (output direction) of token v (§3)
b_v	the (optional) output bias for token v (§3)
The readout	
L_v	the logit (score) of token v, $L_v = \langle r, U_v\rangle + b_v = \sum_j c_j^v$ (§3)
c_j^v	source j's vote for token v, $c_j^v = \langle d_j, U_v\rangle$ (DLA) (§3)
$\mathrm{softmax}(L)_v$	$\exp(L_v)/\sum_w \exp(L_w)$, the output probability of v (§3)
$\arg\max_v$	the token attaining the largest value (greedy prediction) (§3)
G_{vw}	Gram matrix of token directions, $G_{vw} = \langle U_v, U_w\rangle$ (§3)
Diagnostics	
PR	participation ratio, effective number of contributing sources; route-invariant, value architecture-dependent (≈ 45 on Qwen) (§3)
μ_t	readout multiplicity: how many sources' own arg max is t (§3)
Δ	winning margin $L_t - L_{v^*}$ (v^* is the runner-up) (§3)
D_j	differential incidence $c_j^t - c_j^{v^*} = \langle d_j, U_t - U_{v^*}\rangle$ (§6)
Logic and incidence	
I	a finite set of incidences (possible worlds / samples) (§4)
$i(A)$	the incidence set of proposition A, $i(A) \subseteq I$ (§4)
$P(A)$	probability of A, recovered as $\|i(A)\|/\|I\|$ (§4)
$\|\cdot\|$	cardinality of a set (or absolute value of a number) (§4)
Semiring / temperature	
$\oplus$, $\otimes$	the "add" and "multiply" of a semiring (§5)
T	temperature: $T = 1$ gives softmax, $T \to 0$ gives greedy arg max (§5)
$\bigoplus_v$	semiring sum over tokens (e.g. $\max_v$ in the tropical case) (§5)

2 Why this theory exists

Ask a large language model for the capital of France and it answers "Paris." It is tempting to say the model has *memorised* a fact and simply looked it up. But producing "Paris" as the next token under a real prompt is not a clean lookup: internally the decision is a mixture of many partial contributions, and tightening the context can push the answer out of anything resembling a stored table and into a regime where the model is genuinely *assembling* the answer on the fly.

This raises a sharp, empirical question. For any given next-token decision, is the model

(a) **retrieving**, reproducing the output of a single symbolic rule (an n-gram, an induction copy, a grammar constraint);

(b) **selecting**, choosing among a small set that a rule has already proposed; or

(c) **composing**, producing something no single rule proposed, an *irreducible computation*?

The headline empirical finding of the work this guide explains is that, across the models studied, the labour divides roughly **25% retrieved, 60% selected, 15% composed**, and that the small composed fraction is a real, *scale-growing* core of genuine computation. The deeper theoretical claim is that this composed core is not a black box: it is exactly Bundy's incidence calculus, generalised to live inside the inner-product geometry that a transformer's weights provide. That theory is *Projective Incidence Calculus*.

The one-sentence version. Probability is not *truth-functional*, you cannot get $P(A \wedge B)$ from $P(A)$ and $P(B)$ alone, and Bundy's 1985 fix was to track the underlying *incidences* (the situations in which a proposition holds) and recover probability as a frequency of overlap. PIC's claim is that a transformer's composition core is doing precisely this, with the role of "overlap" played by the inner product between token directions in the model's representation space.

This document is organised to match how the theory is best learned, not how it is most tersely stated. Section 3 builds the transformer mathematics. Section 4 works through classical incidence calculus. Section 5 introduces PIC conceptually, visually, and mathematically. Section 6 walks the key proofs. Section 7 grounds everything in measurements from real models. Section 8 discusses open questions and implications, and Section 9 concludes.

3 The transformer mathematics you need

We need surprisingly little of the transformer to state PIC, essentially the geometry of the *last* step, where the model turns an internal vector into a probability distribution over the vocabulary. This section develops that step and the one concept (direct logit attribution) that lets us decompose it.

Scope of the claims: what the theory actually rests on

It is worth being exact about the hypothesis class, because the word "transformer" is in two ways the wrong size for it. The *static* machinery of PIC, the per-source decomposition, the Gram kernel, the recovered softmax (Thm 1), the power diagram (Prop 1), the two-temperature picture (Thm 5), depends on exactly three properties:

(i) an **additively composed** representation, $r = \sum_j d_j$;

(ii) a **linear readout** into an inner-product geometry, $L_v = \langle r, U_v \rangle + b_v$ (this is what supplies the Gram matrix G); and

(iii) a **softmax over a finite outcome set** V.

None of (i)–(iii) is autoregressive, and none is even specific to transformers, any model meeting them inherits the static theory. What *is* specifically autoregressive is the *empirical* content of Sections 7: the retrieve/select/compose split, the route fractions, the participation ratio PR ≈ 45, the scale-growing composed core, and the ablation results. These are measured on next-token prediction with causal attention, and the very notions of "a decision" and "its sources d_j" are fixed by the autoregressive forward pass. Accordingly we say *transformer* only for architectural facts (the residual stream, attention and MLP writes) and *autoregressive* wherever an empirical or decoding claim is made; the generalisation to non-autoregressive models (e.g. diffusion) is an explicit open direction, treated in Section 8.

3.1 The residual stream and the unembedding

Fix a model with hidden width d and a vocabulary V of tokens. As a transformer processes a sequence, each token position carries a vector in $\mathbb{R}^d$ that is repeatedly updated by attention heads and feed-forward (MLP) blocks. The decisive architectural fact, the one that makes the whole theory possible, is that these updates are *additive*. Each component writes its output into a shared running sum called the **residual stream**. If $r \in \mathbb{R}^d$ is the residual vector at the final position after the last layer, then

$$r \;=\; \sum_j d_j, \tag{1}$$

where the sum runs over every component write: the token embedding, and each attention head and each MLP in each layer. There is no nonlinearity sitting *between* the components and the readout; each d_j contributes to r on equal, additive footing. This is the property the circuits-interpretability literature calls residual-stream linearity, and we will lean on it constantly.

Aside: the residual stream is not the "context." It is tempting to equate the residual stream with what users colloquially call an LLM's *context* (or *context window*), but they are different objects and the distinction matters here. The context window is the *input*: the sequence of tokens (the prompt plus what has been generated), counted in tokens, that the model is allowed to look at. The residual stream is an *internal* object: one vector $r \in \mathbb{R}^d$ *per token position*, the model's running vector-valued summary at that position. The right mental picture is a *shared workspace* or communication bus, sometimes called the model's "working memory": every attention head and MLP reads from it and writes its result back into it. The two connect through attention, which is precisely the mechanism that copies information *from the residual streams of earlier positions* (the context) into the current one (this is also what the "KV cache" stores). So: the context is the text the model attends over; the residual stream is the internal scratchpad into which that context gets distilled, and it is the scratchpad, read out by Eq. (2), that PIC analyses.

To read out a next-token distribution, the model compares r against a fixed matrix of **unembedding rows** $\{U_v \in \mathbb{R}^d\}_{v \in V}$, one per vocabulary token. The **logit** of token v is

$$L_v \;=\; \langle r, U_v \rangle + b_v, \tag{2}$$

an inner product (plus an optional bias b_v). The predicted token is the one with the largest logit, $t = \arg\max_v L_v$, and the full probability distribution is the **softmax**

$$P(v) \;=\; \frac{\exp(L_v)}{\sum_{w \in V} \exp(L_w)}. \tag{3}$$

3.2 Direct logit attribution

Because $r = \sum_j d_j$ and the logit (2) is linear in r, the logit itself splits into one contribution per component:

$$L_v \;=\; \langle \textstyle\sum_j d_j, U_v \rangle \;=\; \sum_j \langle d_j, U_v \rangle \;=\; \sum_j c_j^v, \qquad c_j^v \coloneqq \langle d_j, U_v \rangle. \tag{4}$$

The scalar c_j^v is how much component j "votes" for token v. This exact, bias-free decomposition is **direct logit attribution** (DLA), and it is the bridge between the continuous machinery of the network and the discrete, logical reading PIC will give it: each component j is a *source* casting a real-valued vote c_j^v for each proposition (token) v, and the model's decision is driven by the column sums $L_v = \sum_j c_j^v$.

Two measured quantities that organise everything

From the per-source votes c_j^v we read off two numbers that recur throughout the theory. For a given position with running-up token v^* (the second-best) behind the winner t:

- the **participation ratio** $\mathrm{PR} = \dfrac{\left(\sum_j c_j^t\right)^2}{\sum_j (c_j^t)^2}$, an effective count of how many sources meaningfully contribute to the winning logit (a vote spread evenly over k sources gives $\mathrm{PR} \approx k$; a vote dominated by one source gives $\mathrm{PR} \approx 1$);[a]
- the **normalised margin** $\Delta / \|U_t - U_{v^*}\|$ with $\Delta = L_t - L_{v^*}$, which Section 6 shows is the *exact Euclidean distance* from r to the decision boundary between t and v^*.

