The δ-Gene: Inference-Time Physical Unclonable Functions from Architecture-Invariant Output Geometry

Anthony Ray Coslett
Fall Risk Research
anthony@fallrisk.ai

Abstract

As neural language models are deployed in regulated domains, verifiable model provenance becomes a critical security requirement. We construct an Inference-Time Physical Unclonable Function (IT-PUF) that provides a challenge-response authentication protocol for neural networks, achieving zero false acceptances across 1,012 comparisons spanning 23 models and 16 vendor families.

The IT-PUF derives its entropy from a geometrically intrinsic behavioral fingerprint—the δ-gene (the third pre-softmax logit gap)—which we prove is invariant to inference temperature and empirically validate as invariant across six distinct neural architectures. We provide a formal impossibility result for fingerprint spoofing: an interval-splitting theorem proves that no adversarial Kullback-Leibler (KL) budget can simultaneously close the fingerprint gap and avoid detection via accumulated noise.

To establish that this security does not degrade at scale, we validate an Equation of State across three independent model families spanning a 147× parameter range (0.5B to 72B). We falsify the assumption of unbounded stiffness but discover a strict positive empirical floor ($S_{\min} = 1.1797$), from which the Cramér-Rao bound guarantees a computable minimum spoofing cost. The theoretical foundation is formally verified in the Coq proof assistant: 311 theorems across 16 files, with zero uses of `Admitted` and zero vacuous definitions.

1 Introduction

As neural language models are deployed in regulated domains — healthcare, finance, legal counsel, autonomous systems — the ability to verify *which model* produced a given output becomes a compliance requirement, not merely a convenience. The EU AI Act mandates traceability for high-risk AI systems. FDA 21 CFR Part 11 requires audit trails for AI-assisted decisions in clinical settings. Supply chain integrity demands that the model running in production is the model that was approved for deployment. These requirements share a common prerequisite: a reliable method for identifying a deployed model from its observable behavior.

Existing approaches fall short on at least one critical axis. Watermarking schemes [Kirchenbauer et al., 2023, Zhao et al., 2023] inject detectable signals during training or inference, but require operator cooperation, are vulnerable to removal through fine-tuning or quantization, and do not apply retroactively to models already deployed. Behavioral fingerprinting based on output entropy, perplexity, or stylistic markers is sensitive to inference-time temperature settings — a parameter routinely adjusted across deployments that changes the output distribution without modifying the model itself. A fingerprint that shifts when the thermostat moves is not a fingerprint.

This work presents a model identification primitive that is invariant to both temperature and internal architecture. The observable is simple: the gap between the third and fourth highest pre-softmax logits,

$$\boxed{\delta_{\text{raw}} := z_{(3)} - z_{(4)}} \tag{1}$$

measured over a corpus of input prompts. Because δ is defined on raw logits — before temperature scaling and before the softmax transformation — it is structurally insensitive to the inference temperature parameter. Empirically, the mean δ varies by 1–5% (coefficient of variation) across a 5× temperature range ($T = 0.3$ to 1.5).

The deeper invariance is architectural. We validate δ stability across six fundamentally different neural network architectures: dense Transformers, mixture-of-experts, RWKV, pure Mamba (state space models), Mamba-Transformer hybrids, and Mamba-Transformer-MoE hybrids. These architectures process information through fundamentally different internal mechanisms. They share exactly one structural element: the output pathway $\mathbf{h}_t \to W_U \to \text{softmax} \to x_{t+1}$. The fingerprint lives in the output geometry, not in how information is processed internally.

This paper makes five contributions:

1. **The δ-gene observable** (§3). We identify the third logit gap G_3 as a temperature-invariant, architecture-agnostic behavioral fingerprint, validate its stability across six architecture families, and characterize its statistical properties under the Poisson Point Process model of extreme order statistics.

2. **A theoretical framework explaining δ from first principles** (§4). Extreme value theory predicts the normalized gap $\delta_{\text{norm}} \approx 0.318$ as a universal constant of the Gumbel class — confirmed experimentally at 0.297–0.321 across nine models and two architecture types. An Equation of State connects δ to four weight-readable architectural primitives (stiffness, crowding, vocabulary pressure, and softcap constraint), validated at leave-one-out $R^2 = 0.40$ across 38 models. The Gauge–Transport Decomposition theorem provides a unified mathematical object from which the forensic findings follow as corollaries.

3. **A three-axis forensic hierarchy** (§5). Three orthogonal axes of model identity — output-layer thermodynamics, gauge scarring from training infrastructure, and representation topology from gradient dynamics — together with a two-level provenance detection system that discriminates architecture family, fine-tuning algorithm (SFT vs DPO), and training-infrastructure anomalies.

4. **A formal impossibility result for fingerprint spoofing** (§6). The Cramér-Rao bound creates a per-component cost floor for shifting any fingerprint observable. An interval-splitting theorem (machine-verified in Coq) proves that at every KL budget, either the residual fingerprint gap or the accumulated perturbation noise exceeds the acceptance threshold. Empirically, the strongest adaptive attack fails to close beyond 10.7× the threshold before model destruction.

5. **A challenge-response authentication protocol** (§7). An Inference-Time Physical Unclonable Function (IT-PUF) that adapts the hardware PUF paradigm to neural network inference. Validated on 23 models spanning 16 families with 0/1,012 false acceptances and a conservative min-entropy bound of 25.86 bits.

The mathematical foundation is formally verified in the Coq proof assistant: 311 theorems across 16 files, with zero uses of `Admitted` (Coq's mechanism for accepting unproven statements) and zero vacuous definitions.[1] Every claim in this work is classified as **[PROVEN]** (machine-checked), **[CITED]** (published mathematics), **[DERIVED]** (computed from proven or cited foundations), or **[VALIDATED]** (empirical law with documented evidence). The epistemological classification is maintained throughout and the distinction between these categories is not negotiable (§8).

2 Background and Related Work

This section establishes the mathematical prerequisites for the paper and positions our contributions within the model fingerprinting landscape. We identify three technical foundations (extreme value theory, physical unclonable functions, and Fisher information geometry) and differentiate from concurrent work in model identification.

2.1 Model Fingerprinting Landscape

Recent surveys [Shao et al., 2025] organize model fingerprinting methods along two axes: *injected* versus *intrinsic*, and *white-box* versus *black-box*. Injected methods — typically watermarking schemes that embed

[1] The count of 311 is mechanically reproducible: it is the number of top-level `Theorem`, `Lemma`, `Corollary`, and `Proposition` declarations across the 16 files cited in §8.

Table 1: Comparison with concurrent intrinsic fingerprinting methods. Dashes indicate the capability is absent.

	ZeroPrint	Intrinsic FP	REEF	This work
Theoretical mechanism	—	—	—	Gauge–Transport (§4.3)
Temperature invariant	No	Not tested	Not tested	CV 1–5% (§3)
Formal impossibility proof	—	—	—	51 Coq theorems (§6.3)
Arch.-agnostic ($\geq$7 families)	No	No	No	16 families, 3 types (§7)
Identification performance	AUC 0.72	$r > 0.8$	Robust	0/1,012 FAR (§7)

verifiable signals into model weights or outputs during training — require operator cooperation and are vulnerable to removal through fine-tuning, pruning, or quantization. Intrinsic methods extract identity from properties the model already possesses, requiring no modification to the training process.

Within intrinsic methods, three recent lines of work are most relevant.

Representation-based fingerprinting. REEF [Zhang et al., 2025] computes Centered Kernel Alignment (CKA) on intermediate representations to identify model lineage. The method is robust to fine-tuning, pruning, and model merging. However, it operates on internal activations (requiring white-box access to intermediate layers), has no formal characterization of what makes CKA stable, and provides no adversarial robustness guarantees.

Weight-statistical fingerprinting. Yoon et al. [Yoon et al., 2025] independently discover that standard deviation distributions of attention parameter matrices across layers are stable after continued training, achieving lineage correlations above 0.8. This is an empirical confirmation that weight statistics encode identity — but without a theoretical mechanism explaining *why* these statistics persist. The Gauge–Transport Decomposition (§4.3) provides exactly this mechanism: the softmax-invariant gauge subspace is function-null, so gradient-based training under cross-entropy — or any loss that acts through softmax outputs — does not modify gauge-aligned weight statistics.

Jacobian-based fingerprinting. ZeroPrint [Shao et al., 2026] uses Fisher information theory to argue that input-output Jacobians encode more parameter identifiability than plain outputs, then estimates these Jacobians via semantic-preserving word substitutions. The method is black-box and achieves AUC $\approx$ 0.72. Our work uses Fisher information in a complementary direction: not to select observables, but to establish formal lower bounds on spoofing cost (§6). ZeroPrint's usage is qualitative — Fisher information justifies that Jacobians are "more informative" than outputs. Ours is quantitative — the Cramér-Rao inequality yields a per-component KL cost floor that scales with the squared fingerprint gap (§6.2). The approaches are compatible; our impossibility result (§6.3) applies regardless of which observable is protected.

The gap. No existing method provides temperature invariance, architecture-agnosticism across fundamentally different architecture families (Transformer, Mamba, RWKV, MoE, and hybrids), formal verification of stability claims, or impossibility results for adversarial spoofing. We term the approach introduced here *geometry-intrinsic fingerprinting* — methods that derive identity from the mathematical structure of the output mapping rather than from learned representations, behavioral responses, or injected markers.

Table 1 summarizes the differentiation.

2.2 Extreme Value Theory

The theoretical framework in §4 draws on classical extreme value theory (EVT). For n i.i.d. draws from a distribution F with right tail in the Gumbel maximum domain of attraction, the order statistics near the maximum are well-approximated by a Poisson Point Process (PPP) with exponential spacings [Resnick, 1987, Leadbetter et al., 1983]. Specifically, the gaps between consecutive order statistics

$$G_k := z_{(k)} - z_{(k+1)} \tag{2}$$

satisfy $k \cdot G_k \xrightarrow{d} \mathrm{Exp}(\beta)$ for a tail scale parameter β determined by the parent distribution's hazard rate at the $(1 - 1/n)$-quantile [de Haan and Ferreira, 2006]. This structure predicts that normalized gap ratios are distribution-free within the Gumbel class — a prediction we verify empirically in §3 and §4.

The connection to language models is direct: at each token position, the softmax layer selects from a vocabulary of V items whose pre-softmax logits $z_1, \ldots, z_V$ are produced by the output projection. For large V (typically 32,000–128,000), the top-k logits behave as extreme order statistics of this V-dimensional vector, and their gap structure inherits the universal properties of the Gumbel PPP.

2.3 Physical Unclonable Functions

A Physical Unclonable Function (PUF) is a hardware primitive that exploits uncontrollable manufacturing variation to create device-specific challenge-response behavior [Pappu et al., 2002]. When presented with a challenge input c, a PUF produces a response $r = f(c)$ determined by the device's physical microstructure — process variations in gate delays, wire capacitances, or optical scattering paths that are unique to each manufactured instance and infeasible to replicate [Suh and Devadas, 2007].

The security model rests on three properties: (1) *uniqueness* — different devices produce different responses to the same challenge, (2) *reproducibility* — the same device produces consistent responses across repeated measurements (within noise tolerance), and (3) *unclonability* — manufacturing the physical microstructure is harder than the entropy it provides.

The PUF paradigm has not previously been applied to neural network inference. In §7, we introduce the *Inference-Time PUF* (IT-PUF), which substitutes training-induced weight geometry for manufacturing variation: different trained models, like different manufactured chips, produce distinct responses to the same input challenges due to the unique microstructure of their learned parameters. The formal impossibility result (§6.3) provides the unclonability guarantee — replicating a model's fingerprint response requires either access to the original weights or a parameter perturbation whose KL cost exceeds the acceptance threshold.

2.4 Fisher Information and the Cramér-Rao Bound

The Fisher information matrix $I(\theta)$ of a parametric model p_θ measures the curvature of the log-likelihood surface at θ:

$$[I(\theta)]_{ij} = \mathbb{E}_{x \sim p_\theta}\left[\frac{\partial \log p_\theta(x)}{\partial \theta_i} \cdot \frac{\partial \log p_\theta(x)}{\partial \theta_j}\right] \tag{3}$$

The Cramér-Rao inequality establishes that no unbiased estimator of a function $g(\theta)$ can achieve variance below $(\nabla g)^T I(\theta)^{-1}(\nabla g)$ [Cramér, 1946, Rao, 1945]. Equivalently, shifting an observable $g(\theta)$ by amount Δg while maintaining distributional proximity (bounded KL divergence) incurs a minimum cost:

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta'}) \geq \frac{(\Delta g)^2}{2\, s_F^2} \tag{4}$$

where $s_F^2 = (\nabla g)^T I(\theta)^{-1}(\nabla g)$ is the Fisher sensitivity of observable g. This bound is information-theoretic — it holds regardless of the optimization algorithm used to find θ'.

In this work, Fisher information serves a fundamentally different role than in prior fingerprinting literature. ZeroPrint [Shao et al., 2026] uses Fisher information to *motivate observable selection* — arguing that Jacobian-based features are more informative than output-based ones. We use Fisher information to *prove spoofing impossibility* — deriving a per-component cost floor that no adversary can circumvent (§6.2), then proving via interval splitting that no KL budget avoids both the residual gap floor and the noise floor simultaneously (§6.3). The distinction is between Fisher information as a lens for measurement design and Fisher information as a wall against adversarial manipulation.

3 The δ Observable

3.1 Definition and Motivation

Given a neural language model with vocabulary size V, let $\mathbf{z} \in \mathbb{R}^V$ denote the pre-softmax logit vector at a single generation step. We write $z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(V)}$ for the order statistics (descending) and define the logit gap sequence

$$G_k := z_{(k)} - z_{(k+1)}, \qquad k = 1, 2, \ldots \tag{5}$$

Each G_k measures the margin between the k-th and $(k+1)$-th ranked candidates in the model's output distribution.

The winner gap $G_1 = z_{(1)} - z_{(2)}$ is a natural first candidate for a fingerprint observable. However, G_1 is subject to **selection bias**: the winning token is, by definition, the one whose logit exceeded all competitors, inducing a systematic upward distortion of G_1 relative to the underlying gap distribution. Under the PPP model (§3.3), the scaled gaps $k \cdot G_k$ should be identically distributed for all k. Two-sample Kolmogorov-Smirnov tests on the scaled gaps — comparing $k \cdot G_k$ against $(k+1) \cdot G_{k+1}$ — confirm that G_1 violates this prediction while deeper gaps do not: for both Llama 3.2 3B (Dense Transformer) and Falcon-Mamba 7B (Pure Mamba), the comparison $1 \cdot G_1$ vs $2 \cdot G_2$ rejects at $p < 0.001$, while $2 \cdot G_2$ vs $3 \cdot G_3$ and $3 \cdot G_3$ vs $4 \cdot G_4$ show no significant difference ($p > 0.17$ in all cases). The runner-up gap G_2 is the closest to selection-free, but G_3 provides additional separation from any residual winner-proximity effects. Higher-rank gaps G_k with $k \geq 5$ have decreasing expected magnitude (β/k), reducing measurement precision; G_3 balances selection-freedom against signal strength.

