OCTA Research – 365 Days of Learning, Theorizing, and Building AGI

Attention as a Stratified Map: Micro-Singularities in Query–Key Space

Week 1 · Day 2 · January 2, 2026

OCTA Research Internal Theory Program

Version 1.3 – Living Document Series (Expanded)

OCTA Research 365 Program Note. This is Week 1, Day 2 of the OCTA Research 365-day program. Day 1 distinguished macro-level *singularity* from micro-level *singularities* in Transformer architectures and introduced basic visualizations. Day 2 zooms in on the *attention mechanism* itself, treating it as a *stratified, piecewise-smooth map* whose tie manifolds and rank properties define a concrete class of micro-singularities. Version 1.3 extends earlier Day 2 drafts with causal/masked attention, multi-head stratification, measure-theoretic and topological analysis, *parameter-space* singularities, *compositional* stratification with MLPs, and an information-geometric pre-bridge into Day 3.

Abstract

This Day 2 note develops a mathematically explicit view of self-attention as a stratified map on query–key space. We show that for fixed parameters, the attention operator decomposes input space into polyhedral regions within which the map is smooth (often analytic) and shares a common combinatorial pattern of *argmax assignments*. The boundaries between these regions are *tie manifolds*, which are natural micro-level singular loci for attention: they mark changes in which tokens attend to which others, and in the hard-attention limit they become genuine discontinuities.

We: (i) formalize the stratification of query–key space by argmax patterns; (ii) work out low-dimensional examples (two tokens, three tokens, one head; both scalar and vector queries/keys) with explicit inequalities and OCTA-branded TikZ visualizations; (iii) analyze the Jacobian structure layer-wise inside a stratum versus on its boundaries; (iv) extend the analysis to *causal/masked* attention and to *multi-head* architectures, showing how masks and multiple heads induce constrained and product stratifications; (v) study measure-theoretic and topological properties of these strata, including almost-everywhere regularity and the structure of the tie set; (vi) analyze *parameter-space* degeneracies where W_Q, W_K themselves satisfy tie constraints and collapse strata; (vii) describe *compositional* stratification when attention is followed by piecewise-linear MLPs, yielding a finer cell decomposition for an entire Transformer block; (viii) outline a preliminary information-geometric view (Fisher concentration near tie manifolds) as a bridge toward singular learning theory on Day 3; (ix) and define experimental protocols for mapping attention strata in trained models and connecting them to interpretability, emergent circuits, and robustness.

The goal is to establish attention stratification as one of the canonical micro-singular geometries in OCTA-style AGI systems, to be used by subsequent days as a building block for attractor design, scaling-law analysis, and safety-relevant control of internal singular structure.

Contents

List of Figures

1 Recap and Day 2 Objective

Day 1 introduced:

- macro-level capability dynamics and technological singularity toy models;
- micro-level singularities in parameter, function, and information geometry;
- singular learning theory and the idea of RLCT as a complexity measure;
- OCTA-flavored conjectures and protocols around Jacobian and Fisher spectra.

Day 2 narrows focus to a single module: *self-attention*. The objective is:

OCTA Principle 1.1 (Day 2 OCTA Principle: Attention Stratification). Self-attention, viewed as a map on query–key space, admits a natural stratification into regions of constant argmax pattern. These strata define the fundamental micro-singular geometry of attention: boundaries between strata correspond to tie manifolds, where infinitesimal changes in query–key configuration produce discrete changes in which tokens attend to which others.

We aim to:

(a) give a precise definition of this stratification;

(b) compute it explicitly in low-dimensional examples;

(c) connect it to Jacobian rank and conditioning;

(d) extend it to causal/masked attention and multi-head architectures;

(e) connect it to parameter-space singularities and block-wise composition;

(f) and frame it as a reusable object in later OCTA Research work.

2 Self-Attention as a Map on Query–Key Space

2.1 Setup

As in Day 1, consider a single-head attention block:

$$Q = XW_Q \in \mathbb{R}^{n\times d_h}, \tag{1}$$

$$K = XW_K \in \mathbb{R}^{n\times d_h}, \tag{2}$$

$$V = XW_V \in \mathbb{R}^{n\times d_h}, \tag{3}$$

with scores

$$S_{ij} = \frac{1}{\sqrt{d_h}}\langle q_i, k_j\rangle, \tag{4}$$

and softmax weights (with temperature $\tau > 0$):

$$A_{ij}^{(\tau)} = \frac{\exp(S_{ij}/\tau)}{\sum_{k=1}^{n}\exp(S_{ik}/\tau)}. \tag{5}$$

For Day 2, we separate *query–key geometry* from the rest:

- treat Q, K as independent variables in $\mathbb{R}^{n\times d_h}$,
- treat V as a fixed array of value vectors (or absorb it later).

Define:

$$\mathcal{QK} := \{(Q,K) \in \mathbb{R}^{n\times d_h} \times \mathbb{R}^{n\times d_h}\},$$

and the attention map:

$$\mathcal{A}^{(\tau)} : \mathcal{QK} \to \mathbb{R}^{n\times n}, \quad (Q,K) \mapsto A^{(\tau)}.$$

2.2 Hard-attention limit and argmax patterns

For each row i, define

$$j_i^\star(Q,K) := \arg\max_{j\in\{1,\dots,n\}} S_{ij}.$$

Definition 2.1 (Argmax pattern)**.** An *argmax pattern* is a function

$$\sigma : \{1,\dots,n\} \to \{1,\dots,n\}, \quad i \mapsto \sigma(i),$$

interpreted as "token i attends to token $\sigma(i)$" in the hard-attention limit.

The set of all patterns is $\Sigma_n := \{1,\dots,n\}^{\{1,\dots,n\}}$, of size n^n.

Definition 2.2 (Pattern region)**.** For a pattern $\sigma \in \Sigma_n$, define its *pattern region*:

$$\mathcal{R}_\sigma := \{(Q,K) \in \mathcal{QK} : j_i^\star(Q,K) = \sigma(i)\ \forall i\}.$$

Within $\mathcal{R}_\sigma$, the hard-attention limit is:

$$A_{ij}^{(0)}(Q,K) = \begin{cases} 1, & j = \sigma(i), \\ 0, & \text{otherwise.} \end{cases}$$

The complement of the union of interiors of these regions is the *tie set*, introduced more carefully next.

