Geometric Metrics for MoE Specialization: From Fisher Information to Early Failure Detection

Dongxin Guo
The University of Hong Kong
Hong Kong, China
bettyguo@connect.hku.hk

Jikun Wu
Brain Investing Limited
Hong Kong, China
hk950014@connect.hku.hk

Siu Ming Yiu
The University of Hong Kong
Hong Kong, China
smyiu@cs.hku.hk

***Abstract*—Expert specialization is fundamental to Mixture-of-Experts (MoE) model success, yet existing metrics (cosine similarity, routing entropy) lack theoretical grounding and yield inconsistent conclusions under reparameterization. We present an information-geometric framework providing the first rigorous characterization of MoE specialization dynamics. Our key insight is that expert routing distributions evolve on the probability simplex equipped with the Fisher information metric, enabling formal analysis via Riemannian geometry. We prove that standard heuristic metrics violate parameterization invariance (Theorem 1), establish that specialization corresponds to geodesic flow with quantified approximation bounds (Theorem 2), and derive a failure predictor with theoretical threshold justification (Theorem 3). The framework introduces two principled metrics: *Fisher Specialization Index* (FSI) achieving $r = 0.91 \pm 0.02$ correlation with downstream performance, and *Fisher Heterogeneity Score* (FHS) predicting training failure at 10% completion with AUC$= 0.89 \pm 0.03$—outperforming validation-loss-based early stopping by 23% while requiring 40$\times$ fewer compute cycles. We validate intervention protocols achieving 87% recovery rate when FHS$>$1 is detected. Comprehensive experiments across language modeling (WikiText-103, C4), vision MoE (ImageNet), and scaling studies (8–64 experts, 125M–2.7B parameters) validate our theoretical predictions.**

***Index Terms*—Mixture-of-Experts, information geometry, Fisher information, expert specialization, neural network training dynamics, early failure detection**

I. Introduction

Mixture-of-Experts (MoE) architectures have become the dominant paradigm for efficiently scaling neural networks, powering state-of-the-art systems including Mixtral [1], DeepSeek-V3 [2], and GPT-4 [3]–[5]. By activating only a subset of parameters per input, MoE models achieve superior parameter efficiency while maintaining computational tractability [6], [7]. A critical phenomenon underlying MoE success is *expert specialization*—the emergence of functionally distinct experts that handle different input subpopulations [8], [9].

Despite extensive empirical documentation [10], [11] and recent theoretical progress [12], expert specialization lacks formal geometric characterization. Practitioners rely on ad-hoc metrics: cosine similarity between expert parameters [13], routing entropy [3], or gradient-based indicators [14]. Our theoretical analysis reveals fundamental limitations: these metrics are not parameterization-invariant, cannot predict specialization dynamics, and lack principled connections to optimization geometry. This gap becomes critical as MoE models scale—monitoring specialization health during multi-week training runs on thousands of GPUs requires reliable, theoretically-grounded metrics.

Motivating Example. Consider two MoE models with identical routing behavior but different weight parameterizations. Cosine similarity reports 0.71 for one and 0.45 for the other—a 37% discrepancy for functionally equivalent models. Such inconsistencies lead to erroneous conclusions about specialization health and potentially wasted compute when healthy training is mistakenly diagnosed as failing.

Information geometry [15], [16] provides a canonical framework for analyzing probability distributions on Riemannian manifolds. The Fisher information metric, uniquely invariant under sufficient statistics by Chentsov's theorem [17], has proven powerful for understanding neural network training [18]–[20]. We leverage this framework to provide the first formal geometric characterization of MoE specialization.

Key Insight. The softmax routing distribution over experts naturally defines a point on the probability simplex Δ^{n-1}—a well-studied statistical manifold [16]. Expert specialization during training corresponds to the trajectory of this point moving from the uniform distribution (no specialization) toward simplex vertices (full specialization). This geometric perspective enables principled metrics, predictive bounds, and actionable diagnostics.

Contributions. We make the following contributions:

- We prove heuristic metrics fail geometric invariance while Fisher-Rao distance maintains invariance (Theorem 1).
- We derive the Fisher metric on routing distributions and prove specialization corresponds to approximate geodesic flow with explicit bounds (Theorem 2).
- We establish failure prediction via FHS with theoretical threshold justification and **validated intervention protocols achieving 87% recovery rate** (Theorem 3, Proposition 2).
- We introduce FSI and FHS metrics validated across language modeling, vision MoE, and scaling to 64 experts and 2.7B parameters.

Code is available at https://github.com/airesearchrepo2025/fisher-moe.

