Towards Transformer Understanding: Transformer dynamics as gradient-shaped additive transport

AI Slop assembled from notes by joshuah.rainstar@gmail.com

December 28, 2025

Abstract **– Mainstream analysis of Transformer architectures typically relies on functional metaphors, such as retrieval, mixing, or modular processing, that fail to account for the emergent geometric behaviors of these systems. This article argues that the Transformer is better understood as a continuous dynamical system performing learned additive coordinate transport. We posit that the behavior of the model is not programmed but discovered via the teleological attractor of the loss function (i.e., end-loss gradients determine where structure is cheapest to realize). The architecture functions through a dimensional lift, where embeddings serve as an expansion of scalar indices into high-dimensional space, providing the degrees of freedom necessary to satisfy the competing demands of synthesis, relevance, and transport. The resulting dynamics are governed by gradient-shaped operator geometry: simplex-constrained mixing induces diffusive transport across T, while norms and residual dynamics determine which motions remain stable and learnable.**

1. DEFINITIONS & NOTATION

To avoid ambiguity regarding the specific mechanical operations discussed, we define the following notation and terms:

Let the state tensor be $X \in \mathbb{R}^{B\times T\times C}$, where B is batch size, T is the sequence length (time/position axis), and C is the channel dimension. Let the layer index be $\ell = 1 \dots L$.

The attention mechanism is an operator that proposes a temporal update ΔX^ℓ. The fundamental operation is:

$$\Delta X^\ell_{:,t,:} = \sum_{t'} A^\ell_{t,t'}\, V^\ell(X^\ell_{:,t',:}) \tag{1}$$

$$X^{\ell+1} = X^\ell + \Delta X^\ell + \mathrm{MLP}(X^\ell + \Delta X^\ell) \tag{2}$$

Here, the operator @ refers strictly to the additive mixing (weighted summation) defined by the matrix multiplication $A \times V$. The residual connection refers to the additive accumulation of these updates across depth ℓ.

We distinguish two simplex constraints. First, the kernel simplex constraint applies to the mixing operator A^ℓ: rows are often normalized (e.g., by softmax) so mixing coefficients form a simplex point, making each update a convex combination of available value vectors. Second, the terminal simplex boundary condition applies at the output: after the LM head projection, logits are normalized into a vocabulary probability distribution to satisfy cross-entropy. The former directly shapes the local transport operator; the latter defines the task boundary the global dynamics must ultimately satisfy. We do not assume the internal state itself lies on a simplex; rather, the learning dynamics are shaped by operating under these constraints and by the need to land on the terminal boundary.

Spine. The Transformer's core computation is a repeated two-step dynamical update on X: (i) a simplex-constrained, additive mixing operator proposes a temporal transport update across the sequence axis T, and (ii) residual accumulation carries the proposed update forward in depth, where subsequent operators can reinforce, cancel, or reinterpret it. Under simplex constraints, the generic stable regime of the mixing operator is diffusive (convex-hull transport). Expressivity is recovered not by making the transport sharp everywhere, but by interleaving transport with reconditioning operators (MLPs and norms) that rebuild a usable coordinate basis for future transport decisions.

2. INTRODUCTION: THE EPISTEMICS OF BEHAVIOR

The prevailing mental model of the Self-Attention mechanism is one of discrete routing: token A "looks at" token B and aggregates information based on similarity. This discrete, relational view, while intuitive, obscures the mechanical structure. It treats the architecture as a set of logical operations rather than a system constrained by a global objective.

To understand why Transformers actually function, and why they exhibit specific pathologies like early-layer fragility and late semantic crystallization, we must ask: who decides how the model behaves? The answer lies at the end of the process. The loss function is the ultimate arbiter. What exists after projection through the `LM_head` is the target the model works to produce: a probability distribution on the simplex.

A clean mathematical lens for the "teleology of the

loss" is that attention dynamics can be idealized as an interacting particle system whose continuum limits relate to Wasserstein-type gradient-flow behavior, including clustering and synchronization [1]. The important conceptual move is: the architecture does not specify "reasoning modules"; it specifies a class of admissible flows, and the loss selects which flows are stable and cheap to realize under the constraints of the operator family.

From this perspective, the Transformer is not a machine that "thinks"; it is a state space being shaped by end-loss gradients. Attention is not retrieval. It is a **learned additive coordinate transport** occurring via simplex-conditioned mixing weights acting on representations. The specific behaviors we witness, such as diffusion or semantic collapse, are not architectural choices but inevitable consequences of optimizing transport in this constrained space.

