OCTA Research

The Perfect Attractor Principle: Advantage-Based Routing, Emergent Bayesian Manifolds, and Gradient Dynamics in Transformer Systems

OCTA Research Group

research@octa.systems

OCTA Research White Paper Series

Abstract

Large transformer models exhibit emergent probabilistic reasoning, calibration, and specialization despite being trained via simple gradient descent on cross-entropy objectives. Recent analysis of attention gradients shows that a single transformer head implements an implicit expectation–maximization (EM) procedure: attention weights behave as soft responsibilities, while value vectors update as responsibility-weighted prototypes. This induces an *advantage-based routing law*: attention increases toward value vectors whose contribution to loss reduction exceeds the average, creating a positive feedback loop that couples routing and representation.

Building on recent work characterizing these dynamics as sculpting low-dimensional Bayesian manifolds inside transformer representations, we propose the *Perfect Attractor Principle*: under sustained predictive pressure and finite capacity constraints, learning systems converge toward stable low-dimensional manifolds—*Perfect Attractors*—on which routing flows along advantage gradients and internal representations adopt Bayesian semantics. We show that the EM-like gradient dynamics of transformer attention constitute an explicit realization of this principle, thereby unifying optimization, geometry, and function in neural sequence models.

We formalize advantage flow as a replicator equation on the attention simplex, analyze its fixed points and stability, connect training to information geometry, outline multi-head interactions as a game over routing–representation space, and propose falsifiable empirical predictions. The framework explains phase transitions, specialization, and calibration phenomena observed in practice, while suggesting a broader law-like view of intelligence as a geometric attractor sculpted by gradient descent.

1 Introduction

Transformer architectures trained with stochastic gradient descent display structured internal organization, including attention head specialization, latent role assignment, and emergent probabilistic reasoning, despite being optimized only for next-token prediction under cross-entropy loss. These behaviors arise *without explicit architectural bias* toward Bayesian inference or latent variable modeling, raising a central question:

> Why do trained transformers behave as if they are performing structured probabilistic inference?

Aggarwal, Dalal, and Misra analyze the gradient of cross-entropy loss with respect to attention scores and value vectors in a transformer head, showing that gradient descent induces an *advantage-based routing law* for attention and *responsibility-weighted prototype updates* for value vectors [1]. In companion work, they demonstrate that these dynamics contract internal states onto low-dimensional manifolds whose geometry matches that of approximate Bayesian inference mechanisms, with attention heads exhibiting frame–precision dissociation and entropy-ordered value geometry [2]. Preliminary scaling evidence suggests that the same Bayesian manifold geometry persists and sharpens in larger models [3].

Motivated by these results, we propose the **Perfect Attractor Principle**:

> Under sustained predictive pressure and capacity constraints, gradient descent drives learning systems toward low-dimensional manifolds on which routing follows advantage gradients and representations encode latent structure with Bayesian semantics.

We argue that the EM-like dynamics identified in transformer attention are a concrete instantiation of this principle, and that this perspective provides a unifying account of optimization, geometry, and function in large language models.

2 Transformer Attention and EM-Like Dynamics

We consider a single attention head in a transformer layer. For tokens $i = 1, \dots, T$, let

$$q_i, k_i, v_i \in \mathbb{R}^d$$

denote queries, keys, and values. Define scores and attention weights:

$$s_{ij} = \frac{1}{\sqrt{d}} q_i^\top k_j, \tag{1}$$

$$\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{m=1}^{T} \exp(s_{im})}. \tag{2}$$

The head output is

$$h_i = \sum_{j=1}^{T} \alpha_{ij} v_j. \tag{3}$$

Let L be the cross-entropy loss and

$$u_i = \frac{\partial L}{\partial h_i}$$

be the upstream error signal at position i. Define the bilinear alignment

$$b_{ij} = u_i^\top v_j. \tag{4}$$

Let $\mathbb{E}_{\alpha_i}[b] = \sum_m \alpha_{im} b_{im}$.

Aggarwal et al. [1] show that

$$\frac{\partial L}{\partial s_{ij}} = \alpha_{ij}\big(b_{ij} - \mathbb{E}_{\alpha_i}[b]\big), \tag{5}$$

and that value vectors update as

$$\Delta v_j \propto -\sum_{i=1}^{T} \alpha_{ij} u_i, \tag{6}$$

i.e., as responsibilities-weighted prototypes of the error signals.

Equation (5) is an *advantage-based routing law*: attention scores are updated in proportion to the difference between a local "advantage" b_{ij} and its attention-weighted mean. Values that are above-average in reducing loss for a given position receive increased attention; those that are below-average are suppressed. Equation (6) is a *responsibility-weighted prototype update*: each value v_j is updated by aggregating upstream gradients from the positions that attend to it, weighted by their attention weights.