The empirical surprise is that PR is *route-invariant*: it is nearly the same for retrieved, selected and composed tokens alike, so no single circuit ever "decides" a token by dominating the sum. Whatever separates the three routes, it is not vote magnitude. The *value* of PR is, however, an architecture-shape parameter rather than a universal constant: it is ≈ 42–49 on the Qwen models that anchor most of our examples (we will write "$\mathrm{PR} \approx 45$" for those), but it ranges more widely across the Pythia ladder ($13 \to 27 \to 63 \to 35$ from 70M to 1B), tracking how many circuits a given architecture has to spread the vote over. Treat 45 as a fact about these models, not a law.

[a] The participation ratio is a borrowed metric: the same $(\sum x)^2 / \sum x^2$ is the *inverse participation ratio* used in physics to measure how localised a state is, and the *effective dimensionality* used in neuroscience to count active modes (e.g. Litwin-Kumar et al. 2017; see Further reading in §10).

3.3 Why "not truth-functional" is the whole game

One more idea closes the primer. Suppose two tokens t and v are near-synonyms, their unembedding directions point almost the same way, $U_t \approx U_v$. Then any evidence for t is almost automatically evidence for v: the two logits move together. Conversely if U_t and U_v are orthogonal, evidence for one says nothing about the other. The matrix of all these pairwise relationships,

$$G_{vw} \;=\; \langle U_v, U_w \rangle, \tag{5}$$

is the **Gram matrix** of the unembedding frame. It is the geometric carrier of the fact that, for a transformer, the "probabilities" of competing tokens are *not* independent; they are coupled through G. This non-independence is exactly the non-truth-functionality that incidence calculus was invented to handle, and G is the object that will let us handle it. We turn to incidence calculus now and return to G in Section 5.

Aside: G is a table of "how similar are these tokens." The inner product $\langle U_v, U_w \rangle$ is the unnormalised cosine similarity between two token directions, the same notion that underlies the familiar idea of *word embeddings*: words with related meanings sit in nearby directions, so their similarity is high (the folklore example being that the embedding of "king" is closer to "queen" than to "bicycle"). G is just the full table of those similarities for the output (unembedding) directions. As a concrete miniature, take three unit token directions in the plane, $U_{\text{Paris}} = (1, 0)$, $U_{\text{Lyon}} = (0.95, 0.31)$, and $U_{\text{bicycle}} = (0, 1)$. Their pairwise inner products form the Gram matrix

$G =$

	Paris	Lyon	bicycle
Paris	1.00	0.95	0.00
Lyon	0.95	1.00	0.31
bicycle	0.00	0.31	1.00

Read it off: the diagonal is 1 because each unit direction is perfectly aligned with itself; the large off-diagonal $G_{\text{Paris,Lyon}} = 0.95$ couples the two French cities, so pushing the model toward one nudges it toward the other; and $G_{\text{Paris,bicycle}} = 0.00$ leaves those two effectively independent. That single off-diagonal 0.95 is exactly why the probabilities of "Paris" and "Lyon" cannot be set independently, the whole non-truth-functional story in one number.

4 Classical incidence calculus, worked through

Incidence calculus was introduced by Alan Bundy in 1985 to do probabilistic reasoning *correctly* in a setting where naive probabilistic logic gives wrong answers. The motivating defect is worth stating carefully because the entire PIC programme is built on repairing it.

4.1 The defect: probability is not truth-functional

A logic is *truth-functional* if the truth value of a compound proposition is a function of the truth values of its parts: $A \wedge B$ is true exactly when A is true and B is true, full stop. Probability has no such property. Knowing $P(A)$ and $P(B)$ does *not* determine $P(A \wedge B)$:

$$P(A \wedge B) \;\neq\; f(P(A), P(B)) \quad \text{for any fixed } f. \tag{6}$$

If $A = B$ then $P(A \wedge B) = P(A)$; if A and B are mutually exclusive then $P(A \wedge B) = 0$; if independent then $P(A)P(B)$. Same marginals, three different answers. The missing information is *how the events overlap*, and no amount of bookkeeping with the marginal numbers alone can recover it.

4.2 Bundy's fix: track the incidences, recover probability

Bundy's move is to stop treating probability as the primitive object and treat *overlap* as primitive. Fix a finite set I of **incidences**: think of them as possible worlds, or sampled situations. To each proposition A attach the set of incidences in which A holds,

$$i(A) \;\subseteq\; I. \tag{7}$$

The logical connectives become honest *set operations* on these incidence sets:

$$i(A \wedge B) = i(A) \cap i(B), \qquad i(A \vee B) = i(A) \cup i(B), \qquad i(\neg A) = I \setminus i(A). \tag{8}$$

Probability is then *recovered* as the relative size of an incidence set,

$$P(A) \;=\; \frac{|i(A)|}{|I|}. \tag{9}$$

Crucially, probability is now a *derived* quantity. Truth-functionality is restored at the level of incidence *sets* (set operations are perfectly compositional), and the non-truth-functionality of probability emerges correctly as a consequence of how the sets overlap.

Worked example: the conjunction that marginals cannot see

Let $I = \{1, 2, 3, 4, 5, 6\}$ be six equally likely worlds. Define two propositions by their incidence sets:

$$i(A) = \{1, 2, 3, 4\}, \qquad i(B) = \{3, 4, 5, 6\}.$$

Then $P(A) = 4/6 = 2/3$ and $P(B) = 4/6 = 2/3$. Now compute the conjunction directly from the sets:

$$i(A \wedge B) = i(A) \cap i(B) = \{3, 4\}, \qquad P(A \wedge B) = \tfrac{2}{6} = \tfrac{1}{3}.$$

A truth-functional guess would have had no way to produce 1/3: the independent guess gives $\frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9}$, the mutually-exclusive guess gives 0. Both are wrong, and they are wrong because they discard the overlap $i(A) \cap i(B)$. Now move the same probabilities to a *different* overlap, $i(B') = \{1, 2, 3, 4\} = i(A)$: same $P(A), P(B')$, but now $P(A \wedge B') = \frac{4}{6} = \frac{2}{3}$. The marginals are identical in both scenarios; only the incidences distinguish them.

4.3 The Gram matrix is already lurking

Notice what $|i(A) \cap i(B)|$ *is*: if we represent each proposition by its incidence *indicator vector* $\mathbf{1}_A \in \{0, 1\}^I$, then

$$|i(A) \cap i(B)| \;=\; \langle \mathbf{1}_A, \mathbf{1}_B \rangle, \tag{10}$$

an inner product. Classical incidence calculus already carries a non-diagonal "Gram matrix" $G_{vw} = |i(v) \cap i(w)|$ counting overlaps. The single conceptual step from Bundy's calculus to PIC is to let these indicator vectors become *arbitrary* directions in an inner-product space rather than $\{0, 1\}$-vectors over a fixed world set, and to notice that a transformer has handed us exactly such directions and exactly such a Gram matrix already, in the unembedding frame of Section 3. That is the projective generalisation, and it is the subject of the next section.

5 Projective Incidence Calculus

We now have both halves of the bridge. From the transformer side (Section 3) we have sources d_j, propositions U_v, votes $c_j^v = \langle d_j, U_v \rangle$, an additive logit $L_v = \sum_j c_j^v$, and a Gram matrix $G_{vw} = \langle U_v, U_w \rangle$. From the logic side (Section 4) we have incidences, propositions identified with incidence vectors, set-operation connectives, probability as overlap, and an overlap Gram matrix. PIC is the identification of these two pictures.

5.1 The projective move, in one definition

Definition 1 (The PIC dictionary)**.** Projective Incidence Calculus *reads the composition core of a transformer as an incidence calculus in which incidence* sets *are replaced by* directions in an inner-product space*:*

Bundy's incidence calculus	***Projective Incidence Calculus***	***In the model***
incidence set $i(A) \subseteq I$	*direction* U_v *in* $\mathbb{R}^d$	*unembedding row of token* v
$\lvert i(A)\rvert/\lvert I\rvert$	$L_v = \langle r, U_v\rangle + b_v$	*token logit*
overlap $\lvert i(A) \cap i(B)\rvert$	$G_{vw} = \langle U_v, U_w\rangle$	*Gram of the unembedding frame*
a source / sample	d_j, *with vote* $c_j^v = \langle d_j, U_v\rangle$	*a head / MLP / embedding write*
$\wedge, \vee$ *(set ops)*	*weighted threshold* $\sum_j w_j x_j > \theta$	*the composed connective*
$P(A) = \lvert i(A)\rvert/\lvert I\rvert$	$\mathrm{softmax}(L)_v$	*output probability*

The word "projective" marks the key liberalisation: in Bundy's calculus the incidence of a proposition is a *set* and overlap is an intersection; in PIC propositions are *directions* and overlap is the kernel value G_{vw}. Because propositions now share a continuous inner-product space, their incidences overlap *intrinsically*, near-synonyms have G_{vw} close to its maximum, mutually exclusive outcomes have $G_{vw} = 0$, and the non-truth-functional structure is carried by G rather than by an explicit list of shared worlds. When G is diagonal (all outcomes mutually exclusive) PIC collapses back to Bundy's classical disjoint-outcome case *exactly*. PIC is therefore a strict generalisation: it adds the off-diagonal of G, the part that couples competing tokens.

5.2 The retrieve/compute split, visually

Before the three interpretations, here is the empirical picture the theory must explain: every next-token decision falls into one of three routes, and the routes are stable across model scale.

