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

Singularity vs. Singularities in Transformer Architectures

Week 1 · Day 1 · January 1, 2026

OCTA Research Internal Theory Program

Version 1.3 – Living Document Series (with Visualizations)

OCTA Research 365 Program Note. This document is the Week 1, Day 1 entry in a 365-day internal series dedicated to building a rigorous theoretical foundation for OCTA-style AGI systems. Each day extends, refines, and operationalizes a specific concept. Today's focus is: the precise relationship between *singularity* (macro-scale capability blow-up) and *singularities* (micro-scale mathematical/statistical structures) in Transformer architectures, with explicit visualizations in Figures 1–4.

Abstract

The term *singularity* is used in at least two very different ways in the context of modern artificial intelligence. On one hand, it denotes a hypothetical *technological singularity*, a macro-scale regime where AI-driven feedback loops produce unbounded or qualitatively discontinuous growth in capabilities, as in the toy model of Equation (10). On the other, it refers to *mathematical* and *statistical singularities*: points or regions where maps become non-smooth, non-invertible, ill-conditioned, or where statistical models lose regularity in the sense of classical asymptotic theory, as formalized in Definitions 3.1 and 6.1.

This article provides a scientifically rigorous LaTeX formulation connecting these notions in the specific context of Transformer architectures, framed as Day 1 of a year-long OCTA Research program. We formalize a Transformer as a parameterized map f_θ between finite-dimensional vector spaces (Section 2), define several classes of singularities (in function space, parameter space, probabilistic output space, and training dynamics; see Sections 3-7), and contrast them with dynamical models of macro-scale technological singularity (Section 4). We then draw on singular learning theory and information geometry (Section 6) to characterize Transformers as *singular statistical models*, analyzing how local singular structures (e.g. rank deficiencies, symmetry-induced degeneracies, sharp phase transitions in training, and scaling-law breakpoints) can contribute to emergent macro-level phenomena that are sometimes informally described as "approaching a singularity."

To anchor the abstractions, we provide OCTA-branded TikZ visualizations: Figure 1 illustrates the conceptual split between macro- and micro-level singularities; Figure 2 depicts finite-time blow-up in a capability model; Figure 3 renders a stratified view of attention singularities; and Figure 4 sketches the singular information geometry of a Transformer family. The overarching goal is to separate rhetoric from structure while identifying concrete, testable mathematical questions about singularities *inside* Transformer models and their relation to macro-level capability dynamics, and to establish a baseline conceptual frame for subsequent days of the OCTA Research 365 program.

Contents

1 OCTA Research Context: Day 1 Framing 3
- 1.1 Programmatic objective 3
- 1.2 Deliverables for Week 1, Day 1 4

2 Preliminaries: Transformers as Maps and Probabilistic Models 4
- 2.1 Deterministic core map 4
- 2.2 Self-attention layer 5
- 2.3 Probabilistic output layer 6

3 Mathematical Notions of Singularity 6
- 3.1 Singular points of a deterministic map 6
- 3.2 Non-smoothness and piecewise linearity 6
- 3.3 Dynamical singularities 7

4 Technological Singularity as a Macro-Scale Dynamical Phenomenon 7
- 4.1 Capability growth models 7
- 4.2 Feedback from Transformers to capability growth 8

5 Singularities in Transformer Parameter and Function Space 8
- 5.1 Parameter symmetries and identifiability singularities 8
- 5.2 Attention kernels and rank collapse 9
- 5.3 Generic full rank inside activation regions 9

6 Transformers as Singular Statistical Models 10
- 6.1 Regular vs. singular models 10
- 6.2 Fisher information and degeneracy 10
- 6.3 Real log canonical threshold (RLCT) 11

7 Singularities in Training Dynamics and Loss Landscapes 11
- 7.1 Gradient flow and critical points 11
- 7.2 Effective dynamical singularities 11

8 Toy Example: Singularity Structure in Minimal Attention 12

9 Information Geometry of Transformers 12

10 Singularity vs. Singularities: Conceptual Distillation 12

11 OCTA Conjectures and Experimental Protocols 12

12 Scaling Laws, Phase Transitions, and Singular Structures 13
- 12.1 Scaling laws as smooth regimes 13
- 12.2 Scaling-law singularities: breakpoints and regime changes 14
- 12.3 Connecting RLCT and scaling exponents 14

13 Coupled Macro–Micro Dynamical Template for OCTA 15
13.1 State variables 15
13.2 Generic coupled system 15
13.3 Singular geometries as control surfaces 15

14 OCTA Perfect Attractor Interpretation 16
14.1 Attractors in the coupled system 16
14.2 Perfect Attractor as constrained singular geometry 16

15 Roadmap for the Next 364 Days 17

16 Conclusion 17

List of Figures

1 Macro vs. micro singularities in the OCTA framing. Macro-level singularity is modeled as a property of capability dynamics $C(t)$ (Section 4), while micro-level singularities arise as geometric and statistical structures internal to Transformer architectures (Sections 3–6). 4
2 Toy macro-level capability trajectory from Equation (10). The vertical dashed line at $t = T^\star$ indicates finite-time blow-up in the toy model. This is a macro-level "singularity" conceptually distinct from the micro-level singularities internal to Transformer architectures (cf. Figures 3 and 4). 8
3 Stratified view of attention singularities in a minimal two-token, one-head case (Section 8). The cyan curve shows the softmax attention weight $A_{11}^{(\tau)}$ as a function of score difference Δ for $\tau > 0$, while the magenta step function shows the hard-attention limit $A_{11}^{(0)}$ with a discontinuity at the tie manifold $\Delta = 0$. This illustrates how micro-level non-smooth singularities arise directly from the attention mechanism. 10
4 Schematic of the singular information geometry of a Transformer family (Section 6). The cyan region represents an information manifold with metric given by the Fisher information $I(\theta)$, while the magenta curve represents a singular stratum where $I(\theta)$ is rank-deficient (Definition 6.1), often arising from symmetries as in Definition 5.1. . 11