3. THE CORE MECHANISM: ADDITIVE MIXING + RESIDUAL ACCUMULATION

The engine of the Transformer is defined by two fundamental operations: additive mixing across the sequence axis T and additive accumulation across the depth axis ℓ. The attention mechanism functions as a mixing operator that proposes a temporal update ΔX^ℓ by computing a weighted sum of value vectors. Crucially, because the mixing weights are constrained to a simplex (via Softmax), this update is a convex combination.

The residual connection then integrates this proposal into the evolving state $X^{\ell+1} = X^\ell + \Delta X^\ell$. This structure implies that transport is not a discrete movement of tokens but a continuous integration of diffusive updates. The ill-conditioning arises because repeated additive mixing in a high-dimensional space tends to wash out distinct features unless actively counteracted.

4. THE DIMENSIONAL LIFT & OPERATOR ROLES

The central structural tension in the Transformer block begins with the embedding layer. In standard nomenclature, this is treated as a lookup table. Epistemically, it is a **dimensional lift**.

The discrete index carries identity but no continuous geometry; the embedding is the lift that supplies continuous degrees of freedom. This expansion is strictly necessary to create a geometric volume capable of supporting complex interference patterns. Within the block, this lift is re-manifested via the Q, K, V projections. Mainstream analysis treats these as symmetric peers. Mechanically, they are distinct epistemic objects.

The Kernel Set (K): K should not be viewed as "keys" to a lock, but as the **kernel** defining the metric space of the attention mechanism. It constitutes the library of addressable locations. The calculation of attention scores (QK^T) depends on the geometric fidelity of K. Consequently, K is fragile; modifying, normalizing, or distorting K erodes the underlying metric structure, leading to catastrophic degradation.

The Probe (Q): Q is better understood as a **probe** issued into the metric space defined by K. It is inherently more resilient. A noisy or convolved query issued against a precise library still yields a valid (if diffuse) result, whereas a precise query issued against a distorted library yields nonsense. This explains why query-side normalizations (e.g., QK-Norm) are permissible while key-side manipulations are dangerous.

The Teleological Payload (V): The lifted space defined by V bears the heaviest burden. It is not merely "content"; it is the material that must satisfy the teleological demand of the loss. V must effectively assemble the pre-image of the terminal simplex. It serves three incompatible roles:

- **The Assembly Role:** It acts as the substrate for the iterative construction of the final output state.
- **The Relevance Role:** It must produce signals that remain interpretable by future layers.
- **The Transport Role:** It acts as a generator for the residual update.

The "alphabet soup" of representations arises because V is forced to incorporate all these demands into a single linear basis using the same additive composition operator.

5. DIFFUSION AND RECONDITIONING

Diffusion is not merely an "oversmoothing pathology"; it can be the generic stable mode of simplex-constrained transport. Mean-field analyses predict long-lived metastable multi-cluster phases followed by eventual global clustering (collapse) under broad conditions [1]. This matches the empirical feel that early layers maintain broad possibility-mass while later layers force sharper commitments.

Because the transport is additive and the manifold is ill-conditioned, sharp, high-frequency updates are unstable. They carry a high risk of catastrophic misalignment. By contrast, diffusive updates, which are smudges in the general direction of the gradient, are conservative. They preserve the "center of mass" of the probability distribution without committing to a precise, brittle coordinate.

5.1. MLPs as Reconditioning and Basis-Restoration Operators

The diffusive character of simplex-constrained mixing creates a structural problem: additive convex combinations tend to suppress high-frequency distinctions. The MLP sublayer is the architectural coun-

terweight.

Structurally, the MLP is not merely "more capacity." It is the primary mechanism by which the model reconditions the post-transport state. It reintroduces separations that convex-hull mixing destroys and rebuilds a basis that makes later relevance signals legible. Critically, because the MLP acts on the residual stream, it functions as a resteering update: it learns to interpret the diffused proposal ΔX^ℓ relative to X^ℓ, and to convert a conservative smudge into a usable, stable coordinate change for subsequent layers.

6. TEMPORAL TRANSPORT AND THE STABILITY TRADEOFF

Real transport occurs **across** T; depth ℓ is the axis of assembly and refinement. The Transformer learns a policy for moving information from past indices via iterative residual injection.

This dynamic creates a fundamental tradeoff between grammar invention and ablation robustness. Early layers are fragile because they are **inventing the transport alphabet**—building the coordinate system in which later relevance is expressed. Ablating them removes the ability to speak. Later layers in wide models are robust because they operate in a regime of **deferred semantic commitment**, transporting ambiguous possibility mass rather than hard decisions.

We thus observe a distinct phase transition:

- **Fragile/Rigid Phase (Early):** The model must establish a grammar. Errors here propagate catastrophically.
- **Robust/Diffuse Phase (Late):** The model transports superpositions. Ablation here merely forces the remaining layers to resolve the ambiguity differently.