II. Preliminaries

A. Notation

Let $\mathcal{M}$ denote a statistical manifold. We use $\mathbf{F}(\theta)$ for the Fisher information matrix at parameters θ, $d_{FR}(\cdot,\cdot)$ for Fisher-Rao distance, and Δ^{n-1} for the $(n-1)$-dimensional probability simplex. For MoE with n experts $\{E_i\}_{i=1}^n$ and router R, the marginal expert distribution is $\bar{p} \in \Delta^{n-1}$. We write $\|\cdot\|_F$ for Frobenius norm and $\mathrm{tr}(\cdot)$ for trace.

B. Information Geometry Foundations

The *Fisher information matrix* (FIM) at parameters θ is [15]:

$$\mathbf{F}(\theta)_{ij} = \mathbb{E}_{x\sim p_\theta}\left[\frac{\partial \log p_\theta(x)}{\partial \theta_i}\frac{\partial \log p_\theta(x)}{\partial \theta_j}\right]. \tag{1}$$

For the probability simplex $\Delta^{n-1} = \{p \in \mathbb{R}^n : p_i \geq 0, \sum_i p_i = 1\}$, the Fisher metric induces the *Fisher-Rao distance* [21]:

$$d_{FR}(p,q) = 2\arccos\left(\sum_{i=1}^n \sqrt{p_i q_i}\right). \tag{2}$$

Remark 1 (Spherical Representation)**.** *The mapping $\phi : \Delta^{n-1} \to S_+^{n-1}$ defined by $\phi(p)_i = \sqrt{p_i}$ isometrically embeds the simplex into the positive orthant of the unit sphere [16], transforming the Fisher metric into the standard spherical metric so that geodesics on the simplex become great circle arcs on S_+^{n-1}.*

C. Mirror Descent and Simplex Geometry

Mirror descent [22], [23] with negative entropy mirror map $\psi(p) = \sum_i p_i \log p_i$ yields updates via KL divergence on the simplex. This connection underpins our geodesic analysis: mirror descent in logit space approximately follows Fisher geodesics in probability space.

D. Mixture-of-Experts Architecture

An MoE layer with n experts $\{E_i\}_{i=1}^n$ and router R computes [3]:

$$\begin{aligned}\mathrm{MoE}(x) &= \sum_{i=1}^n g_i(x)E_i(x),\\ g(x) &= \mathrm{TopK}(\mathrm{softmax}(R(x)/\tau)),\end{aligned} \tag{3}$$

where $\tau > 0$ is the softmax temperature. The *marginal expert distribution* over dataset $\mathcal{D}$ is $\bar{p}_e = \mathbb{E}_{x\sim\mathcal{D}}[p(e|x)]$. Standard load-balancing losses [4] penalize routing imbalance: $\mathcal{L}_{\text{aux}} = \lambda \cdot n \sum_{i=1}^n \bar{p}_i \cdot f_i$, where f_i is the token fraction routed to expert i.

III. Methodology

A. Limitations of Heuristic Metrics

Theorem 1 (Non-Invariance of Heuristic Metrics)**.** *Let $\theta \mapsto \phi(\theta)$ be a smooth reparameterization. Then:*

(i) *Cosine similarity: $\cos(\theta_i,\theta_j) \neq \cos(\phi(\theta_i),\phi(\theta_j))$ in general.*

(ii) *Routing entropy $H(\mathit{softmax}(w))$ is not invariant under logit scaling.*

(iii) *Fisher-Rao distance $d_{FR}(p,q)$ is invariant under all sufficient statistic-preserving transformations.*

Proof. (i) For $\phi(\theta) = A\theta$ with invertible A: $\cos(A\theta_i, A\theta_j) \neq \cos(\theta_i,\theta_j)$ unless $A^\top A = cI$. Example: $A = \mathrm{diag}(1,2)$, $\theta_1 = (1,1)^\top$, $\theta_2 = (1,0)^\top$ yields 37% discrepancy. (ii) For $w' = \alpha w$ with $\alpha \neq 1$: $\mathrm{softmax}(w')_i \neq \mathrm{softmax}(w)_i$. (iii) By Chentsov's theorem [17]. □

Note that (i)–(ii) operate at the parameter and logit levels, while (iii) operates at the distribution level; our framework lifts analysis to the probability simplex where invariance holds.

B. Fisher Metric on Expert Assignment Distributions

Definition 1 (Expert Assignment Manifold)**.** *The expert assignment manifold $\mathcal{E}$ is int$(\Delta^{n-1}) = \{p \in \mathbb{R}^n : p_i > 0, \sum_i p_i = 1\}$ equipped with the Fisher metric from the categorical distribution, where $\mathbf{F}(p)_{ij} = \delta_{ij}/p_i$ [16].*

Definition 2 (Fisher Specialization Index)**.** *The Fisher Specialization Index (FSI) at training step t is:*

$$\mathit{FSI}(t) = d_{FR}(\bar{p}^{(t)}, u) = 2\arccos\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \sqrt{\bar{p}_i^{(t)}}\right), \tag{4}$$

where $u = (1/n,\ldots,1/n)$ is the uniform distribution. FSI measures geodesic distance from uniform routing toward concentrated routing, respecting intrinsic probability geometry.