3 Stratification by Argmax Patterns

3.1 Tie hyperplanes and polyhedral decomposition

For each triple (i,j,k) with i a query index and $j \neq k$ key indices, define the tie hyperplane in score space:

$$H_{i,j,k} := \{(Q,K) : S_{ij} = S_{ik}\}.$$

Using $S_{ij} = \frac{1}{\sqrt{d_h}}\langle q_i, k_j\rangle$, this becomes a linear equation in the entries of q_i, k_j, k_k:

$$\langle q_i, k_j - k_k\rangle = 0.$$

Definition 3.1 (Tie set)**.** The *tie set* is

$$\mathcal{T} := \bigcup_{i,j\neq k} H_{i,j,k}.$$

Proposition 3.2 (Polyhedral stratification)**.** *The complement $\mathcal{QK} \setminus \mathcal{T}$ is partitioned into open regions $\{\mathcal{R}_\sigma^\circ\}_{\sigma \in \Sigma_n}$, where*

$$\mathcal{R}_\sigma^\circ := \operatorname{int}(\mathcal{R}_\sigma),$$

and each $\mathcal{R}_\sigma^\circ$ is a finite intersection of strict linear inequalities in the entries of Q and K.

Sketch. Fix σ and i. The condition $j_i^\star(Q, K) = \sigma(i)$ is equivalent to

$$S_{i,\sigma(i)} > S_{ij} \quad \forall j \neq \sigma(i).$$

Each inequality $S_{i,\sigma(i)} > S_{ij}$ is linear in the entries of Q, K. Intersecting over all $j \neq \sigma(i)$ and over $i = 1, \ldots, n$ gives a finite intersection of strict linear inequalities: an open polyhedral region. Distinct patterns correspond to disjoint regions because argmax is unique away from ties. □

This yields a natural stratification:

Definition 3.3 (Attention stratification)**.** The *attention stratification* of $\mathcal{QK}$ is the decomposition

$$\mathcal{QK} = \bigsqcup_{\sigma \in \Sigma_n} \mathcal{R}_\sigma^\circ \;\sqcup\; \mathcal{T},$$

where $\mathcal{R}_\sigma^\circ$ are open polyhedral strata and $\mathcal{T}$ is a union of hyperplanes (and their intersections).

3.2 Interpretation

Inside a region $\mathcal{R}_\sigma^\circ$:

- the *combinatorial structure* of attention is fixed: each query index i has a unique top key index $\sigma(i)$;
- in the hard-attention limit, $A^{(0)}$ is *constant* across the region;
- in the soft-attention case with $\tau > 0$, the weights $A^{(\tau)}$ are smooth functions of Q, K whose qualitative pattern (largest entry in each row) is fixed.

On the tie set $\mathcal{T}$:

- the argmax is non-unique; multiple patterns are compatible;
- in the hard-attention limit, $A^{(0)}$ is discontinuous;
- the soft-attention map $A^{(\tau)}$ remains smooth for $\tau > 0$, but its local behavior is often ill-conditioned as $\tau \to 0$.

4 Two-Token, One-Head Example Revisited

4.1 Scalar queries and keys

Let $n = 2$, $d_h = 1$, so $q_i, k_j \in \mathbb{R}$. Then

$$S_{ij} = \frac{1}{\sqrt{1}} q_i k_j = q_i k_j.$$

For $i = 1$, the relevant tie hyperplane is

$$S_{11} = S_{12} \quad \Longleftrightarrow \quad q_1(k_1 - k_2) = 0.$$

We define $\Delta := q_1(k_1 - k_2)$ as in Day 1.

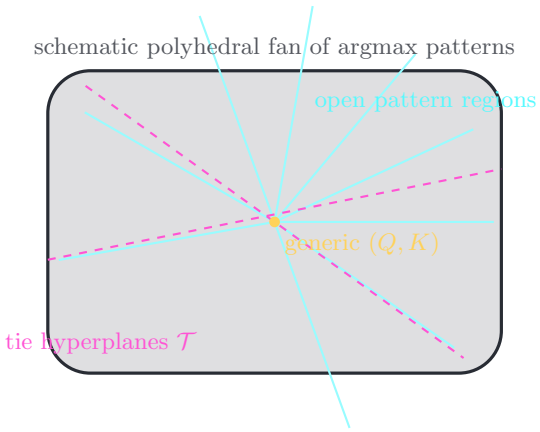

Figure 1: OCTA schematic of attention stratification as a polyhedral fan in (Q, K)-space. Rays represent directions along which the argmax pattern is constant; dashed magenta lines represent tie hyperplanes where patterns change. In high dimensions, each pattern region is a high-dimensional polyhedral cone (locally), and $\mathcal{T}$ is a union of codimension-1 faces.

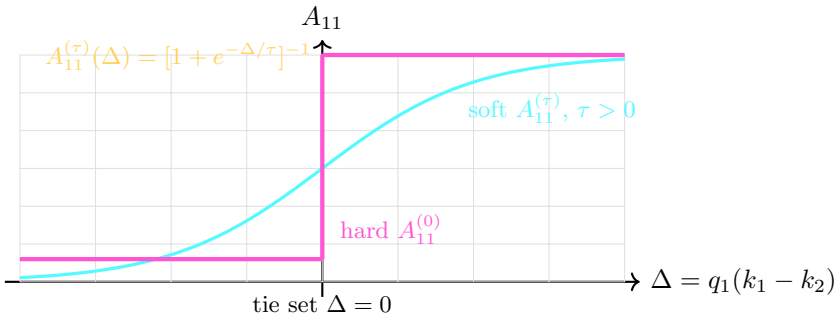

Figure 2: Day 2 refinement of the one-dimensional scalar two-token attention singularity. For $\tau > 0$, the map is smooth in Δ. As $\tau \to 0$, the hard-attention limit becomes piecewise constant with a discontinuity at the tie set $\Delta = 0$.

- If $\Delta > 0$, then $S_{11} > S_{12}$ and $\sigma(1) = 1$.
- If $\Delta < 0$, then $S_{11} < S_{12}$ and $\sigma(1) = 2$.
- If $\Delta = 0$, we are on the tie hyperplane $H_{1,1,2}$.