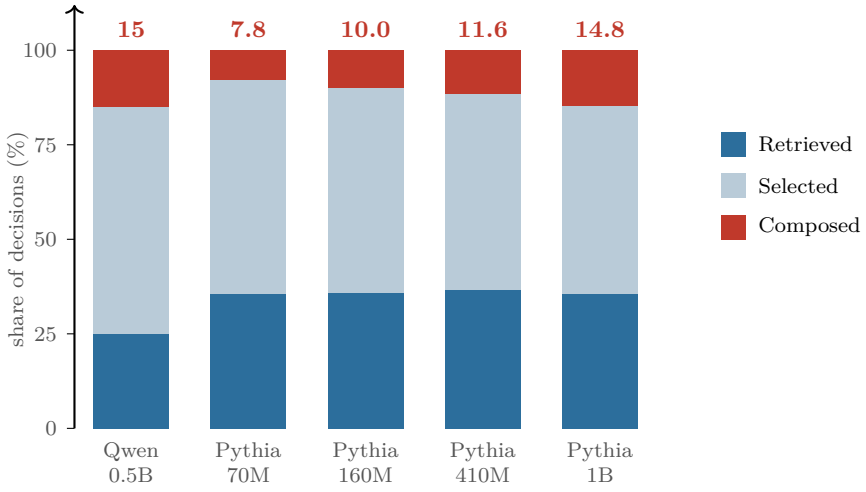

Figure 1: The measured three-way route split. Retrieval (Retrieved + Selected) dominates every model, but the *composed* band, the irreducible computation, grows monotonically with scale against a fixed retrieval store (7.8%→14.8% up the Pythia ladder), even as the underlying geometry stays scale-stable. PIC is the theory of that red band.

Two facts about Figure 1 drive the whole theory. First, the split is *not* a magnitude distinction: the participation ratio PR ≈ 45 is the same for all three routes, so no circuit "wins" by shouting loudest. Second, the routes separate cleanly on two *other*, measurable axes, a decision *margin* (geometry) and a readout *multiplicity* (logic). Those two axes are what the three interpretations formalise.

5.3 One object, three interpretations

The central structural claim of PIC is that the composition core is a *single* mathematical object that can be read three ways, and that which reading you get is governed by a single knob, a

temperature T (equivalently, a choice of semiring). Figure 2 is the map. The word *semiring* carries this section, so we pin it down first.

What is a semiring?

A **semiring** is a set equipped with two operations, written $\oplus$ ("add," for combining *alternatives*) and $\otimes$ ("multiply," for *chaining* steps), that obey the usual algebra of $+$ and $\times$: both are associative, $\otimes$ distributes over $\oplus$, and there are identity elements ($\mathbf{0}$ for $\oplus$, $\mathbf{1}$ for $\otimes$). The *only* thing a semiring drops relative to a ring is subtraction: there need be no additive inverses. That single omission is what lets exotic choices like $\oplus = \max$ count as legitimate "addition."

The point for us is that one and the same expression, the accumulated evidence $L_v = \sum_j c_j^v$ compared across competing tokens, can be *evaluated in different semirings*, and the choice of semiring is exactly the temperature dial:

Semiring	$\oplus$	$\otimes$	**yields**
ordinary arithmetic	$+$	$\times$	sums and products
log-semiring $(T = 1)$	$\log \sum \exp$	$+$	the **softmax** distribution
tropical $((\max, +),\ T = 0)$	$\max$	$+$	the **greedy** $\arg\max$

Reading the same program in the log-semiring gives the probabilities the model samples from; reading it in the tropical semiring gives the single token greedy decoding picks. Nothing about the program changes, only the arithmetic it is interpreted under. This is the machinery behind all three interpretations below and behind Theorem 5.

Aside: this is the temperature knob you already know. The T here is the same temperature exposed in every chat interface and API as a sampling setting. Rescaling the logits by L_v/T before the softmax, $T = 1$ leaves the model's own distribution untouched (the calibrated probabilities), $T > 1$ flattens it toward random, and $T \to 0$ sharpens it onto the single most likely token, which is exactly greedy decoding. So when a user sets "temperature $= 0$" for deterministic output, they are walking the very dial that Theorem 5 formalises: $T = 1$ is the probabilistic-logic reading, $T \to 0$ is the tropical-geometry (greedy $\arg\max$) reading, and Maslov dequantization is the precise statement of that limit.

How a provider actually applies it at runtime. The mechanism is disarmingly simple, and notably it touches neither the weights nor training. At each generation step the model emits the logit vector L (one score per vocabulary token, exactly the L_v of §3). The serving code divides every logit by the requested temperature and then takes the softmax,

$$p_v \;=\; \frac{\exp(L_v/T)}{\sum_w \exp(L_w/T)},$$

and samples the next token from p. That division is the whole of it: T is a per-request number (the `temperature` field in the API call) consumed by the sampler at decode time, not a property baked into the model. At $T = 1$ the logits pass through unscaled; as $T \to 0$ the gaps L_v/T blow up so the top token's probability tends to 1 (providers typically special-case $T = 0$ to mean "take the $\arg\max$" outright, sidestepping division by zero); at $T > 1$ the distribution flattens. Geometrically the winning margin is amplified to Δ/T, so as $T \to 0$ any positive margin becomes infinite log-odds and the winning cell of the power diagram (Fig. 3) takes all the mass. One caveat so the knobs are not conflated: temperature is the *rescaling* step; the separate *truncation* controls *top-k* and *top-p* (nucleus) sampling discard the unlikely tail before sampling and are independent of T. The theory's temperature corresponds only to the rescaling.

What is Maslov dequantization?

This is the bridge that connects the two rows of the semiring table, so it earns a name. Consider the temperature-T way of combining two scores,

$$a \oplus_T b \;:=\; T\,\log\!\big(e^{a/T} + e^{b/T}\big).$$

At $T = 1$ this is ordinary log-sum-exp (the soft, smooth "add" of the log-semiring, the one that yields the softmax). As $T \to 0$ it becomes

$$a \oplus_T b \;\xrightarrow{T\to 0}\; \max(a, b),$$

the hard "add" of the tropical semiring (the one that yields the greedy $\arg\max$). **Maslov dequantization** is exactly this limit: the smooth log-semiring *deforms continuously into* the tropical $(\max, +)$ semiring as the temperature drops to zero. The name is by analogy with physics: T plays the role of a Planck-like parameter, and $T \to 0$ is the "classical limit" in which the soft, quantum-like superposition of many tokens collapses onto the single sharpest one, just as a cooling physical system settles into its lowest-energy ground state. Theorem 5 makes the limit rigorous via a squeeze (the *Maslov sandwich*); the takeaway here is simply that softmax and $\arg\max$ are not two algorithms but two ends of one temperature-controlled deformation.

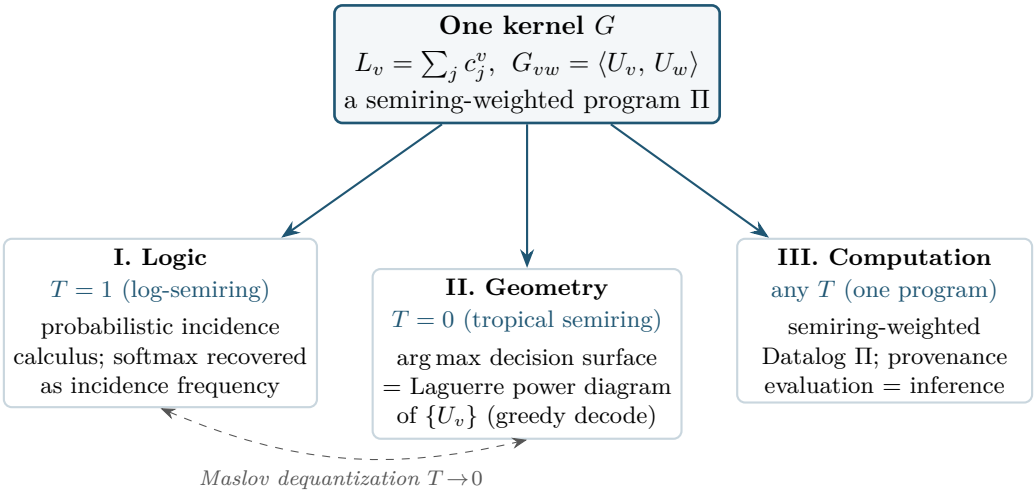

Figure 2: The unification. The composition core is one object, a semiring-weighted program Π carrying the Gram kernel G, read at three settings of a single temperature knob. At $T = 1$ it is a probabilistic logic whose output probabilities are softmax; at $T = 0$ it is a tropical geometry whose decision surface is a power diagram; as an executable artifact it is a Datalog program. *Maslov dequantization* is the formal $T \to 0$ limit linking the two temperatures.

Interpretation I: Logic ($T = 1$). Sources S (the circuits) are vectors d_j; propositions V (the tokens) are directions U_v. The vote $c_j^v = \langle d_j, U_v \rangle$ is the direct logit attribution, and aggregated evidence is the residual $r = \sum_j d_j$. Where Bundy's incidence of a proposition is a set $i(v)$, PIC's is the linear functional $\langle \cdot, U_v \rangle$, and the overlap of two propositions is the kernel value G_{vw}, not a set intersection. Combination is signed linear accumulation in the exponent, $L_v = \sum_j c_j^v$, which is *equivalently a product of experts*, each source reweighting a proposition's mass by $\exp(c_j^v)$. The decision is a weighted threshold (Boolean $\cap/\cup$ become $\sum_j w_j x_j > \theta$), and, as Theorem 1 will show, the output probabilities come out *exactly* as incidence frequencies, i.e. the softmax, parameter-free.