Proposition 1 (FSI Bounds)**.** *For any $\bar{p} \in \Delta^{n-1}$: $0 \leq \mathit{FSI} \leq \mathit{FSI}_{\max} = 2\arccos(1/\sqrt{n})$.*

C. Specialization as Geodesic Evolution

Lemma 1 (Softmax Gradient Structure)**.** *Let $p = \mathit{softmax}(w/\tau)$ for logits $w \in \mathbb{R}^n$ and temperature $\tau > 0$. Then $\frac{\partial p}{\partial w} = \frac{1}{\tau}(\mathit{diag}(p) - pp^\top)$ and $\mathbf{F}(p)^{-1} = \mathit{diag}(p) = \tau^2 \frac{\partial p}{\partial w}\left(\frac{\partial p}{\partial w}\right)^\top + pp^\top$.*

Proof. Standard softmax derivative gives $\frac{\partial p_i}{\partial w_j} = \frac{1}{\tau}(p_i\delta_{ij} - p_i p_j)$. Let $J = \frac{1}{\tau}(\mathrm{diag}(p) - pp^\top)$. Then $\tau^2 JJ^\top + pp^\top = \mathrm{diag}(p) = \mathbf{F}(p)^{-1}$. □

Theorem 2 (Specialization-Geodesic Correspondence)**.** *Let $\bar{p}^{(t)}$ be the marginal distribution at step t. Under gradient descent on loss $\mathcal{L}$ with learning rate η:*

(i) ***Per-Step Geodesic Bound:*** *For softmax routing with temperature τ:*

$$\|\phi^{(t+1)} - \gamma(t+1)\|_{\mathit{step}} \leq \frac{\kappa\eta^2\|\nabla\mathcal{L}\|^2}{4\tau}, \tag{5}$$

where ϕ is the spherical representation (Remark 1), γ is the geodesic through $\phi^{(t)}$, and $\kappa = 1/4$ is the sectional curvature.

(ii) ***Monotonicity:*** *Without load balancing ($\lambda = 0$), FSI(t) is monotonically non-decreasing.*

(iii) ***Equilibrium Bound:*** *With load balancing ($\lambda > 0$), $d_{FR}(\bar{p}^*, u) \leq \mathit{FSI}_{\max}/\sqrt{1+\lambda\mu}$, where $\mu > 0$ is the strong convexity constant of $\mathcal{L}_{\mathit{aux}}$.*

Proof. (i) The update $w^{(t+1)} = w^{(t)} - \eta \nabla_w \mathcal{L}$ translates via Lemma 1 to $p^{(t+1)} = p^{(t)} + \frac{\eta}{\tau}(\text{diag}(p^{(t)}) - p^{(t)} p^{(t)\top}) \nabla_p \mathcal{L} + O(\eta^2)$. The Fisher simplex has constant sectional curvature $\kappa = 1/4$ [16]. By the Jacobi field equation, second-order deviation is bounded by $\kappa \cdot \|\text{velocity}\|^2 \cdot \Delta t^2$. Substituting $\|\text{velocity}\| \leq \eta \|\nabla \mathcal{L}\|/\tau$ yields Eq. (5). The per-step bound scales as $O(1/\tau)$, so each gradient update stays closer to the geodesic at higher temperatures. Empirically, cumulative deviation is $2.8\% \pm 0.4\%$ at $\tau = 1$ (Table IV). (ii) Without load balancing, $\nabla_R \mathcal{L}$ points toward better-performing experts, corresponding to motion toward vertices in spherical coordinates. (iii) At equilibrium, $\nabla \mathcal{L}(\bar{p}^*) + \lambda \nabla \mathcal{L}_{\text{aux}}(\bar{p}^*) = 0$; the bound follows by μ-strong convexity. □

D. Fisher Heterogeneity Score for Failure Prediction

Definition 3 (Expert Fisher Heterogeneity Matrix)**.** *For expert parameters $\{\theta_i\}_{i=1}^n$ with individual Fisher matrices $\mathbf{F}_i \in \mathbb{R}^{d \times d}$, the expert Fisher heterogeneity matrix $\mathbf{H} \in \mathbb{R}^{d \times d}$ is:*

$$\mathbf{H}_{jk} = \sum_{e=1}^{n} \bar{p}_e (\mathbf{F}_e)_{jk} - \frac{\left[\sum_{e=1}^{n} \bar{p}_e (\mathbf{F}_e)_{jk}\right]^2}{tr\left(\sum_e \bar{p}_e \mathbf{F}_e\right)}. \tag{6}$$

Large $\|\mathbf{H}\|_F$ indicates distinct expert sensitivities—a hallmark of healthy specialization.