The corresponding soft attention weight is

$$A_{11}^{(\tau)}(\Delta) = \frac{1}{1 + \exp(-\Delta/\tau)},$$

and the hard limit is the step function illustrated in Figure 2, which refines and reuses the Day 1 visual.

4.2 Jacobian with respect to queries/keys

In this minimal case, we can explicitly differentiate:

$$\frac{\partial A_{11}^{(\tau)}}{\partial \Delta} = \frac{1}{\tau} \frac{e^{-\Delta/\tau}}{(1 + e^{-\Delta/\tau})^2} = \frac{1}{\tau} A_{11}^{(\tau)}(1 - A_{11}^{(\tau)}).$$

As $\tau \to 0$:

- For $\Delta \neq 0$, $A_{11}^{(\tau)} \to 1$ or 0 and the derivative tends to 0.
- Near $\Delta = 0$, the derivative becomes sharply peaked with height $\sim 1/(4\tau)$.

Thus the Jacobian of $A^{(\tau)}$ with respect to Δ is:

- small away from the tie locus (robust mapping),
- large near the tie locus as τ decreases (potentially ill-conditioned).

Attention strata therefore come with a natural conditioning profile: interiors are stable; boundaries can be extremely sensitive, especially in low temperature regimes.

5 Three-Token Example and 2D Stratification

We now compute the stratification explicitly in the next non-trivial case, refining the geometric picture.

5.1 Score space and tie lines

Fix a particular query row q_i, and let

$$s_j := S_{ij} = q_i k_j, \quad j = 1, 2, 3,$$

be the three scores for that query. The argmax pattern for row i is determined by which of s_1, s_2, s_3 is largest.

Working in score space $(s_1, s_2, s_3) \in \mathbb{R}^3$, define tie planes:

$$H_{12} = \{s_1 = s_2\}, \tag{6}$$
$$H_{23} = \{s_2 = s_3\}, \tag{7}$$
$$H_{13} = \{s_1 = s_3\}. \tag{8}$$

These partition $\mathbb{R}^3$ into six regions corresponding to the six strict orderings of three real numbers, e.g.:

$$s_1 > s_2 > s_3, \quad s_1 > s_3 > s_2, \quad \ldots$$

For attention, we collapse each region to the information "which s_j is largest." That yields three main regions:

$$\mathcal{R}_1^\circ = \{s_1 > s_2, s_1 > s_3\}, \quad \mathcal{R}_2^\circ = \{s_2 > s_1, s_2 > s_3\}, \quad \mathcal{R}_3^\circ = \{s_3 > s_1, s_3 > s_2\}.$$

The tie set is the union of the three planes $H_{12} \cup H_{23} \cup H_{13}$ (and their intersections).

5.2 2D visualization via differences

To draw this, we project onto the plane of score differences:

$$u = s_1 - s_3, \quad v = s_2 - s_3.$$

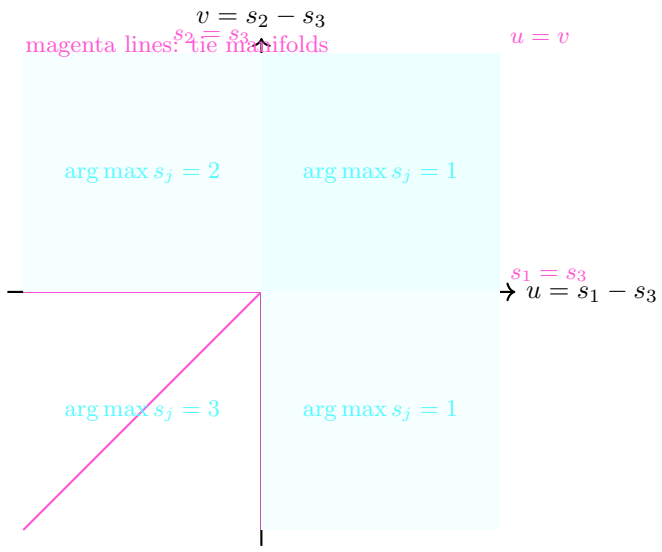

Figure 3: Day 2 visualization of three-token attention stratification for a single query row in the (u, v)-plane of score differences $u = s_1 - s_3$, $v = s_2 - s_3$. The plane is partitioned into regions where each s_j is strictly largest; boundaries (magenta) are tie manifolds where two scores are equal. This is a low-dimensional shadow of the general argmax-pattern stratification.

In (u, v)-coordinates, the conditions become:

$$\begin{aligned} s_1 > s_2 &\iff u > v, \\ s_1 > s_3 &\iff u > 0, \\ s_2 > s_3 &\iff v > 0. \end{aligned}$$

This yields a wedge partition of the (u, v)-plane with three main regions (Figure 3).
This illustrates:

- the polyhedral nature of pattern regions;
- the codimension-1 tie manifolds at which argmax changes;
- the inherent piecewise-constant structure of hard attention across these regions.

6 Jacobian Structure Inside and Across Strata

6.1 General form of the Jacobian

For fixed V, attention yields outputs

$$Y_i = \sum_{j=1}^{n} A_{ij}^{(\tau)} V_j.$$

Treating (Q, K) as inputs, the Jacobian of the map $(Q, K) \mapsto Y$ encodes how outputs change with respect to perturbations in queries/keys.

Using the chain rule, for a given row i:

$$\frac{\partial Y_i}{\partial q_i} = \sum_{j=1}^{n} \frac{\partial A_{ij}^{(\tau)}}{\partial q_i} V_j,$$

and similarly for k_j. We can factor:

$$\frac{\partial A_{ij}^{(\tau)}}{\partial S_{il}} = A_{ij}^{(\tau)}(\delta_{jl} - A_{il}^{(\tau)}),$$

and

$$\frac{\partial S_{il}}{\partial q_i} = \frac{1}{\sqrt{d_h}} k_l, \quad \frac{\partial S_{il}}{\partial k_l} = \frac{1}{\sqrt{d_h}} q_i.$$

Thus the sensitivity of Y_i to q_i and k_l depends on:

- the current attention distribution $A_{i\cdot}^{(\tau)}$,
- the geometric configuration of q_i, k_l,
- and the value vectors V_j.

