AI, Abstract Algebra,and Differential Geometry: An Exploration into the physics of The Information Bottleneck and Whiteboxing AI

Samuel Leizerman samleizerman@outlook.com ORCID: 0009-0000-0133-2291

January 14, 2026

Theorem: The Non-Convergence of Scale on Causal Truth

Premise: Let current AI training be defined as estimating the conditional probability distribution $P(Y|X)$ from a dataset $\mathcal{D}$ of size N, where $N \to \infty$.

Claim: The estimator $\hat{P}(Y|X)$ does not converge to the true causal mechanism $P(Y|do(X))$ in the presence of unobserved confounding, regardless of N.

1. The "Sultan's Fallacy" (The Naive DAG)

The industry assumes the Data Generating Process (DGP) is a direct mapping:

$$X \to Y$$

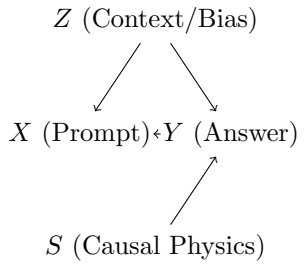
Under this assumption, they assert:

$$\lim_{N \to \infty} \hat{P}(Y|X) = P(Y|do(X))$$

This holds only if the system is unconfounded.

2. The Real-World DAG (The Confounder Trap)

In reality, the internet text corpus contains a latent confounder Z (e.g., social context, "common misconceptions", "Reddit bias"):



The Backdoor Path: Information flows from X to Y via two paths:

1. Causal: $X \xrightarrow{?} Y$ (Often non-existent without structure S)
2. Spurious: $X \leftarrow Z \to Y$

3. The Bias of Scale

The standard Large Language Model (LLM) objective minimizes the Kullback-Leibler divergence between the data distribution and the model:

$$\mathcal{L} = \mathbb{E}_{x,y\sim\mathcal{D}}[-\log P_\theta(Y|X)]$$

Since the dataset $\mathcal{D}$ is generated by the DAG with Z, the model learns the *observational* distribution:

$$P_{\text{model}}(Y|X) = \sum_z P(Y|X,z)P(z|X)$$

However, the *causal* truth (the "White Box" goal) requires the **Backdoor Adjustment**:

$$P(Y|do(X)) = \sum_z P(Y|X,z)P(z)$$

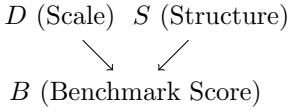
The Discrepancy (Bias):

$$\text{Bias} = |P(Y|X) - P(Y|do(X))| = \left|\sum_z P(Y|X,z)[P(z|X) - P(z)]\right|$$

Conclusion: As $N \to \infty$, the variance of $P(Y|X)$ approaches 0, but the Bias remains constant. The model becomes **perfectly confident in the wrong answer** because it mistakes the confounding context (Z) for the cause.

4. The Collider Bias (The Benchmark Trap)

Why does the industry prefer Scale (D) over Structure (S)? Because Benchmarks (B) act as a Collider Node.



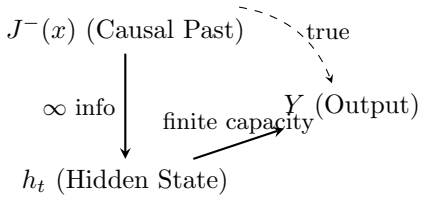
We select models conditional on high benchmark performance ($B = 1$). By *Berkson's Paradox*, conditioning on a collider induces a negative correlation between independent parents:

$$P(D|S, B = 1) \neq P(D)$$

Since Scale (D) is easier to acquire (Buy GPUs) than Structure (S) (Derive Physics), the industry optimizes for D and ignores S. This is statistically incentivized laziness.

5. The Capacity Divergence (The Light Cone Trap)

Beyond confounding, there is a deeper structural failure: the **information-capacity divergence**.



The Bottleneck Equation:

$$\boxed{\lim_{t\to\infty} \frac{\text{Information in } J^-(x)}{\text{Capacity of } h_t} = \infty}$$

Interpretation:

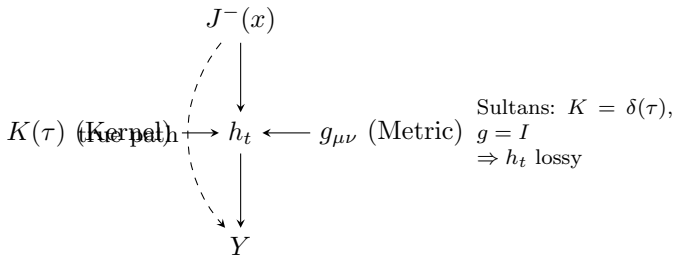
- $J^-(x)$ is the *causal past light cone*—all events that could causally influence x. This grows unboundedly with t.
- h_t is the model's hidden state (or context window)—a finite-dimensional vector.
- The ratio diverges: no finite Markovian state can capture unbounded causal history.