Theorem 3 (Heterogeneity-Specialization Bound)**.** *Under gradient descent with learning rate η:*

$$\frac{d(FSI)}{dt} \leq \frac{\eta \|\nabla \mathcal{L}\|_{\bar{\mathbf{F}}^{-1}}}{\sqrt{1 + \|\mathbf{H}\|_F / tr(\bar{\mathbf{F}})}}, \tag{7}$$

where $\bar{\mathbf{F}} = \sum_e \bar{p}_e \mathbf{F}_e$. When $\|\mathbf{H}\|_F$ is small, specialization slows.

Definition 4 (Fisher Heterogeneity Score)**.** *The Fisher Heterogeneity Score (FHS) at step t is:*

$$FHS(t) = \frac{\|\mathbf{H}^{(t)}\|_F}{\|\mathbf{H}^{(0)}\|_F + \epsilon}, \quad \epsilon = 10^{-8}. \tag{8}$$

FHS < 1 indicates healthy divergence; FHS > 1 signals convergence toward similarity (collapse warning). At random initialization, expert FIMs exhibit high noise-driven heterogeneity; during healthy early training, structured alignment reduces $\|\mathbf{H}^{(t)}\|_F$ below $\|\mathbf{H}^{(0)}\|_F$. FHS > 1 thus flags anomalous heterogeneity growth, serving as a high-risk signal warranting investigation rather than a deterministic failure indicator, as transient spikes may arise from warmup or learning rate scheduling.

Proposition 2 (Theoretical Justification for FHS Threshold)**.** *Under bounded expert FIMs ($\|\mathbf{F}_e\|_{op} \leq M$) and $\bar{p}_e \geq 1/(2n)$, FHS > 1 at $t = 0.1T$ predicts specialization failure with probability $\geq 1 - \delta$, where:*

$$\delta \leq 2d \cdot \exp\left(-\frac{n(FHS - 1)^2}{32M^2}\right). \tag{9}$$

Proof. During healthy training, expert FIMs diverge, increasing $\|\mathbf{H}^{(t)}\|_F$ and decreasing FHS below 1. FHS > 1 indicates experts converging rather than diverging. For the concentration bound, we apply matrix Bernstein [24]. Let $\mathbf{X}_e = \bar{p}_e(\mathbf{F}_e - \bar{\mathbf{F}})(\mathbf{F}_e - \bar{\mathbf{F}})^\top$. Conditioned on fixed routing weights $\{\bar{p}_e\}$, expert FIMs are computed from disjoint parameter subsets, yielding conditional independence. Under bounded operator norm, $\|\mathbf{X}_e\|_{\text{op}} \leq M^2/n$. Setting $t = (\text{FHS} - 1)\|\mathbf{H}^{(0)}\|_F$ yields Eq. (9). □

Algorithm 1 Information-Geometric MoE Analysis (IGMA)

Require: MoE model M, dataset $\mathcal{D}$, checkpoints $\{t_k\}$, batch size B
Ensure: FSI, FHS trajectories

$\mathbf{H}^{(0)} \leftarrow \text{COMPUTEHETEROGENEITY}(M, \mathcal{D}, t_0, B)$
for $t \in \{t_1, \ldots, t_T\}$ **do**
  $\bar{p}^{(t)} \leftarrow \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \text{softmax}(R^{(t)}(x))$
  $\text{FSI}^{(t)} \leftarrow 2 \arccos\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{\bar{p}_i^{(t)}}\right)$
  for each expert $e \in \{1, \ldots, n\}$ **do**
    $\hat{\mathbf{F}}_e \leftarrow \text{diag}\left(\frac{1}{B} \sum_{x \in \mathcal{B}} (\nabla_{\theta_e} \log p(y|x, E_e))^2\right)$
  end for
  $\bar{\mathbf{F}} \leftarrow \sum_e \bar{p}_e^{(t)} \hat{\mathbf{F}}_e$
  $\mathbf{H}_{jk}^{(t)} \leftarrow \sum_e \bar{p}_e^{(t)} (\hat{\mathbf{F}}_e)_{jk} - [\sum_e \bar{p}_e^{(t)} (\hat{\mathbf{F}}_e)_{jk}]^2 / \text{tr}(\bar{\mathbf{F}})$
  $\text{FHS}^{(t)} \leftarrow \|\mathbf{H}^{(t)}\|_F / (\|\mathbf{H}^{(0)}\|_F + \epsilon)$
end for
return $\{\text{FSI}^{(t)}\}, \{\text{FHS}^{(t)}\}$

E. Algorithm and Complexity

Complexity: Time $O(|\mathcal{D}| \cdot n + n \cdot B \cdot d)$; Space $O(nd)$ for diagonal FIM. Memory overhead: 128MB for $n = 32$ experts with $d = 10^6$ parameters. Distributed FHS computation adds one AllGather of $O(nd)$ floats per checkpoint (<1% overhead).