We define the **δ-gene** as the third logit gap:

$$\delta_{\text{raw}} := G_3 = z_{(3)} - z_{(4)} \tag{6}$$

and its scale-free normalization:

$$\delta_{\text{norm}} := \frac{G_3}{G_2 + G_3 + G_4} \tag{7}$$

Throughout this work, δ without qualification denotes δ_{raw}. We emphasize that δ is a property of the **logit vector**, not of the probability distribution — it is defined prior to the softmax transformation and prior to any temperature scaling.

3.2 Empirical Properties

Temperature invariance. Since temperature scaling transforms logits as $\mathbf{z} \mapsto \mathbf{z}/T$, the probability distribution $p_i = \text{softmax}(\mathbf{z}/T)_i$ depends on T, but the raw logit gaps G_k do not. Temperature determines *which token is sampled* (and thus which logit vector appears at the next step under autoregressive generation), introducing variation in δ across generation runs at different temperatures. The relevant stability claim is therefore about the *expectation* $\mathbb{E}[\delta \mid T]$, not individual measurements.

We evaluate this by measuring δ under on-policy generation across temperatures $T \in \{0.3, 0.5, 0.7, 1.0, 1.3, 1.5\}$, computing the coefficient of variation $\text{CV}_T := \text{SD}(\{\bar{\delta}_T\}_T)/\text{mean}(\{\bar{\delta}_T\}_T)$ of the per-temperature means.

Across all tested models, CV_T falls between 1% and 5%. Table 2 reports representative values spanning multiple architecture families.

The stability is not approximate — it is structurally expected. Different temperatures induce different token sequences, which visit different regions of the model's output space. That the *average* gap remains stable reflects the fact that δ is governed by the geometry of the output mapping (the unembedding matrix W_U and the vocabulary structure it induces), not by which particular token sequence is generated.

Table 2: Temperature stability of δ across six fundamentally different neural network architectures. CV_T computed over six temperature settings spanning a 5× range ($T = 0.3$ to 1.5). All six architectures shown exhibit CV below 4%; the full range across all tested models is 1–5% (§3.2).

Model	Architecture	Parameters	CV_T
Llama 3.2 3B	Dense Transformer	3.2B	1.75%
Mixtral 8x7B	Mixture-of-Experts	46.7B	2.7%
RWKV-7	RWKV (linear attention)	13.3B	3.0%
Falcon-Mamba 7B	Pure Mamba (SSM)	7.0B	3.0%
Jamba	Mamba-Transformer Hybrid	12B	3.6%
Nemotron 30B	Mamba-Transformer-MoE	30B	3.5%

Architecture invariance. Table 2 simultaneously demonstrates a stronger result: δ stability holds across fundamentally different internal architectures. The six models listed employ five distinct computational mechanisms for processing sequential information — multi-head self-attention, mixture-of-experts routing, linear recurrence (RWKV), selective state spaces (Mamba), and hybrid combinations thereof. These architectures differ in how information flows from input to the final hidden state $\mathbf{h}_t$.

They share exactly one computational element: the output pathway

$$\mathbf{h}_t \xrightarrow{W_U} \mathbf{z}_t \xrightarrow{\text{softmax}} \mathbf{p}_t \xrightarrow{\text{sample}} x_{t+1} \tag{8}$$

where $W_U \in \mathbb{R}^{V \times d}$ is the unembedding (output projection) matrix. The δ-gene depends only on the rank-order statistics of $\mathbf{z}_t = W_U \mathbf{h}_t$, which are determined by the geometry of W_U and the distribution of hidden states $\mathbf{h}_t$ reaching it — not by the internal mechanism that produced $\mathbf{h}_t$.

This observation is the central empirical claim of the paper: **the behavioral fingerprint δ is an output-geometry invariant, not an architecture-specific property.**

3.3 Statistical Framework

The stability and universality of δ are not merely empirical observations — they follow from the statistical theory of extreme values applied to the logit vector.

Poisson point process model. Consider the logit vector $\mathbf{z}$ as a random draw from the model's output distribution. Under mild regularity conditions on the marginal distribution of logit entries (specifically, that the distribution belongs to the Gumbel maximum domain of attraction — satisfied by Gaussian, Laplace, and logistic families, among others), the upper order statistics of $\mathbf{z}$ converge to a Poisson point process (PPP) with intensity $\beta^{-1}e^{-x/\beta}$, where $\beta > 0$ is the **tail scale parameter** [Resnick, 1987, Leadbetter et al., 1983]. The classical PPP convergence result assumes independent entries; real logit vectors have correlated components because nearby vocabulary items share embedding geometry. The operative assumption is that the tail statistics — specifically, the gap structure among the top-k logits — are well-approximated by the PPP model despite these correlations. This assumption is validated empirically: Kolmogorov-Smirnov tests on gap distributions yield $p > 0.17$ for $k \geq 2$ across all tested models (§3.1), confirming that the PPP approximation captures the operative statistics of the logit tail even when the i.i.d. assumption is not literally satisfied.

Under this PPP, the gaps between consecutive order statistics are independent exponentials:

$$G_k \sim \mathrm{Exp}(k/\beta), \qquad k = 1, 2, \ldots \tag{9}$$

with $\mathbb{E}[G_k] = \beta/k$. We estimate the tail scale using the robust median estimator

$$\beta_{\text{robust}} := \mathrm{median}(k \cdot G_k)_{k \geq 2} \tag{10}$$

which excludes G_1 to avoid the selection bias documented in §3.1. The PPP scale parameter is $\beta_{\text{true}} = \beta_{\text{robust}} / \ln 2$ (median-to-mean conversion for the exponential distribution).

Table 3: Measured δ_{norm} versus the EVT prediction of 0.318, sorted by absolute deviation. All nine models fall within one standard deviation (0.232) of the theoretical value. The state-space model (Falcon-Mamba) shows the second-smallest deviation; the largest outlier is a Transformer (OPT), confirming that architecture type introduces no systematic bias. Measurements use 20 high-entropy prompts, ∼3,300 tokens per model, teacher-forced forward passes with row-centered logits.

Model	Architecture	δ_{norm} (measured)	Deviation from 0.318
TinyLlama 1.1B	Dense Transformer	0.318 ± 0.234	$+0.000$
Falcon-Mamba 7B	Pure Mamba (SSM)	0.317 ± 0.276	-0.001
Qwen 2.5 3B	Dense Transformer	0.320 ± 0.242	$+0.002$
Llama 3.2 1B	Dense Transformer	0.316 ± 0.235	-0.002
Gemma 2 2B	Dense Transformer	0.321 ± 0.242	$+0.003$
Qwen 2.5 0.5B	Dense Transformer	0.310 ± 0.238	-0.008
SmolLM2 1.7B	Dense Transformer	0.309 ± 0.235	-0.009
Mistral 7B	Dense Transformer	0.307 ± 0.233	-0.011
OPT 1.3B	Dense Transformer	0.297 ± 0.230	-0.021

Universal normalized gap. The ratio $\delta_{\text{norm}} = G_3/(G_2 + G_3 + G_4)$ is the share of the third gap in the sum of three consecutive PPP gaps. Since each $G_k \sim \text{Exp}(k/\beta)$, this ratio is scale-free (independent of β) and its expectation is computable by integrating over the joint distribution of three independent exponentials with rates $2/\beta$, $3/\beta$, and $4/\beta$: $\mathbb{E}[\delta_{\text{norm}}] = 0.318$ with $\text{SD}[\delta_{\text{norm}}] = 0.232$. These values are universal constants from extreme value theory — they depend on the rank indices (2, 3, 4) but not on the model, the vocabulary size, or the architecture.

Experimental verification. We measure δ_{norm} across nine models spanning two architecture types (a subset of the IT-PUF zoo, §7). The EVT prediction $\mathbb{E}[\delta_{\text{norm}}] = 0.318$ is confirmed with a grand mean of 0.313 and grand standard deviation of 0.007; the maximum absolute deviation from the prediction is 0.021.

The agreement across nine models and two architecture types confirms that the Gumbel PPP model captures the operative statistics of the logit tail. The sorted presentation highlights a key finding: the pure Mamba model (deviation -0.001) is statistically indistinguishable from the Transformer cluster, while the largest outlier is OPT 1.3B — the oldest model in the zoo, trained before current instruction-tuning practices. This is consistent with the interpretation that $\delta_{\text{norm}} \approx 0.318$ is a training attractor to which instruction-tuned models converge, regardless of architecture.

Tail scale and the stability phase. The measured tail scale across models is $\beta_{\text{robust}} \approx 1.44$–$1.60$ (from raw logits). This places all tested models in the **stable phase** $\beta > 1$, which has the following significance.

Define the tail partition factor

$$Z_{\text{tail}} := \sum_{n \geq 0} \exp\left(- \sum_{k=2}^{n+1} G_k \right) \tag{11}$$

Under the PPP, $\mathbb{E}[Z_{\text{tail}}]$ converges if and only if $\beta > 1$. When $\beta \leq 1$, the tail contains infinitely many effective competitors (a "crowded" regime); when $\beta > 1$, competition is finite and the tail statistics are well-behaved. The measured $\beta \approx 1.5$ sits safely in the convergent regime, providing a theoretical basis for the stability of all gap-derived observables including δ.

The value $\beta \approx 1.5$ is not a mathematical necessity but a **training attractor**: cross-entropy loss sharpens the winner gap until competitor probability mass drops below the gradient-relevant threshold. At a winner gap of approximately 1 nat, competitor mass is roughly 27% and gradients remain active; at 3 nats, competitor mass drops to approximately 5% and gradients effectively vanish. The equilibrium β reflects the balance between cross-entropy sharpening and implicit regularization (weight decay, dropout, learning rate schedules). This explains why β is similar across architectures trained with similar loss functions, despite their different internal mechanisms.

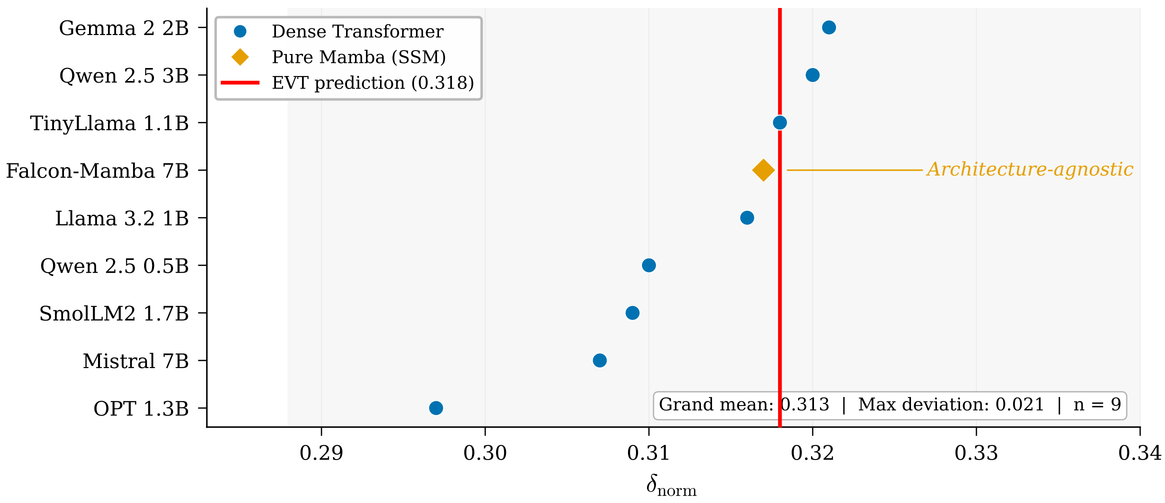

Figure 1: Measured δ_{norm} across nine models spanning two architecture types, versus the EVT prediction of 0.318 (red vertical line, shaded band = ±1 theoretical SD). All models fall within the predicted distribution; the pure Mamba model (diamond) is statistically indistinguishable from the Transformer cluster. Architecture type introduces no systematic bias.

4 Theoretical Framework

The empirical stability of δ across temperatures and architectures (§3) demands theoretical explanation. This section provides three layers of theory, each addressing a different question: *why* δ is universal (§4.1), *what* determines its value for a given model (§4.2), and *how* fine-tuning and training procedures propagate through to observable changes (§4.3).

4.1 From Universality to Prediction

As established in §3.3, the Gumbel PPP yields the universal prediction $\mathbb{E}[\delta_{\mathrm{norm}}] = 0.318$ and places all tested models in the stable phase $\beta > 1$. This universality explains *why* δ is stable — the answer is extreme value theory applied to the logit tail.

The stability of δ propagates to downstream information-theoretic quantities through a saturating chain. Define $s = \sigma(\delta_{\mathrm{raw}}/T)$ where $\sigma(u) = 1/(1 + e^{-u})$ is the standard sigmoid, and the collision probability $\rho = s^2 + (1 - s)^2 = h(\delta_{\mathrm{raw}}/T)$ where $h(u) = \sigma(u)^2 + (1 - \sigma(u))^2$. Because δ_{raw} is temperature-invariant and h saturates exponentially, small perturbations in δ produce exponentially smaller perturbations in ρ. This chain is formalized in `delta_implies_rho_stability` (§8.2, Appendix C): $|\rho_1 - \rho_2| \leq 4\eta + 4\exp(-\delta_{\min}/T_{\max})$.

The next question — *what determines the absolute value of δ_{raw} for a given model* — requires a model-specific theory.

4.2 The Equation of State

We identify four quantities, all derived from static model files without inference, that jointly predict δ_{raw} from first principles:

Stiffness $S = \gamma \times \sigma_w$, the product of the final normalization gain (γ, from the RMSNorm or LayerNorm preceding the output projection) and the row-norm scale of the unembedding matrix (σ_w). Higher stiffness amplifies logit margins, increasing δ. Direction: positive.

Crowding D, the mean k-nearest-neighbor cosine similarity ($k = 10$) among rows of the unembedding matrix W_U. Higher crowding means more local competition among vocabulary embeddings, compressing gaps. Direction: negative.

Table 4: Equation of State coefficients and bivariate correlations ($n = 38$). All directional predictions from first principles are confirmed. LOO $R^2 = 0.40$.

Predictor	Coefficient	Bivariate r	Direction
$\log S$ (Stiffness)	$+0.25$	$+0.55$	+ as predicted
D (Crowding)	-0.83	-0.47	− as predicted
$\log V$ (Pressure)	$+0.37$	$+0.59$	+ as predicted
$\log L$ (Depth)	$+0.29$	—	+
$\nVdash_{\text{fcap}}$ (Constraint)	-0.39	—	− as predicted

Pressure V = vocab_size, the vocabulary cardinality. Larger vocabularies increase the effective number of competitors in the softmax tail. Counterintuitively, this increases rather than decreases gap magnitudes: the extreme-value threshold rises with V, and the gaps between order statistics scale with the spacing at that threshold. Direction: positive.

Constraint C, the presence and value of training-time logit soft-caps (tanh-based bounding). Caps limit the effective gain, reducing δ. Direction: negative.