This structure is directly analogous to the expectation–maximization (EM) algorithm for latent-variable models:

- E-step: compute responsibilities $r_{ij} = p_\theta(z = j \mid x_i)$;
- M-step: update component parameters using r_{ij} as weights.

Identifying attention weights α_{ij} with responsibilities and values v_j with component parameters, the coupling of (5) and (6) defines an implicit EM-like procedure. Aggarwal et al. [2] show that this procedure sculpts low-dimensional Bayesian manifolds inside transformer representation space.

3 The Perfect Attractor Principle

Let θ denote model parameters, $p_\theta(y \mid x)$ the predictive distribution, and $h_\theta(x) \in \mathcal{H}$ internal representations in a high-dimensional space $\mathcal{H}$. Let $\mathcal{M} \subset \mathcal{H}$ be a candidate low-dimensional manifold and π a routing policy (e.g., attention distributions) defined on $\mathcal{M}$.

3.1 Setup

The model is trained to minimize expected loss

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim P_{\text{data}}}\big[\ell(y, p_\theta(\cdot \mid x))\big], \tag{7}$$

typically cross-entropy, under capacity and regularization constraints. Parameters evolve under gradient flow:

$$\dot{\theta}(t) = -\nabla_\theta \mathcal{L}(\theta(t)). \tag{8}$$

Training induces trajectories $h_{\theta(t)}(x)$ in representation space and routing policies $\pi_{\theta(t)}$.

3.2 Definition

[Perfect Attractor] A pair $(\mathcal{M}^\star, \pi^\star)$ is a *Perfect Attractor* of the learning dynamics if:

(i) **Geometric attraction:** for P_{data}-almost all x and almost all initializations in a basin $\mathcal{B}$,

$$\lim_{t\to\infty} \operatorname{dist}\big(h_{\theta(t)}(x), \mathcal{M}^\star\big) = 0;$$

(ii) **Routing convergence:** routing policies converge, $\pi_{\theta(t)} \to \pi^\star$;

(iii) **Advantage optimality:** $(\mathcal{M}^\star, \pi^\star)$ is a stationary point of an advantage functional

$$\mathcal{J}(\mathcal{M}, \pi) = \mathbb{E}_{(x,y)}\Big[\mathbb{E}_{(i,j)\sim\pi(\cdot|h_\theta(x))}\big[A_{ij}(h_\theta(x))\big]\Big] - \lambda\, \mathcal{C}(\mathcal{M}, \pi),$$

where A_{ij} is an advantage signal and $\mathcal{C}$ encodes capacity constraints.

Intuitively, a Perfect Attractor is a low-dimensional manifold of internal states and a routing policy such that:

- gradient descent drives internal states toward the manifold;
- routing flows along directions of positive advantage on the manifold;
- the resulting configuration is variationally optimal given constraints.

4 Gradient Flow and Information Geometry

The model family $\{p_\theta(y \mid x)\}$ induces a Fisher information metric

$$g_{ab}(\theta) = \mathbb{E}_{x\sim P_{\text{data}}}\mathbb{E}_{y\sim p_\theta(\cdot|x)}\Big[\partial_{\theta_a} \log p_\theta(y \mid x)\, \partial_{\theta_b} \log p_\theta(y \mid x)\Big], \tag{9}$$

making parameter space a statistical manifold (Θ, g).

Under continuous-time gradient flow

$$\dot\theta(t) = -\nabla_\theta \mathcal{L}(\theta(t)), \tag{10}$$

training trajectories follow curves inside (Θ, g).

Geometric attractor hypothesis. Gradient flow concentrates trajectories along a geodesically coherent, low-dimensional submanifold corresponding to latent-variable structure and advantage-aligned routing; internal representation space mirrors this as a Bayesian manifold $\mathcal{M}^\star$.

This gives gradient descent a geometric meaning: the model rolls downhill inside a curved information landscape until it falls into a stable inference geometry.

5 Advantage Flow as a Replicator Equation

Define the advantage

$$A_{ij} = b_{ij} - \mathbb{E}_{\alpha_i}[b], \tag{11}$$

with $b_{ij} = u_i^\top v_j$ and $\mathbb{E}_{\alpha_i}[b] = \sum_m \alpha_{im} b_{im}$ as before. Let $\boldsymbol{\alpha}_i \in \Delta^{T-1}$ be the attention simplex for query i.

Gradient descent on scores s_{ij} according to (5) induces an effective flow on $\boldsymbol{\alpha}_i$:

$$\dot\alpha_{ij} = \alpha_{ij}\Big(A_{ij} - \mathbb{E}_{\alpha_i}[A]\Big), \tag{12}$$

where $\mathbb{E}_{\alpha_i}[A] = \sum_m \alpha_{im} A_{im}$.

Thus positive advantage increases mass on component j and negative advantage decreases it, exactly as in classical replicator dynamics.