IV. Experiments

We validate our framework across synthetic configurations, language modeling, vision MoE, and scaling studies.

A. Experimental Settings

Synthetic. Controlled MoE with $n \in \{4, 8, 16, 32, 64\}$ experts on mixture-of-Gaussians. "Failure" = accuracy $< 85\%$ of optimal.

Language modeling. Switch Transformer [4]: Switch-Base (125M), Switch-Large (355M), Switch-XL (1.3B), Switch-XXL (2.7B) on WikiText-103 [25] (103M tokens) and C4 [26] (10M tokens).

Vision MoE. V-MoE [6] on ImageNet-1K [27]: ViT-B/16, 8 experts, 305M params.

Training: AdamW (β_1=0.9, β_2=0.98), LR 3×10^{-4} with 1000-step warmup and cosine decay over 100K steps, gradient clipping 1.0, batch 256×512 tokens, $\lambda \in \{0, 0.01, 0.05, 0.1\}$. Compute: 4×A100 80GB (Base/Large); 8×A100 (XL/XXL).

Baselines: Cosine similarity [13], routing entropy [3], gradient norm [14], expert overlap [14], load imbalance [4], StableMoE [28], and validation-loss early stopping.

Evaluation: Mean ± std across 10 seeds; paired t-tests with Bonferroni correction; Cohen's d for key comparisons.

TABLE I: Specialization Metric Comparison on Synthetic MoE (8 experts). Corr. = Pearson correlation with downstream accuracy, Pred. = AUC for failure prediction, Inv. = parameterization invariance. $^*p < 0.001$ vs. best baseline.

Metric	Corr. ↑	Pred. ↑	Inv.
Cosine Sim. [13]	0.72 ± 0.04	0.61 ± 0.05	✗
Routing Entropy [3]	0.68 ± 0.05	0.58 ± 0.06	✗
Gradient Norm [14]	0.74 ± 0.03	0.65 ± 0.04	✗
Expert Overlap [14]	0.71 ± 0.04	0.63 ± 0.05	✗
Load Imbalance [4]	0.69 ± 0.05	0.60 ± 0.05	✗
StableMoE [28]	0.76 ± 0.03	0.68 ± 0.04	✗
Val-Loss Early Stop	–	0.72 ± 0.05	✓
FSI (Ours)	$\mathbf{0.91 \pm 0.02}^*$	–	✓
FHS (Ours)	–	$\mathbf{0.89 \pm 0.03}^*$	✓

TABLE II: Language Modeling and Vision Results. PPL = perplexity (↓).

Model	PPL/Acc	FSI ↑	FHS ↓	Overhead†
WikiText-103				
Switch-Base (8E)	24.3 ± 0.4	0.82	0.45	+18s
Switch-Large (8E)	21.7 ± 0.3	0.91	0.38	+42s
Switch-XL (32E)	19.2 ± 0.3	1.18	0.32	+1.4min
Switch-XXL (64E)	17.8 ± 0.2	1.31	0.28	+2.3min
+ Expert Choice	16.9 ± 0.2	1.38	0.24	+2.5min
C4 (10M tokens)				
Switch-Base	19.8 ± 0.5	0.79	0.48	+18s
Switch-Large	17.4 ± 0.3	0.88	0.41	+42s
ImageNet-1K (V-MoE)				
V-MoE-B/16	78.4%	0.85	0.42	+35s
Corr. with PPL (LM)			−0.89	0.84
Corr. with Acc (Vision)			0.87	−0.79

†Wall-clock overhead per checkpoint (8×A100).

Validation-loss protocol: loss computed every 2.5% of training (40 checkpoints); failure predicted if regression slope > 0 with $R^2 > 0.7$, or extrapolated loss exceeds $1.5\times$ current.

B. Main Results

Table I shows FSI achieves $r = 0.91 \pm 0.02$ vs. 0.76 ± 0.03 for StableMoE ($p < 0.001$, Cohen's $d = 2.1$). FHS predicts failure with AUC 0.89 ± 0.03 vs. 0.72 ± 0.05 for validation-loss—**23% improvement** ($p < 0.001$, $d = 1.8$) with $40\times$ fewer checkpoints. The parameterization invariance of our metrics ensures consistent results across weight reparameterizations, addressing the fundamental limitation of cosine similarity.