6.2 Behavior within a stratum

Inside a pattern region $\mathcal{R}_\sigma^\circ$:

- the ordering of scores in each row is fixed;
- the attention vector $A_{i\cdot}^{(\tau)}$ remains dominated by the same index $\sigma(i)$ as τ is small but positive;
- the derivatives vary smoothly and stay bounded away from the tie manifolds.

Proposition 6.1 (Smoothness inside a stratum)**.** *Fix $\tau > 0$. On each open region $\mathcal{R}_\sigma^\circ$, the attention map $(Q, K) \mapsto A^{(\tau)}$ is smooth (infinitely differentiable), and so is $(Q, K) \mapsto Y$.*

Sketch. $A^{(\tau)}$ is obtained from S by finite compositions of smooth operations (linear maps and exponentials normalized by finite sums) without division by zero. Inside a region where the ordering of S_{ij} values is fixed, these operations have no singularities. □

6.3 Behavior near tie manifolds

Near the tie set $\mathcal{T}$, derivatives become large for small τ, as seen explicitly in the two-token example:

$$\frac{\partial A_{11}^{(\tau)}}{\partial \Delta} = \frac{1}{\tau} A_{11}^{(\tau)}(1 - A_{11}^{(\tau)}),$$

peaking at $\Delta = 0$ with height $\sim 1/(4\tau)$.

In higher dimensions, similar peak behaviors occur along directions normal to tie hyperplanes; along directions tangent to tie hyperplanes, derivatives can remain moderate.

Remark 6.2 (Directional conditioning). For small τ:

- perturbations that move (Q, K) *across* a tie hyperplane can cause large changes in $A^{(\tau)}$;
- perturbations that move *along* the hyperplane may cause smaller changes.

This anisotropy is a key part of attention's micro-geometry and will matter for robustness and interpretability.

7 Causal and Masked Attention as Constrained Stratification

Real Transformer architectures rarely use unconstrained attention over all positions; instead, attention is typically *masked* (e.g. causal masks for autoregressive modeling, padding masks, or structured sparsity patterns). Masks prune argmax patterns and modify the stratification.

7.1 Masking via $-\infty$ scores

Let $M \in \{-\infty, 0\}^{n \times n}$ be a mask, applied additively to the score matrix:

$$\tilde{S}_{ij} = S_{ij} + M_{ij}.$$

Typical cases:

- **Causal mask:** For $i < j$, set $M_{ij} = -\infty$; otherwise $M_{ij} = 0$.
- **Padding mask:** For padded positions j, set $M_{ij} = -\infty$ for all i.
- **Structured locality:** For $|i - j| > w$, set $M_{ij} = -\infty$.

The effective scores $\tilde{S}$ inherit the same linear dependence on (Q, K) where finite, but masked entries are fixed at $-\infty$ and never win an argmax.

7.2 Masked argmax patterns and reduced pattern set

Definition 7.1 (Masked argmax)**.** Given mask M, define

$$j_i^{\star,M}(Q, K) := \arg \max_{j : M_{ij} > -\infty} \tilde{S}_{ij}.$$

Only keys j with $M_{ij} > -\infty$ are eligible per row i, reducing the candidate set.

Definition 7.2 (Masked pattern set)**.** The set of *masked argmax patterns* is

$$\Sigma_n^M := \left\{ \sigma : \{1, \ldots, n\} \to \{1, \ldots, n\} \;\middle|\; M_{i,\sigma(i)} > -\infty \; \forall i \right\}.$$

All previous definitions go through with Σ_n replaced by Σ_n^M, and tie hyperplanes now only involve admissible pairs (i, j, k) with $M_{ij}, M_{ik} > -\infty$.

7.3 Effect on stratification

Masking:

- reduces the pattern set from Σ_n to Σ_n^M;
- removes tie hyperplanes involving masked indices;
- can increase the volume of some regions (since fewer competitors exist).

From an OCTA perspective, masks carve out a *sub-fan* of the original stratification, imposing temporal or structural constraints on which micro-patterns are ever realized.

8 Multi-Head Attention as Product Stratification

Real Transformer layers use multi-head attention, each head with its own (W_Q, W_K, W_V).

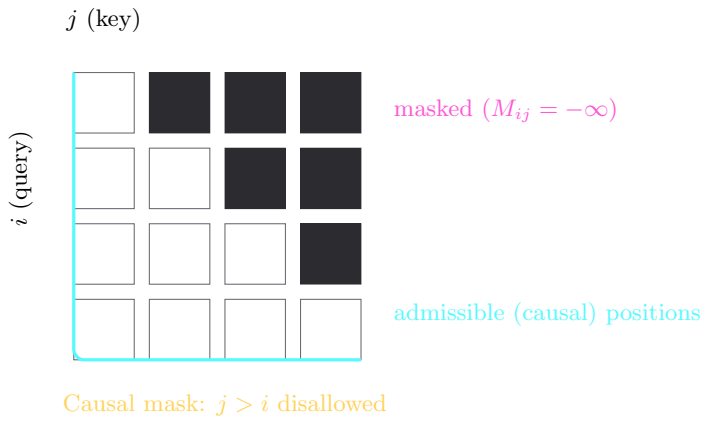

Figure 4: Causal masking as a constraint on argmax patterns. Only entries with $j \leq i$ (cyan-outline region) are admissible; mask entries with $j > i$ (dark region) are fixed to $-\infty$ and excluded from the stratification. The pattern set shrinks from Σ_n to Σ_n^M, simplifying the polyhedral fan in (Q, K)-space.

8.1 Independent heads

For H heads, head h has queries and keys:

$$Q^{(h)} = XW_Q^{(h)}, \quad K^{(h)} = XW_K^{(h)}.$$

Each head has its own attention map

$$\mathcal{A}^{(\tau,h)} : (Q^{(h)}, K^{(h)}) \mapsto A^{(\tau,h)},$$

and its own stratification into pattern regions $\mathcal{R}_{\sigma_h}^{(h)}$ with tie sets $\mathcal{T}^{(h)}$.

Definition 8.1 (Multi-head pattern)**.** A *multi-head pattern* is a tuple

$$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_H) \in \Sigma_n^H,$$

where each σ_h is a single-head pattern.