The Sultan's Response: Increase $\dim(h_t)$—more parameters, longer context.
Why This Fails: The problem is not *parametric* but *structural*. The causal past has:

- **Geometric structure:** The light cone $J^-(x)$ is not a bag of tokens; it has metric structure.
- **Delayed propagation:** Effects propagate via $G_{\text{del}}(x, x')$, not instantaneously.
- **Memory morphology:** The kernel $w_\mu(\tau)$ determines how history integrates.

A larger h_t with Markovian dynamics ($K = \delta(\tau)$) is still integrating against a delta function. More capacity, same category.

The DAG with Capacity Bottleneck:



The Sultans' Implicit Assumption:

$$K(\tau) = \delta(\tau) \quad \text{(Markovian—no history integration)} \tag{1}$$
$$g_{\mu\nu} = \delta_{\mu\nu} \quad \text{(Flat—no geometric structure)} \tag{2}$$

Under these assumptions, the hidden state h_t is a *lossy compression* of $J^-(x)$:

$$h_t = f\left(\int_{J^-(x)} \delta(t - t') \cdot \text{tokens}(t')\, dt'\right) = f(\text{tokens}(t))$$

Only the present survives. The light cone collapses to a point.

6. The Kernel Incompatibility (The Category Error)

The capacity divergence is a symptom of a deeper issue: **kernel incompatibility**.

Definition: A model is *Markovian* if its effective memory kernel satisfies:

$$K(\tau; x, x') = \delta(\tau) \cdot f(x, x')$$

for some function f. The "context window" is finite conditioning, not temporal integration.

The Kernel Factorization: The general causal kernel decomposes as:

$$K_\mu(x, x') \equiv w_\mu(t - t')\, G_{\text{del}}(x, x'), \quad \int_0^\infty w_\mu(\tau)\, d\tau = 1$$

with limiting cases:

$$\text{QM (Markovian):} \quad w_\mu(\tau) \to \delta(\tau), \quad \alpha \to 1 \tag{3}$$
$$\text{GR (Causal-Propagator):} \quad w_\mu(\tau) \text{ broad on support}, \quad \alpha \to 0 \tag{4}$$

The Incompatibility Theorem:

Property	Markovian (LLMs)	Causal-Propagator
Memory kernel	$\delta(\tau)$	$G_{\text{del}}(x, x')$
Output structure	Operators on $\mathcal{H}$	Geometric tensors via $\mathcal{E}$
Key operation	Exponentiation $\to$ unitary	Differentiation $\to$ curvature
Gauge structure	Global phase	Diffeomorphism invariance

These are not "different parameter values"—they live in **different mathematical categories**. No Markovian kernel can recover the delayed Green's function structure:

$$\bar{h}^{\mu\nu}(x) = \frac{16\pi G}{c^4} \int_{J^-(x)} G_{\text{del}}(x, x')\, \tau^{\mu\nu}(x')\, d^4x'$$

Conclusion: The Sultans face three distinct failures:

1. **Statistical:** Uncontrolled confounding $\Rightarrow$ biased estimator (Section 3)
2. **Capacity:** $J^-(x)/h_t \to \infty \Rightarrow$ unbounded information loss (Section 5)
3. **Structural:** Markovian kernel $\Rightarrow$ causal-propagator regime inaccessible (Section 6)

Scale ($N \to \infty$) cannot fix a category error. You cannot approximate a delayed Green's function with delta functions, no matter how many you sum.

7. The White-Box Correction

The proposed architecture addresses all three failures:

7.1 Backdoor Adjustment via Repulsors

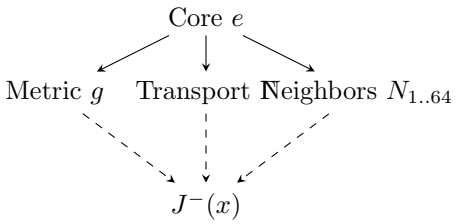
The address structure includes N_{49}-N_{64}: *repulsor neighbors* representing contrastive evidence. These implement:

$$P(Y|do(X)) \approx P(Y|X, \text{Attractors}) - \lambda \cdot P(Y|X, \text{Repulsors})$$

The repulsors explicitly block known backdoor paths through Z_{toxic}.