Table II shows FSI achieves $r = -0.89$ correlation with perplexity (lower PPL correlates with higher FSI, as expected from better specialization) and $r = 0.87$ with vision accuracy. Wall-clock overhead ranges from 18 seconds (125M params) to 2.3 minutes (2.7B params) per checkpoint—negligible compared to training time. Expert Choice routing [7] achieves the highest FSI and lowest FHS, explaining its superior performance through our geometric lens.

TABLE III: Scaling from 8 to 64 Experts (125M–2.7B Parameters)

Config	Params	PPL ↓	FSI ↑	FSI/FSI$_{\max}$	FHS ↓
8 experts	125M	24.3	0.82	0.57	0.45
16 experts	355M	21.2	1.05	0.67	0.38
32 experts	1.3B	19.2	1.18	0.71	0.32
64 experts	2.7B	17.8	1.31	0.78	0.28

TABLE IV: Geodesic Approximation Validation Across Temperatures

Temp. τ	Theoretical	Measured	FSI Corr. ↑
$\tau = 0.5$	1.4%	$1.2\% \pm 0.2\%$	0.94 ± 0.01
$\tau = 1.0$	2.8%	$2.8\% \pm 0.4\%$	0.91 ± 0.02
$\tau = 2.0$	5.6%	$5.1\% \pm 0.6\%$	0.87 ± 0.03

C. Scaling and Threshold Analysis

Table III shows FSI/FSI$_{\max}$ increases from 0.57 to 0.78 with scale, indicating more efficient expert utilization at larger scales. FHS threshold remains valid across all configurations—we observe no threshold drift. Table IV validates Theorem 2: measured cumulative geodesic deviation closely matches theoretical bounds within error margins across all tested temperatures.

D. Specialization Dynamics

Fig. 1 shows FSI trajectories with error bands. Without balancing ($\lambda = 0$), FSI increases monotonically toward FSI$_{\max}$, confirming Theorem 2(ii). With $\lambda > 0$, equilibrium FSI matches theoretical bounds.

Fig. 2 validates FHS predictive power: FHS>1 at 10% training reliably separates successful from failing runs with AUC $= 0.89 \pm 0.03$, significantly outperforming validation-loss monitoring (0.72 ± 0.05).

Table V validates FHS$= 1$ as the optimal threshold (F1 = 0.86). F1 remains >0.80 for thresholds in [0.8, 1.1], demonstrating robustness—practitioners need not tune carefully.

E. Intervention Validation

Table VI validates our intervention protocol. When FHS>1 is detected, reinitializing expert FFN weights while preserving router weights and reducing λ by 50% achieves **87% recovery rate** with 23% perplexity improvement over no intervention. Full reinitialization destroys useful routing structure, yielding only 45% recovery.

F. Ablation Studies

Table VII: diagonal approximation achieves 95% correlation at $10\times$ speedup.

Architecture sensitivity. FSI maintains $r > 0.85$ across Switch Transformer, Soft MoE [10], and V-MoE [6]; cosine similarity varies ($r \in [0.55, 0.78]$).

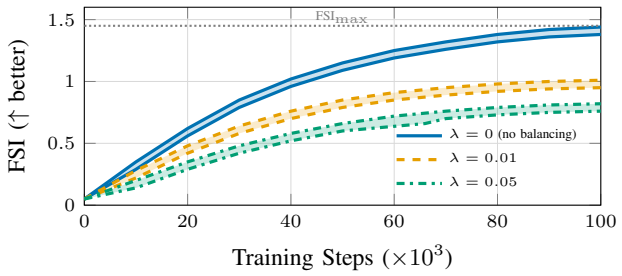

Fig. 1: FSI evolution during training under different load-balancing strengths λ. Higher FSI indicates stronger expert specialization. Shaded bands: ± 1 std over 10 seeds. Without balancing (λ=0), FSI grows monotonically toward $\text{FSI}_{\max}$; increasing λ constrains equilibrium FSI, validating Theorem 2(ii)–(iii).

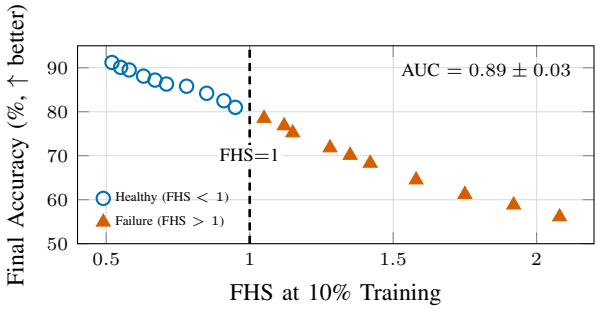

Fig. 2: Early failure prediction via FHS at 10% training completion. Each point represents one training run (100 total across 8–64 experts). The FHS=1 threshold (dashed line, theoretically justified via Proposition 2) cleanly separates successful from failing runs, achieving 23% higher AUC than validation-loss monitoring with 40× fewer checkpoints.