1 OCTA Research Context: Day 1 Framing

1.1 Programmatic objective

The OCTA Research 365 program is designed as a daily cadence of theory, experimental design, and system-building steps toward AGI-level systems with transparent, mathematically grounded behavior. Day 1 sets up a critical distinction between macro-level and micro-level singularities, which we visualize in Figure 1.

We explicitly separate:

- **Macro-level *singularity*:** a regime in which self-improving AI systems induce capability dynamics that appear discontinuous on human time scales (Section 4).
- **Micro-level *singularities*:** precise mathematical and statistical structures—non-invertibility, degeneracies, bifurcations, ill-conditioned Jacobians/Fisher matrices—inside concrete architectures (here, Transformers; Sections 3–7).

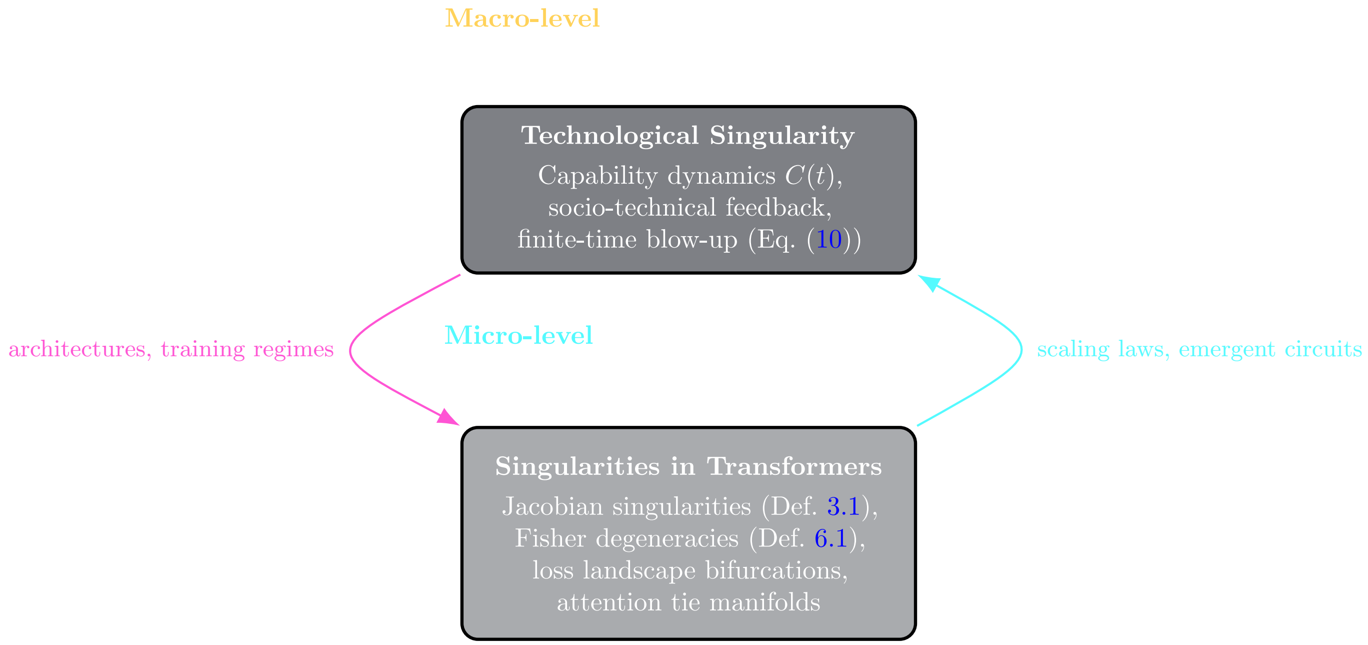
Macro-level
Technological Singularity
Capability dynamics $C(t)$,
socio-technical feedback,
finite-time blow-up (Eq. (10))
architectures, training regimes
Micro-level
scaling laws, emergent circuits
Singularities in Transformers
Jacobian singularities (Def. 3.1),
Fisher degeneracies (Def. 6.1),
loss landscape bifurcations,
attention tie manifolds

Figure 1: Macro vs. micro singularities in the OCTA framing. Macro-level singularity is modeled as a property of capability dynamics $C(t)$ (Section 4), while micro-level singularities arise as geometric and statistical structures internal to Transformer architectures (Sections 3–6).

OCTA Principle 1.1 (Day 1 OCTA Principle: Singularity Separation). For any OCTA-class system composed of Transformer-like components, we must maintain a clear separation between:

(a) *Local geometric/statistical singularities* in the parameter, input, and representation spaces of the model; and

(b) *Global capability singularity hypotheses* about feedback loops in the AI+human+world system.

Linking (a) to (b) requires explicit multi-scale models, not analogy.