This stability-learning tradeoff implies that a model can either learn a rigid, efficient grammar (fast convergence, high fragility) or maintain a diffuse, robust superposition (slow convergence, low fragility). The gradient navigates this tradeoff based on the available width and capacity.

Consequently, the "heads" in Multi-Head Attention are **parallel transport hypotheses**. Each head proposes a different kernel for what constitutes a survivable signal. Shared recombination via the output map W_O forces competition among these hypotheses, filtering for kernels that remain non-collapsing under contraction pressures [1].

7. ATTENTION SINKS AS BASIS ANCHORS

A critical, often overlooked mechanism is the "attention sink"—the tendency of models to dump high attention scores on specific tokens (often the start token) to seemingly "opt out" of transport. Epistemically, this is not merely a waste bin; it is the provision of a **static basis vector**.

Since the attention weights must sum to 1 ($\sum A = 1$), every update is a convex combination. If a head wishes to perform a "no-op," it cannot output zero; it must output a vector. By assigning mass to a sink token, the head effectively adds a learned, constant bias vector V_{sink} to the update:

$$Y = (1 - \lambda)V_{context} + \lambda V_{sink}$$

This V_{sink} provides a fixed origin or anchor point in the vector space. It stabilizes the additive composition by providing a static basis against which dynamic context updates are measured. Without sinks, the V projection would be forced to construct a stable basis purely from dynamic context, which is combinatorially difficult. Sinks allow the model to learn an affine transformation geometry within a strictly linear mixing operator.

8. OPERATOR GEOMETRY

The "lens" view becomes unusually literal in a geometric framework where query–key interactions induce an effective metric and attention acts as a discrete connection implementing parallel transport across the sequence [2]. This pins down "coordinate transport" precisely: attention is the learned discrete connection; stacked blocks are discrete time-slices through which representations evolve under a geometry shaped by training.

8.1. Linear Layers as Spectral Filters

A linear layer is fundamentally a learned filter bank. In structured cases (convolution, circulant, Toeplitz), spectral multiplication becomes literal; in dense layers, the "filter bank" analogy is heuristic. Nevertheless, these layers reshape the spectral density of the signal, focusing energy from the input distribution into specific subspaces.

8.2. Norms as Proscriptive Manifold Constraints

LayerNorm and RMSNorm are typically justified via numerical stability. Epistemically, they function as proscriptive geometric constraints. By normalizing the vector, these layers annihilate an entire degree of freedom (the radial direction). The manifold picture clarifies why norm constraints feel like proscriptive geometry: normalization does not merely prevent blow-ups; it forces updates to live in tangent spaces of an induced constraint surface [3].

8.3. Residuals as Interference Mechanisms

The residual connection is an interference mechanism establishing a **phase relationship** between the current state and the update. Destructive interference corresponds to updates that are orthogonal or

cancelling. A concrete empirical prediction is that trajectories exhibit meaning-consistent bending under controlled context edits [2].

8.4. Caustics

We use the term **caustics** to describe regions where semantic ambiguity must collapse into a decision. The final logits are the ultimate caustic. A robust model avoids early collapse to maintain the transportability of the signal, keeping the representation in a superposition of potential meanings until the geometry of the final layers forces a resolution.

9. EMPIRICAL VERIFICATION

To validate the Dimensional Lift Hypothesis, we moved beyond standard benchmarks—which conflate capacity with mechanism—and instead constructed "mechanical isolation" experiments. These experiments were designed to explicitly falsify competing models of attention (e.g., Barycentric Mixing or Associative Retrieval) in favor of the Affine Coordinate Transport model.

9.1. The Probe-Metric Asymmetry

Theory: Standard analysis treats Queries (Q) and Keys (K) as symmetric peers in a similarity search. Our operator geometry suggests a fundamental asymmetry: K defines the metric space (the manifold curvature), while Q is merely a probe (a geodesic). Consequently, distortions to the metric (K) should be catastrophically more damaging than distortions to the probe (Q).

Experiment: We injected Gaussian noise $\mathcal{N}(0, \sigma^2)$ independently into the Q and K projections of a pretrained GPT-2 model and measured the KL Divergence (D_{KL}) of the output distribution relative to the unperturbed baseline.

Results: As predicted, the architecture exhibits a distinct hyper-sensitivity to Kernel distortion. At low noise regimes ($\sigma = 0.01$), the metric distortion (K) causes nearly double the divergence of probe distortion (Q).

Table 1: The Metric Fragility Profile

Noise Scale (σ)	$D_{KL}(Q)$	$D_{KL}(K)$	Ratio (K/Q)
0.01	0.0007	0.0013	1.86x
0.05	0.0342	0.0296	0.86x
0.10	0.1237	0.1581	1.28x

Formal Implication: The high fragility of K confirms it functions as the constitutive geometry of the layer. The model does not "compare" Q and K; it projects Q into the geometry defined by K.