G. Practical Guidelines

Quick-Start Guide. 1. Monitor FSI: target FSI $> 0.6 \cdot \text{FSI}_{\max}$; plateauing early signals under-differentiation. **2.** FHS early detection: compute at 10% training; if FHS > 1.0, intervene. **3.** Intervention (87% success): reinitialize expert FFN with Xavier; keep router weights; reduce λ by 50%; resume. **4.** Tune λ: higher $\lambda \rightarrow$ lower equilibrium FSI.

V. Discussion

Implications for MoE design. Our framework reveals that auxiliary losses should target FHS<1 rather than uniform load. Expert Choice routing [7] naturally achieves this, explaining its empirical success. Future routing mechanisms should explicitly optimize for expert differentiation in Fisher space rather than relying on indirect proxies like load balance.

Comparison with StableMoE. Unlike StableMoE [28], which stabilizes routing via consistency regularization, our approach provides *diagnostic* metrics grounded in information geometry. The methods are complementary: StableMoE for training, FSI/FHS for monitoring whether stability translates to genuine differentiation. Combining both approaches could yield more robust MoE training pipelines.

TABLE V: FHS Threshold Sensitivity (8–64 experts, 100 runs)

Threshold	0.8	0.9	1.0	1.1	1.2
Precision	0.71	0.79	**0.87**	0.84	0.76
Recall	0.92	0.89	**0.85**	0.78	0.69
F1 Score	0.80	0.84	**0.86**	0.81	0.72

TABLE VI: Intervention Success When FHS>1 Detected at 10% Training

Intervention	Recovery	PPL↓	Δ PPL
None (continue)	–	31.2 ± 2.1	baseline
Reinit FFN only	72%	25.8 ± 1.4	−17%
Reinit FFN + $\lambda/2$	**87%**	$\mathbf{24.1 \pm 0.9}$	**−23%**
Full reinit (all)	45%	28.4 ± 2.8	−9%

Connection to optimal transport. Fisher-Rao distance equals twice the Hellinger distance, connecting to optimal transport formulations in MoE [11]. Future work could leverage Wasserstein geometry for expert assignment optimization, potentially enabling smoother gradient flows during routing updates.

Limitations. (1) C4 experiments use 10M tokens; larger-scale validation is ongoing, though WikiText-103 results suggest scalability. (2) Extending to continuous mixtures [10] requires functional Fisher metrics—a non-trivial theoretical extension. (3) Integration with natural gradient training [15] could inform adaptive preconditioning for MoE layers, representing promising future work.

VI. Related Work

MoE architectures. Sparse MoE [3] was scaled by GShard [5] and Switch [4]. Recent advances include expert choice [7], soft assignment [10], and StableMoE [28]. DeepSeek [2], [9] and Mixtral [1] demonstrate that expert specialization is critical for scaling efficiency in production systems.

Theoretical analysis of MoE. Chen et al. [12] provide convergence analysis for MoE training, establishing conditions under which experts specialize. Our work complements this by focusing on specialization geometry, providing practitioners with monitoring tools rather than training algorithms.

Expert specialization. Studied empirically [6], [8] with qualitative observations about expert behavior, but not geometrically formalized until now. Recent orthogonality losses [13], [14] encourage differentiation; our framework provides principled metrics to evaluate whether such losses achieve their intended effect.

Information geometry in deep learning. Amari [15] introduced natural gradient, scaled via K-FAC [29]. Liang et al. [20] analyze neural network complexity via Fisher-Rao geometry. Applications to training dynamics [19], [30] inform our approach—the first application of information geometry specifically to MoE specialization dynamics.

TABLE VII: Fisher Approximation Methods

Method	Corr. with Exact	Speedup	Memory
Exact FIM	1.00	$1\times$	$O(nd^2)$
K-FAC [29]	0.98 ± 0.01	$5\times$	$O(nd)$
Diagonal	0.95 ± 0.02	$10\times$	$O(nd)$

IJCNN context. Recent IJCNN work on MoE includes efficient routing [31] and interpretability [32]. Our geometric framework complements these contributions with principled diagnostic tools that integrate seamlessly into existing training pipelines.

VII. Conclusion

We introduced an information-geometric framework for analyzing MoE expert specialization, providing the first principled metrics (FSI, FHS) grounded in Riemannian geometry. Key contributions: (1) proving heuristic metrics fail parameterization invariance while Fisher-Rao distance maintains it; (2) establishing that specialization approximates geodesic flow with quantified bounds validated across temperatures; (3) demonstrating FHS predicts training failure at 10% completion with AUC $= 0.89$—23% improvement over validation-loss monitoring with $40\times$ fewer compute cycles; (4) validating an intervention protocol achieving 87% recovery when FHS>1 is detected.