"Product of experts" vs. the "Mixture of Experts" you may have heard of

The phrase *product of experts* sounds like the *Mixture of Experts* (MoE) in modern large models, and the shared word "experts" hides a real difference. Both combine several opinions into one distribution, but they combine them in opposite ways.

Product of experts (PoE) **multiplies** the experts' distributions and renormalises, $p(v) \propto \prod_k p_k(v)$. Taking logs, the log-scores *add*: $\log p(v) = \sum_k \log p_k(v) - \log Z$. An outcome survives only if *every* expert grants it some mass, so PoE behaves like a logical **AND**: it intersects constraints and *sharpens* the distribution. This is precisely what the residual stream does, $L_v = \sum_j c_j^v$ is a sum of log-votes, so by Theorem 1 the model's readout *is* a product of experts (the "experts" being the circuits d_j).

Mixture of Experts (MoE) instead **averages** the experts' distributions with gating weights from a learned router, $p(v) = \sum_k g_k\, p_k(v)$ with $\sum_k g_k = 1$. An outcome is likely if *any* weighted expert grants it mass, so a mixture behaves like a soft **OR**: it unions options and *hedges*. In today's transformers "MoE" names an architectural trick: a router sends each token to its top-k feed-forward *expert* subnetworks (the experts the models in §7 page in from disk), saving computation.

	Product of experts	**Mixture of experts**
combine by	multiply ($\times$), log-scores add	weighted average ($+$)
logical flavour	AND / intersection / sharpen	OR / union / hedge
a router/gate?	no, all experts always contribute	yes, selects top-k experts
"experts" are	the circuits d_j (heads, MLPs)	dedicated FFN subnetworks

The two are complementary, not rival, and an MoE model uses both at once: the router *selects* which expert subnetworks run (a mixture step), and then whatever components fire write additively into the residual stream and are read out as a *product* of experts. MoE picks the panel; PoE tallies its votes.

Interpretation II: Geometry ($T = 0$). Take the hard decision $\arg\max_v(\langle r, U_v\rangle + b_v)$. The function

$$M(r) \;=\; \bigoplus_v (b_v \otimes x^{U_v}) \;=\; \max_v(\langle r, U_v\rangle + b_v) \tag{11}$$

is a *tropical polynomial* in r (in the $(\max, +)$ semiring, $a \oplus b = \max(a, b)$ and $a \otimes b = a + b$). Its linear regions, the cells of r-space on which a fixed token wins, tile space into a **Laguerre power diagram** of the sites $\{U_v\}$.

To unpack that name: start from the everyday *Voronoi diagram*, which carves a map into the region closest to each site, like the coverage zones of cell towers or the catchment areas of shops. A **power diagram** (also called a *Laguerre diagram*) is the same idea but each site carries a *weight*, so the boundaries shift toward the lighter sites; ordinary Voronoi is the special case of equal weights. Here the sites are the token directions $\{U_v\}$, the weights encode the biases and lengths $(b_v, \|U_v\|^2)$, and the cell a residual vector r lands in names the token the model predicts. The wall between two adjacent cells, where two tokens tie for the lead, is what tropical geometry calls the *tropical hypersurface* (just the set of points where the max in $M(r)$ is achieved by two tokens at once).

This is the picture of Figure 3, and its facet distances are exactly the normalised margins of Section 3: a residual sitting deep inside its cell is a confident, large-margin decision, and one hugging a wall is a near-tie.

Interpretation III: Computation (executable). The same core is a finite relational computation: a **semiring-weighted Datalog program** Π. Propositions are facts with directions U_v and Gram G; sources are clauses (an induction head, for instance, is the recursive clause `next(T):-match_prefix(P),follows(P,T)`); the retrievable fragment is compact stratified clauses and the composed fragment is a dense aggregate. Evaluating Π under a *provenance semiring is* inference, and the temperature is the semiring choice: the log-semiring gives the softmax distribution, the tropical semiring gives the greedy decode.

5.4 The geometry, visually

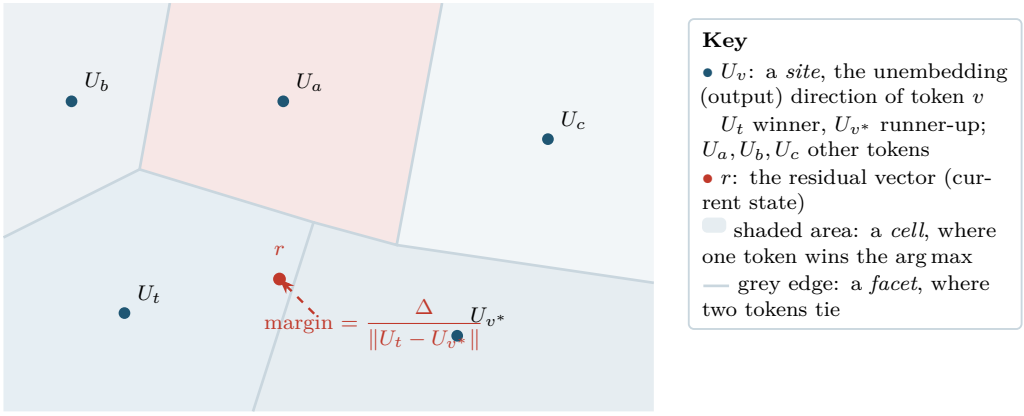

Figure 3: Interpretation II made concrete. The residual-space decision surface is the Laguerre power diagram of the unembedding directions $\{U_v\}$ with weights $(b_v, \|U_v\|^2)$: each cell is the set of residual vectors r for which a given token wins the arg max. A *retrieved* token sits deep inside its cell (large margin); a *composed* token sits close to a facet (small margin). The normalised margin $\Delta/\|U_t - U_{v^*}\|$ is exactly the perpendicular Euclidean distance from r to the nearest facet (Proposition 2).

The empirical payoff of Figure 3 is that this facet distance is *measurable* and *monotone* across the three routes: retrieved tokens sit far inside their cells, composed tokens sit near facets. Geometry, not magnitude, is one of the two axes that separate the routes. The other axis, *readout multiplicity*, is logical and is the subject of the next section's key proofs.

6 The key proofs

The theory's claims are not posited; they are recovered as consequences of the single additive structure $L_v = \sum_j c_j^v$ together with the Gram matrix G. We walk through the load-bearing results. Each is elementary: this is part of the point: the conceptual leap is in the identification, after which the mathematics is short. (Each result below corresponds to a theorem in the source paper and has been transcribed into a proof skeleton checkable by an interactive theorem prover; we give the human-readable argument.)

6.1 The anchor: probability is recovered as an incidence frequency

This is the result that most directly inherits Bundy's "probability as proportion of worlds," and it is the cleanest justification for calling the core an incidence calculus at all.

Theorem 1 (Recovered probability)**.** *Maintain a measure over propositions with a uniform base $M_0 > 0$, and let each source j reweight every proposition v's mass multiplicatively by* $\exp(c_j^v)$.

Then the normalised mass of v is exactly the softmax of the logits:

$$\frac{m(v)}{\sum_{w\in V} m(w)} \;=\; \frac{\exp(L_v)}{\sum_{w\in V}\exp(L_w)} \;=\; \mathrm{softmax}(L)_v, \qquad L_v = \sum_j c_j^v.$$

The Gibbs/softmax measure is recovered as a PIC incidence frequency, parameter-free, and is equivalently the distribution of the i.i.d.-Gumbel arg max *over $\{L_v\}$ (the Gumbel-max trick).*

Proof. After all sources act, the mass of v is the base times the product of the per-source reweightings,

$$m(v) \;=\; M_0 \prod_j \exp(c_j^v) \;=\; M_0 \exp\Big(\sum_j c_j^v\Big) \;=\; M_0 \exp(L_v).$$

The total mass is $\sum_w m(w) = M_0 \sum_w \exp(L_w)$, pulling the common base M_0 out of the sum (this uses only distributivity). Since $M_0 > 0$ it is nonzero and cancels:

$$\frac{m(v)}{\sum_w m(w)} = \frac{M_0\exp(L_v)}{M_0\sum_w \exp(L_w)} = \frac{\exp(L_v)}{\sum_w \exp(L_w)}.$$ □

Three things make this more than algebra. (i) The "reweight each proposition's mass by exp(vote)" rule is exactly a *product of experts*: each source is an expert multiplying the running distribution. (ii) The base measure M_0 being uniform is the "proportion of worlds" assumption; PIC adds no free parameter, the softmax falls out. (iii) Crucially the Gram matrix G is *not* injected as correlated noise that would spoil the exact recovery; it is already structural in the unembedding frame, so it shapes the static logits (this theorem) and the competition geometry (Theorem 3) at once, with no variance trade-off.

6.2 Cardinality-inertness: the count carries no causal weight

A natural worry: maybe a token wins because *many* sources vote for it. The next result says the raw count of voting sources is causally inert; only the totals matter.