1.2 Deliverables for Week 1, Day 1

For this entry, the deliverables are:

- A mathematically clean definition suite for singularities relevant to Transformers (Section 3).
- A singular-learning-theoretic view of Transformers as singular statistical models (Section 6).
- OCTA-flavored conjectures and protocols for mapping singular structures empirically (Section 11).
- A roadmap for how this document will be extended/reused over the rest of the year (Section 15).

2 Preliminaries: Transformers as Maps and Probabilistic Models

2.1 Deterministic core map

We adopt a continuous vector-space view, abstracting away discrete tokens until we introduce probabilistic outputs.

- Let d denote the model (embedding) dimension.
- Let n denote the sequence length.
- An input sequence is represented as a matrix $X \in \mathbb{R}^{n \times d}$, where each row is a token embedding.
- A Transformer with L layers and parameters θ induces a map

$$f_\theta : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d}.$$

- We write $f_\theta = f_\theta^{(L)} \circ \cdots \circ f_\theta^{(1)}$, where $f_\theta^{(\ell)}$ is the ℓ-th layer.

Vectorizing X, we define:

$$x = \operatorname{vec}(X) \in \mathbb{R}^N, \qquad N := nd,$$

and regard f_θ as a map $f_\theta : \mathbb{R}^N \to \mathbb{R}^N$.

2.2 Self-attention layer

We follow the standard architecture described informally in [7].

For one head h, with parameters $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_h}$ and $W_O \in \mathbb{R}^{d_h \times d}$, the forward computation is:

$$Q = XW_Q, \tag{1}$$
$$K = XW_K, \tag{2}$$
$$V = XW_V, \tag{3}$$

and the attention scores and weights are:

$$S = \frac{1}{\sqrt{d_h}} QK^\top \in \mathbb{R}^{n \times n}, \tag{4}$$
$$A = \operatorname{softmax}(S) \in \mathbb{R}^{n \times n}, \tag{5}$$

with row-wise softmax:

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

The head output is

$$H = AV \in \mathbb{R}^{n \times d_h}, \tag{6}$$
$$Y = HW_O \in \mathbb{R}^{n \times d}. \tag{7}$$

With multiple heads and residual/normalization structure, the overall layer is

$$X' = X + \mathrm{MHA}(X) + \mathrm{FFN}(X + \mathrm{MHA}(X)), \tag{8}$$

where MHA is the multi-head attention block and FFN is a position-wise feedforward network.

2.3 Probabilistic output layer

In language modeling, the Transformer core produces hidden states $H \in \mathbb{R}^{n \times d}$, which are mapped to token logits via a readout matrix $W_{\text{out}} \in \mathbb{R}^{d \times V}$:

$$Z = HW_{\text{out}} \in \mathbb{R}^{n \times V}.$$

For each position i, we obtain a distribution over a vocabulary of size V:

$$p_\theta(y_i = v \mid X) = \frac{\exp(Z_{iv})}{\sum_{u=1}^{V} \exp(Z_{iu})}.$$

Collecting all positions, the conditional distribution $p_\theta(\cdot \mid X)$ lies in the product of simplices:

$$p_\theta(\cdot \mid X) \in \Delta^{V-1} \times \cdots \times \Delta^{V-1} \quad (n \text{ copies}),$$

where $\Delta^{V-1} := \{p \in \mathbb{R}^{V}_{\geq 0} : \sum_v p_v = 1\}$.

Thus, the Transformer defines a parametric family of conditional distributions

$$\mathcal{P} = \{p_\theta(\cdot \mid X) : \theta \in \Theta\}.$$

3 Mathematical Notions of Singularity

3.1 Singular points of a deterministic map

We recall the classical definition used throughout this document and referenced in Figures 3 and 4.

Definition 3.1 (Regular and singular points of a map)**.** Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be differentiable at $x \in \mathbb{R}^N$.

- The point x is *regular* if $\operatorname{rank}(J_f(x)) = \min(N, M)$, where $J_f(x)$ is the Jacobian.
- Otherwise x is a *singular point* of f.

The *singular set* of f is

$$\Sigma_f := \{x \in \mathbb{R}^N : \operatorname{rank}(J_f(x)) < \min(N, M)\}.$$

In our setting, typically $N = M = nd$, so a point x is singular if and only if $J_f(x)$ is not full rank.

Remark 3.2 (Measure-zero typicality). Under mild conditions (e.g. f is a generic polynomial or analytic map with independent parameters), the singular set Σ_f has Lebesgue measure zero. However, in highly structured models such as Transformers with symmetries and shared parameters, singular sets can be systematically induced and aligned with architectural constraints.

3.2 Non-smoothness and piecewise linearity

Transformers often use piecewise linear activations (ReLU, variants of GELU) and non-smooth operations, so the function is typically only *piecewise differentiable*:

Definition 3.3 (Piecewise smooth map)**.** A function $f : \mathbb{R}^N \to \mathbb{R}^M$ is *piecewise smooth* if there exists a finite or countable partition of $\mathbb{R}^N$ into regions $\{R_\alpha\}_\alpha$ such that f is smooth on the interior of each R_α. The boundaries between regions are *non-smooth loci.*

In Figure 3 we visualize this as strata in the attention map space.