Replicator interpretation. Equation (12) is the standard replicator equation on the probability simplex, with A_{ij} in the role of fitness:

$$\dot\alpha_{ij} = \alpha_{ij}\big(A_{ij} - \bar A_i\big), \quad \bar A_i = \mathbb{E}_{\alpha_i}[A].$$

Attention distributions evolve as if components with higher-than-average advantage replicate, while others die out.

5.1 Support-Maximization Fixed Points

[Support-Maximization for Advantage Replicator Dynamics] For a fixed query position i, consider the replicator dynamics (12) on the simplex Δ^{T-1}. Then:

(a) Any fixed point $\boldsymbol{\alpha}_i^\star$ satisfies

$$\alpha_{ij}^\star > 0 \;\Rightarrow\; A_{ij} = \lambda_i$$

for some scalar λ_i, and all indices outside the support have $A_{ij} \le \lambda_i$.

(b) The set of asymptotically stable fixed points consists of distributions whose support is contained in the set of maximal-advantage indices:

$$\operatorname{supp}(\boldsymbol{\alpha}_i^\star) \subseteq \arg\max_j A_{ij}.$$

If advantage gaps are nonzero, stable fixed points have support *equal* to the maximal-advantage set.

Proof Sketch. For fixed i, define the Lyapunov function

$$V(\boldsymbol{\alpha}_i) = \max_j A_{ij} - \mathbb{E}_{\alpha_i}[A].$$

$V(\boldsymbol{\alpha}_i) \ge 0$ with equality if and only if all probability mass lies on indices achieving maximal advantage and those indices share equal advantage. Differentiating V along trajectories of the replicator dynamics yields $\dot{V} \le 0$, so V is non-increasing and trajectories converge to the set where $V = 0$. Linearizing (12) about a fixed point shows that perturbations in non-support coordinates decay at a rate proportional to pairwise advantage gaps $A_{ij} - A_{ik}$, establishing asymptotic stability. □

Thus, at convergence, attention support coincides with the maximal-advantage structure and non-maximal components are suppressed. This formalizes the intuition that advantage flow "hardens" soft assignments into a specialized routing structure and explains empirical attention head specialization.

6 Multi-Head and Multi-Layer Coupling

Real transformer layers contain multiple heads and are stacked into deep networks. Let heads be indexed by $h = 1, \dots, H$. Each head has its own scores, attention, and values:

$$s_{ij}^{(h)}, \quad \alpha_{ij}^{(h)}, \quad v_j^{(h)},$$

but receives shared upstream gradients through the residual pathway.

Each head therefore:

- competes with other heads for advantage-weighted routing mass;
- modifies the representation space that other heads see;
- and cooperates in minimizing a shared global loss.

This induces a **multi-population replicator system**: each head acts like a population evolving under fitness $A^{(h)}$, and the joint behavior can be understood as a game in routing–representation space. Head specializations correspond to *evolutionary stable strategies (ESS)* under the shared loss landscape: a configuration in which no head can unilaterally change its routing policy to achieve higher long-term advantage given the others' strategies.

Empirically observed head specialization (e.g., to syntactic, positional, or semantic roles) naturally follows from this perspective, without explicit architectural bias toward such roles.

7 Phase Transitions and Attractor Capture

As training progresses, the representational manifold and advantage structure co-evolve. When capacity or data exposure crosses certain thresholds, the advantage landscape can change rank, opening new high-advantage directions.

From the Perfect Attractor perspective, abrupt capability jumps correspond to *attractor capture events*, where trajectories enter the basin of a new Perfect Attractor $(\mathcal{M}^\star, \pi^\star)$. This predicts that phase transitions in capabilities should coincide with qualitative reorganizations of:

- attention support and advantage alignment;
- value geometry (e.g., entropy ordering [2]);
- and representation curvature along the emerging manifold.

8 Relation to Physical and Biological Attractors

Many adaptive systems display similar attractor geometry:

- neural population codes organize on low-dimensional manifolds that govern behavior;
- thermodynamic systems minimize free energy under constraints;
- ecological systems evolve toward competitive equilibria under fitness gradients, described by replicator equations;
- predictive coding theories view cortex as minimizing a variational free energy bound.

The Perfect Attractor Principle suggests a common pattern:

> Under sustained pressure to reduce surprise or prediction error, adaptive systems converge to low-dimensional manifolds where routing of influence aligns with expected advantage.

Transformer attention is one instance, with advantage-defined replicator dynamics playing the role of fitness-driven selection over internal communication channels.

9 Experimental Program

We outline concrete tests of the Perfect Attractor framework and the advantage replicator view, closely aligned with the "Bayesian wind-tunnel" setups of Aggarwal et al. [1, 2].