Proposition 8.2 (Product stratification (disentangled heads))**.** *Assume the heads are independent in $(Q^{(h)}, K^{(h)})$ (no cross-head coupling in Q, K). Then the overall query–key space $\mathcal{QK}^{(1)} \times \cdots \times \mathcal{QK}^{(H)}$ is stratified by the product partition*

$$\mathcal{R}_{\boldsymbol{\sigma}}^{\circ} = \mathcal{R}_{\sigma_1}^{(1)\circ} \times \cdots \times \mathcal{R}_{\sigma_H}^{(H)\circ},$$

with tie set

$$\mathcal{T}^{multi} = \bigcup_{h=1}^{H} \left(\mathcal{QK}^{(1)} \times \cdots \times \mathcal{T}^{(h)} \times \cdots \times \mathcal{QK}^{(H)} \right).$$

Even when Q, K across heads share a base X (as in practice), this product picture remains approximately valid in the space of head-specific scores.

From the OCTA viewpoint, multi-head patterns $(\sigma_1, \ldots, \sigma_H)$ are primary atoms of micro-circuit structure: each tuple defines a distinct wiring configuration for information flow within a layer.

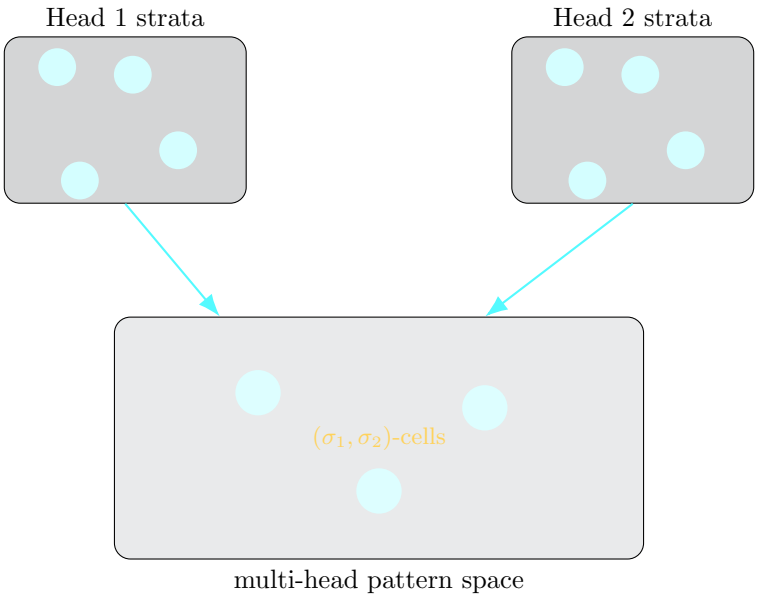

Figure 5: Schematic of multi-head attention stratification. Each head has its own stratification in $(Q^{(h)}, K^{(h)})$-space; the overall multi-head pattern space is (roughly) a product of these stratifications, with cells labeled by tuples $(\sigma_1, \ldots, \sigma_H)$. In practice, shared X couples heads but the combinatorial picture remains a useful approximation.

9 Measure-Theoretic and Topological Properties

We briefly record basic properties of the stratification that will be useful for singular learning theory (Day 3) and for connecting to RLCT.

9.1 Measure-zero of tie set

Under mild genericity assumptions on (W_Q, W_K):

Proposition 9.1 (Tie set has measure zero)**.** *Assume W_Q, W_K are fixed and X has a continuous density over $\mathbb{R}^{n\times d}$. Then the induced distribution over (Q, K) is absolutely continuous on $\mathcal{QK}$, and the tie set $\mathcal{T}$ has Lebesgue measure zero.*

Sketch. Each tie hyperplane $H_{i,j,k}$ is a codimension-1 linear subspace of $\mathcal{QK}$ (unless degenerate choices of W_Q, W_K collapse it). A finite union of codimension-1 linear subspaces has measure zero in $\mathbb{R}^m$. Absolute continuity of (Q, K) implies that the probability of landing on $\mathcal{T}$ is zero. □

Thus in the probabilistic sense, almost all inputs fall in some open pattern region $\mathcal{R}_\sigma^\circ$.

9.2 Almost-everywhere differentiability

Since $A^{(\tau)}$ is smooth on each $\mathcal{R}_\sigma^\circ$ and $\mathcal{T}$ has measure zero:

Corollary 9.2 (Almost-everywhere regularity)**.** *For fixed $\tau > 0$, the attention map $(Q, K) \mapsto A^{(\tau)}$ (and hence $(Q, K) \mapsto Y$) is differentiable almost everywhere with respect to any absolutely continuous input distribution.*

This aligns with the general picture of deep networks as piecewise-smooth maps: non-smooth behavior is confined to a measure-zero union of boundaries.

9.3 Topology of pattern regions

Each $\mathcal{R}_\sigma^\circ$ is:

- non-empty for many σ (although some patterns may be unreachable under architectural constraints);
- open and convex (intersection of strict linear inequalities);
- contractible (topologically a ball or polyhedral cone).

Their closures intersect along faces of lower dimension (e.g. tie hyperplanes and their intersections), forming a polyhedral complex.

Remark 9.3 (Polyhedral complex viewpoint). The family $\{\overline{\mathcal{R}_\sigma^\circ}\}_{\sigma\in\Sigma_n}$, with intersections along faces, forms a polyhedral complex. The tie set $\mathcal{T}$ is the union of non-maximal cells (faces). This polyhedral structure is what makes attention stratification amenable to algebraic and combinatorial analysis.

10 Parameter-Space Stratification and Degeneracies

So far, stratification has lived in input space (Q, K) with parameters (W_Q, W_K) fixed. However, the same tie equations define *parameter-space* singularities.

10.1 Parameterization of queries and keys

Recall:

$$Q = XW_Q, \quad K = XW_K.$$

For a given dataset X, the scores become functions of parameters:

$$S_{ij}(W_Q, W_K) = \frac{1}{\sqrt{d_h}} \langle x_i W_Q, x_j W_K \rangle .$$

Tie conditions $S_{ij} = S_{ik}$ then define algebraic constraints on (W_Q, W_K).