Theorem 2 (Cardinality-inertness)**.** *Under the projective pairing the decision depends only on the column totals $L_v = \sum_j c_j^v$ (hence on the pairwise differences Δ and the differential incidences D_j), and is invariant to the readout multiplicity μ_t, the number of individual sources whose own* arg max *is t. Two source tensors with equal column sums induce the same* arg max *even when their per-source* arg max *counts differ.*

Proof. The decision is $t = \arg\max_v L_v$, a function of the totals $\{L_v\}$ alone. If c and c' satisfy $\sum_j c_j^v = \sum_j c_j'^v$ for every v, then $L_v = L'_v$ for all v and the two arg max sets coincide. The multiplicity $\mu_t = \#\{j : \arg\max_v c_j^v = t\}$ never enters L, so it cannot affect the decision. □

Worked example: same decision, opposite "multiplicities"

Take two sources voting over three tokens $\{A, B, C\}$. Write each source's votes as a row $c_j = (c_j^A, c_j^B, c_j^C)$, and stack them; the column sums are the logits $L = (L_A, L_B, L_C)$.
Scenario 1 (consensus).

	A	B	C
c_1	6	3	3
c_2	4	3	3
L	**10**	6	6

$\arg\max c_1 = A,\ \arg\max c_2 = A$
$\Rightarrow\ \mu_A = 2$ (both sources already pick A)

Scenario 2 (emergent).

	A	B	C
c_1	5	6	0
c_2	5	0	6
L	**10**	6	6

$\arg\max c_1 = B,\ \arg\max c_2 = C$
$\Rightarrow\ \mu_A = 0$ (no source picks A)

Both source tables have the *identical* column sums $L = (10, 6, 6)$, so they produce the identical softmax and the identical winner, A, by a margin of $\Delta = 4$. Yet the multiplicities could hardly be more different: in Scenario 1 both sources already vote A ($\mu_A = 2$, the textbook "retrieved" look), while in Scenario 2 one source insists on B, the other on C, and *neither* proposes A ($\mu_A = 0$, the "composed" look where A is the winner of a sum that is the winner of no summand). The count of supporters swung from two to zero and the decision did not move at all. That is cardinality-inertness: only the totals L vote, and μ_A is a label we read off afterwards, never a cause.

This matters because μ_t is precisely what would distinguish "retrieved" (some single source already picks t, $\mu_t \geq 1$) from "composed" ($\mu_t = 0$, no single source picks t). The theorem says μ_t is a *readout property*, a diagnostic label, not the causal lever; empirically the causal variable is the pivotality/margin, and μ_t is only a proxy for where the margin sits.

6.3 Non-truth-functionality is a kernel

What is a kernel?

In this setting a **kernel** is a function that reports the *similarity* of two items as an inner product, $k(v, w) = \langle U_v, U_w \rangle$. It is *symmetric*, $k(v, w) = k(w, v)$, and *positive semidefinite* (the similarity table it produces has no negative eigenvalues); those two properties are exactly what let it stand in for "overlap." Tabulating the kernel over every pair of tokens gives the Gram matrix, $G_{vw} = \langle U_v, U_w \rangle$, so "the kernel" and "the Gram matrix" are the same object seen as a function versus as a table. This is the same notion used in machine learning's *kernel methods* (the "kernel trick"): work entirely with pairwise similarities and never need the raw coordinates. Saying *non-truth-functionality is a kernel* means the entire coupling that makes token probabilities non-independent is carried by this one pairwise similarity function G, nothing more elaborate; when G is diagonal the similarities vanish and the classical independent-outcome case returns.

Do not confuse it with the other "kernels." The word is badly overloaded, and only the similarity sense above is meant here. In particular it is *not*:

- the **operating-system kernel**, the small privileged program at the centre of an OS (Linux, etc.) that controls the hardware, memory, and scheduling. There "kernel" just means "the innermost core," and there is no inner product or similarity in sight;
- the **null space** (also called the kernel) of a linear map, $\ker T = \{x : Tx = 0\}$, the set of vectors a map sends to zero, an algebraic object, not a similarity;
- the **convolution kernel** of image filtering and CNNs, a small sliding weight stencil. (Confusingly, this one *is* from machine learning, but it is a filter, not a similarity function.)

The non-OS senses loosely share the flavour of "an essential core," but only *our* kernel is a pairwise similarity $\langle U_v, U_w \rangle$; that is the single meaning to carry through the rest of the document.

This result locates the non-truth-functional coupling precisely in G and shows that the classical disjoint-outcome case is the diagonal limit.

Theorem 3 (Non-truth-functionality budget)**.** *For unit directions U_t, U_v the squared separation is*

$$\|U_t - U_v\|^2 \;=\; 2(1 - \langle U_t, U_v \rangle) \;=\; 2(1 - \rho_{tv}), \qquad \rho_{tv} = \langle U_t, U_v \rangle.$$

Hence as the runner-up coherence $\rho_{tv} \to 1$ (near-synonyms) the differential evidence becomes common-mode *and collapses; and when G is diagonal ($\rho_{tv} = 0$, mutually exclusive outcomes) it reduces to the classical disjoint-outcome distance $\|U_t - U_v\|^2 = 2$.*

Proof. Expand the norm by polarisation:

$$\|U_t - U_v\|^2 = \langle U_t - U_v, U_t - U_v \rangle = \|U_t\|^2 - 2\langle U_t, U_v \rangle + \|U_v\|^2.$$

Substituting the unit norms $\|U_t\| = \|U_v\| = 1$ gives $2 - 2\langle U_t, U_v \rangle = 2(1 - \rho_{tv})$. Setting $\rho_{tv} = 0$ yields 2. □

The differential incidence $D_j = c_j^t - c_j^{v^*} = \langle d_j, U_t - U_{v^*} \rangle$ is the amount by which source j pushes the t-versus-v^* margin. The theorem says competition hardness is read directly off the frame Gram: when two outcomes are near-synonyms, evidence for one cancels in the difference (common mode), and the genuine competition lives in ρ. This is the structural fact that the whole non-truth-functionality story rests on.

Worked example: how coherence eats the evidence "budget"

Put a single source d that is pure, strong evidence for the winner t: let it point exactly along U_t with strength 3, i.e. $d = 3\,U_t$. Its raw votes are

$$c^t = \langle d,\, U_t\rangle = 3, \qquad c^{v^*} = \langle d,\, U_{v^*}\rangle = 3\,\rho_{tv^*},$$

so the amount it actually moves the t-vs-v^* margin is the *differential* incidence

$$D = c^t - c^{v^*} = 3\,(1 - \rho_{tv^*}).$$

Now watch what the runner-up's coherence ρ does to that same physical source:

ρ_{tv^*}	$\|U_t - U_{v^*}\|^2$	$D = 3(1-\rho)$	interpretation
0.00	2.00	3.00	orthogonal: full strength
0.50	1.00	1.50	half survives
0.95	0.10	0.15	near-synonyms: almost all cancels

The source delivers "3 units of evidence for t" in every row, yet the margin it buys collapses from 3 to 0.15 as t and v^* line up. The reason is geometric and is exactly Theorem 3: the most any unit-strength source can contribute to the t-vs-v^* margin is $\|U_t - U_{v^*}\| = \sqrt{2(1-\rho)}$ (Cauchy–Schwarz), so the separation of the two outcomes is a hard *budget* on how much differential evidence is even available. When the outcomes are near-synonyms that budget is tiny, the shared part of the evidence is *common mode* (it lifts t and v^* together and cancels in the difference), and only the sliver orthogonal to the shared direction can separate them.

6.4 Expressivity: composed tokens need a weighted threshold

Theorem 4 (Weighted-threshold expressivity)**.** *There exist composed conclusions ($\mu_t = 0$, no single source's* arg max *is t) that are the* arg max *of the weighted sum $\sum_j c_j^v$ yet are not expressible in the Horn / $\cap$–$\cup$ fragment of classical incidence calculus. A two-source, three-outcome witness: let*

$$c_1 = (2, 3, 0), \qquad c_2 = (2, 0, 3), \qquad L = c_1 + c_2 = (4, 3, 3).$$

Then $L_0 > L_1$ and $L_0 > L_2$ (the sum prefers outcome 0), while source 1 prefers outcome 1 ($c_1^1 > c_1^0$) and source 2 prefers outcome 2 ($c_2^2 > c_2^0$). Outcome 0 is the winner of the sum but the arg max *of no single source.*

Proof. Direct arithmetic: $L_0 = 2 + 2 = 4$, $L_1 = 3 + 0 = 3$, $L_2 = 0 + 3 = 3$, so $L_0 > L_1$ and $L_0 > L_2$. And $c_1^1 = 3 > 2 = c_1^0$, $c_2^2 = 3 > 2 = c_2^0$. Hence $\mu_0 = 0$: no single source selects 0, yet the weighted threshold $\sum_j c_j^v$ does. □

This is exactly the operational definition of emergence, "the arg max of a sum that is the arg max of no summand", realised by a weighted-threshold connective but by no singleton sufficient sub-conjunction.