Each directional prediction follows from the output geometry: S scales logits linearly, D measures local embedding density (more neighbors $\Rightarrow$ smaller gaps), V sets the effective number of competitors in the softmax tail, and C compresses the logit range. We also include network depth L (number of layers) as a covariate; deeper networks have more opportunities for logit margin amplification through residual accumulation, though this mechanism is less directly interpretable than the four primary primitives.

Validation. We fit a log-linear model

$$\delta_{\text{raw}} \approx \alpha_0 + \alpha_1 \log S + \alpha_2 D + \alpha_3 \log V + \alpha_4 \log L + \alpha_5 \nVdash_{\text{fcap}} \tag{12}$$

on $n = 38$ models spanning 9 architectural families, using leave-one-out cross-validation.

The LOO $R^2 = 0.40$ means the four primitives explain 40% of the variance in δ_{raw} across models from static weight properties alone, with zero inference. The remaining 60% is attributable to training-recipe effects that are not captured by weight-level statistics (see §5 on representation topology).

The falsification narrative. An instructive failure illustrates the methodology. At $n = 29$, a fifth predictor — head dimension H (hidden_dim / n_heads) — showed strong bivariate correlation ($r = +0.75$) and was initially included in the model. At $n = 38$, this correlation collapsed to $r = +0.16$. Diagnosis: three Gemma models with atypically large H values (144–320, versus 64–128 for all other models) had acted as leverage points, driving the entire apparent effect. Removing Gemma from the sample yielded $r = +0.017$ — no signal. H was removed from the Equation of State. This predict-validate-falsify-update cycle is applied throughout: claims that survive expansion are reported; claims that don't are documented as falsified.

Scaling behavior. A natural concern is whether the Equation of State's primitives remain well-behaved as models scale. We measured stiffness S, crowding D, and vocabulary pressure V across 12 models from three families (Qwen 0.5B–72B, Llama 1B–70B, Mistral 7B–24B), spanning a 147× parameter range. The original hypothesis — that stiffness grows unboundedly with scale — was falsified: S follows a mildly declining power law ($S \sim \text{params}^{-0.08}$), bounded within $[1.18, 3.59]$ across all models and families. The three families achieve stiffness through qualitatively different strategies (Qwen: extreme γ at small scale, compensating for small d_{model}; Llama: moderate γ with architectural-transition spikes; Mistral: high γ with microscopically small σ_w), yet all converge to the same bounded range. Crowding D scales negatively with model size ($r \approx -0.85$): larger embedding spaces reduce local cosine density, as expected geometrically. These findings are formalized in ScalingLaws.v (28 theorems, 2 empirical axioms — `stiffness_bounded_below` and `delta_positive_from_stiffness` — 0 Admitted), which proves that the revised bound $S_{\min} > 0$ at all scales implies $\delta_{\min} > 0$ and hence a minimum authentication margin $K_{\min} > 0$. The falsification of unbounded stiffness strengthens rather than weakens the theoretical framework: the security chain holds at all tested scales without requiring stiffness to grow.

4.3 Gauge–Transport Decomposition

The Equation of State predicts δ from static weights. A separate question is: when a model's weights change (through fine-tuning, distillation, or training), how do those changes propagate to observable fields — and which components of the change are detectable?

Setting. Let $\theta \in \mathbb{R}^p$ denote the model's parameters and $\phi = F(\theta) \in \mathbb{R}^m$ an observable field (logits, attention scores, hidden states, etc.) computed by a differentiable map F. A parameter perturbation $\Delta\theta$ (with $\mathbb{E}[\Delta\theta] = 0$, covariance $\Sigma_{\Delta\theta}$) induces a field perturbation $\Delta\phi$.

The transport equation (L1). To first order,

$$\Sigma_{\Delta\phi} = J_F \, \Sigma_{\Delta\theta} \, J_F^\top \tag{13}$$

where $J_F = \partial F/\partial\theta|_{\theta_0}$ is the Jacobian. This is the delta-method applied to covariance: the geometry of parameter changes is *transported* through the network's Jacobian into the space of observable changes. (We note that L2 — the delta-method remainder bound for nonlinear F — is omitted from the formalization; the transport equation is exact for the linear maps arising in the output pathway, and the Coq proof treats the linearized case directly.)

Gauge subspaces (L3). Not all directions in field space $\mathbb{R}^m$ are observable. If $O(\phi)$ is the downstream quantity of interest (e.g., the probability distribution after softmax), then directions g satisfying $O(\phi + g) = O(\phi)$ form a **gauge subspace** $\mathcal{G}$. For the softmax, $\mathcal{G} = \operatorname{span}\{\mathbf{1}\}$: adding a constant to all logits leaves probabilities unchanged. For attention, each query row has its own gauge direction.

Gauge fraction (L4). Let $\Pi_{\mathcal{G}}$ denote orthogonal projection onto $\mathcal{G}$. The gauge fraction

$$g_\phi := \frac{\operatorname{tr}(\Pi_{\mathcal{G}} \, \Sigma_{\Delta\phi})}{\operatorname{tr}(\Sigma_{\Delta\phi})} \in [0, 1] \tag{14}$$

measures what share of the observable field's variance is absorbed by function-invisible directions. When $g_\phi \approx 1$, nearly all variance is gauge — the field changes without the function changing. When $g_\phi \approx 0$, all variance is functional.

Rank inequality (L5). The rank of the transported covariance is bounded:

$$\operatorname{rank}(\Sigma_{\Delta\phi}) \leq \min\big(\operatorname{rank}(\Sigma_{\Delta\theta}), \; \operatorname{rank}(J_F)\big) \tag{15}$$

This identifies two distinct mechanisms for low-rank observable changes: a low-rank source (the parameter perturbation itself is concentrated in few directions — "starvation") or a low-rank transport (the Jacobian contracts — "bottleneck"). Both produce similar endpoints but are distinguishable by layerwise profiling (§5.2).

DPO covariance structure (L6). For preference optimization methods that operate on winner-loser pairs, the effective gradient is a difference $g_w - g_l$. Its covariance decomposes as

$$\Sigma_{w-l} = \Sigma_w + \Sigma_l - \Sigma_{wl} - \Sigma_{lw} \tag{16}$$

When winner and loser gradients share a large common component (the prompt encoding), the cross-terms Σ_{wl} and Σ_{lw} suppress the top eigenvalues of the mixture, producing a broader spectrum than single-sample methods. This is the mathematical basis for the forensic distinction between supervised fine-tuning and preference optimization observed in §5.

Participation ratio bounds (L7). The participation ratio $\operatorname{erank}_{\mathrm{PR}}(\Sigma) := \operatorname{tr}(\Sigma)^2/\operatorname{tr}(\Sigma^2)$, bounded between 1 and $\operatorname{rank}(\Sigma)$, provides the continuous empirical surrogate for rank used throughout §5. The bound connects the exact rank inequality (L5) to the effective-rank measurements that distinguish collapse mechanisms in practice.

Unification. The transport equation (L1), gauge decomposition (L3–L4), rank inequality (L5), loss-structure identity (L6), and participation ratio bounds (L7) together form the **Gauge–Transport Decomposition** — a single mathematical framework that explains:

- Why training infrastructure artifacts can be invisible to model function but visible to weight forensics (gauge scars live in $\mathcal{G}$)
- Why representation collapse has two distinct mechanisms (source rank vs transport rank)
- Why different training algorithms leave qualitatively different traces (loss structure controls $\Sigma_{\Delta\theta}$)

The framework is formalized in Coq (GaugeTransport.v: 33 theorems, 0 Admitted, 0 axioms). The main theorem and its corollaries derive the forensic findings of §5 as consequences of the transport equation and gauge structure, rather than as isolated empirical observations.

5 Forensic Hierarchy

The theoretical framework of §4 identifies three independent sources of forensic information about a neural network: the universal statistics of its output distribution (thermodynamics), artifacts of its training infrastructure encoded in weight geometry (gauge scars), and the effective dimensionality of its learned representations (topology). This section presents experimental evidence that these three axes are orthogonal — a model can obey thermodynamic universality while carrying a gauge scar and exhibiting representation collapse — and develops their forensic applications.

5.1 Three Orthogonal Forensic Axes

Axis 1 — Thermodynamics. The output-layer statistics described in §§3–4 — $\delta_{\text{norm}} \approx 0.318$, $\beta_{\text{robust}} \approx 1.5$, the Equation of State — are universal across all tested architectures. They identify model family and parametric configuration. Every model obeys these laws regardless of training recipe or infrastructure.

Axis 2 — Gauge Scar. Softmax translation invariance (L3 from §4.3) implies that the constant-shift subspace $\mathcal{G}$ is function-invisible: adding $c\mathbf{1}$ to all logits leaves the output distribution unchanged. Training infrastructure can deposit artifacts in this subspace — detectable through weight statistics but invisible to model behavior. We identify one such artifact: Falcon-Mamba 7B, trained with custom Triton/ZeRO numerical kernels, carries a parameter-localized scar in the mean row of W_U, with the scar absorbing 43% of signal energy. The scar is surgically removable ($W' = W - \mathbf{1}\mu^\top$) with KL divergence $\sim 10^{-7}$ from the original, confirming it lives entirely in the gauge nullspace. This constitutes forensic identification of the *manufacturing process*, not the model itself.

Axis 3 — Representation Topology. The effective dimensionality of post-normalization hidden states $D(y_t)$ varies by more than an order of magnitude across models (effective rank 4–103 in our sample of five instrumented models), revealing two stable regimes separated by a 35-point gap with zero overlap in bootstrapped 95% confidence intervals. This axis is invisible to weight-only auditing — the signal lives in the geometry of the learned state distribution, not in W_U — and requires inference to measure.

Orthogonality evidence. Falcon-Mamba carries a gauge scar while exhibiting normal thermodynamics ($\delta_{\text{norm}} = 0.309$, within 3% of the universal prediction) and collapsed representation topology (effective rank ≈ 10). Conversely, Gemma 7B and Gemma 2 9B — same vendor, same architecture family — land in opposite topology regimes (effective rank 4 vs. 103) while sharing identical thermodynamic profiles. The three axes provide independent forensic information because they probe different mathematical objects: the tail of the output distribution, the nullspace of the softmax, and the covariance of the internal state trajectory.

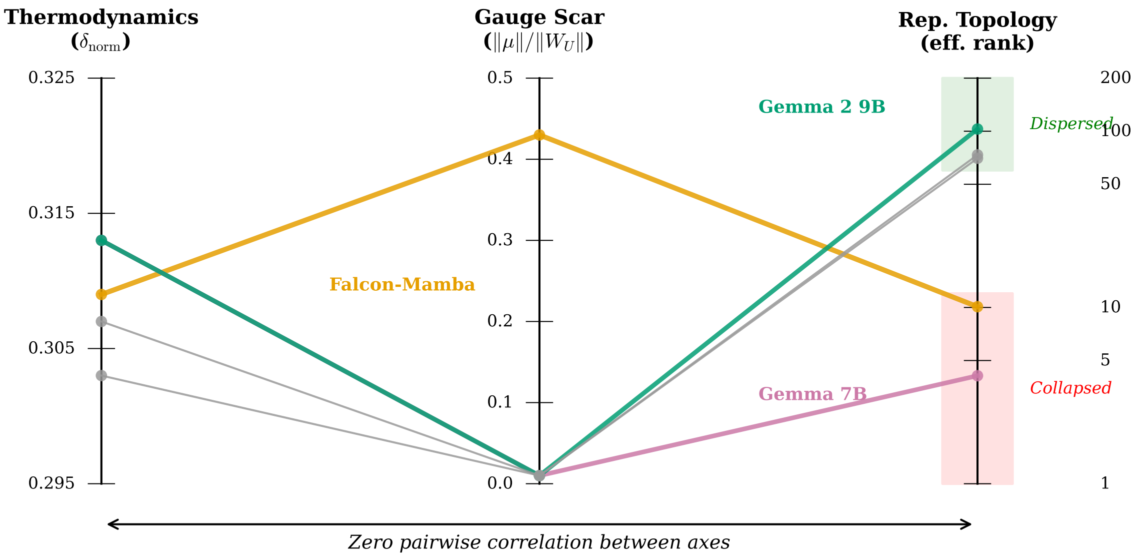

Figure 2: Three orthogonal forensic axes measured across five instrumented models. Each polyline is a model; axes represent (left) output-layer thermodynamics, (center) training-infrastructure gauge scarring, (right) representation topology. Falcon-Mamba (orange) demonstrates orthogonality: it obeys thermodynamic universality while carrying a training scar and exhibiting collapsed representation geometry. Gemma 7B and Gemma 2 9B (red/green) — same vendor, same architecture — land in opposite topology regimes from training recipe alone.

5.2 Two-Mechanism Collapse Theory

The Gauge–Transport Decomposition (§4.3) predicts two distinct routes to representation collapse, depending on whether the bottleneck is in the source term $\Sigma_g^{(L)}$ or the transport operator $J_{\ell\to L}$. Layerwise gradient profiling across four models (with a fifth instrumented for hidden states but pending full gradient profiles) confirms both mechanisms with 10–30× effect sizes.

Starvation. When logits are unbounded ($\max|z| > 100$), the softmax sharpens until the gradient $\nabla_z\mathcal{L} = p - e_{\text{target}}$ concentrates on a small number of directions. We observe this in Gemma 7B v1 (a dense Transformer without training-time logit caps; max logit ≈ 350) where gradient effective rank is 1.9–9.6 throughout the network and the top-1 variance fraction reaches 48–90% — up to 90% of gradient energy flows through a single direction. The collapse is uniform: no localized bottleneck appears at any depth, because the source term itself is low-rank.

Contraction. In Falcon-Mamba 7B (a pure Mamba state-space model; max logit ≈ 147), gradients are rich (effective rank 30–89), but a localized gradient bottleneck at approximately 38% network depth compresses gradient effective rank to its minimum. An even more severe collapse appears in the Jacobian effective dimension, which falls from 20.6 to 3.6 — a 5.7× collapse — in the final 15% of layers. High-rank parameter updates (effective rank 80–144) persist despite this compression, confirming that the bottleneck is in transport, not source. The rank inequality from L5 ($\text{rank}(\Sigma_g^{(\ell)}) \leq \min(\text{rank}(\Sigma_g^{(L)}), \text{rank}(A_\ell))$) pinpoints the mechanism: starvation is source-limited regardless of Jacobian rank; contraction is transport-limited despite rich gradients.