3.3 Dynamical singularities

When training dynamics are modeled as differential equations, we use:

Definition 3.4 (Dynamical singularity)**.** Let $\dot{z} = g(z)$ define a dynamical system on $\mathbb{R}^K$. A point $z^\star$ is a *fixed point* if $g(z^\star) = 0$. A *dynamical singularity* (for our purposes) is a point or parameter value at which:

- the Jacobian $Dg(z^\star)$ changes stability type (e.g. an eigenvalue crosses the imaginary axis, leading to a bifurcation), or
- solution trajectories exhibit finite-time blow-up or loss of existence/uniqueness.

This concept recurs in the training-focused protocols in Section 11.

4 Technological Singularity as a Macro-Scale Dynamical Phenomenon

4.1 Capability growth models

To formalize the technological singularity minimally, consider a scalar "capability" variable $C(t) \geq 0$ evolving according to a self-reinforcing growth law:

$$\frac{dC}{dt} = F(C; \alpha, \beta, \ldots). \tag{9}$$

A standard toy model of super-exponential growth is:

$$\frac{dC}{dt} = \alpha C^\gamma, \quad \alpha > 0,\ \gamma > 1. \tag{10}$$

Lemma 4.1 (Finite-time blow-up for $\gamma > 1$)**.** *For* (10) *with initial condition* $C(0) = C_0 > 0$ *and* $\gamma > 1$*, the solution blows up in finite time, i.e. there exists* $T^\star < \infty$ *such that*

$$\lim_{t \uparrow T^\star} C(t) = +\infty.$$

Proof. Separate variables:

$$\int_{C_0}^{C(t)} C^{-\gamma}\, dC = \alpha \int_0^t dt \quad \Longrightarrow \quad \frac{C^{1-\gamma} - C_0^{1-\gamma}}{1-\gamma} = \alpha t.$$

Thus

$$C^{1-\gamma} = C_0^{1-\gamma} + (1-\gamma)\alpha t.$$

Since $1 - \gamma < 0$, there exists finite

$$T^\star = \frac{C_0^{1-\gamma}}{\alpha(\gamma - 1)}$$

such that the right-hand side vanishes. As $t \uparrow T^\star$, $C^{1-\gamma} \downarrow 0$ and therefore $C(t) \uparrow \infty$. □

This blow-up is illustrated in Figure 2.

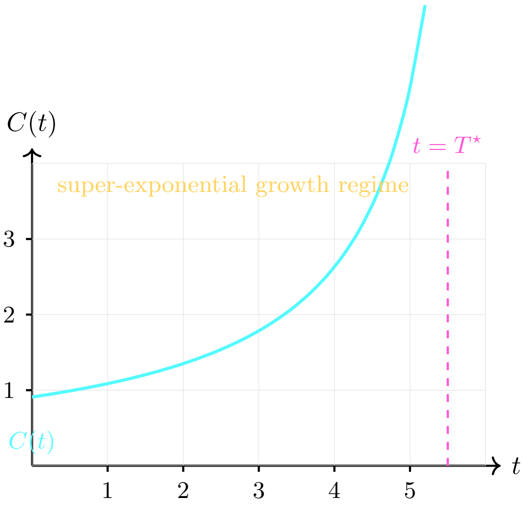

Figure 2: Toy macro-level capability trajectory from Equation (10). The vertical dashed line at $t = T^\star$ indicates finite-time blow-up in the toy model. This is a macro-level "singularity" conceptually distinct from the micro-level singularities internal to Transformer architectures (cf. Figures 3 and 4).

4.2 Feedback from Transformers to capability growth

Large-scale Transformer systems:

- accelerate research and development of new architectures, training algorithms, and engineering workflows,
- increase productivity in software, science, and design,
- and may themselves be used to design more efficient Transformers and deployment pipelines.

This creates feedback loops where $C(t)$ depends on the state of deployed models, the efficiency of scaling laws, and the structure of loss/capability landscapes, as sketched conceptually in Figure 1.

5 Singularities in Transformer Parameter and Function Space

5.1 Parameter symmetries and identifiability singularities

Transformers possess inherent parameter symmetries:

- permutation symmetries across attention heads,
- scaling symmetries involving layer normalization and residual paths,
- rotational symmetries in embedding space under some initializations or constraints,
- neuron permutation symmetries in feedforward layers.

These induce *identifiability singularities* in parameter space: multiple parameter vectors correspond to the *same* functional map f_θ and hence the same conditional distribution family $\{p_\theta(\cdot \mid X)\}$.

Let Θ denote parameter space and let $\mathcal{F}$ be the space of functions from $\mathbb{R}^N$ to $\mathbb{R}^N$. The model defines a map

$$\Phi : \Theta \to \mathcal{F}, \quad \theta \mapsto f_\theta.$$

Definition 5.1 (Identifiability singularity)**.** A parameter $\theta \in \Theta$ is an *identifiability singularity* if there exists a non-trivial neighborhood U of θ such that

$$\dim\{\tilde{\theta} \in U : f_{\tilde{\theta}} = f_\theta\} \geq 1.$$

Equivalently, the differential $D\Phi(\theta)$ has rank deficiency.

Example 5.2 (Permutation symmetry of attention heads)*.* Consider a Transformer layer with H attention heads, each head h having parameters θ_h and output $Y_h(X)$, and a final projection W_O that aggregates concatenated head outputs. If W_O is block-permutation equivariant (or the architecture effectively symmetrizes over heads), then any permutation π of heads yields parameters $(\theta_{\pi(1)}, \dots, \theta_{\pi(H)})$ encoding the same function. The set of permutations forms a finite group S_H acting on Θ, and parameter points related by this group action are identifiability singularities of Φ as in Definition 5.1.