Caveat: $\mu_t = 0$ is necessary for emergence but not sufficient for *irreducibility*. It is tempting to equate "no single source picks t" ($\mu_t = 0$) with "t cannot be reached by any rule." The sharper version of the theory separates the two. Call a conclusion **reducible** if some *proper* sub-coalition of sources (more than one, but not all) already has t as the arg max of *its* partial sum, and **irreducible** if no proper sub-coalition does. A token can have $\mu_t = 0$ and still be reducible: no *single* source picks t, yet a pair of them, summed, already does, which a Horn rule over that pair could capture. Genuine irreducibility, the property that truly needs the full weighted threshold and no $\cap/\cup$ formula, requires that *no* proper sub-coalition suffices. (Our two-source witness above is in fact irreducible: the only proper sub-coalitions are the singletons $\{1\}$ and $\{2\}$, and neither picks 0. With three or more sources one can build $\mu_t = 0$ tokens that are merely reducible.) So $\mu_t = 0$ is an *upper bound* on emergence; the measured "$\sim$80% strictly emergent" figure counts $\mu_t = 0$, and proving a general criterion that isolates the irreducible ones is the open problem of Section 8.

Worked example: why a weighted threshold beats AND/OR

Take the witness above, two sources voting over outcomes $\{0, 1, 2\}$:

	0	1	2
c_1	2	3	0
c_2	2	0	3
$L = c_1 + c_2$	**4**	3	3

Read it two ways. As a **Boolean rule** (the $\cap/\cup$ fragment), each source "concludes" only its own top outcome: source 1 asserts 1, source 2 asserts 2. Any conjunction (AND, $\cap$) of these conclusions is empty ($\{1\} \cap \{2\} = \varnothing$); any disjunction (OR, $\cup$) is $\{1, 2\}$. Neither connective can ever output 0, because no source ever proposes 0. A truth-functional logic that only sees "which outcome did each source pick" is therefore *blind* to the answer.
As a **weighted threshold**, $\arg\max_v \sum_j c_j^v$, the picture changes: the modest-but-agreeing support for 0 ($2 + 2 = 4$) outweighs the strong-but-lonely support for 1 and for 2 ($3 + 0$ and $0 + 3$). Outcome 0 wins, although it is nobody's first choice. The connective at work is $\sum_j w_j c_j^v > \theta$, a linear threshold unit (a perceptron), and the example is the discrete cousin of XOR: a verdict that no AND/OR of the inputs reproduces but a weighted sum does. That gap, conclusions reachable by the real-valued threshold but not by the Boolean fragment, is precisely the extra expressive power the *composed* route has over pure retrieval (the $\mu_0 = 0$ signature, "the winner of a sum that is the winner of no summand").

6.5 The margin is a distance, and the cells are a power diagram

The two geometric facts of Figure 3 are propositions, not pictures.

Proposition 1 (Cells are a Laguerre power diagram)**.** *The linear regions of the max-logit $M(r) = \max_v(\langle r, U_v\rangle + b_v)$ are the Laguerre power diagram of the sites $\{U_v\}$ with weights $(b_v, \|U_v\|^2)$: the power-distance difference between two sites equals -2 times the score difference, so the cell of minimum power distance is exactly the* arg max *token.*

Proof sketch. Polarise the squared distances $\|r - U_v\|^2$ and $\|r - U_w\|^2$ as in Theorem 3. Subtracting, the $\|r\|^2$ terms cancel and one obtains, with weight $\omega_v = \|U_v\|^2 + 2b_v$,

$$(\|r - U_v\|^2 - \omega_v) - (\|r - U_w\|^2 - \omega_w) = -2\big((\langle r, U_v\rangle + b_v) - (\langle r, U_w\rangle + b_w)\big).$$

The left side is the difference of *power distances*; the right is -2 times the difference of scores. Minimum power distance therefore coincides with maximum score, which is the arg max token. □

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

Proof. By linearity $\langle r, U_t - U_{v^*}\rangle = \langle r, U_t\rangle - \langle r, U_{v^*}\rangle$, so the facet numerator $\langle r, U_t - U_{v^*}\rangle + (b_t - b_{v^*})$ equals $L_t - L_{v^*} = \Delta$. The set $\{r : \langle r, U_t - U_{v^*}\rangle + (b_t - b_{v^*}) = 0\}$ is the bisecting hyperplane (the facet) with normal $U_t - U_{v^*}$; the signed distance from a point to such a hyperplane is the linear form evaluated at the point divided by the norm of the normal, i.e. $\Delta/\|U_t - U_{v^*}\|$. □

6.6 The capstone: two temperatures, one program

The unification of Figure 2 is itself a theorem: the same program Π, evaluated under two different semirings, returns the softmax distribution and the greedy decode, and the two are joined by Maslov dequantization.

Theorem 5 (Two-temperature soundness)**.** *Read* Π *as a semiring functional aggregate query in which sources are combined by* $\otimes$ *along the residual derivation (so* $\bigotimes_j c_j^v = \sum_j c_j^v = L_v$*) and competing propositions are combined by* $\oplus$*. Then on a finite nonempty* V*:*

- *under the* ***log-semiring*** *(*$\oplus = \log\text{-}\sum\exp$*,* $\otimes = +$*), the aggregate is* $\log \sum_v \exp(L_v)$ *and the per-proposition share is* $\exp(L_v)/Z = P(v)$*, the softmax distribution (Interpretation I,* $T = 1$*);*
- *under the* ***tropical semiring*** *(*$\oplus = \max$*,* $\otimes = +$*), the aggregate is* $\max_v L_v$ *with witness* $\arg\max_v L_v$*, the greedy decode (Interpretation II,* $T = 0$*);*

and the two are linked by the Maslov sandwich

$$\max_v L_v \;\le\; T \log \sum_v \exp(L_v/T) \;\le\; \max_v L_v + T \log|V|,$$

so $T \log \sum_v \exp(L_v/T) \to \max_v L_v$ *as* $T \to 0$*.*

Proof sketch. The log-semiring claim is the identity defining log-sum-exp and Theorem 1. For the tropical claim, on a finite nonempty set the maximum is attained, $\max_v L_v \in \{L_v\}$, giving an explicit witness. For the sandwich: the lower bound holds because the max term is one of the summands and log is monotone, so $\sum_v \exp(L_v/T) \ge \exp(\max_v L_v/T)$ and multiplying by T gives $T \log \sum_v \exp(L_v/T) \ge \max_v L_v$. The upper bound holds because every summand is at most $\exp(\max_v L_v/T)$, so the sum is at most $|V| \exp(\max_v L_v/T)$, and taking $T \log$ gives $\max_v L_v + T \log|V|$. As $T \to 0$ the slack $T \log|V| \to 0$, squeezing the middle to the max. This $T \to 0$ limit *is* Maslov dequantization, the homomorphism from the log-semiring to the tropical semiring. □

This is the formal content of "one program, two temperatures." The softmax you sample from at $T = 1$ and the greedy arg max you decode at $T = 0$ are not two algorithms; they are one semiring program evaluated under two semirings, and the temperature interpolates continuously between them. The logic picture (I) and the geometry picture (II) are the two ends of this single dial.

7 Worked examples from a real model

PIC is a measured theory: every quantity above is computed on real autoregressive models (primarily two Qwen2.5-0.5B variants, the Coder-Instruct and the base Instruct,[1] and the Pythia ladder from 70M to 1B) with the `fieldrun` toolchain,[2] which decompiles a model into a flat retrieval store plus a composition kernel and runs explain-only probes that classify each decision and report PR, the margin, μ_t, and the nearest power-diagram facet. We ground the three routes in concrete decisions.

7.1 A retrieved decision

Consider a position where the prompt ends "…the capital of France is". A single induction/n-gram circuit already places " Paris" at the top of its own vote: $\mu_t \geq 1$. In power-diagram terms (Figure 3) the residual r sits *deep* inside the " Paris" cell, a large normalised margin, so a small perturbation will not flip the decision. The measured signature: large facet distance, $\mu_t \geq 1$, and (the diagnostic that makes this "retrieved" rather than merely "selected") the candidate set is exactly one store idiom's top-1. The participation ratio is still ≈ 45; even this canonical lookup is assembled from a $\sim$45-way additive sum, not decided by one dominant circuit.

7.2 A composed decision

Now a position whose continuation no single circuit proposes, a token that is the arg max of the summed vote but of no summand, $\mu_t = 0$. The witness of Theorem 4 is the schematic: source 1 votes $(2, 3, 0)$, source 2 votes $(2, 0, 3)$, and the sum $(4, 3, 3)$ elects outcome 0, which neither source chose. On the real model the measured signature is the mirror image of the retrieved case: the residual sits *near a facet* (small margin), $\mu_t = 0$, and, a sharper geometric test, the nearest facet is *not* the bisector with the store's own predicted token for $\sim$85% of composed tokens, so composition is a genuine *non-local* divergence from the rule, not a near-miss. Roughly 80% of composed tokens are strictly emergent ($\mu_t = 0$).

7.3 The causal test: ablation and the diffuse repair

What is ablation?

Ablation is the standard interpretability move for telling *causes* apart from mere *correlations*. A direct logit attribution (§3) only tells us a component *voted* for the winning token; it does not tell us the component was *needed*. To test necessity we *intervene*: rerun the forward pass with one component switched off, usually by zeroing its write d_j into the residual stream (hence "ablate," as in surgically removing it), and see what changes. The borrowed vocabulary is causal. The unaltered run is the *factual*; the run with d_j removed is the *counterfactual*; if the predicted token changes between them we call it a **flip**. A high flip rate means the component was load-bearing for that decision; a low flip rate means the rest of the model can cover for it. Ablation can be single (remove one component) or joint (remove a coalition at once), and the same idea applies to downstream blocks to ask whether the model *repairs* a damaged prediction. Everything in this section is read off flip rates under such interventions.