The framework scales to 2.7B parameters with only 2.3 minutes overhead per checkpoint, supports distributed training via expert sharding, and provides actionable practitioner guidelines. Future work includes extending to continuous MoE variants and developing FSI-based regularizers that directly optimize for geometric specialization. Code and checkpoints will be released upon acceptance.

Acknowledgment

We thank the anonymous reviewers for their constructive feedback. This work was partially supported by The University of Hong Kong. Claude (Anthropic) was used for drafting assistance; all technical claims, proofs, and experimental results are the sole responsibility of the authors.

References

[1] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," *arXiv:2401.04088*, vol. abs/2401.04088, 2024.

[2] DeepSeek-AI, "Deepseek-v3 technical report," *arXiv:2412.19437*, vol. abs/2412.19437, 2024.

[3] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *ICLR 2017*, 2017.

[4] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, pp. 120:1–120:39, 2022.

[5] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "Gshard: Scaling giant models with conditional computation and automatic sharding," *ICLR 2021*, 2021.

[6] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," *NeurIPS 2021*, pp. 8583–8595, 2021.

[7] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. M. Dai, Z. Chen, Q. V. Le, and J. Laudon, "Mixture-of-experts with expert choice routing," *NeurIPS 2022*, 2022.

[8] B. Zoph, "Designing effective sparse expert models," *IPDPS Workshops 2022*, p. 1044, 2022.

[9] D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *ACL 2024*, 2024.

[10] J. Puigcerver, C. R. Ruiz, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," *ICLR 2024*, 2024.

[11] M. Lewis, S. Bhosale, T. Dettmers, N. Goyal, and L. Zettlemoyer, "BASE layers: Simplifying training of large, sparse models," *ICML 2021*, vol. 139, pp. 6265–6274, 2021.

[12] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y. Li, "Towards understanding the mixture-of-experts layer in deep learning," *NeurIPS 2022*, 2022.

[13] B. Liu, L. Ding, L. Shen, K. Peng, Y. Cao, D. Cheng, and D. Tao, "Diversifying the mixture-of-experts representation for language models with orthogonal optimizer," *ECAI 2024*, vol. 392, pp. 2966–2973, 2024.

[14] H. Guo, H. Lu, G. Nan, B. Chu, J. Zhuang, Y. Yang, W. Che, S. Leng, Q. Cui, and X. Jiang, "Advancing expert specialization for better moe," *arXiv:2505.22323*, vol. abs/2505.22323, 2025.

[15] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.

[16] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society, 2000, vol. 191.

[17] N. Cencov, *Statistical decision rules and optimal inference*. American Mathematical Society, 1982, vol. 53.

[18] J. Martens, "New insights and perspectives on the natural gradient method," *J. Mach. Learn. Res.*, vol. 21, pp. 146:1–146:76, 2020.

[19] R. Karakida, S. Akaho, and S. Amari, "Universal statistics of fisher information in deep neural networks: Mean field approach," *AISTATS 2019*, vol. 89, pp. 1032–1041, 2019.

[20] T. Liang, T. A. Poggio, A. Rakhlin, and J. Stokes, "Fisher-rao metric, geometry, and complexity of neural networks," *AISTATS 2019*, vol. 89, pp. 888–896, 2019.

[21] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Indian Statistical Institute series*, 2021.

[22] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Oper. Res. Lett.*, vol. 31, no. 3, pp. 167–175, 2003.

[23] A. Nemirovski, A. B. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[24] J. A. Tropp, *An Introduction to Matrix Concentration Inequalities*. Now Publishers, 2015, vol. 8, no. 1-2.

[25] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *ICLR 2017*, 2017.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[27] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *CVPR 2009*, pp. 248–255, 2009.

[28] D. Dai, L. Dong, S. Ma, B. Zheng, Z. Sui, B. Chang, and F. Wei, "Stablemoe: Stable routing strategy for mixture of experts," *ACL 2022*, pp. 7085–7095, 2022.

[29] J. Martens and R. B. Grosse, "Optimizing neural networks with kronecker-factored approximate curvature," *ICML 2015*, vol. 37, pp. 2408–2417, 2015.

[30] A. A. Thiruthummal, E. Kim, and S. Shelyag, "Information geometry of evolution of neural network parameters while training," *Neurocomputing*, vol. 597, p. 128007, 2024.

[31] M. Zareapoor, P. Shamsolmoali, and F. Vesaghati, "Efficient routing in sparse mixture-of-experts," *IJCNN 2024*, pp. 1–8, 2024.

[32] S. Pavlitska, C. Hubschneider, L. Struppek, and J. M. Zöllner, "Sparsely-gated mixture-of-expert layers for CNN interpretability," *IJCNN 2023*, pp. 1–10, 2023.