5.2 Attention kernels and rank collapse

Introduce a temperature $\tau > 0$ and define:

$$S^{(\tau)} = \frac{1}{\tau\sqrt{d_h}} QK^\top, \tag{11}$$

$$A^{(\tau)} = \operatorname{softmax}(S^{(\tau)}). \tag{12}$$

Definition 5.3 (Hard-attention limit)**.** The *hard-attention limit* is the map

$$A^{(0)}(X) := \lim_{\tau \downarrow 0} A^{(\tau)}(X)$$

where the limit exists pointwise.

For generic X, within each row i there is a unique maximizer over j:

$$A^{(0)}_{ij}(X) = \begin{cases} 1, & j = \arg\max_k S^{(1)}_{ik}(X), \\ 0, & \text{otherwise.} \end{cases}$$

Proposition 5.4 (Non-smooth limit of attention)**.** *The map $X \mapsto A^{(\tau)}(X)$ is smooth for each fixed $\tau > 0$. As $\tau \downarrow 0$, the limiting map $X \mapsto A^{(0)}(X)$ is piecewise constant with non-differentiable boundaries along codimension-*1 *manifolds where max-scores tie.*

These tie manifolds and their stratification are visualized in Figure 3.

5.3 Generic full rank inside activation regions

Theorem 5.5 (Generic full-rank Jacobian for a linearized layer)**.** *Consider a simplified Transformer layer*

$$f_\theta(x) = x + Wx,$$

where $W \in \mathbb{R}^{N \times N}$ depends polynomially on parameters θ. Assume:

(i) the map $\theta \mapsto W(\theta)$ is polynomial with real coefficients,

(ii) there exists at least one parameter setting θ_0 such that $I + W(\theta_0)$ is invertible.

Then the set of parameters θ such that $I + W(\theta)$ is singular has Lebesgue measure zero in parameter space.

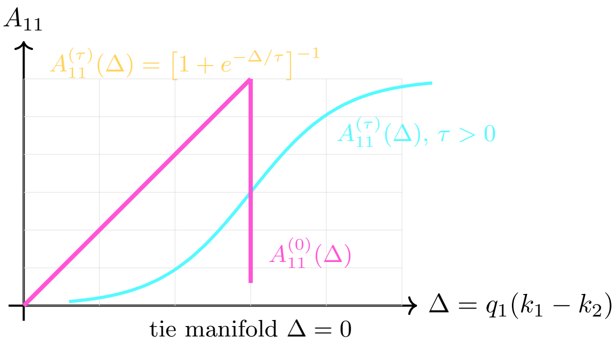

Figure 3: Stratified view of attention singularities in a minimal two-token, one-head case (Section 8). The cyan curve shows the softmax attention weight $A_{11}^{(\tau)}$ as a function of score difference Δ for $\tau > 0$, while the magenta step function shows the hard-attention limit $A_{11}^{(0)}$ with a discontinuity at the tie manifold $\Delta = 0$. This illustrates how micro-level non-smooth singularities arise directly from the attention mechanism.

6 Transformers as Singular Statistical Models

6.1 Regular vs. singular models

Classical parametric statistics considers a family of distributions $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^K\}$ with regularity assumptions (smooth, injective parameter map; positive-definite Fisher information) [1].

Neural networks, including Transformers, typically violate these assumptions due to:

- parameter symmetries (Definition 5.1),
- overparameterization,
- piecewise-linear activations creating non-analytic boundaries.

Definition 6.1 (Singular statistical model (informal))**.** A parametric model $\{p_\theta : \theta \in \Theta\}$ is *singular* if the Fisher information matrix $I(\theta)$ fails to be positive definite at some θ corresponding to the true distribution, or if the parameter-to-distribution map is not locally bi-analytic at that point; see [8].

6.2 Fisher information and degeneracy

Let $p_\theta(y \mid x)$ be the conditional distribution defined by the Transformer for input x and output y, with data distribution $\pi(x, y)$. The Fisher information is:

$$I(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log p_\theta(y \mid x)\, \nabla_\theta \log p_\theta(y \mid x)^\top \right].$$

Proposition 6.2 (Information degeneracy under symmetries (informal))**.** *Suppose the model has a continuous symmetry group G acting on Θ such that $p_{g\cdot\theta}(\cdot \mid x) = p_\theta(\cdot \mid x)$ for all $g \in G$, and the action is non-trivial near $\theta^\star$. Then $I(\theta^\star)$ has at least* $\dim G$ *zero eigenvalues: Fisher information is singular along directions tangent to the group orbit through $\theta^\star$.*

The resulting information geometry is illustrated schematically in Figure 4.

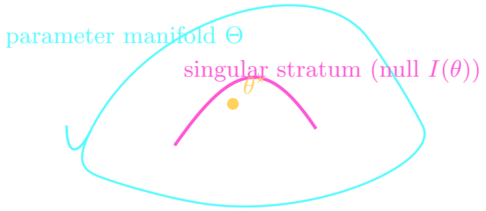
parameter manifold Θ
singular stratum (null $I(\theta)$)
$\theta^\star$

Figure 4: Schematic of the singular information geometry of a Transformer family (Section 6). The cyan region represents an information manifold with metric given by the Fisher information $I(\theta)$, while the magenta curve represents a singular stratum where $I(\theta)$ is rank-deficient (Definition 6.1), often arising from symmetries as in Definition 5.1.