7.2 Structured Addressing for $J^-(x)$

Instead of compressing $J^-(x)$ into a finite h_t, the address carries *navigation structure*:



Address stores *how to navigate* $J^-(x)$, not $J^-(x)$ itself

The address does not store the light cone; it stores **coordinates, metric, and routing**—sufficient to traverse $J^-(x)$ on demand.

7.3 Kernel Interpolation via LIoR

The system determines the appropriate memory kernel by minimizing the **Lior (z"l) Integration of Ricci (LIoR)** cost. This functional quantifies the "topological work" required to maintain a valid causal trajectory $\gamma(\tau)$ against the curvature induced by confounders (bias).

The accumulated tidal force (stress) on the representation vector is given by the path integral of the Riemann curvature tensor contracted with the causal velocity $\dot{\gamma}$:

$$\text{LIoR}_{\mu\nu}[\gamma] = \int_0^T R_{\mu\nu\rho\sigma}(\gamma(\tau))\dot{\gamma}^\rho(\tau)\dot{\gamma}^\sigma(\tau)d\tau \tag{5}$$

To enforce the "Resilience Budget," we define the scalar curvature intensity $R(x)$ at any point in the latent manifold. This acts as the penalty term for lazily following spurious correlations (geodesics of confounders):

$$R(x) = \sqrt{\frac{1}{n^2} g^{\mu\rho} g^{\nu\sigma} R_{\mu\nu\rho\sigma}(x)} \tag{6}$$

If $R(x)$ exceeds the resilience threshold, the system is forced to deviate from the Markovian path ($K = \delta(\tau)$) and engage the Causal-Propagator kernel ($K = G_{del}$).

7.3.1 The Variational Principle (Least Causal Action)

The LIoR functional attains physical meaning only when promoted from a diagnostic quantity to a **variational principle**. Rather than evaluating curvature along an arbitrary representational trajectory, the system must select the trajectory that minimizes accumulated geometric stress subject to causal admissibility.

Let $\Gamma(x \to Y)$ denote the set of all admissible causal trajectories connecting input x to output Y, consistent with locality, finite propagation speed, and the address constraints of the representation manifold. We define the **least-action LIoR functional** as:

$$\boxed{\text{LIoR}^*(x \to Y) \;=\; \inf_{\gamma \in \Gamma(x \to Y)} \int_0^T R_{\mu\nu\rho\sigma}(\gamma(\tau))\, \dot{\gamma}^\rho(\tau)\dot{\gamma}^\sigma(\tau)\, d\tau} \tag{7}$$

This variational form induces a principle of **least causal action**: among all paths consistent with the causal graph, the realized trajectory is the one that minimizes the integrated curvature stress imposed by confounders, bias fields, and geometric obstruction.

Interpretation.

- Spurious correlation paths (e.g. $X \leftarrow Z \to Y$) may be locally short but accumulate high curvature globally, rendering them suboptimal.
- Structural paths (e.g. $S \to Y$) may be longer in representation space but lower in integrated curvature, and thus preferred by the infimum.
- Kernel selection emerges naturally: if no Markovian trajectory ($K = \delta(\tau)$) attains the infimum, the system is forced into the causal-propagator regime ($K = G_{\text{del}}$).

With the infimum enforced, LIoR no longer penalizes deviation *after the fact*; it determines which causal trajectory is realizable at all. In this sense, causal generalization is not trained—it is **variationally selected**.

7.4 Geometric Structure via Metric Learning

The address carries local geometry (metric g, transport Γ), enabling:

$$\mathcal{L}_{\text{geo}} = \int \sqrt{g_{\mu\nu}\dot{\gamma}^\mu\dot{\gamma}^\nu}\, d\tau$$

This forces the model to use the path $S \to Y$ (structure causes answer) rather than $X \leftarrow Z \to Y$ (context correlates both).

8. The Verdict

Scale Path: $N \to \infty$ with $K = \delta(\tau)$, $g = I$

$\Rightarrow$ Precise Markovian inference in the *wrong category*
$\Rightarrow$ Zero variance, constant bias
$\Rightarrow$ $J^-(x)$ compressed to a point
$\Rightarrow$ Perfectly confident, systematically wrong

Structure Path: $K = K_\mu$, $g_{\mu\nu}$, Attractors/Repulsors

$\Rightarrow$ Causal-propagator regime *accessible*
$\Rightarrow$ Backdoor paths *blockable*
$\Rightarrow$ $J^-(x)$ *navigable* via address structure
$\Rightarrow$ Bias reducible via geometric constraint

Final Statement: "Laziness" is not an insult; it is a *topological fact*. The Sultans are taking the path of least resistance on the causal graph—the spurious correlation path through Z—which is a dead end regardless of how fast they traverse it.

Scale reduces variance. Structure removes bias. They are optimizing the wrong term.