E1: Low-Rank Advantage Field. In a small controlled transformer, compute A_{ij} throughout training and perform PCA on flattened advantage tensors (across batch and time). Prediction: the effective rank of A remains much smaller than the ambient dimension, indicating a low-dimensional advantage manifold.

E2: Early Routing Freeze. Freeze attention scores or patterns after an early fraction of training (e.g., 5%). Continue training values and downstream layers. Prediction: calibration and ELBO-proxy metrics continue to improve, as value geometry and downstream parameters align to a fixed routing policy on the emerging manifold.

E3: Advantage-Preserving Perturbations. Perturb attention scores by monotone transformations that preserve the advantage ordering within each row (e.g., sorting scores plus small noise, or applying monotone rescaling). Prediction: performance degrades minimally, indicating that relative advantage, not absolute score magnitude, carries the functional signal for routing.

E4: Advantage-Ordering Violation. Conversely, scramble advantage ordering while preserving marginal statistics (e.g., permuting scores across positions or values within each row). Prediction: performance drops sharply, consistent with the causal role of advantage ordering in routing.

These experiments provide quantitative tests of the Perfect Attractor predictions and the replicator interpretation of attention training dynamics.

10 Figures

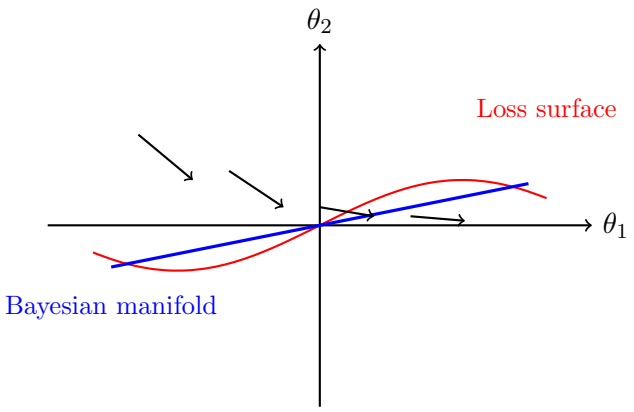

Figure 1: Schematic: gradient flow on the loss surface contracts trajectories onto a low-dimensional Bayesian manifold in parameter space, interpreted as a Perfect Attractor.

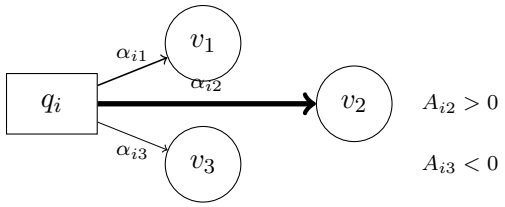

Figure 2: Advantage-based routing: arrow thickness indicates attention weights. Mass flows toward value vectors with positive advantage A_{ij} and away from those with negative advantage, implementing a replicator dynamic over internal communication channels.

11 Limitations and Open Questions

The Perfect Attractor Principle, as formulated here, remains a hypothesis. Key limitations and open questions include:

- The EM analogy is local and approximate; global convergence guarantees for deep, nonlinear networks under stochastic optimization are not yet available.
- The scaling laws governing the intrinsic dimensionality of the attractor manifold as a function of model and data size are unknown.
- The interaction between optimization noise, regularization, and advantage-based dynamics requires further theoretical and empirical study.
- The connection between attractor capture and observed double-descent phenomena in generalization remains to be clarified.

Nonetheless, the framework yields concrete, falsifiable predictions and aligns with a growing body of empirical evidence on transformer training dynamics.

12 Conclusion

We have articulated the Perfect Attractor Principle: under predictive pressure and capacity constraints, gradient descent drives transformer systems toward low-dimensional manifolds on which routing flows along advantage gradients and internal representations adopt Bayesian semantics.

Recent attention-gradient analyses show that transformer attention realizes an implicit EM-like procedure and that routing dynamics obey a replicator law over internal communication channels. This unifies optimization (gradient flow), geometry (attractor manifolds), and function (in-context probabilistic reasoning).

The framework predicts—and helps explain—attention specialization, calibration gains under frozen routing, low-dimensional advantage structure, and phase transitions via attractor capture. It suggests that intelligence across adaptive systems may be best understood as geometry sculpted by advantage-based flows.

References

[1] N. Aggarwal, S. R. Dalal, and V. Misra. Gradient Dynamics of Attention: How Cross-Entropy Sculpts Bayesian Manifolds. *arXiv preprint arXiv:2512.22473*, 2025.

[2] N. Aggarwal, S. R. Dalal, and V. Misra. The Bayesian Geometry of Transformer Attention. *arXiv preprint arXiv:2512.22471*, 2025.

[3] N. Aggarwal, S. R. Dalal, and V. Misra. Scaling Laws for Bayesian Geometry in Transformer Attention. *arXiv preprint arXiv:2512.23752*, 2025.