6.3 Real log canonical threshold (RLCT)

Define the KL divergence between true and model distributions:

$$K(\theta) = \mathbb{E}_{(x,y)\sim\pi}\left[\log \frac{q(y \mid x)}{p_\theta(y \mid x)}\right].$$

When q is in the closure of the model, the set of θ minimizing $K(\theta)$ is typically singular. The *real log canonical threshold* (RLCT) λ characterizes model complexity and generalization in singular learning theory [8].

OCTA Principle 6.3 (OCTA RLCT View). Instead of treating “model size” as a one-dimensional knob, we treat the geometry of the KL-minimizer set and its RLCT as a primary object: changes in architecture that alter λ are first-class interventions in capability trajectories.

7 Singularities in Training Dynamics and Loss Landscapes

7.1 Gradient flow and critical points

Given loss $\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\theta(x), y)]$, gradient descent is

$$\theta_{k+1} = \theta_k - \eta\nabla\mathcal{L}(\theta_k),$$

and gradient flow is

$$\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta).$$

Definition 7.1 (Critical point)**.** A parameter $\theta^\star$ is a *critical point* of $\mathcal{L}$ if $\nabla\mathcal{L}(\theta^\star) = 0$.

7.2 Effective dynamical singularities

Definition 7.2 (Effective dynamical singularity region)**.** A region $U \subset \Theta$ is an *effective dynamical singularity region* if:

- the condition number of $H(\theta)$ exceeds a large threshold for all $\theta \in U$, or
- the spectrum of $H(\theta)$ crosses stability boundaries as a function of a control parameter (e.g. learning rate, batch size, model width).

These regions will be targeted explicitly in the experimental protocols of Section 11.

8 Toy Example: Singularity Structure in Minimal Attention

We consider the minimal attention example (two tokens, one head, $d_h = 1$) and identify tie manifolds and Jacobian conditioning, as visualized in Figure 3.

Let $X = [x_1^\top; x_2^\top] \in \mathbb{R}^{2\times d}$, $W_Q, W_K \in \mathbb{R}^{d\times 1}$, and define scalar queries and keys:

$$q_i = x_i^\top W_Q, \quad k_i = x_i^\top W_K, \quad i = 1, 2.$$

The score matrix is

$$S = \begin{pmatrix} q_1 k_1 & q_1 k_2 \\ q_2 k_1 & q_2 k_2 \end{pmatrix}.$$

For row $i = 1$, define $\Delta = q_1(k_1 - k_2)$. The softmax attention weight is:

$$A_{11} = \frac{1}{1 + \exp(-\Delta/\tau)}, \quad \tau > 0,$$

consistent with the logistic form annotated in Figure 3. As $\tau \to 0$, we obtain the hard limit with a discontinuity at $\Delta = 0$.

9 Information Geometry of Transformers

The Fisher information defines a (possibly degenerate) Riemannian structure on Θ (Figure 4). Natural gradient methods [1] invert $I(\theta)$ on regular strata, while singular directions require generalized inverses or quotienting out symmetries.

10 Singularity vs. Singularities: Conceptual Distillation

We distinguish:

- **Micro-level:** functional, parametric, dynamical, and statistical singularities inside Transformers (Sections 3–7).
- **Macro-level:** technological singularity as a property of global capability dynamics (Section 4).

Micro structure can catalyze macro regime changes (Figure 1), but the implication is not automatic; it is mediated by scaling laws, feedback loops, and external constraints.

11 OCTA Conjectures and Experimental Protocols

Day 1 outputs concrete conjectures and experimental blueprints that future days will refine.

Conjecture 11.1 (Jacobian Spectrum Phase Transition). For a family of Transformer models $\{f_{\theta_N}\}$ with increasing size N (parameters, width, or depth), there exist scaling thresholds N_c at which:

- the empirical distribution of singular values of $J_{f_{\theta_N}}(x)$ (averaged over data) undergoes a qualitative change, and
- these thresholds align with observed emergent capabilities (e.g. few-shot generalization, tool use).

Conjecture 11.2 (Fisher Rank and Grokking). In algorithmic tasks exhibiting grokking [5], the onset of generalization corresponds to a topological change in the rank structure of $I(\theta)$ near the minimizer set: directions that were previously null (or nearly so) gain positive curvature, reflecting a reorganization of the parameter manifold around the data manifold.

Experimental Protocol 11.3 (Day 1.1: Jacobian Spectrum Tracking).

(a) Fix a family of Transformers of increasing size.

(b) During training, at logarithmically spaced steps, sample a batch of inputs x and compute (or approximate) the singular value spectra of $J_f(x)$ for each layer.

(c) Fit empirical spectral densities and track:

- bulk shape (e.g. fit to random matrix baselines),
- tail behavior (power-law exponents),
- fraction of near-zero singular values.

(d) Correlate spectrum features with emergent behaviors and scaling-law regime changes.

Experimental Protocol 11.4 (Day 1.2: Fisher Geometry Probing).

(a) For a fixed trained model, approximate $I(\theta)$ or its action on random vectors using finite differences or Kronecker-factored approximations.