Definition 10.1 (Parameter tie varieties)**.** For each (i, j, k), define the *parameter tie variety*

$$\mathcal{V}_{i,j,k} := \{(W_Q, W_K) : S_{ij}(W_Q, W_K) = S_{ik}(W_Q, W_K)\}.$$

The global parameter tie set is

$$\mathcal{V} := \bigcup_{i,j\neq k} \mathcal{V}_{i,j,k}.$$

10.2 Degenerate heads and collapsed strata

On $\mathcal{V}$, distinct input configurations may map to identical score differences, reducing the effective richness of attention strata.

Extreme examples:

- $W_Q = 0$ or $W_K = 0$: all scores vanish, $S_{ij} = 0$, attention is uniform and pattern structure collapses.
- rank-1 W_Q or W_K: all queries or keys lie in a 1D subspace, severely restricting possible patterns.

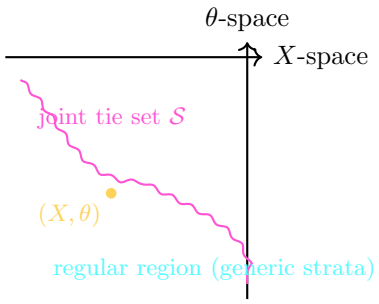

Figure 6: Schematic of joint stratification in input space X and parameter space $\theta = (W_Q, W_K)$. The magenta curve represents the joint singular set $\mathcal{S}$ where certain score differences vanish. Away from $\mathcal{S}$, both input-space and parameter-space behavior is regular; on $\mathcal{S}$, degeneracies and model-class singularities appear.

Remark 10.2 (Model-class singularities). From the standpoint of the full model class (parameters + inputs), points where $(W_Q, W_K) \in \mathcal{V}$ are *singular* in the sense that multiple parameter settings can induce the same function on X (non-identifiability). Day 3 will connect this to RLCT and singular learning theory.

10.3 Joint stratification in input and parameter space

We can formally treat (X, θ) with $\theta = (W_Q, W_K)$ as a joint variable.

Define:

$$\mathcal{M} := \{(X, \theta) : X \in \mathbb{R}^{n \times d}, \theta \in \Theta \subset \mathbb{R}^p\},$$

with Θ a parameter domain.

Tie constraints become:

$$g_{i,j,k}(X, \theta) := S_{ij}(X, \theta) - S_{ik}(X, \theta) = 0.$$

Definition 10.3 (Joint singular set)**.** The joint singular set is

$$\mathcal{S} := \bigcup_{i,j \neq k} \{(X, \theta) : g_{i,j,k}(X, \theta) = 0\}.$$

From an OCTA perspective, $\mathcal{S}$ is where training dynamics can stall, bifurcate, or concentrate, and where RLCT is determined. Day 3 will re-express these constraints as algebraic varieties in parameter space and analyze their contribution to learning exponents.

11 Compositional Stratification in a Transformer Block

Attention does not operate in isolation. A standard Transformer block composes attention with residuals and an MLP with piecewise-linear activations (e.g. ReLU, GELU approximations).

11.1 ReLU-region stratification

Consider a single MLP with ReLU:

$$h = \mathrm{ReLU}(W_1 y + b_1), \tag{9}$$

$$z = W_2 h + b_2, \tag{10}$$

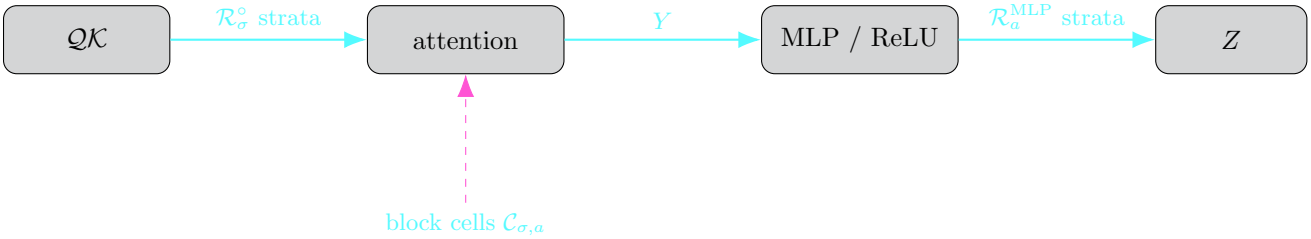

Figure 7: Compositional stratification in a Transformer block. Attention induces a stratification $\{\mathcal{R}_\sigma^\circ\}$ in (Q, K)-space; the subsequent MLP induces ReLU-region strata $\{\mathcal{R}_a^{\mathrm{MLP}}\}$ in Y. Their pullback defines block cells $\mathcal{C}_{\sigma,a}$ on which the entire block behaves as a single smooth (often affine) map.

where y is the attention output.

As is well known, the MLP is:

- affine on each *ReLU region* (fixed activation pattern $a = \mathbb{1}\{W_1 y + b_1 > 0\}$),
- with region boundaries given by hyperplanes $(W_1 y + b_1)_k = 0$.

Let $\mathcal{R}_a^{\mathrm{MLP}}$ denote the subset of y-space where activation pattern a holds.

11.2 Block-wise composition of stratifications

A Transformer block maps (Q, K) to Y (attention), then to Z (MLP output):

$$(Q, K) \xrightarrow{\text{attention}} Y \xrightarrow{\text{MLP}} Z.$$

Attention provides a stratification $\{\mathcal{R}_\sigma^\circ\}$ in (Q, K)-space; MLP provides a stratification $\{\mathcal{R}_a^{\mathrm{MLP}}\}$ in Y-space. Composing them yields:

Definition 11.1 (Block cell)**.** A *block cell* is a non-empty set of the form

$$\mathcal{C}_{\sigma,a} := \{(Q, K) : (Q, K) \in \mathcal{R}_\sigma^\circ,\ Y(Q, K) \in \mathcal{R}_a^{\mathrm{MLP}}\}.$$

On each $\mathcal{C}_{\sigma,a}$:

- the attention pattern σ is fixed,
- the MLP activation pattern a is fixed,
- the overall block map is a *single smooth (often affine) function* of (Q, K) for fixed τ.