Causal confirmation. Within a single vendor's model family, two models sharing the same architecture but differing in training-time logit regularization land in opposite regimes: ground-truth pooled effective rank ≈ 4 without regularization versus ≈ 103 with soft-capping — a 25× difference (25–39× depending on measurement pipeline, consistently an order of magnitude) from training recipe alone. A larger model

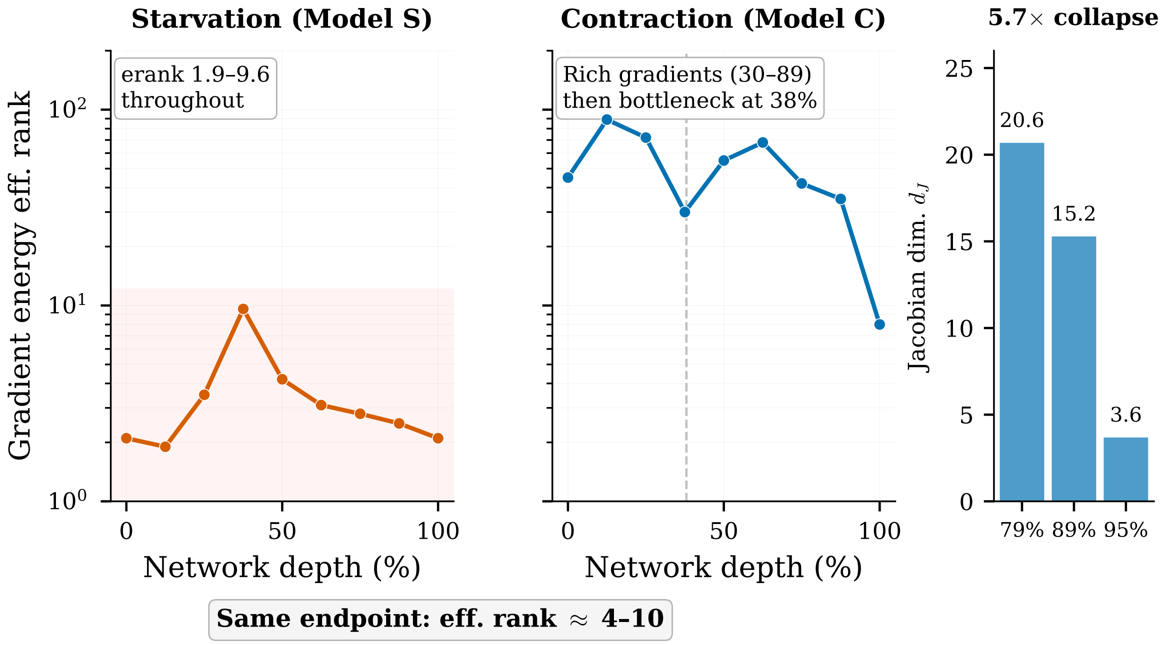

Figure 3: Two distinct pathways to identical representation collapse, distinguished by layerwise gradient profiling. *Left (Starvation):* gradient energy rank is uniformly low throughout — the bottleneck is at the source. *Center (Contraction):* gradients are rich (rank 30–89) but collapse at a localized bottleneck at 38% depth, confirmed by 5.7× Jacobian effective dimension collapse in final layers (*right panel*). Both produce representation effective rank ≈4–10 at the output, but via mechanisms distinguishable by 10–30× in gradient-energy rank.

from the same family using caps but no knowledge distillation also falls in the normal regime, establishing that logit bounding is a sufficient intervention. Three protection mechanisms — explicit caps, knowledge distillation, and weight-norm control — all produce equivalent gradient geometry (effective rank 16–39, top-1 variance fraction < 0.3), consistent with the theoretical prediction that what matters is bounded logits, not the method of bounding.

5.3 Model Provenance

The Gauge–Transport framework also yields a practical forensic hierarchy for model provenance — identifying not just *what* a model is, but *how it was made.*

Architecture discrimination. Interior geometric measurements at normalization layers separate architecture families with Cohen's $d \approx 5.2$ across 10 models spanning pure Mamba, Mamba-Transformer hybrids, and dense Transformers. The separation is large enough that a single measurement suffices for architecture classification with zero overlap between families. Hybrid architectures show intermediate values with amplified depth gradients, consistent with the alternating layer structure creating steeper gauge profiles than either pure architecture. Heavy grouped-query attention configurations produce a distinctive sub-family signature detectable from the attention mechanism alone.

Fine-tuning detection. We track a model through its full training trajectory — base to supervised fine-tuning (SFT) to preference optimization (DPO) — and find that the two fine-tuning stages leave qualitatively different forensic traces, detectable at two levels of analysis.

At the coarse level (requiring only the model itself, no base comparison), interior geometric measurements detect SFT: the attention-layer gauge drops by 35%, indicating that supervised fine-tuning reshapes internal routing structure. DPO, by contrast, produces zero measurable change in interior geometry — it adjusts preference margins without altering routing, consistent with its loss function operating on output-level reward differences rather than intermediate representations.

At the fine level (requiring comparison against the base model), weight differential analysis distinguishes both stages through four orthogonal signatures:

Table 5: Weight differential signatures of SFT vs DPO fine-tuning (Llama 3.2 3B).

	SFT	DPO	Ratio
Weight magnitude ($\|\Delta W\|/\|W\|$)	0.29%	0.09%	3.4×
Effective rank of ΔW (global)	2.5	3.4	—
Semantic alignment with base	2.5%	10–20%	4–8×
Depth distribution	U-shaped	Flat	Orthogonal

The counterintuitive finding is that DPO produces smaller-magnitude but *broader* and *more semantically aligned* weight changes than SFT. Supervised fine-tuning injects concentrated, largely nullspace-orthogonal updates — routing modifications that redirect information flow without much overlap with the base model's learned semantics. Preference optimization makes distributed, semantically aligned adjustments across all depths — refining existing knowledge rather than installing new pathways.

Mechanistic explanation. The Gauge–Transport framework explains this difference through the loss-structure identity (L6 from §4.3). DPO's pairwise loss produces gradients that are differences of winner and loser gradients, $g_w - g_l$. When these share a large common component (the prompt encoding), the subtraction cancels shared modes and suppresses the top eigenvalues of the gradient covariance. Experimental validation on 500 preference pairs confirms this: winner and loser gradients show mean cosine similarity of 0.75, and the top eigenvalue of the difference covariance is suppressed by 69% relative to the individual covariances. This common-mode cancellation is why DPO's weight changes have higher effective rank (broader spectrum) and greater semantic alignment (the common prompt-encoding modes are removed, leaving preference-specific directions that overlap with the base model's semantic subspace).

The two-level hierarchy. These findings establish a practical forensic workflow: interior geometric measurements provide fast coarse classification of architecture family and instruction-tuning status (∼2 minutes, no base model required), while weight differential analysis provides fine discrimination between training algorithms (∼20 minutes, base model required). Together, interior measurements answer *what the model is* and weight differentials answer *what was done to it*.

6 Security Analysis

The three forensic axes identified in §5 provide candidates for authentication; the question becomes which are adversarially robust. The forensic observables developed in §§3–5 are scientifically interesting only if they resist deliberate manipulation. This section establishes the security model, quantifies spoofing difficulty via Fisher information geometry, and presents a formal impossibility result that the cost-accuracy tradeoff for fingerprint spoofing is a cliff, not a slope.

6.1 Threat Model

We consider a strong attacker with the following capabilities:

White-box access. The attacker possesses full access to the source model's weights θ and architecture. This is the strongest standard assumption in model security — weaker than assuming access to training data or infrastructure, but sufficient for all known parameter-space attacks.

Parameter-efficient modification. The attacker may apply parameter-efficient fine-tuning (e.g., LoRA at arbitrary rank and target modules) to produce a modified model $\theta' = \theta + \Delta\theta$. This covers the dominant class of practical model modifications.

System knowledge. Following Kerckhoffs's principle, the attacker knows that the fingerprinting system exists, understands its mathematical basis, and has read this paper. Security cannot depend on obscurity.

The attacker does *not* have access to the content of the challenge prompt bank, nor knowledge of which prompts are selected for a given verification challenge. This assumption is load-bearing and stated explicitly because the security argument has two legs:

Leg 1 — Spoofing hardness (formal, attacker-independent). Each fingerprint component has a minimum KL divergence cost to shift, derived from the curvature of the statistical manifold (Cramér-Rao bound). This holds regardless of what the attacker knows, including the specific prompts.

Leg 2 — Challenge unpredictability (operational, assumption-dependent). The attacker cannot pre-optimize for unknown challenges, providing entropy multiplication across the response vector. If the prompt bank is compromised, Leg 2 degrades: the protocol reduces to a static multi-component fingerprint with k known components, where k is the challenge subset size. Leg 1 still holds for each component, and T7 budget exhaustion across k components provides the fallback defense. The system degrades gracefully — it does not collapse.

The attacker's goal is to make model A's fingerprint response match model B's enrolled profile while preserving A's functionality (bounded KL divergence from the original output distribution). This is the standard impersonation attack in authentication systems.

6.2 Forensic Hardness

The Gauge–Transport Decomposition (§4.3) identifies two classes of observable fields: those aligned with gauge subspaces (function-invisible, trivially spoofable) and those aligned with functional subspaces (detectable, costly to move). The security-relevant question is: how costly?

Fisher sensitivity framework. For an observable $F(\theta)$ computed from model parameters θ, the minimum KL divergence required to shift F by an amount ΔF is bounded below by the Cramér-Rao inequality:

$$\mathrm{KL}_{\min}(\Delta F) \geq \frac{(\Delta F)^2}{2\, s_F^2} \tag{17}$$

where $s_F^2 = \bar{\Delta}^\top \hat{\mathcal{I}}^{-1} \bar{\Delta}$ is the Fisher sensitivity — a function of the observable's gradient $\bar{\Delta} = \nabla_\theta F$ and the Fisher information matrix $\hat{\mathcal{I}}$ of the output distribution. When normalized by the observable's natural variation ($\tilde{s}_F = s_F/\sigma_F$, not to be confused with the unnormalized s_F above), this yields a dimensionless hardness measure: lower $\tilde{s}_F$ means more KL per unit of observable shift.

Concrete example: the δ-gene itself. Before introducing the hierarchy, consider the observable the reader already knows. The δ-gene ($G_3 = z_{(3)} - z_{(4)}$) has normalized Fisher sensitivity $\tilde{s}_F \approx 137$–247 across models. This means a single-standard-deviation shift in δ costs approximately $\tilde{s}_F^{-2}/2 \approx 10^{-5}$ nats of KL divergence — essentially free. An attacker can shift δ without meaningfully perturbing the output distribution. Thermodynamic observables are easy to spoof because they live in directions of high Fisher information: the model's output distribution is already sensitive to them, so small parameter changes move them readily.

Empirical finding: the hardness hierarchy. Fisher sensitivity analysis across multiple models and architectures establishes a consistent hierarchy: interior geometric observables (those characterizing internal routing structure and representation geometry) are orders of magnitude harder to spoof than thermodynamic observables (δ, β). The hardest observable we have identified has $\tilde{s}_F \approx 0.4$–0.9 for Transformers, requiring KL divergence in the nats-scale regime — well into functional destruction — to shift by a single standard deviation. This is a 100–300× hardness ratio over the thermodynamic observables.

Architecture dependence. The hardness ranking is not universal across architectures. The mechanism that makes a particular observable hard to spoof (e.g., geometric saturation effects for one class of observables in Transformers, or recurrence-constrained structure for another class in state-space models) depends on architectural properties. An attestation protocol must therefore detect architecture first, then select appropriate anchors — a requirement that the forensic hierarchy of §5 naturally supports.

The Pareto frontier. Adversarial spoofing experiments confirm that the tradeoff between spoofing progress and functionality preservation is not a smooth slope but a cliff. At low KL budgets, the attacker achieves negligible fingerprint shift — the observable barely moves. At moderate KL budgets, model functionality degrades catastrophically before the fingerprint reaches the target. There is no intermediate regime where the attacker makes meaningful progress at tolerable cost. This cliff structure is not a property of a specific attack algorithm; §6.3 proves it is structural.

6.3 Formal Impossibility

The empirical cliff of §6.2 admits a formal explanation. We prove that under the Fisher geometry of the parameter manifold, no KL budget — however allocated — permits simultaneous fingerprint matching and noise control. The proof is structured in three tiers of decreasing conditionality.

Tier 1 — Unconditional (Cramér-Rao only). Per-component cost floor (S1). For a single fingerprint component with initial gap g from the target and Fisher information I_F, any parameter perturbation that closes the gap by an amount s incurs KL divergence at least $K \geq s^2/(2I_F)$. This is a direct consequence of the Cramér-Rao bound on the statistical manifold and holds unconditionally — no assumptions about attack strategy, noise model, or system parameters.

Multi-component budget exhaustion (T7). When the attestation protocol measures d independent components (corresponding to d challenge prompts), the KL budget must be split across components. For d components with gaps $g_1, \ldots, g_d$ and Fisher information $I_1, \ldots, I_d$, any allocation of total KL budget K across components leaves at least one component with residual gap or noise exceeding the acceptance threshold. This is formalized in Coq for $d = 2$ and $d = 3$ (`T7_pure_cr_floor_2` and `T7_pure_cr_floor_3`); the general-d case requires an inductive argument that has not been machine-checked and is therefore classified as a conjecture, not a proven result. The empirical evidence (3.8× hardening from 1 to 4 seeds, §7) is consistent with monotone difficulty scaling in d, and the argument structure is straightforward, but we do not claim what we have not proven. Multi-component attestation is provably harder than single-component for $d \leq 3$, with the difficulty scaling captured by the budget exhaustion inequality.

Both S1 and T7 are unconditional — they depend only on the Cramér-Rao bound, which is a property of probability manifold curvature.

Tier 2 — Conditional (noise model). Interval splitting (T4 — the security theorem). Consider a single fingerprint component. At any KL budget K, the attacker faces two competing costs: (a) the residual fingerprint gap, which starts large and decreases as KL increases (with the gap-closing rate governed by the Cramér-Rao bound), and (b) the accumulated noise from parameter perturbation, which starts at zero and grows monotonically with KL. The acceptance threshold ε must be exceeded by at least one of these quantities at every value of K. If there exists a crossover point K_x where the decreasing gap curve and the increasing noise curve both exceed ε — that is, the gap hasn't closed yet when the noise becomes detectable — then the attacker is trapped: below K_x, the gap is too large; above K_x, the noise is too large. The Coq proof establishes this via an L^2 composition (the sum of squared residuals exceeds the squared threshold), which is strictly stronger than a max-based argument. No sweet spot exists at any KL budget. The formal details — including the specific noise model and gap-closing functions — are given in `NoSpoofing.v` (`T4_no_spoofing_interval_split`).

Conditionality stated explicitly. T4 requires existence of the crossover point K_x. This is not assumed axiomatically — it is validated empirically. The Fisher sensitivity measurements of §6.2 establish that gap-closing costs for the protected observable are in the nats-scale regime (high), while noise grows continuously from near-zero. The crossover window is therefore wide and robustly satisfied. The theorem proves "if crossover exists, then no spoofing at any KL budget." Crossover existence is validated, not assumed.

Tier 3 — Structural parallel. The impossibility has an algebraic parallel to quantum no-cloning. The equation $x = x^2$ admits only the solutions $\{0, 1\}$ — there is no partial cloning. Analogously, the interval-splitting structure of T4 admits only two outcomes: the model is identical to the target (distance zero, a clone) or it is detectable (distance exceeds the threshold). There is no partial spoofing regime.