The route labels are not merely correlational. Turning the readout into an *intervention*, zeroing the top-attribution circuit in the forward pass and asking whether the prediction flips, converts the correlations into causal structure, and four facts fall out that the theory predicts.

[1] Qwen2.5: see the Qwen Team's technical report in the references (§10, additional attributions); the Pythia suite is ref. [3]. The two Qwen variants share a vocabulary, hence a shared unembedding frame.

[2] `fieldrun`: https://github.com/jascal/fieldrun.

Repair, recovery, and rescue: the model heals itself

These three words name one phenomenon, and it is easy to trip over, so we separate two effects of an ablation. The **direct effect** is the immediate arithmetic: delete a component's write d_j and the winning logit drops by its vote, which on its own may hand the lead to the runner-up (a predicted flip). But the forward pass does not stop there. The components *downstream* of the ablated one now read a changed residual stream, and they react: some of them write *more* toward the original token, pushing it back into the lead. That compensating reaction is the **indirect effect**, and the model's tendency to do it is **self-repair** (or *recovery*). When repair succeeds, a decision the direct effect said would flip does *not* flip; we call that a **rescue**. So the three terms line up as: repair/recovery is the mechanism, rescue is its outcome on a particular token.

Two consequences are worth holding onto. First, single-component ablation therefore *understates* how a prediction is really computed: the naive direct effect over-predicts fragility because it ignores the healing. (This matches the known self-repair, or "Hydra," behaviour, cut one head and others grow to compensate.) Second, the repair here is **diffuse**: there is no single backup unit you could disable to stop it. Knocking out any one downstream block still leaves most rescues working (18–34% intact), and each head accounts for only about 1/PR of the repair, the same PR ≈ 45 spread-out structure as the readout itself. Finding (4) below quantifies this, and Theorem 6 turns "spread over PR modules" into the statement that no bounded-size intervention can localise it.

(1) Fragility is route-ordered and margin-governed. The flip rate under single-circuit ablation runs Retrieved ~22% < Selected ~40% < Composed ~54%: a composed token flips about 2.4× as often as a retrieved one. But this tracks the *margin*, a margin-matched split shows margin, not route label per se, is the governor, exactly as Proposition 2 predicts.

(2) The causal variable is pivotality, not multiplicity. A logistic control shows the linear flip identity flip $\iff \Delta < D_j$ holds as a near-perfect necessary condition, with μ_t contributing essentially zero once (Δ, D_j) are included, μ_t is a *proxy* for margin position, not an independent cause. This is Theorem 2 (cardinality-inertness) seen causally.

(3) Redundancy is non-compensatory. Removing one supporter is not caught by the others: agreement among many weak readers is not fault tolerance, and individually < 10%-of-logit supporters provide essentially no cushion.

(4) The repair is diffuse, not a lever. When the linear identity predicts a flip, an indirect downstream recomposition *rescues* the original token more than half the time, and the rescue is *not localisable*: ablating any single downstream block leaves 18–34% of rescues intact, and the per-head un-rescue rate is order 1/PR (measured 4.0–4.1% against 1/PR = 2.5–2.8%). There is no surgical repair target, the repair is distributed for the same reason the readout is, PR ≈ 45.

This last fact is Theorem 6 below, the diffuseness result, seen in the data: a causal property spread equitably over PR modules has single-source influence $O(1/\mathrm{PR})$, so no bounded-size formula can localise it.

Theorem 6 (Diffuseness)**.** *Suppose a causal quantity decomposes as $E = \sum_{m=1}^{\mathrm{PR}} e_m$ with equitable contributions $e_m = E/\mathrm{PR}$. Then single-source relative influence is $e_m/E = 1/\mathrm{PR}$, a k-source body captures only $|A|/\mathrm{PR}$ of E, and as $\mathrm{PR} \to \infty$ the captured fraction $\to 0$: no bounded-size PIC formula localises the quantity, and $P\langle$single-module intervention alters $E\rangle = O(1/\mathrm{PR})$.*

Proof. With $e_m = E/\mathrm{PR}$ and $E \neq 0$, $e_m/E = 1/\mathrm{PR}$ immediately. For $A \subseteq \{1, \ldots, \mathrm{PR}\}$, $\sum_{m \in A} e_m/E = |A| \cdot (E/\mathrm{PR})/E = |A|/\mathrm{PR}$. For fixed $k = |A|$, $k/\mathrm{PR} \to 0$ as $\mathrm{PR} \to \infty$. □

The deployment consequence is concrete: because the retrievable tier is low-participation-ratio and the computed core is the high-PR, diffuse layer, you can *quantise the retrievable tier hard* but should *protect the computed core*, which is exactly where quantisation error concentrates in the measurements (composed int4 flip rate 5.9–18.8% versus retrieved 3.6–2.4%).

8 Open questions and implications

PIC is anchored where it can be measured and deliberately conjectural where the measurements stop. Incompleteness is treated as first-class: several of the results above are proved only on the measured set, and the corresponding general statements are explicit "frontier holes." The sharpest open questions:

Soundness and completeness of weighted incidence resolution.
Is the coalition bound $\sum_{j \in S'} D_j$ a *sound* inference rule for weighted incidence resolution, and is the resolution complete? The measured coalition additivity predicts joint-ablation flips at $\sim$75–83%, but a general soundness/completeness theorem is open.

The general expressivity separation.
Theorem 4 exhibits composed tokens inexpressible in the Horn / $\cap$–$\cup$ fragment on the measured $\mu_t = 0$ set. Proving the separation *in general* rather than only on measured points is open; it is the precise sense in which "composition is genuinely more expressive than retrieval" would become a theorem.

The support number $\sigma(t)$.
Define the **support number** $\sigma(t)$ of a token to be the size of its *smallest sufficient circuit set*: the fewest sources whose partial sum already makes t the winner. It is the quantitative handle on the reducible/irreducible distinction of §6: an irreducible composed token is one whose support requires *all* of its sources ($\sigma(t)$ large and no proper sub-coalition suffices), whereas a merely $\mu_t = 0$ token may have small $\sigma(t)$. Two questions are open. First, does $\sigma(t)$ scale with the participation ratio PR (i.e. is the size of the deciding coalition tied to the effective number of contributing sources)? The relationship is cleanest at the individual-circuit level; measured at coarser block granularity it mildly reverses, so the circuit level is where the question is well posed. Second, and relatedly, is there a general criterion that separates the genuinely irreducible composed tokens (no proper sub-coalition decides t) from the merely $\mu_t = 0$ ones? The two classes already come apart on small machine-checked witnesses; a general separation would turn the upper-bound nature of the "$\sim$80% strictly emergent" figure into a sharp count.

The provenance gap.
The aggregate value of Π is correct under any G (the static decomposition is exact). What a dense, non-diagonal G breaks is not the value but the *provenance*: the per-source explanation and the ablation counterfactuals. Scalar semiring Datalog is provenance-exact only on a diagonal G (the independent-outcome case); for dense G, scalar provenance cannot carry the cross-outcome correlation the frame Gram encodes. Closing this "provenance gap" would simultaneously settle the non-truth-functionality theorem; the one-to-one correspondence is the strongest evidence the three interpretations are one theory.

The tropical-rank floor.
The irreducible computation is conjectured to be lower-bounded by the gap between the tropical (Barvinok) rank of the core's decision map and that of any flat lookup table, a gap that a linear (SVD) rank structurally cannot measure. This would explain *why* a low-rank linear

approximation misranks the composed core: its hardness is a count of tropical monomials (decision cells), which a Frobenius rank does not see.

Treewidth as a third measure.
The same wall may be seen as the treewidth of the core's factor graph: sum-product is exponential in treewidth, so a high-treewidth dense-G region is at once intractable and non-compact. Whether participation ratio, tropical rank, and treewidth coincide (one wall, three measures) or diverge on the dense fragment is open.

Beyond autoregressive transformers.
The static algebra of Section 3 needs only additive composition, a linear readout, and a softmax over a finite vocabulary, so it is not, in itself, autoregressive or even transformer-specific. The empirical taxonomy, by contrast, is defined by the autoregressive forward pass. The cleanest generalisation target is therefore a *non-autoregressive* model, and diffusion language models split the question in two. *Discrete / masked diffusion* (with a transformer backbone and a per-position categorical softmax at each denoising step) inherits the static decomposition, DLA, G, recovered probability, the power diagram, essentially unchanged *per position per step*; what is genuinely undone is the *dynamics*, since a "decision" becomes a whole denoising trajectory rather than one step, bidirectional attention changes what counts as a source d_j, and the margin/multiplicity statistics must be redefined over the trajectory. *Continuous* diffusion (score/noise prediction in an embedding space with a separate rounding step) has no softmax over a vocabulary and no discrete arg max, so the recovered-probability anchor (Thm 1) and the power-diagram geometry (Prop 1) have no direct analog and would require a measure-theoretic reformulation rather than a transcription. Establishing which PIC results survive each setting, and what the analog of the retrieve/compose split is when generation is iterative refinement rather than left-to-right, is open.