(b) Estimate the effective rank (number of significant eigenvalues) and principal subspaces.

(c) Repeat across training snapshots to observe how the Fisher geometry evolves from initialization to convergence, especially near grokking transitions.

Experimental Protocol 11.5 (Day 1.3: Scaling Law Breakpoint Detection).

(a) Collect scaling curves (loss vs. model size, data size, compute) across multiple architectural families, as in [3].

(b) Fit piecewise models with potential breakpoints in exponents.

(c) For each detected breakpoint, cross-reference with:

- architectural changes (new modules),
- Jacobian/Hessian/Fisher statistics from Protocols 11.3 and 11.4,
- emergence of qualitatively new behaviors.

12 Scaling Laws, Phase Transitions, and Singular Structures

12.1 Scaling laws as smooth regimes

Empirically, many language models obey power-law scaling relations of the form [3]

$$\mathcal{L}(N) \approx aN^{-\alpha} + b, \tag{13}$$

where N is a measure of model size or compute, $\mathcal{L}$ is loss, and (a, α, b) are fitted constants within a regime.

Within such a regime, the mapping $N \mapsto \mathcal{L}(N)$ is smooth and admits a well-defined derivative, and the underlying geometry (Jacobian spectrum, Fisher spectrum, RLCT) can be approximately treated as stable, in the sense that small changes in N do not induce qualitative changes in the loss landscape or emergent capabilities.

12.2 Scaling-law singularities: breakpoints and regime changes

In practice, scaling curves often exhibit:

- *exponent changes* (different α above/below some N_c),
- *double descent* patterns as a function of data, width, or regularization,
- *emergent behaviors* that appear above certain thresholds (e.g. in-context learning).

We treat such transitions as *scaling-law singularities*:

Definition 12.1 (Scaling-law singularity (informal))**.** A value N_c of a scaling parameter (model size, data size, compute) is a *scaling-law singularity* if, in a neighborhood of N_c, at least one of the following holds:

- the effective scaling exponent $\alpha(N)$ changes non-smoothly,
- higher-order derivatives of $\mathcal{L}(N)$ are discontinuous or undefined,
- qualitative capabilities appear or disappear.

Under the OCTA lens, such N_c are candidate points where the underlying singular geometry (e.g. Fisher rank, RLCT, Jacobian spectrum) changes structure.

12.3 Connecting RLCT and scaling exponents

In singular learning theory, the RLCT λ controls asymptotic generalization error: roughly,

$$\mathbb{E}[\mathcal{L}_{\text{test}}] \approx \mathcal{L}_{\text{train}} + \frac{\lambda}{n} + \text{(lower-order terms)},$$

for sample size n in an appropriate regime, where λ depends on the singularity structure of the KL minimizer set [8].

Informally:

- smooth, regular models correspond to simple λ tied to parameter dimension;
- singular models exhibit effective λ that reflects a lower-dimensional "essential" manifold plus singular directions.

Conjecture 12.2 (RLCT shift at scaling-law singularities). When a Transformer family crosses a scaling-law singularity at N_c, there is a corresponding shift in the estimated RLCT $\lambda(N)$ of the family:

- below N_c, $\lambda(N)$ reflects one effective singular structure;
- above N_c, $\lambda(N)$ reflects a different one;
- the break in effective scaling exponent $\alpha(N)$ is co-located (up to noise) with this RLCT shift.

In other words, the "kink" in scaling laws is interpreted as a macroscopic shadow of a microscopic change in singular geometry.

13 Coupled Macro–Micro Dynamical Template for OCTA

To connect internal singular structures to macro-level capability dynamics, we sketch a minimal coupled system that OCTA Research can instantiate and refine.

13.1 State variables

We introduce coarse-grained variables:

- $C(t)$: scalar capability index (e.g. benchmark-aggregated performance, economic productivity proxy).
- $M(t)$: model scale index (parameters, compute, or an effective measure combining them).
- $S(t)$: singular-structure index (e.g. a functional of Jacobian/Fisher spectra, RLCT estimate, or a proxy such as fraction of near-zero singular values).
- $R(t)$: resource index (available compute, energy, data, etc.).

The full microscopic state includes parameters $\theta(t)$, data distribution, and environment state; here we work with these coarse aggregates.

13.2 Generic coupled system

A minimal coupled ODE template is:

$$\frac{dC}{dt} = f_C(C, M, S, R), \tag{14}$$

$$\frac{dM}{dt} = f_M(C, M, R), \tag{15}$$

$$\frac{dS}{dt} = f_S(\theta(t), \text{training regimen}) \approx \tilde{f}_S(C, M, S, R), \tag{16}$$

$$\frac{dR}{dt} = f_R(C, M, R). \tag{17}$$

In this picture:

- f_C captures how capability changes as a function of model scale, singular geometry, and resources.
- f_M captures how quickly larger models are built when capabilities and resources are at certain levels (e.g. more capability leads to more investment).
- f_S captures how internal singular geometry evolves with training and architecture choices.
- f_R captures resource replenishment and depletion dynamics.

13.3 Singular geometries as control surfaces

We can think of S as a vector of scalar metrics:

$$S = (S_{\text{Jac}}, S_{\text{Fisher}}, S_{\text{RLCT}}, \dots),$$

where each component measures a different aspect of singular geometry:

- S_{Jac}: some statistic of Jacobian singular values (e.g. fraction near zero).
- S_{Fisher}: effective Fisher rank or condition number.
- S_{RLCT}: current RLCT estimate.