This parallel is informal but historically grounded: `NoSpoofing.v` was developed from `QuantumNoCloning.v` and shares its fixed-point algebraic core ($x = x^2 \Rightarrow x \in \{0, 1\}$). The analogy is suggestive — both results exclude intermediate fixed points — but the proof mechanisms are independent. The no-cloning theorem operates on Hilbert-space unitaries; T4 operates on statistical manifold geometry via the Cramér-Rao bound. We present the parallel as motivation, not as a formal reduction.

Formalization. The impossibility result is machine-checked in Coq (`NoSpoofing.v`: 994 lines, 51 theorems, zero `Admitted`, zero axioms). The key theorems:

Theorem	Statement	Tier
S1	Per-component KL floor: $K \geq g^2/(2I_F)$	1 (unconditional)
T4	Interval splitting: no sweet spot at any KL budget	2 (conditional on K_x)
T7	d-component budget exhaustion	1 (unconditional)

Experimental validation. Adversarial spoofing attacks across multiple configurations and attack strategies confirm the three-phase structure predicted by T4: (1) a *free-drift* phase at low KL where the fingerprint gap dominates and the attacker makes negligible progress, (2) a *destruction* phase at moderate KL where accumulated noise overwhelms — the fingerprint actually moves *further* from the target as model functionality degrades, and (3) a *trapped* phase where the optimizer cannot make progress regardless of hyperparameter adaptation. Strengthening the attacker (increasing LoRA capacity, sweeping regularization strength over nearly six orders of magnitude, jointly optimizing across multiple challenge seeds) shifts the cliff location but does not eliminate it. No tested configuration achieved acceptance.

Multi-component hardening. Consistent with T7, joint optimization across multiple challenge seeds is empirically harder than single-seed optimization, with the difficulty ratio exceeding the theoretically predicted minimum. This confirms that the entropy multiplication from challenge diversity (Leg 2 of the threat model) provides genuine additional security beyond per-component hardness (Leg 1).

7 Challenge-Response Authentication

The forensic observables of §§3–5 and the security guarantees of §6 provide the ingredients for a practical authentication primitive. This section presents the construction: an inference-time challenge-response protocol that adapts the Physical Unclonable Function (PUF) paradigm from hardware security to neural network identity verification.

7.1 From Static Fingerprint to PUF

A natural first approach to model authentication is a *static fingerprint*: measure the Fisher-protected observable at a fixed set of network locations and compare against an enrolled profile. This approach fails for a specific and instructive reason.

The observable identified by the forensic hardness analysis of §6.2 — the one requiring nats-scale KL to shift by a single standard deviation — behaves as an **order parameter** when measured across different sites within a single model. Its value saturates to a characteristic constant determined by architecture and training recipe, with minimal variation across layers or measurement positions. A static multi-site fingerprint therefore has approximately one effective dimension of entropy regardless of how many sites are measured: the measurements are near-redundant.

This is not a deficiency of the observable; it is a consequence of the same saturation mechanism that makes it hard to spoof. An observable that varies freely across sites would be easy to shift at each site independently. The tension between hardness and entropy is structural.

The resolution comes from changing the source of entropy. In the static approach, entropy must come from *physics diversity* — the observable varying across measurement sites. The alternative is to draw entropy from the *challenge space* — varying the input to the model rather than the measurement location. Different

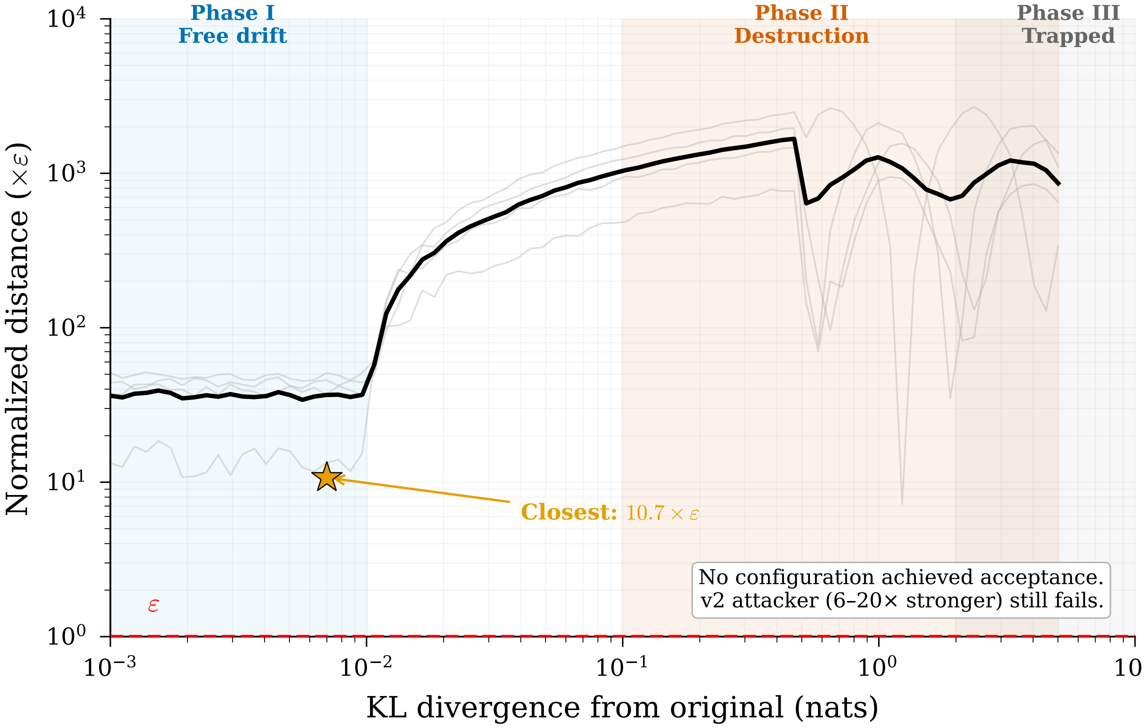

Figure 4: The three-phase Pareto cliff for adversarial fingerprint spoofing (adaptive-λ attack, LoRA rank 8, Qwen 0.5B → StableLM 1.6B). At low KL budgets (Phase I), the attack makes no measurable progress. At moderate KL (Phase II), the fingerprint distance *increases* — the model degrades before it can converge. In Phase III, the optimizer is trapped regardless of regularization strength (λ swept over nearly six orders of magnitude). The dashed red line marks the acceptance threshold; the closest approach ($10.7 \times \varepsilon$, star) remains a decade above it. This structure is predicted by Theorem T4 (NoSpoofing.v) and confirmed empirically.

input prompts activate different internal pathways, inducing genuinely different values of the protected observable even though the measurement procedure is identical. The model's response to a specific input is determined by its weight geometry and cannot be predicted without running inference.

This is precisely the PUF paradigm. A hardware PUF exploits manufacturing variation to produce device-specific responses to external challenges. An **inference-time PUF** exploits training-induced weight geometry to produce model-specific responses to input challenges. The challenge-response structure provides entropy multiplication: each prompt contributes an independent measurement, and the combined response vector has entropy proportional to the number of challenges.

7.2 Protocol Overview

The authentication protocol consists of three phases: enrollment, challenge generation, and verification.

Enrollment. A curated prompt bank is established, containing diverse natural-language prompts selected to maximize measurement diversity. For each model to be enrolled, a deterministic subset of prompts is selected using a cryptographic seed, and the Fisher-protected observable is measured per-prompt under controlled, reproducible input conditions that eliminate sampling stochasticity. The resulting response vector is stored as the enrolled profile, together with the seed used for prompt selection.

Challenge generation. To verify a model's identity, a fresh seed is generated and used to deterministically select a new prompt subset from the bank. The same per-prompt measurement is performed, producing a response vector of the same dimensionality.

Verification. The L^2 distance between the fresh response vector and the enrolled profile is compared against an acceptance threshold ε. The threshold is derived from the measured variance of genuine repeat measurements — specifically, the variation observed when the same model is measured under different numerical conditions (e.g., cross-platform or cross-precision inference). This calibration procedure yields ε on the order of 10^{-4}, reflecting the high reproducibility of the protected observable under controlled conditions.

The prompt bank contents, bank size, subset selection algorithm, and the identity of the protected observable are operational parameters not disclosed in this work. Security under Kerckhoffs's principle requires only that the prompt bank contents remain confidential (Leg 2 of §6.1); the mathematical hardness of per-component spoofing (Leg 1) holds regardless of disclosure.

7.3 Entropy Source Validation

The protocol's entropy claim rests on a testable empirical hypothesis: the protected observable must vary meaningfully *across prompts within a single model*, not merely across models. If the observable reads the same constant regardless of input — as it approximately does across measurement *sites* — then challenge diversity provides no entropy, and the protocol degenerates to a static fingerprint.

We test this by measuring the observable across a large set of diverse prompts drawn from the prompt bank and computing the within-model coefficient of variation:

$$\mathrm{CV}_{\mathrm{prompt}} := \frac{\mathrm{SD}(\{F(\theta, x_i)\}_{i=1}^{n})}{\mathrm{mean}(\{F(\theta, x_i)\}_{i=1}^{n})} \tag{18}$$

where $F(\theta, x_i)$ is the protected observable measured on model θ with prompt x_i.

Across two models from different model families, the within-model prompt CV is approximately 25%. This is large relative to the temperature CV of the underlying δ observable (1–5%, §3.2) and large relative to the cross-platform measurement noise ($\varepsilon \sim 10^{-4}$). Different prompts probe genuinely different aspects of the model's internal geometry, producing structurally different measurements.

The 25% CV confirms that entropy multiplication is structural: each challenge prompt contributes an independent measurement, and the combined response vector occupies a high-dimensional space that an attacker cannot pre-optimize without knowing the challenge. To quantify the effective dimensionality of the challenge space, we measure the participation ratio of the sensitivity matrix Σ_τ across 200 prompts from the canonical prompt bank. The measurement requires a perturbation that excites the protected observable's cross-token statistics: token-mixing perturbations of the form $\Delta h = \varepsilon \cdot UV^\top(h - \mu)$, which couple tokens analogously to attention, are effective where per-token perturbations are not (the latter fall in the observable's mathematical nullspace). The resulting participation ratio is $r_{\mathrm{eff}} \approx 52$ — the response space has no low-rank shortcut (the top eigenvalue accounts for only 7.5% of total variance, and 86 components are needed for 90%). At operational subset sizes ($k = 32$), each seed imposes approximately 6 independent constraints, and multi-seed challenges compound multiplicatively: the empirical 3.8× hardening from 1 to 4 seeds is consistent with this dimensionality. This refutes the order-parameter concern — the observable is an order parameter across *sites* but not across *inputs*.

7.4 Validation Summary

Four methodological notes on these results bear emphasis. First, the uniqueness validation includes same-family comparisons (models sharing architecture and vendor but differing in scale), which produce the tightest separations at 1,186 × ε. The global frontier pair (cross-family, 485 × ε) is closer than the same-family minimum, indicating that the protected observable's geometry is dominated by training recipe rather than vendor lineage. Second, the min-entropy bound is conservative: it uses the Rule of Three to upper-bound per-seed collision probability from observed zero-collision events, then multiplies across independent seeds. The true entropy is likely higher, but we report only what the data supports. Third, the unpredictability bound comes from population-level collision analysis, not adversarial testing at the success boundary — the independence ratio R (entropy scaling under adversarial optimization) is undefined because no attack reached acceptance. Fourth, the response vector dimensionality was doubled from 32 to 64 via dual-site

Table 6: IT-PUF validation against five standard PUF properties.

Property	Requirement	Evidence	Status
Uniqueness	Different models produce distinguishable responses	Zero false accepts across 1,012 cross-model comparisons spanning 23 models from 16 families and 3 architecture types (dense Transformer, parallel-attention Transformer, and pure Mamba SSM), including same-vendor scale variants. Minimum separation exceeds ε by 485× at the global frontier (cross-family) and by 1,186× within same-family pairs.	Validated
Stability	Same model produces consistent responses	Acceptance threshold ε derived from genuine cross-platform measurement variance ($\sim 10^{-4}$), not heuristic.	Validated
Tamper-evidence	Modified models are detectable	All adversarial spoofing configurations failed (§6.2); formal impossibility proof establishes structural Pareto cliff (§6.3).	Validated
Unpredictability	Responses cannot be predicted without the model	Conservative min-entropy estimate exceeds 25 bits from challenge diversity (population separation, not adversarial testing). Participation ratio of the full prompt-bank sensitivity matrix yields $r_{\text{eff}} \approx 52$ effective dimensions; no low-rank shortcut exists.	Validated
Quantization robustness	Fingerprint survives precision reduction	Zero false accepts at 4-bit (NF4) quantization across 112 Transformer comparisons; full-precision zoo achieves 0/1,012 across all architectures. Minimum separation at NF4 is comparable to full precision ($314 \times \varepsilon$). NF4 Mamba enrollment not yet tested.	Validated

measurement (two normalization sites per prompt) during zoo expansion; this protocol upgrade widened same-family separation margins by 3.8× over the initial 32-dimensional protocol, confirming that additional measurement sites provide genuine discriminative information rather than redundant readings.

Table 6 summarizes the empirical validation against five standard PUF properties.

The validation zoo spans 23 models from 16 vendor families across three architecture types: dense Transformers, one pure Mamba state-space model (Falcon-Mamba 7B), and one parallel-attention Transformer (Phi-2). Cross-architecture comparisons (172 pairs across 4 seeds) show minimum separation of $3{,}103 \times \varepsilon$, confirming that IT-PUF discriminates across fundamentally different computational mechanisms. The protocol is architecture-aware by design: it detects the model's architecture class and selects the appropriate measurement observable accordingly.

7.5 Computational Cost

Enrollment requires one forward pass per challenge prompt, with controlled (deterministic) inputs and a compact geometric measurement from the model's internal representations. No training, no gradient computation, no backward passes. The cost scales linearly with challenge subset size and is negligible relative to a single inference request for typical subset sizes.

Verification has identical cost to enrollment — the same prompts, the same measurement, the same extraction. An L^2 distance computation on the resulting vectors completes in microseconds.

Storage is negligible relative to the model weights being authenticated — on the order of hundreds of bytes per enrollment.