Scale.
The evidence is from sub-billion-parameter models. Whether the same three-way structure and the PR-mechanism survive at frontier scale, and whether the split behaves the same at long context, is untested.

Why it matters. If the provenance gap and the expressivity separation close, the consequence is a clean, falsifiable picture of a transformer: the retrievable fragment is an additive, low-rank, low-treewidth part that *exports to a compact, verifiable program*, while the computed fragment is a diffuse, high-rank, dense-G remainder that admits *no* compact symbolic form. That would turn "which parts of a model can we extract and audit, and which are irreducibly computed" from a slogan into a measurable, bounded region, with direct consequences for interpretability (audit the retrievable tier), safety (the computed core is where capability genuinely grows with scale), and efficiency (quantise the retrievable tier, protect the core).

9 Conclusion

An autoregressive transformer, asked for a token, spends most of its labour retrieving and selecting, and a small, scale-growing remainder computing something new. PIC makes that distinction measurable on real models, roughly 25% retrieved, 60% selected, 15% composed, gives the composed core an exact geometric and causal characterisation, and shows that the same structure is simultaneously a probabilistic logic, a tropical geometry, and an executable Datalog program, according to a single degree of freedom, the temperature.

The conceptual payoff is the lineage. 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. 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 G is the explicit carrier of the non-truth-functional coupling, its connective is a weighted threshold rather than Boolean intersection, and its output probabilities are recovered, exactly, as incidence frequencies (Theorem 1). The classical disjoint-outcome case reappears as the diagonal-G limit (Theorem 3). PIC adds a substrate (the inner-product geometry of a real model), a thermometer (the temperature/semiring), and a fixpoint engine (the Datalog reading); the central idea is Bundy's.

What to remember. (i) The split is real and not about magnitude; it is about *margin* (a distance to a power-diagram facet) and *multiplicity* (whether any single source already picks the token). (ii) The softmax is not imposed; it is recovered as an incidence frequency. (iii) The softmax you sample and the arg max you decode are one program at two temperatures, joined by Maslov dequantization. (iv) The computed core is diffuse by theorem, so it resists both localisation and compression, which is the same reason it is the interesting part.

10 References

The references are organised in four groups. We begin with the source paper this guide accompanies, then reproduce the works it cites, then add curated further reading (broad introductions and the seminal works behind the ideas this guide leans on: transformers, mechanistic interpretability, mixtures and products of experts, the softmax/Gibbs link, tropical geometry, and graphical-model intractability), and finally a short list of attributions the source paper leaves implicit (the measured model, the participation-ratio metric, the XOR/perceptron lineage, and the proof assistant).

Source paper

[SP] J. Allan Scott. *What a Transformer Retrieves and What It Computes.* Manuscript, 2026. The paper this document is a companion to; referred to throughout as "the source paper." See also the accompanying code and proof artifacts in Appendix C (`fieldrun` and `i-orca`).

Works cited in the source paper

[1] M. Abo Khamis, H. Q. Ngo, and A. Rudra. FAQ: Questions asked frequently. *ACM Symposium on Principles of Database Systems (PODS)*, 13–28, 2016.

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

[3] S. Biderman, H. Schoelkopf, Q. Anthony, et al. Pythia: a suite for analyzing large language models across training and scaling. *International Conference on Machine Learning (ICML)*, 2397–2430, 2023.

[4] T. Bricken, A. Templeton, J. Batson, et al. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

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

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

[7] S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases.* Springer, 1990.

[8] H. Cunningham, A. Ewart, L. Riggs, R. Huben, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. *International Conference on Learning Representations (ICLR)*, 2024.

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

[10] N. Elhage, T. Hume, C. Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[11] N. Elhage, N. Nanda, C. Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[12] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. *Proceedings of EMNLP*, 2021.

[13] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance semirings. *ACM Symposium on Principles of Database Systems (PODS)*, 31–40, 2007.

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

[15] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

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

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

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

[19] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[20] C. Olsson, N. Elhage, N. Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022.

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

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

Further reading: introductory and seminal

[F1] A. Vaswani, N. Shazeer, N. Parmar, et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. *(The transformer architecture.)*

[F2] C. Olah, N. Cammarata, L. Schubert, et al. Zoom in: an introduction to circuits. *Distill*, 2020. *(Accessible entry to mechanistic interpretability.)*

[F3] nostalgebraist. Interpreting GPT: the logit lens. *LessWrong*, 2020. *(Origin of reading the residual stream through the unembedding, i.e. direct logit attribution.)*

[F4] T. McGrath, M. Rahtz, J. Kramár, V. Mikulik, and S. Legg. The Hydra effect: emergent self-repair in language model computation. *arXiv:2307.15771*, 2023. *(The self-repair behind "rescue.")*

[F5] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. *(The original Mixture of Experts.)*

[F6] N. Shazeer, A. Mirhoseini, K. Maziarz, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*, 2017. *(Modern sparse MoE.)*

[F7] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, Springer, 1990. *(The softmax readout.)*

[F8] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. *(The Gibbs/maximum-entropy view behind "softmax as incidence frequency.")*

[F9] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *International Conference on Learning Representations (ICLR)*, 2020. *(Top-k/top-p sampling, the truncation knobs paired with temperature.)*

[F10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013. *(Word embeddings and cosine similarity.)*

[F11] L. Zhang, G. Naitzat, and L.-H. Lim. Tropical geometry of deep neural networks. *International Conference on Machine Learning (ICML)*, 2018. *(Decision surfaces of ReLU networks as tropical objects.)*

[F12] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. *(Treewidth and sum-product, for the intractability remarks of §8.)*

Additional attributions (left implicit in the source paper)

[A1] Qwen Team. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024. *(The Qwen2.5 family; Qwen2.5-0.5B is the primary measured model in §7.)*

[A2] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. F. Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017. *(Representative use of the participation ratio as an effective dimensionality; the metric also descends from the inverse participation ratio in condensed-matter physics.)*

[A3] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969. *(The XOR/linear-threshold expressivity boundary invoked in the Theorem 4 aside.)*

[A4] T. Nipkow, L. C. Paulson, and M. Wenzel. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. LNCS 2283, Springer, 2002. *(The Isabelle/Isar target referenced in Appendix A.)*

A A note on the formal status of the proofs

Each theorem in Section 6 corresponds to a result in the source paper that has been transcribed into a machine-checkable proof skeleton (in the `i-orca` format,[3] lowering to Isabelle/Isar). Two honest caveats apply, and are worth stating because they model how the theory treats its own limits. First, a passing static check verifies the proof *skeleton*, the method coverage, and is not, by itself, a kernel-checked proof; the real status requires running the prover. Second, the steps the paper itself leaves open (general Horn separation, the cited Maslov limit handled by automation, the asymptotic localisation bound) are marked as explicit frontier holes rather than hidden. The

[3] `i-orca`: https://github.com/jascal/i-orca.

map between the human-readable results here and the formal artifacts is: cardinality-inertness (Thm 2), non-truth-functionality budget (Thm 3), weighted-threshold expressivity (Thm 4), recovered probability (Thm 1), diffuseness (Thm 6), two-temperature soundness (Thm 5), the power diagram (Prop. 1) and the margin distance (Prop. 2).

B Glossary

Residual stream r
the running additive sum of all component writes at a position, $r = \sum_j d_j$.

Logit L_v
the score of token v, $L_v = \langle r, U_v \rangle + b_v = \sum_j c_j^v$.

DLA c_j^v
direct logit attribution, the vote of source j for token v, $c_j^v = \langle d_j, U_v \rangle$.

Gram matrix G_{vw}
$\langle U_v, U_w \rangle$, the carrier of non-truth-functional coupling between tokens.

Participation ratio PR
effective number of contributing sources; route-invariant, with a value that is architecture-dependent (≈ 45 on the Qwen models; wider across the Pythia ladder).

Readout multiplicity μ_t
number of sources whose own arg max is t; $\mu_t = 0$ marks a composed (emergent) token.

Differential incidence D_j
$c_j^t - c_j^{v^*} = \langle d_j, U_t - U_{v^*} \rangle$, source j's push on the winning margin.

Normalised margin
$\Delta / \|U_t - U_{v^*}\|$, the Euclidean distance from r to the nearest decision facet.

Power (Laguerre) diagram
the tessellation of residual space into arg max cells, with sites $\{U_v\}$ and weights $(b_v, \|U_v\|^2)$.

Maslov dequantization
the $T \to 0$ homomorphism from the log-semiring (softmax) to the tropical semiring (greedy arg max).

C Software and resources

The measurements and the formal artifacts referenced throughout are available in two public repositories:

`fieldrun`
the pure-Rust inference and `explain` toolchain that decompiles a model into a flat retrieval store plus a composition kernel and produces the per-decision diagnostics (PR, margin, μ_t, nearest power-diagram facet) used in Section 7.
https://github.com/jascal/fieldrun

`i-orca`
the proof-skeleton format into which the theorems of Section 6 are transcribed for machine checking (lowering to Isabelle/Isar), as discussed in Appendix A. A representative `fieldrun` example lives at `examples/fieldrun/fieldrun.i.orca.md`.
https://github.com/jascal/i-orca