OCTA Principle 13.1 (OCTA Singular Surface Control). Instead of controlling C and M directly, an OCTA-class system exercises control at the level of S, steering training and architecture choices to remain within singular-geometry regions that yield stable, interpretable macro-level dynamics.

This aligns with the idea that the macro-level singularity question (finite-time blow-up of C) should be constrained by deliberately managing micro-level singularities encoded in S.

14 OCTA Perfect Attractor Interpretation

14.1 Attractors in the coupled system

In the coupled system of Section 13, an *attractor* is a subset $\mathcal{A}$ of the (C, M, S, R) state space such that:

- trajectories starting in a neighborhood of $\mathcal{A}$ converge to it (in the sense of dynamical systems),
- the restriction of the dynamics to $\mathcal{A}$ is stable under perturbations.

We can distinguish:

- **Capability-safe attractors**, where $C(t)$ grows but remains bounded or saturates.
- **Runaway attractors**, where $C(t)$ exhibits super-exponential or finite-time blow-up behavior.

14.2 Perfect Attractor as constrained singular geometry

In an OCTAcontext, a *Perfect Attractor* can be interpreted as an attractor $\mathcal{A}_\star$ such that:

- internally, the singular geometry $S(t)$ lies within a carefully engineered manifold $\mathcal{S}_\star$ (e.g. RLCT and Fisher metrics kept in a desired regime),
- externally, $C(t)$ exhibits strong growth but is constrained by design to avoid runaway modes.

OCTA Principle 14.1 (OCTA Perfect Attractor Principle (Day 1 Sketch)). A Perfect Attractor is an attractor $\mathcal{A}_\star$ in the coupled (C, M, S, R) system where:

(a) $S(t)$ remains within a singular-geometry manifold $\mathcal{S}_\star$ that is stable under training and scaling interventions;

(b) $C(t)$ grows sub-singularly (no finite-time blow-up) while continuously improving capabilities;

(c) deviations in S that would push C toward runaway modes are dynamically corrected by control loops (architecture choice, regularization, training curricula).

Day 1 does not attempt to formalize the full Perfect Attractor mathematics; instead, it establishes the vocabulary to treat the attractor as a joint constraint on macro-level dynamics and micro-level singular geometry.

15 Roadmap for the Next 364 Days

To make OCTA Research 365 concrete, Day 1 ends with a minimal roadmap for how this document branches:

- **Week 1 (Foundations):**
  - Day 1: Singularity vs. singularities (this document).
  - Day 2: Detailed analysis of attention as a piecewise-smooth, stratified map; explicit computation for low-dimensional cases, extending Figure 3.
  - Day 3: Singular learning theory mini-primer specialized to Transformers and sequence models, expanding on Section 6.
  - Day 4: Experimental design for Jacobian and Fisher estimation in large-scale models (Protocols 11.3–11.4).
  - Day 5–7: Implementation notes, code skeletons, and first small-scale experiments inspired by Figures 2 and 4, and by the coupled system in Section 13.
- **Weeks 2–4 (Geometry and Attractors):**
  - Connect singular sets to attractor structures in representation space.
  - Make the Perfect Attractor notion precise: specify candidate $\mathcal{S}_\star$ manifolds and their stability properties.
- **Months 2–6 (Scaling and Emergence):**
  - Empirically map singular structures across multiple model families.
  - Tie these to emergent behaviors and scaling-law transitions (Section 12, Protocol 11.5).
- **Months 7–12 (Macro Dynamics and Safety):**
  - Embed empirical scaling and singularity data into macro-level capability dynamics (Section 4).
  - Analyze scenarios that resemble macro-level singularity behavior vs. Perfect Attractor regimes.
  - Use this to constrain and design safe OCTA-class AGI trajectories, in line with Principles 1.1 and 6.3.

16 Conclusion

On January 1, 2026 (Week 1, Day 1 of the OCTA Research 365 program), we have:

- distinguished macro-level *singularity* from micro-level *singularities* (Figure 1),
- grounded the latter in explicit mathematics and singular learning theory (Sections 3 and 6, Figures 3 and 4),
- introduced a macro-level capability model (Section 4, Figure 2),
- framed scaling-law phase transitions as signatures of shifting singular geometry (Section 12),

- sketched a coupled macro–micro dynamical template (Section 13),
- and provided a first, high-level interpretation of the OCTA Perfect Attractor in this language (Section 14),
- alongside concrete OCTA-flavored conjectures and experimental protocols (Section 11).

Subsequent days will:

- take specific slices of this document (e.g. attention stratification, RLCT estimation, Jacobian spectra, coupled dynamics),
- convert them into executable code and experiments,
- and fold the resulting empirical insights back into both architecture design and macro-level dynamic models.

In this sense, Day 1 is a *singularity separator*: its purpose is to ensure that, as OCTA Research pushes toward increasingly powerful systems over the next 364 days, we preserve a clear, mathematically tractable distinction between local singularities inside our models and global claims about capability blow-up, while creating the scaffolding needed to formalize and eventually realize OCTA-style Perfect Attractors.

References

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

[2] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. arXiv:1703.04933, 2017.

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

[4] S. Mallat. Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A*, 374(2065), 2016.

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

[6] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. CRC Press, 2018.

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

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