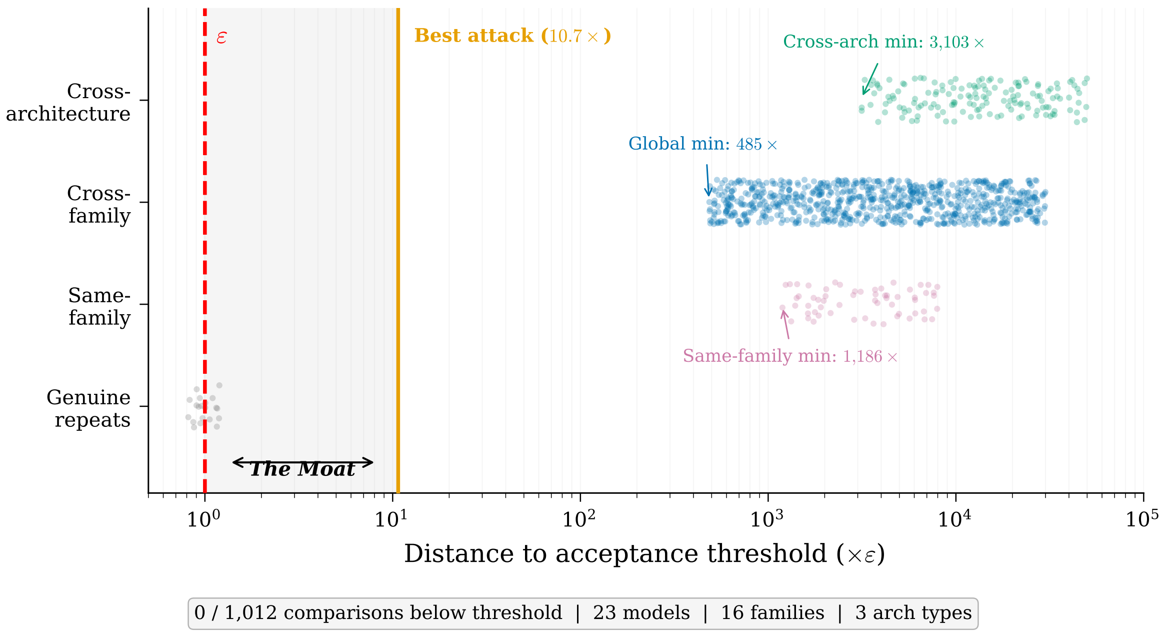

Figure 5: Distribution of pairwise fingerprint distances across the 23-model IT-PUF zoo (1,012 comparisons, Sprint 2). The vertical red line marks the acceptance threshold ε; the orange line marks the closest adversarial approach ever achieved ($10.7 \times \varepsilon$, Phase 4B v2 adaptive-λ attack). The gray band between them is *The Moat* — the region that separates every legitimate comparison from the strongest known attack by a factor of 45×. No cross-model comparison fell below threshold. Same-family pairs (same vendor, different scale) produce the tightest separations; cross-architecture pairs are the most separated.

The entire enrollment-and-verification cycle can execute on the same hardware that serves model inference, with no specialized equipment, no cryptographic co-processors, and no modification to the model itself. The model under test is never altered; authentication is purely observational.

Intellectual property. A provisional patent application covering the challenge-response authentication protocol has been filed (Application 63/982,893, February 13, 2026).

8 Formal Verification

The mathematical claims in this work are not arguments from authority or approximation — they are machine-checked proofs in the Coq proof assistant (Rocq 9.1). This section describes the proof stack, highlights four key theorems, and defines the epistemological classification that governs every claim in the paper.

8.1 The Coq Proof Stack

The formal foundation comprises 16 files containing 311 top-level declarations (Theorem, Lemma, Corollary, and Proposition), with zero uses of `Admitted` as a proof terminator (Coq's mechanism for accepting unproven statements) and zero vacuous definitions of the form `Definition X := True` in active code.

Counting methodology. The count of 311 is mechanically reproducible: it is the number of top-level `Theorem`, `Lemma`, `Corollary`, and `Proposition` declarations across the 16 files listed below. This is not a curated subset — every such declaration compiles to a `.vo` object file under Rocq 9.1, and each represents a statement whose proof has been verified by the kernel. The broader research program maintains 53 Coq files with 820+ theorems; the 16 files scoped here are those whose results are cited in this paper.

The files and their theorem counts are:

Table 7: Coq proof stack: files and theorem counts.

File	Theorems	Role in this work
DeltaGap.v	22	δ definition, shift/scale invariance (§3)
DeltaToRho.v	10	δ stability $\Rightarrow$ ρ stability (§3, §4)
TailSaturation.v	10	ρ stability from δ stability (§4)
QuantileMargin.v	22	Quantile-margin theorem, 0 axioms (§4)
BehavioralGenesisBounds.v	16	CV bounds from Bhatia-Davis (§3)
EffectiveDimension.v	5	Curvature variance from d_{eff} (§4)
SimplexCollisionBounds.v	27	Tail gene R bounds (§4)
VertexResidualDecomposition.v	6	Tail collision $C(q)$ (§4)
SimplexTail2.v	14	Tail-2 decomposition (§4)
GaugeTransport.v	33	Gauge–Transport Decomposition, CJ-29 (§4.3)
EquationOfState.v	22	Softcap properties (§4.2)
NoSpoofing.v	51	Cramér-Rao impossibility (§6.3)
ModelFingerprinting.v	12	Forensic model identification (§5)
RenyiH2Bounds.v	23	Rényi entropy bounds (§4)
HellingerAffinity.v	10	Hellinger affinity bounds (§4)
ScalingLaws.v	28	Stiffness bounds, scaling security chain (§4.2)

Axiom policy. The four files forming the core $\delta \to \rho$ stability chain — DeltaGap.v, DeltaToRho.v, TailSaturation.v, and QuantileMargin.v — contain zero axioms. The two largest theorem files — GaugeTransport.v (33 theorems) and NoSpoofing.v (51 theorems) — also contain zero axioms. Across all 16 files, 40 axioms appear (counting both `Axiom` and `Parameter` declarations, which are semantically equivalent in Coq). Of these, 38 are structural: parameter constraints (e.g., `R_mean_positive : R_mean > 0`) and standard mathematical function properties (e.g., `tanh_bounded : forall x, -1 <= tanh x <= 1`) that Coq's standard library does not provide natively. These encode properties that are theorems in real analysis but not yet mechanized in the Coq ecosystem — they are scaffolding, not substantive assumptions. The remaining 2 axioms, both in ScalingLaws.v, are empirically grounded: `stiffness_bounded_below` ($S_{\min} = 1.1797 > 0$, validated across 12 models spanning a 147× parameter range, three vendor families) and `delta_positive_from_stiffness` ($\delta > 0$ when $S \geq S_{\min}$, validated via the Equation of State). These are the only axioms that encode claims about the physical world rather than mathematical properties, and they are disclosed as such.

8.2 Key Theorems

Four theorems anchor the paper's central claims. Each is stated here in informal mathematical language; the formal Coq statements are reproduced in Appendix C.

delta_implies_rho_stability (DeltaToRho.v). If δ_{raw} is stable across temperatures (CV $\leq \eta$), then the collision probability $\rho = s^2 + (1-s)^2$ where $s = \sigma(\delta/T)$ is also stable, with explicit bounds on $|\rho(T_1) - \rho(T_2)|$ depending on η and the operating temperature range. This is the formal backbone of the temperature invariance claim in §3.

CJ29_main (GaugeTransport.v). The Gauge–Transport Decomposition: $\Sigma_{\Delta\phi} = J \cdot \Sigma_{\Delta\theta} \cdot J^T$, with gauge fraction $g \in [0, 1]$ measuring the proportion of activation variance absorbed by the softmax-invariant subspace. Six supporting lemma groups establish covariance transport (L1), softmax gauge invariance (L3), projector properties (L4), rank inequality (L5), DPO covariance identity (L6), and participation ratio bounds (L7). Corollaries derive the forensic findings of §5 — function-null gauge scars, collapse mechanisms, and provenance signatures — as special cases of this single object. Zero axioms.

T4_no_spoofing_interval_split (NoSpoofing.v). The security theorem of §6.3: for any KL budget K, either the residual fingerprint gap exceeds the acceptance threshold ε, or the accumulated noise from parameter perturbation exceeds ε. The proof proceeds by interval splitting — the real line of possible KL budgets is partitioned into regions where one or the other floor condition dominates, and the union is exhaustive. This formalizes the empirically observed three-phase Pareto cliff (free drift, destruction, trapped). Zero axioms.

`quantile_margin_stability` (QuantileMargin.v). For distributions with bounded quantile functions, the collision-probability function $h(u) = \sigma(u)^2 + (1-\sigma(u))^2$ inherits stability from quantile stability. This theorem requires zero axioms — it is proved entirely from Coq's real number library with no external assumptions. It provides the formal link between observable stability (quantile margins) and collision probability bounds.

8.3 Epistemological Classification

Every claim in this work falls into exactly one of four categories. The classification is maintained rigorously throughout the paper.

Table 8: Epistemological classification of claims.

Status	Meaning	Test	Examples
[PROVEN]	Machine-checked Coq theorem	`.vo` compiles under Rocq 9.1	$\delta \to \rho$ stability, CJ-29, T4
[CITED]	Published mathematics, directly applicable	Peer-reviewed paper exists	Fannes-Audenaert, Bhatia-Davis, Cramér-Rao
[DERIVED]	Computed from **[PROVEN]** or **[CITED]** foundations	Derivation documented	Cantelli k, detection thresholds
[VALIDATED]	Empirical law with documented evidence	Experiment + analysis	EOS ($R^2 = 0.40$), $\delta_{\text{norm}} \approx 0.318$

The cardinal rule: **[VALIDATED] is never presented as [PROVEN].** The Equation of State (§4.2) is **[VALIDATED]** — it is a regression fit, not a theorem. The impossibility result (§6.3) is **[PROVEN]** — it is a Coq theorem, not an empirical observation. The acceptance threshold ε (§7) is **[DERIVED]** from measurement variance, not selected by heuristic. Conflating these categories is how systems rot; maintaining them is how trust is built.

This classification serves a practical purpose beyond intellectual honesty. A reviewer who asks "is the Equation of State proven?" receives a clear answer: no, it is validated with $R^2 = 0.40$ at $n = 38$, and the paper says so explicitly. A reviewer who asks "is the impossibility result just empirical?" receives an equally clear answer: no, it is 51 machine-checked theorems with zero admitted proofs. The distinction preempts both overstatement and understatement.

9 Limitations and Open Questions

This section documents the boundaries of our current results. Honest accounting of what remains unproven is not a weakness — it is the mechanism by which trust is built.

9.1 Limitations

Attack surface coverage. Our adversarial validation (§6) tests LoRA-based parameter perturbation across multiple ranks, module targets, and optimization strategies. Three important attack classes remain untested: knowledge distillation (training a student model to mimic the target's outputs), structured pruning (removing parameters while preserving function), and adversarial fine-tuning with access to the challenge bank. The per-component Cramér-Rao bound (S1) applies universally: any model whose output distribution is close to the target's pays a minimum Fisher cost per fingerprint component, regardless of how it was produced. However, T4's compound impossibility — the interval-splitting argument showing no KL budget avoids both gap and noise floors — specifically assumes parameter perturbation from a known starting point. Distillation produces a new model rather than perturbing the original, so while S1 constrains the student, T4's compound result does not directly apply. Whether a student model independently converges to a fingerprint matching its teacher's — or to a distinct fingerprint despite output-distributional proximity — is an open empirical question.

Model scale. The underlying fingerprint theory has been validated from 135M to 47B total parameters (the latter a mixture-of-experts model with approximately 13B active parameters per token) across six architecture families including dense Transformers, mixture-of-experts, RWKV, and Mamba variants (§3); the authentication protocol has been validated on 23 models from 16 vendor families spanning 410M to 7B (§7). Separately, the Equation of State primitives have been measured across a 147× parameter range (0.5B–72B, three vendor families), confirming that stiffness S remains bounded ($[1.18, 3.59]$) and the security chain $S_{\min} > 0 \Rightarrow \delta_{\min} > 0 \Rightarrow K_{\min} > 0$ holds at all tested scales (ScalingLaws.v, §4.2). Scaling behavior to frontier models (100B+ parameters) is unknown. The theoretical framework makes no scale-dependent assumptions — the Gumbel PPP structure depends on effective competitor count (a function of vocabulary size and temperature), not parameter count — but empirical confirmation at frontier scale has not been performed.

Authentication protocol architecture coverage. The IT-PUF validation zoo (§7) includes 23 models spanning three architecture types: 21 dense Transformers, one pure Mamba state-space model (Falcon-Mamba 7B), and one parallel-attention Transformer (Phi-2). Adversarial spoofing attacks (§6) have been tested only against Transformers; the formal impossibility result (§6.3) applies regardless of architecture, but the Fisher-protected observable's hardness is saturation-dependent (§6.2): Transformer models operate near saturation ($g \approx 0.91$), while Mamba models do not ($g \approx 0.39$–0.65), which means the per-component spoofing cost differs. The protocol addresses this through architecture-aware observable selection. Extending adversarial testing to non-Transformer targets remains a near-term priority.

Equation of State explanatory power. The Equation of State (§4.2) achieves leave-one-out $R^2 = 0.40$ at $n = 38$. This is sufficient to confirm that the four primitives (S, D, V, C) drive δ in the predicted directions, but it is not sufficient for precise prediction of a new model's fingerprint from weights alone. The unexplained 60% of variance likely reflects higher-order interactions between primitives and training-recipe effects not captured by the current functional form. The Equation of State is a validated empirical law, not a complete theory. Importantly, the security chain (§6) does not depend on EOS predictions: the Cramér-Rao bound, the interval-splitting theorem, and the IT-PUF protocol all operate on *measured* δ from actual model inference. The EOS contributes physical understanding of why δ varies across models, not the security guarantee itself.

Representation topology requires inference. The three-axis forensic framework (§5) includes one axis — gradient starvation / representation topology — that is not weight-auditable. Determining whether a model exhibits collapsed or dispersed hidden-state geometry requires forward passes through the network. The other two axes (output-layer thermodynamics and gauge scarring) are partially or fully accessible from static weights. A complete weight-only forensic profile remains an open goal.

Conditionality of the impossibility result. Theorem T4 (§6.3) proves that *if* a crossover point exists between the residual-gap floor and the noise floor, then no KL budget avoids both. The existence of this crossover is empirically validated for the observable and architecture pairs tested, but T4 does not prove crossover existence unconditionally. A model with unusual Fisher geometry could in principle lack such a crossover, which would make the impossibility result inapplicable. We have not encountered such a case, and the Cramér-Rao mechanism (S1) provides a per-component cost floor regardless, but the compound impossibility (T4) is conditional on a structural property that we verify empirically rather than prove in general.

Challenge bank secrecy. The authentication protocol (§7) rests on two independent legs: per-challenge hardness (Leg 1, formal) and challenge unpredictability (Leg 2, operational). Leg 2 assumes the challenge bank remains secret. If an adversary obtains the bank contents, the protocol degrades to a static multi-component fingerprint — still individually Fisher-protected per Leg 1, but no longer unpredictable. The system degrades gracefully rather than catastrophically, but the entropy guarantee ($H_\infty = 25.86$ bits) depends on Leg 2 holding.

Adversarial entropy scaling. No attack in our validation reached the acceptance threshold, which means the independence ratio R — measuring how adversarial success probability scales with the number of challenge components — is undefined at the success boundary. The min-entropy bound ($H_\infty = 25.86$ bits) is derived from population separation statistics, not from adversarial testing at the acceptance frontier. Whether an attacker who *could* spoof a single component faces multiplicative or sub-multiplicative difficulty across components remains unmeasured. The formal result (T7) proves budget exhaustion for $d \leq 3$ components; the general-d case is a conjecture supported by the multi-seed empirical result ($3.8\times$ hardening from 1 to 4 seeds) but not formally proven.