Remark 11.2 (OCTA-block cells). From an OCTA design perspective, block cells $\mathcal{C}_{\sigma,a}$ are candidate "micro-atoms" of computation. Higher-level attractors will correspond to recurrent trajectories that stay within or cycle over small families of such cells.

12 OCTA View: Attention Strata as Micro-Singular Geometry

In the OCTA program, attention strata and block cells play several roles:

- **Micro-regimes of computation:** each pattern region $\mathcal{R}_\sigma^\circ$ and block cell $\mathcal{C}_{\sigma,a}$ corresponds to a discrete wiring pattern of the computation graph (which tokens and neurons feed into which).
- **Implicit state machines:** transitions between strata as tokens or parameters vary can be seen as transitions in an implicit automaton, with tie manifolds and ReLU boundaries as decision surfaces.
- **Singularity scaffolding:** the union of tie manifolds $\mathcal{T}$ and ReLU boundaries provides a concrete, algebraically simple singular set whose geometry we can explicitly probe and control.
- **Mask-structured dynamics:** in causal/masked settings, the reachable pattern set is constrained, which reduces the reachable micro-dynamics and may simplify attractor analysis.

OCTA Principle 12.1 (Attention Strata and Block Cells as Primitive Cells). For OCTA-style AGI, attention strata and their compositions with MLP regions are treated as primitive "cells" of micro-dynamics. Higher-level attractors and behaviors (including Perfect Attractors) are designed, in part, by controlling how trajectories in (Q, K) and representation space move across these cells and how often they approach or cross tie manifolds and activation boundaries.

13 Interpretability and Circuits on Stratified Attention

Attention strata give a clean way to talk about interpretability at the level of micro-circuits.

13.1 Pattern-conditioned circuits

For a fixed pattern σ and head h, the head's contribution to the output at row i is:

$$Y_i^{(h)} = \sum_j A_{ij}^{(\tau,h)} V_j^{(h)}.$$

Within $\mathcal{R}_\sigma^\circ$, the index of the dominant key is $\sigma(i)$; for small τ,

$$A_{ij}^{(\tau,h)} \approx \begin{cases} 1, & j = \sigma(i), \\ 0, & \text{otherwise}, \end{cases}$$

so

$$Y_i^{(h)} \approx V_{\sigma(i)}^{(h)}.$$

Thus, conditioned on being in $\mathcal{R}_\sigma^\circ$:

- the head implements a near-deterministic routing from i to $\sigma(i)$;
- interpretability can focus on the value path $i \to \sigma(i)$ and downstream effects.

13.2 Pattern ensembles and behavior

Let $\mathcal{D}$ be a data distribution over inputs. Each datum induces a pattern σ for each head and layer. We can define:

$$\pi(\sigma) = \mathbb{P}_{(X\sim\mathcal{D})}\left[(Q(X), K(X)) \in \mathcal{R}_\sigma^\circ\right].$$

From an interpretability standpoint:

- high-probability patterns under π correspond to *dominant circuits*;
- rare patterns may correspond to out-of-distribution regimes or fragile behavior;
- pattern distributions can be compared across heads, layers, or training checkpoints.

13.3 OCTA interpretability principle

OCTA Principle 13.1 (Pattern-Level Interpretability). OCTA-style interpretability attaches explanations and semantics not only to individual heads or parameters, but to *pattern-conditioned* regimes:

- **within** a pattern region $\mathcal{R}_\sigma^\circ$, the circuit is effectively linear and stable;
- **across** pattern boundaries, semantics can flip abruptly, and interpretation should account for these micro-discontinuities.

This suggests that attention-based explanations (e.g. "head 7 attends to the previous verb") are only meaningful relative to which pattern region we are in, and how often that region is visited under the actual data distribution.

14 Fisher Geometry and the Pre-Day3 Bridge

Day 3 will formalize singular learning theory for Transformers via RLCT and algebraic geometry of parameter spaces. Here we sketch how attention strata naturally enter the Fisher information.

14.1 Fisher information at the attention layer

Suppose the model defines a conditional distribution $p_\theta(y \mid x)$ with parameters θ including attention parameters. The (expected) Fisher information matrix is:

$$F(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim p_\theta(\cdot|x)}\left[\nabla_\theta \log p_\theta(y \mid x)\, \nabla_\theta \log p_\theta(y \mid x)^\top\right].$$

Gradients $\nabla_\theta \log p_\theta$ depend back through attention:

$$\nabla_\theta \log p_\theta = \frac{\partial \log p_\theta}{\partial Z}\frac{\partial Z}{\partial Y}\frac{\partial Y}{\partial A}\frac{\partial A}{\partial S}\frac{\partial S}{\partial (Q,K)}\frac{\partial (Q,K)}{\partial \theta}.$$

The factor $\frac{\partial A}{\partial S}$ is where attention strata enter:

$$\frac{\partial A_{ij}^{(\tau)}}{\partial S_{il}} = A_{ij}^{(\tau)}(\delta_{jl} - A_{il}^{(\tau)}).$$

Near tie manifolds, this Jacobian can be large in directions that cross $H_{i,j,k}$. Thus:

- contributions to $F(\theta)$ from inputs near $\mathcal{T}$ are potentially large;
- generic inputs (far from $\mathcal{T}$) contribute more modestly.

14.2 Concentration near singular sets

Qualitatively, Fisher eigenvalues associated with directions that *move the model across strata* (e.g. change which token is attended) can behave differently from those that perturb weights without changing patterns.

Conjecture 14.1 (OCTA Fisher–Strata Alignment). Eigen-directions of the Fisher matrix with unusually small or large eigenvalues (near-singular modes) tend to align with parameter directions that either:

- move (Q, K) trajectories close to tie manifolds $\mathcal{T}$ (pattern-switching directions), or
- leave argmax patterns largely unchanged but modulate values within strata (pattern-preserving directions).

This alignment is a key link between:

- the geometric singular set in (X, θ)-space (Day 2),
- and the statistical singularity and RLCT (Day 3).

From an OCTA engineering standpoint, it suggests that:

- controlling how often training data lie near $\mathcal{T}$ can shape the effective Fisher spectrum;
- regularizers or curricula that avoid pathological concentration near tie manifolds could improve stability.

15 Experimental Protocols for Attention Stratification

Day 2 defines concrete experiments to attach this theory to real models.