9.2. Attention Sinks as Affine Bias Anchors

Theory: Linear attention ($\Delta X = \sum AV$) is constrained to the convex hull of the inputs. To perform general computation, the model requires an affine transformation ($\Delta X = WX + b$). We hypothesized that "Attention Sinks" (heads attending to the BOS token) are not discarding information, but are mechanically implementing the bias term b by retrieving a static basis vector V_{sink}.

Experiment: We isolated the Value vector (V_{sink}) retrieved by the primary sink head (Layer 10) across semantically disjoint contexts (Context A vs. Context B) and computed their Cosine Similarity.

Results:

$$\text{Sim}(V_{\text{sink}}^{(A)}, V_{\text{sink}}^{(B)}) = \mathbf{0.994}$$

Formal Implication: The vector retrieved from the sink is effectively constant, independent of the input text. Mathematically, this proves the effective update equation is affine:

$$X_{t+1} = X_t + \underbrace{(1 - \lambda_t)\text{Mix}(X)}_{\text{Linear}} + \underbrace{\lambda_t \mathbf{b}_{\text{sink}}}_{\text{Bias}}$$

where λ_t is the attention weight on the sink. The model learns to "valve" the bias term via the attention mechanism.

9.3. Spectral Signatures of Transport

Theory: If Attention were primarily "Barycentric Mixing" (averaging features), deep networks would suffer from rank collapse (smoothing). If it were "Coordinate Transport" (moving subspaces), it must be isometric (preserving volume). We defined the Spectral Retention Ratio (ρ) as the ratio of the nuclear norm of the output to the input: $\rho = ||\text{Op}(X)||_* / ||X||_*$.

Experiment: We subjected synthetic high-dimensional signals to iterative application of three idealized operators: Mixing (Softmax), Retrieval (Sparse), and Transport (Permutation/Rotation).

Table 2: Operator Spectral Diagnostics

Operator	ρ	κ (Cond. #)	Interpretation
Mixing	0.0358	$10,758$	Collapse: Destroys geomet
Retrieval	0.6261	∞	Unstable: Rank deficient.
Transport	1.0000	1.00	Isometry: Preserves signa

Formal Implication: The deep Transformer exists only because it approximates the Coordinate Transport regime. Any drift toward pure mixing results in an ill-conditioned Jacobian ($\kappa \gg 1$) that forbids gradient propagation. The residual connection $X + \text{Attn}(X)$ is a mechanism to force the mixing operator closer to the identity/transport regime.

9.4. The Breathing Manifold

Theory: The "Dimensional Lift" hypothesis posits a dual process: Attention diffuses (expands entropy to

transport) and MLPs recondition (compress to manifold).

Experiment: We tracked the Intrinsic Dimensionality (ID) of the representation across layers using the Two-NN estimator.

Results: We observed a rigorous "Sawtooth" pattern :

- **Expansion Phase (Attention):** Every attention layer increased the ID (Spike).
- **Compression Phase (MLP):** Every MLP layer decreased the ID (Drop).

Formal Implication: The network does not maintain a constant manifold. It "breathes":

$$\dim(\text{Attn}(X)) > \dim(X) > \dim(\text{MLP}(\text{Attn}(X)))$$

This confirms that the MLP acts as a projection operator $P_{\mathcal{M}}$ restoring the signal to the semantic manifold $\mathcal{M}$ after the entropic operation of transport.

10. CONCLUSION

The behavior of the Transformer is not an architectural choice but a derived consequence. The loss defines the terminal constraint. The gradient, flowing backward, discovers the optimal transport policy within the constraints we have imposed.

Design must be reasoned backward from the loss: modules do not "own" roles; the gradient assigns roles to wherever they are cheapest to realize under constraints. We do not program the model's reasoning. We design the stage: the dimensional lift of the embeddings, the distinct roles of the $Q/K/V$ operators, and the proscriptive constraints of normalization.

Ultimately, the Transformer is a machine for managing uncertainty under the constraint of additive transport. It works not because it is a perfect reasoner, but because the dimensional lift allows it to separate competing demands, and the diffusive transport allows it to smear meaning across time and depth until the loss function forces it to focus.

References

[1] Philippe Rigollet. *The Mean-Field Dynamics of Transformers*. arXiv:2512.01868 (2025).

[2] Riccardo Di Sipio, Jairo Diaz-Rodriguez, Luis Serrano. *The Curved Spacetime of Transformer Architectures*. arXiv:2511.03060 (2025).

[3] Akhil Kedia et al. *Transformers Get Stable: An End-to-End Signal Propagation Theory for Language Models*. arXiv:2403.09635 (2024).