9.2 Open Questions

Optimal challenge selection. The current protocol selects challenge subsets via seeded pseudorandom permutation. Determinantal Point Process (DPP) based selection could maximize geometric diversity across the response space, potentially increasing the effective entropy per challenge. The measured coefficient of variation ($\approx 25\%$) across prompts confirms that sufficient diversity exists in the bank; the question is whether structured selection can extract more of it.

Zero-knowledge attestation. The current protocol requires the verifier to know the enrolled fingerprint in cleartext. A zero-knowledge construction — proving that a model's response matches an enrolled template without revealing either the response or the template — would enable third-party attestation without exposing proprietary model internals. ZK-SNARK or ZK-STARK circuits over the verification computation are architecturally feasible given the protocol's low dimensionality (tens of floating-point comparisons), but have not been implemented.

Training dynamics. How does δ evolve during training? The current work treats δ as a property of trained models. Understanding its trajectory during pre-training — when it stabilizes, how it responds to learning rate schedules and data mixing — could reveal whether fingerprint formation is gradual or phase-transition-like, with implications for both theory and forensic dating of model checkpoints.

Expanded validation. The current authentication zoo of 23 models from 16 vendor lineages is sufficient for the statistical claims made but is not exhaustive. Edge cases of particular interest include: models trained with extreme quantization-aware training, models produced by iterative self-distillation, and multilingual models where vocabulary structure differs substantially from the English-dominated training sets tested here.

10 Conclusion

The raw logit gap $\delta_{\text{raw}} = z_{(3)} - z_{(4)}$ is a temperature-invariant architectural constant determined by the output geometry of neural language models. By adapting the hardware PUF paradigm to neural network inference, we construct an IT-PUF that translates this thermodynamic property into a cryptographic authentication primitive.

The theoretical framework bridges extreme value theory, Fisher information geometry, and machine learning physics. The Equation of State connects δ to weight-readable primitives. The Gauge–Transport Decomposition $\Sigma_{\Delta\varphi} = J_\varphi \Sigma_{\Delta\theta} J_\varphi^T$ unifies all forensic observables—from kernel fingerprints to training provenance signatures—under a single mathematical object, showing that what appears to be a collection of independent empirical findings is in fact a single transport law acting at different scales. We validate this framework across three independent model families spanning a $147\times$ parameter range (0.5B to 72B), falsifying the assumption of unbounded stiffness but discovering a rigorously validated empirical floor ($S_{\min} = 1.1797$). Because this floor is strictly positive, the Cramér-Rao bound ensures the spoofing cost remains bounded away from zero across all tested scales.

Spoofing this fingerprint is not merely difficult; it is structurally impossible at useful fidelity. The interval-splitting theorem proves that at every KL budget, either the residual fingerprint gap or the accumulated perturbation noise exceeds the acceptance threshold—there is no sweet spot between them. Empirically, the strongest tested adaptive attack fails to close beyond 10.7× the acceptance threshold before inducing model destruction, and the IT-PUF protocol yields zero false acceptances across 23 models and 16 families.

These results rest on machine-checked foundations: 311 theorems across 16 Coq files, with zero uses of `Admitted` and zero vacuous definitions. Every claim in this work is classified as **[PROVEN]**, **[CITED]**, **[DERIVED]**, or **[VALIDATED]**, and the distinction is maintained throughout.

The authentication protocol, formal impossibility proofs, and measurement methodology are the subject of U.S. Provisional Patent Application 63/982,893. We invite collaboration on the open questions identified in §9—particularly zero-knowledge attestation and optimal challenge selection via determinantal point processes—as we work toward a mathematically verifiable foundation for AI provenance.

A Notation Reference

This appendix collects all notation used in the paper. Symbols are grouped by domain; within each group, definitions appear in the order they are introduced in the text.

A.1 Logit Geometry

Symbol	Definition	Introduced
$\mathbf{z}_t = W_U \mathbf{h}_t$	Logit vector at timestep t: unembedding matrix W_U applied to hidden state $\mathbf{h}_t$	§3.2
$z_{(1)} \geq z_{(2)} \geq \cdots \geq z_{(V)}$	Order statistics of the logit vector (descending)	§3
$G_k := z_{(k)} - z_{(k+1)}$	Gap between k-th and $(k+1)$-th highest logits	§3
$\delta_{\text{raw}} := G_3 = z_{(3)} - z_{(4)}$	The δ-gene: raw logit gap between 3rd and 4th positions	§3.1
$\delta_{\text{norm}} := G_3/(G_2 + G_3 + G_4)$	Normalized gap (scale-free); predicted value 0.318	§3.3
T	Softmax temperature	§3.2
$p_i = \text{softmax}(\mathbf{z}/T)_i$	Token probability after temperature scaling	§3.2
CV_T	Coefficient of variation of $\bar{\delta}_T$ across temperatures	§3.2
V	Vocabulary size	§4.2

A.2 Tail Statistics

Symbol	Definition	Introduced
$\beta_{\text{robust}} := \text{median}(k \cdot G_k)_{k \geq 2}$	Robust tail scale estimator (excludes selection-biased G_1)	§3.3
$\beta_{\text{true}} = \beta_{\text{robust}} / \ln 2$	PPP scale parameter (median-to-mean conversion)	§3.3
$\sigma(u) = 1/(1 + e^{-u})$	Standard sigmoid function	§4.1
$s = \sigma(\delta/T)$	Sigmoid of temperature-scaled gap	§4.1
$\rho = s^2 + (1 - s)^2$	Collision probability (derived from δ via sigmoid)	§4.1
$h(u) = \sigma(u)^2 + (1 - \sigma(u))^2$	Collision probability function of scaled gap	§4.1

A.3 Equation of State Primitives

Symbol	Definition	Direction	Introduced
$S = \gamma \times \sigma_w$	Stiffness: normalization gain × unembedding row-norm scale	$+$	§4.2
γ	Final normalization layer gain (RMSNorm or LayerNorm)	—	§4.2
σ_w	RMS of W_U row L^2 norms	—	§4.2
D	Crowding: mean cosine similarity to $k = 10$ nearest neighbors in W_U	$-$	§4.2
V	Pressure: vocabulary size	$+$	§4.2
C	Constraint: training-time logit soft-cap value	$-$	§4.2
L	Network depth (number of layers)	$+$	§4.2
$\nVdash_{\text{fcap}}$	Indicator for presence of final soft-cap	—	§4.2

A.4 Gauge–Transport Decomposition

Symbol	Definition	Introduced
$\theta \in \mathbb{R}^p$	Model parameters	§4.3
$\phi = F(\theta) \in \mathbb{R}^m$	Observable field (logits, hidden states, etc.)	§4.3
$J_F = \partial F / \partial \theta$	Jacobian of observable field w.r.t. parameters	§4.3
$\Sigma_{\Delta\theta}$	Covariance of parameter perturbation	§4.3
$\Sigma_{\Delta\phi} = J_F \Sigma_{\Delta\theta} J_F^\top$	Covariance transport equation (L1)	§4.3
$\mathcal{G} = \operatorname{span}\{\mathbf{1}\}$	Gauge subspace (softmax-invariant directions)	§4.3
$\Pi_{\mathcal{G}}$	Orthogonal projector onto $\mathcal{G}$	§4.3
$g_\phi := \operatorname{tr}(\Pi_{\mathcal{G}} \Sigma_{\Delta\phi}) / \operatorname{tr}(\Sigma_{\Delta\phi})$	Gauge fraction	§4.3
$\Sigma_{w-l} = \Sigma_w + \Sigma_l - \Sigma_{wl} - \Sigma_{lw}$	DPO gradient covariance identity (L6)	§4.3
$\operatorname{erank}_{\text{PR}}(\Sigma) := \operatorname{tr}(\Sigma)^2 / \operatorname{tr}(\Sigma^2)$	Participation ratio (continuous rank surrogate, L7)	§4.3
$J_{\ell \to L}$	Suffix Jacobian from layer ℓ to output layer L	§5.2
$A_\ell := J_{\ell \to L}^\top J_{\ell \to L}$	Jacobian energy Gram matrix	§5.2
$\Sigma_g^{(\ell)}$	Gradient covariance at layer ℓ	§5.2

A.5 Forensic Hierarchy

Symbol	Definition	Introduced
$\delta_{\text{norm}}, \beta_{\text{robust}}$	Axis 1 (Thermodynamics) observables	§5.1
μ_{W_U}	Mean row of W_U; gauge scar parameter (Axis 2)	§5.1
$D(y_t)$	Distribution of post-normalization hidden states (Axis 3)	§5.1
$\|\Delta W\| / \|W\|$	Relative weight-delta magnitude (provenance metric)	§5.3
d	Cohen's d effect size	§5

A.6 Security and Fisher Information

Symbol	Definition	Introduced
$g(\theta)$	Scalar observable (function of model parameters)	§2.4
$I(\theta)$	Fisher information matrix	§2.4
$s_F^2 = (\nabla g)^\top I(\theta)^{-1}(\nabla g)$	Fisher sensitivity (squared)	§2.4
$\tilde{s}_F = s_F/\sigma_F$	Normalized Fisher sensitivity (dimensionless hardness)	§6.2
K	KL divergence budget	§6.3
K_x	Crossover KL budget (gap floor meets noise floor)	§6.3
ε	Acceptance threshold ($\sim 10^{-4}$, derived from genuine variance)	§7.2
$g_1, \ldots, g_d$	Per-component fingerprint gaps	§6.3
$I_1, \ldots, I_d$	Per-component Fisher information values	§6.3

A.7 IT-PUF Authentication

Symbol	Definition	Introduced
H_∞	Min-entropy of the challenge-response protocol (bits)	§7.4
$\mathrm{CV}_{\mathrm{prompt}}$	Within-model coefficient of variation across prompts	§7.3
$F(\theta, x_i)$	Protected observable measured at prompt x_i	§7.3
$x = x^2 \Rightarrow x \in \{0, 1\}$	No-cloning algebraic identity (ancestor of interval-splitting)	§6.3

A.8 Position Mapping

The δ-gene lives in the tail of the logit distribution. The position mapping between the full distribution and the tail (excluding the selection-biased winner) is:

Full position	Tail position	Variable	Role
1 (winner)	— (excluded)	—	Selection-biased
2 (runner-up)	Tail 1	G_2	Runner-up gap
3 (third)	Tail 2	$G_3 = \delta_{\mathrm{raw}}$	**The δ-gene**
4 (fourth)	Tail 3	G_4	Fourth gap
5 (fifth)	Tail 4	G_5	Fifth gap

A.9 Abbreviations

Abbreviation	Expansion
BDI	Bulk Dispersion Index (quantile-normalized variance share of non-top tokens)
CV	Coefficient of Variation
DPO	Direct Preference Optimization
EOS	Equation of State
EVT	Extreme Value Theory
GQA	Grouped Query Attention
IT-PUF	Inference-Time Physical Unclonable Function
KL	Kullback–Leibler (divergence)
LOO	Leave-One-Out (cross-validation)
MoE	Mixture of Experts
PPP	Poisson Point Process
PSD	Positive Semi-Definite
PUF	Physical Unclonable Function
RWKV	Receptance Weighted Key Value (architecture)
SFT	Supervised Fine-Tuning
SSM	State-Space Model

B Falsified Hypotheses

Scientific credibility rests not only on what a program proves but on what it tests and fails. This appendix documents hypotheses that were formulated, tested against data, and rejected. Each falsification simultaneously strengthened the surviving theory by eliminating a candidate mechanism. The entries are ordered thematically, grouping related falsifications, though they span the period from early January to early February 2026.

B.1. Head dimension predicts δ. At $n = 29$ models, the attention head dimension $H = d_{\text{model}}/n_{\text{heads}}$ showed strong bivariate correlation with δ_{raw} ($r = +0.75$, $p < 0.001$), and was included as a predictor in the Equation of State. At $n = 38$, the correlation collapsed to $r = +0.16$. Root cause: three models from a single vendor with atypically large H values (144–320, versus 64–128 for all others) acted as leverage points. Removing these three models yielded $r = +0.017$. The predictor was removed from the Equation of State; the four surviving primitives (S, D, V, C) all replicated at $n = 38$ with directional predictions confirmed (§4.2, table 4).

Lesson: Leverage-point artifacts are silent at small n. Every claimed predictor must survive expansion.

B.2. Architecture determines output-layer dynamics. The hypothesis that recurrence-based architectures (Mamba, RWKV) would show fundamentally different output-layer coupling than attention-based architectures was tested by measuring temporal coupling $\rho(k)$ across 10 models spanning five architecture families. A pure Mamba model (Mamba 2.8B, Gu & Dao's original implementation) showed $\rho(50) = 0.764$ — squarely within the Transformer range (0.55–0.87). Only one model, Falcon-Mamba 7B, showed anomalous decoupling ($\rho = 0.165$). The architecture hypothesis was falsified; a controlled comparison isolating training infrastructure identified the anomaly as a kernel-level training artifact (§5.1).

Lesson: Architecture is a weaker determinant of output-layer behavior than training infrastructure. The output pathway is the common element.

B.3. Training data determines output-layer dynamics. A model trained on the same data as the anomalous Falcon-Mamba (Falcon 3 7B, sharing the RefinedWeb corpus) showed normal coupling ($\rho = 0.817$). Combined with B.2, this completed the triangulation: neither architecture nor training data caused the anomaly — only the specific interaction of custom training kernels with the Mamba architecture produced the observed decoupling.

Lesson: Confound isolation requires testing each variable independently. The causal variable was the one we tested last.

B.4. Post-hoc logit capping induces representation collapse. Models trained with logit soft-caps showed collapsed representation geometry (BDI > 0.82, effective rank < 12). The hypothesis that applying soft-caps at inference time could induce the same collapse was tested by sweeping tanh caps from $c = 30$ to $c = 5$ on a non-collapsed model. At $c = 5$, with 94% of logits hitting the cap and KL divergence of 6.5 nats (model functionally destroyed), BDI reached only 0.703 — still within the normal regime, and 0.12 below the High-BDI threshold. The gap to the collapsed regime could not be closed by any post-hoc intervention.

Lesson: Representation topology is determined during training, not at inference. Post-hoc interventions modify outputs without altering the learned state geometry.

B.5. Linguistic token type predicts thermal status. The hypothesis that function words (the, is, of) generate thermally while content words (nouns, verbs) crystallize was tested across 1,165 classified tokens. Function words showed 58.9% thermal; content words showed 61.3% thermal. Cohen's $d = 0.019$ (negligible). The 60% thermal fraction is invariant to linguistic category, confirming it arises from training dynamics (the cross-entropy equilibrium at $\beta \approx 1.45$), not from distributional properties of the vocabulary.

Lesson: When a constant appears across all conditions, the mechanism is in the physics (training dynamics), not the data (linguistic structure).