Experimental Protocol 15.1 (Day 2.1: Empirical Strata Mapping).

(a) Fix a trained Transformer and a particular attention head and layer.

(b) Sample a large batch of input sequences from a data distribution of interest.

(c) For each sequence and each position i:

- compute $Q_i, K_1, \ldots, K_n$ and scores S_{ij},
- record the argmax index $\sigma(i)$ and the margin $m_i = S_{i,\sigma(i)} - \max_{j \neq \sigma(i)} S_{ij}$.

(d) Build:

- histograms of margins m_i (distance to tie manifolds),
- empirical distributions of argmax patterns across tokens and contexts,
- correlation of small margins with mispredictions or robustness issues.

Experimental Protocol 15.2 (Day 2.2: Tie-Manifold Probing).

(a) For a fixed head and position i, select examples where two keys j, k have near-equal scores $S_{ij} \approx S_{ik}$.

(b) Construct controlled perturbations in input embedding space that move across the approximate tie hyperplane $S_{ij} = S_{ik}$ (e.g. by solving a small linear system in embedding space).

(c) Measure:

- sensitivity of logits and outputs to these perturbations;
- changes in attention pattern and downstream activations;
- impact on loss and behavior.

This estimates local conditioning near empirical tie manifolds.

Experimental Protocol 15.3 (Day 2.3: Temperature and Stratification).

(a) Introduce a tunable temperature τ in attention at inference time.

(b) For fixed inputs, vary τ over a range (e.g. geometric grid).

(c) Track:

- how often argmax patterns change as a function of τ;
- stability of outputs and performance metrics;
- effective width of transition layers near tie manifolds.

This probes how strongly the model relies on being near singular hard-attention behavior.

Experimental Protocol 15.4 (Day 2.4: Multi-Head Pattern Profiling).

(a) For a fixed layer with H heads, record for each input and position i the multi-head pattern

$$\boldsymbol{\sigma}(i) = (\sigma_1(i), \ldots, \sigma_H(i)).$$

(b) Estimate the empirical distribution $\pi(\boldsymbol{\sigma})$ over multi-head patterns.

(c) Identify:

- the top-k patterns by probability,
- patterns strongly associated with particular behaviors (e.g. correct vs. incorrect predictions, specific tasks).

This builds a vocabulary of "circuit modes" at the multi-head level.

Experimental Protocol 15.5 (Day 2.5: Masked vs. Unmasked Regimes).

(a) Compare models or layers with different masking schemes (e.g. causal vs. bidirectional attention).

(b) For each, run Protocols 15.1 and 15.4.

(c) Analyze:

- how masking reduces or reshapes the reachable pattern set,
- whether masked models have different margin distributions (distance to tie manifolds),
- whether this correlates with robustness or interpretability advantages.

Experimental Protocol 15.6 (Day 2.6: Parameter-Space Sweeps Across Tie Varieties).

(a) Fix a small dataset X and a head; select a pair (i, j, k) of interest.

(b) Parameterize a 1D or 2D path $\theta(t)$ in (W_Q, W_K)-space that crosses the parameter tie variety $\mathcal{V}_{i,j,k}$ (e.g. linearly interpolate between two trained checkpoints with different argmax behaviors).

(c) For each t:

- compute scores $S_i.(X, \theta(t))$ and record the argmax pattern;
- compute loss, logits, and local Fisher estimates along the path.

(d) Identify:

- how often and how sharply patterns switch as parameters cross $\mathcal{V}_{i,j,k}$;
- whether Fisher eigenvalues spike or vanish near these crossings.

This directly probes parameter-space singularity structure implied by Day 2.

16 Day 2 Roadmap and Links to Future Days

Day 2 creates the attention stratification scaffold that later days will reuse:

- **Day 3 (Singular Learning Primer for Transformers):** connect attention strata, tie manifolds, and parameter varieties to KL minimizer sets and RLCT, building on Day 1's singular model framing.
- **Day 4 (Jacobian/Fisher Estimation in Practice):** implement Protocols 15.1–15.6 and combine them with Jacobian/Fisher-spectrum estimation at or near tie manifolds and parameter varieties.
- **Week 2 (Attractors in Representation Space):** use attention strata and block cells as base cells in which representation trajectories live; study how training shapes which strata are visited and how often, and how this interacts with macro-level capability dynamics.

For OCTA-style AGI, the intent is that this Day 2 structure becomes a standard lens: any time we talk about interpretability, robustness, or emergent capabilities of an attention head or block, we ask:

> *Which strata are we in? How close are we to the tie set or parameter varieties? Which multi-head pattern and block cell are we in? How does this relate to the singular geometry and capability dynamics described on Days 1 and 3?*

17 Conclusion

Day 2 has:

- formalized self-attention as a stratified map on query–key space;
- identified argmax patterns and tie hyperplanes as the core combinatorial and geometric objects;

- provided explicit low-dimensional examples with visualizations in 1D and 2D;
- analyzed Jacobian behavior inside strata and near tie manifolds;
- extended the analysis to masked/causal attention and multi-head architectures;
- established basic measure-theoretic and topological properties (polyhedral complex, measure-zero tie set, almost-everywhere regularity);
- analyzed parameter-space singularities and joint input–parameter tie sets;
- described compositional stratification when attention is followed by ReLU MLPs, yielding block cells as micro-atoms of computation;
- framed attention strata and block cells as primitive micro-singular geometry for OCTA-style AGI and as a base for pattern-level interpretability;
- sketched how Fisher geometry and near-tie behavior prefigure Day 3's singular learning theory;
- and defined practical experimental protocols (Day 2.1–2.6) to observe these structures in real models and to relate them to circuits, robustness, and parameter-space behavior.

Day 2, Version 1.3, is thus the "micro-geometry" companion to Day 1's singularity separator, extended to parameter and block composition space, and ready to be wired into the broader OCTA Research program over the next 363 days.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*. AMS and Oxford University Press, 2000.

[2] J. Kaplan et al. Scaling laws for neural language models. arXiv:2001.08361, 2020.

[3] A. Power et al. Grokking: Generalization beyond overfitting on small algorithmic datasets. arXiv:2201.02177, 2022.

[4] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[5] S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.