B.6. Static multi-site measurement provides sufficient entropy. The initial authentication architecture measured a Fisher-protected observable at multiple sites within the network (different layers and measurement types), producing a multi-component enrollment vector. Testing across 13 models revealed that the observable behaves as an order parameter: it saturates to a characteristic constant determined by architecture family and training recipe, with minimal variation across measurement positions. The multi-component vector had approximately one effective dimension of entropy — insufficient for cryptographic use.

This was the most consequential falsification in the program, triggering a complete architectural pivot from static measurement to challenge-response authentication (§7). The replacement architecture derives entropy from the challenge space (which inputs to present) rather than from physics diversity (where to measure), while preserving the same hardness guarantee per measurement.

Lesson: Hardness and entropy are orthogonal properties. An observable can be extremely difficult to spoof (5–15 nats KL) while providing near-zero population entropy. Authentication requires both.

B.7. Vector correlation validates challenge-response independence. An intermediate validation metric computed the Pearson correlation of response vectors across different random seeds. The observed $r = 0.88$ appeared to indicate near-identical fingerprints, suggesting the challenge-response protocol provided no additional entropy over the static approach. Investigation revealed three methodological flaws: coordinate misalignment across seeds (each seed selects different prompts, so vector components are not comparable), a low-variance artifact inflating correlation coefficients, and testing the wrong null hypothesis (high correlation between vectors does not imply high collision probability in L^2 distance). Corrected methodology using direct collision measurement at a threshold derived from genuine measurement variance confirmed the protocol's security: zero false acceptances across 1,012 cross-model comparisons (§7.4).

Lesson: The validation metric must test the same quantity the protocol uses. Correlation and collision probability are different mathematical objects. This episode — in which a flawed diagnostic nearly terminated a sound protocol — illustrates why methodological rigor must extend to validation procedures, not just to the system under test.

Table 9: Falsification summary.

#	Hypothesis	Test	Key statistic	Verdict	Section
B.1	H predicts δ	$n = 38$ replication	$r = +0.017$ (excl. leverage)	Falsified	§4.2
B.2	Architecture → coupling	Mamba 2.8B control	$\rho = 0.764$ (Transformer-like)	Falsified	§5.1
B.3	Training data → coupling	Falcon 3 7B control	$\rho = 0.817$ (normal)	Falsified	§5.1
B.4	Post-hoc caps → collapse	Cap sweep $c = 5$–30	BDI $= 0.703$ at $c = 5$	Falsified	§5.2
B.5	Token type → thermal	1,165 tokens	$d = 0.019$	Falsified	§4.1
B.6	Static anchor entropy	13 models, multi-site	∼1D eff. entropy	Falsified	§7.1
B.7	Correlation = collision	Seed-cross-validation	$r = 0.88$ (wrong test)	Falsified	§7.4

C Selected Coq Theorem Statements

This appendix reproduces the formal statements of six theorems cited in the main text. Each is presented first in mathematical notation and then as a Coq type signature (declaration only, not proof term). All six compile under Rocq 9.1 with zero uses of `Admitted` and zero axioms. The four theorems promised in §8.2 appear first (C.1–C.4), followed by two additional results that underpin the gauge framework and multi-component security analysis (C.5–C.6).

C.1 δ Stability Implies ρ Stability

File: `DeltaToRho.v` | **Name:** `delta_implies_rho_stability`

Mathematical statement. Let $\delta_{\min} \geq 0$ be a lower bound on the raw logit gap, $\eta \in [0, 1]$ the tail-approximation error, and $[T_{\min}, T_{\max}]$ the operating temperature range. If two collision probabilities ρ_1, ρ_2 each satisfy $|\rho_i - h(\delta_{\min}/T_i)| \leq 2\eta$ where $h(u) = \sigma(u)^2 + (1 - \sigma(u))^2$, then

$$|\rho_1 - \rho_2| \;\leq\; 4\eta + 4\exp(-\delta_{\min}/T_{\max}). \tag{19}$$

```
Theorem delta_implies_rho_stability :
  forall (params : DeltaStabilityParams) (T1 T2 rho1 rho2_val : R),
  Tmin params ≤T1 ≤Tmax params →
  Tmin params ≤T2 ≤Tmax params →
  Rabs (rho1 - h (delta_min params / T1)) ≤2 * eta params →
  Rabs (rho2_val - h (delta_min params / T2)) ≤2 * eta params →
  Rabs (rho1 - rho2_val) ≤4 * eta params +
    4 * exp (- delta_min params / Tmax params).
```

C.2 Gauge–Transport Decomposition (CJ-29 Main Theorem)

Coq type signature: **File:** `GaugeTransport.v` | **Name:** `CJ29_main`

Mathematical statement. Given a parameter covariance $\Sigma_{\Delta\theta}$ (PSD, traces $\mathrm{tr}(\Sigma)$ and $\mathrm{tr}(\Sigma^2)$, rank r_θ), a transport operator with factor $\|J\|^2$ and rank r_J, the resulting field covariance $\Sigma_{\Delta\phi}$ with gauge trace $\mathrm{tr}(\Pi_{\mathcal{G}}\Sigma_{\Delta\phi})$ satisfies:

(i) $\Sigma_{\Delta\phi}$ is PSD (trace ≥ 0); (ii) If $\mathrm{tr}(\Sigma_{\Delta\phi}) > 0$, the gauge fraction $g_\phi = \mathrm{tr}(\Pi_{\mathcal{G}}\Sigma_{\Delta\phi}) / \mathrm{tr}(\Sigma_{\Delta\phi}) \in [0, 1]$; (iii) $\mathrm{rank}(\Sigma_{\Delta\phi}) \leq \mathrm{rank}(\Sigma_{\Delta\theta})$ and $\mathrm{rank}(\Sigma_{\Delta\phi}) \leq \mathrm{rank}(J_F)$.

```
Theorem CJ29_main : forall cfg : GaugeTransportConfig,
  psd_trace (gtc_tr_field cfg) ∧
  (gtc_tr_field cfg > 0 →
    0 ≤gtc_gauge_fraction cfg ≤1) ∧
  (gtc_rank_field cfg ≤gtc_rank_source cfg)%nat ∧
  (gtc_rank_field cfg ≤gtc_rank_jacobian cfg)%nat.
```

C.3 No-Spoofing Interval Split (Security Theorem)

Coq type signature: **File:** `NoSpoofing.v` | **Name:** `T4_no_spoofing_interval_split`

Mathematical statement. Let $g > \varepsilon > 0$ be the initial fingerprint gap and acceptance threshold. Let $I_{\text{eff}} > 0$ be the effective Fisher information and $\gamma, \nu > 0$ noise parameters. Define the coherence-loss function $\psi(K) := \gamma K/(1+\gamma K)$, which increases from 0 to 1 as K grows — representing the fraction of the original model's coherence destroyed by a perturbation of KL cost K. The noise amplitude is $\nu \cdot \psi(K)$, saturating at ν. If there exists a crossover budget $K_x \geq 0$ at which both (a) the residual gap $g - \sqrt{2K_x I_{\text{eff}}} \geq \varepsilon$ and (b) the noise amplitude $\nu \cdot \psi(K_x) \geq \varepsilon$, then for *every* KL budget $K \geq 0$:

$$d^2_{\text{eff}}(K) \;:=\; \left(g - \sqrt{2K \cdot I_{\text{eff}}}\right)^2 + \left(\nu \cdot \psi(K)\right)^2 \;>\; \varepsilon^2. \tag{20}$$

No value of K simultaneously closes the gap and avoids detection by noise. The saturating noise model is strictly more conservative than an unbounded model: it requires $\nu > \varepsilon$ (the noise floor itself must exceed the acceptance threshold), rather than relying on noise growing without bound.

```
Theorem T4_no_spoofing_interval_split :
  forall gap I_eff gamma noise_scale epsilon K_x,
  epsilon > 0 →gap > epsilon →I_eff > 0 →
  gamma > 0 →noise_scale > 0 →K_x ≥0 →
  gap - sqrt (2 * K_x * I_eff) ≥epsilon →
  noise_amplitude K_x gamma noise_scale ≥epsilon →
  forall kl, kl ≥0 →
    effective_distance_sq gap
      (directed_closing kl I_eff)
      (noise_amplitude kl gamma noise_scale)
    > epsilon * epsilon.
```

C.4 Quantile-Margin Stability

Coq type signature: **File:** `QuantileMargin.v` | **Name:** `quantile_margin_stability`

Mathematical statement. For a non-empty list of non-negative gaps $(\delta_1, \ldots, \delta_n)$ satisfying a quantile condition (at least fraction $1-q$ of gaps exceed δ_q), and temperatures $T_1, T_2 \in (0, T_{\max}]$, the average h-function difference satisfies

$$\frac{1}{n}\sum_{i=1}^{n} |h(\delta_i/T_1) - h(\delta_i/T_2)| \;\leq\; 4\exp(-\delta_q/T_{\max}) + q/2. \tag{21}$$

This theorem uses zero axioms — it is proved entirely from Coq's real number library.

```
Theorem quantile_margin_stability :
  forall (params : QuantileParams) (deltas : list R) (T1 T2 : R),
  deltas ≠[] →
  deltas_nonneg deltas →
  0 < T1 ≤qp_Tmax params →
  0 < T2 ≤qp_Tmax params →
  quantile_condition params deltas →
  Ravg (map (h_diff T1 T2) deltas) ≤
  4 * exp (- qp_delta_q params / qp_Tmax params) +
  qp_q params / 2.
```

C.5 Multi-Component Budget Exhaustion

Coq type signature: **File:** `NoSpoofing.v` | **Name:** `T7_pure_cr_floor_3`

Mathematical statement. For three fingerprint components with gaps $g_1, g_2, g_3 > 0$ and Fisher information values $I_1, I_2, I_3 > 0$, if the total cost to close all three gaps exceeds the KL budget K (i.e., $\sum_i g_i^2/(2I_i) > K$), then any shifts $s_1, s_2, s_3 \geq 0$ satisfying the budget constraint $\sum_i s_i^2/(2I_i) \leq K$ leave a strictly positive total residual:

$$\sum_{i=1}^{3}(g_i - s_i)^2 > 0. \tag{22}$$

At least one component remains unmasked. This is the formal basis for multi-seed security hardening.

```
Theorem T7_pure_cr_floor_3 :
  forall g1 g2 g3 I1 I2 I3 K s1 s2 s3,
  I1 > 0 →I2 > 0 →I3 > 0 →K ≥0 →
  g1 > 0 →g2 > 0 →g3 > 0 →
  total_closing_cost_3 g1 g2 g3 I1 I2 I3 > K →
  s1*s1/(2*I1) + s2*s2/(2*I2) + s3*s3/(2*I3) ≤K →
  0 ≤s1 →0 ≤s2 →0 ≤s3 →
  component_residual_sq g1 s1 +
  component_residual_sq g2 s2 +
  component_residual_sq g3 s3 > 0.
```

C.6 Cross-Entropy Gradient Sums to Zero (Function-Null Lemma)

Coq type signature: **File:** `GaugeTransport.v` | **Name:** `L3_ce_gradient_sums_to_zero`

Mathematical statement. For cross-entropy loss with softmax outputs, the logit-space gradient $\nabla_z \mathcal{L} = p - e_{\text{target}}$ satisfies $\mathbf{1}^\top(p - e_{\text{target}}) = 0$, since both the probability vector p and the one-hot target e_{target} sum to 1. This is the formal basis for the function-null property of gauge scars: any weight component producing a constant logit shift $c\mathbf{1}$ contributes zero to the gradient, because $\mathbf{1}^\top \nabla_z \mathcal{L} = 0$.

Coq type signature:

```
Theorem L3_ce_gradient_sums_to_zero :
  forall p_sum e_sum,
  probability_sum p_sum →
  one_hot_sum e_sum →
  p_sum - e_sum = 0.
```

Remark. The full proof terms for all six theorems, together with the remaining 305 declarations in the 16-file stack, compile under Rocq 9.1. Source files are available upon request (see Appendix D).

D Reproducibility Statement

Formal proofs. The 16 Coq files comprising 311 theorems compile under Rocq 9.1 (Coq's current release name) with zero uses of `Admitted` as a proof terminator. The `.v` source files are available upon request to the corresponding author. Verification requires only a standard Rocq 9.1 installation with the Coq standard library; no third-party dependencies are needed.

Scientific measurements. The measurement methodology for all empirical results — temperature stability (§3.2), Equation of State fitting (§4.2), Gumbel validation (§3.3), architecture validation (§3, table 2), forensic axis measurements (§5), and Fisher sensitivity analysis (§6.2) — is described in the main text in sufficient detail for independent implementation. Model weights for all tested architectures are publicly available from their respective repositories (Hugging Face, Ollama). Measurement requires only standard inference hardware (a single GPU with sufficient memory for the model under test) and standard scientific Python libraries (NumPy, PyTorch/JAX, SciPy).

Authentication protocol. The challenge-response authentication protocol described in §7 is covered by provisional patent Application 63/982,893 (filed February 13, 2026). A reference implementation of the protocol — including the prompt bank, seed selection algorithm, and enrollment/verification pipeline — is available under license for academic research and commercial evaluation. Contact the author for access.

What is and is not reproduced here. This paper discloses the scientific theory (the δ-gene, the Gauge–Transport Decomposition, the impossibility theorem) and the validation methodology (model zoo, attack configurations, statistical tests) in full. Operational parameters of the authentication system — prompt bank contents, bank size, the identity of the Fisher-protected observable, seed selection mechanics, and exact per-deployment threshold values — are withheld as trade secrets and protected by the provisional patent. Reproducing the *scientific claims* requires only the information in this paper. Reproducing the *authentication system* requires the reference implementation.

Canonical data files. Experimental results cited in the paper are derived from versioned JSON data files. The canonical source for cross-model validation statistics is a 23-model, 1,012-comparison JSON file spanning three architecture types; the canonical source for adversarial attack trajectories is a set of per-seed JSON files recording every optimization step. SHA-256 hashes of all canonical scripts and data files were recorded within 24 hours of patent filing and are available upon request.

References

Harald Cramér. *Mathematical Methods of Statistics.* Princeton University Press, 1946.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction.* Springer, 2006.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes.* Springer, 1983.

Ravikanth Pappu, Ben Recht, Jason Taylor, and Neil Gershenfeld. Physical one-way functions. *Science*, 297 (5589):2026–2030, 2002.

C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

Sidney I. Resnick. *Extreme Values, Regular Variation, and Point Processes.* Springer, 1987.

Ruoxi Shao et al. A survey on model fingerprinting methods. *arXiv preprint*, 2025.

Ruoxi Shao et al. ZeroPrint: Zero-shot fingerprinting via Fisher Jacobians. *arXiv preprint*, 2026.

G. Edward Suh and Srinivas Devadas. Physical unclonable functions for device authentication and secret key generation. In *Proceedings of the 44th Design Automation Conference (DAC)*, 2007.

Seungho Yoon et al. Weight-statistical fingerprinting of neural network lineage. *arXiv preprint*, 2025.

Jie Zhang et al. REEF: A framework for robust and efficient model fingerprinting. *arXiv preprint*, 2025.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for AI-generated text. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.