WHEN A LOAD-BALANCING LOSS REWARDS DEAD EXPERTS: SHARP GLOBAL GEOMETRY OF SWITCH ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Top-1 mixture-of-experts routers are commonly regularized by the Switch auxiliary loss $L_{\mathrm{sw}} = N \sum_i f_i P_i$, which couples hard expert frequencies f with mean soft routing probabilities P. It is widely believed that the loss attains its minimum value of 1 under uniform routing. We show that this claim is false. We prove the sharp infimum

$$\inf L_{\mathrm{sw}} = \frac{N}{\lfloor (N+1)^2/4 \rfloor} = \frac{4 + o(1)}{N} \qquad (N \geq 3).$$

The bound is approached by routing all tokens evenly among roughly *half* the experts, leaving the rest unused. For each token, the routing probabilities approach a uniform distribution over its assigned expert and all unused experts, with the assigned expert remaining the unique top-1 choice. We characterize all extremal hard loads and prove that hard-load vectors must approach this half-dead family as the loss approaches its infimum. We also construct cases where the loss decreases while expert assignments and routed outputs remain unchanged under top-k routing with post-selection renormalization. We explicitly observe normalized losses below 1 in layers of released Switch Transformer, Qwen, and DeepSeek models. In Switch Transformer and Qwen, such values coexist with unused experts that retain above-uniform soft mass. Controlled language-model training experiments also produce losses below 1 alongside unused experts and show that Switch regularization can induce *negative* hard–soft correlation. These results show that minimizing the auxiliary loss does not necessarily balance expert utilization.

1 INTRODUCTION

Language models based on the Transformer (Vaswani et al., 2017) improve with more parameters, but training and serving larger models requires proportionally more compute (Kaplan et al., 2020; Hoffmann et al., 2022). Mixture-of-experts (MoE) layers address this by routing each token to a small subset of experts, so that capacity grows while per-token compute stays fixed (Fedus et al., 2022; Lepikhin et al., 2020; Shazeer et al., 2017; Du et al., 2022). The efficiency gain depends on routing load: an overloaded expert becomes a compute bottleneck (He et al., 2022; Gale et al., 2023), while an unused expert stops learning. Switch Transformers therefore introduced the now-standard top-1 auxiliary loss

$$L_{\mathrm{sw}} = N\langle f, P\rangle, \qquad f_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{a_t = i\}, \qquad P_i = \frac{1}{T}\sum_{t=1}^{T} p_{t,i}, \tag{1}$$

where p_t is the softmax gate and $a_t = \arg\max_i p_{t,i}$. The original paper states that the loss is minimized under uniform routing (Fedus et al., 2022). The same lower-bound interpretation appears in later work, including ReMoE, CLIP-MoE, and recent theoretical work (Wang et al., 2025; Zhang et al., 2025a; Su & Liu, 2026). In practice, a value near one is monitored as evidence that routing is balanced (Qiu et al., 2025), and decreases in the loss are interpreted as improved load balance (Yang et al., 2025; 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

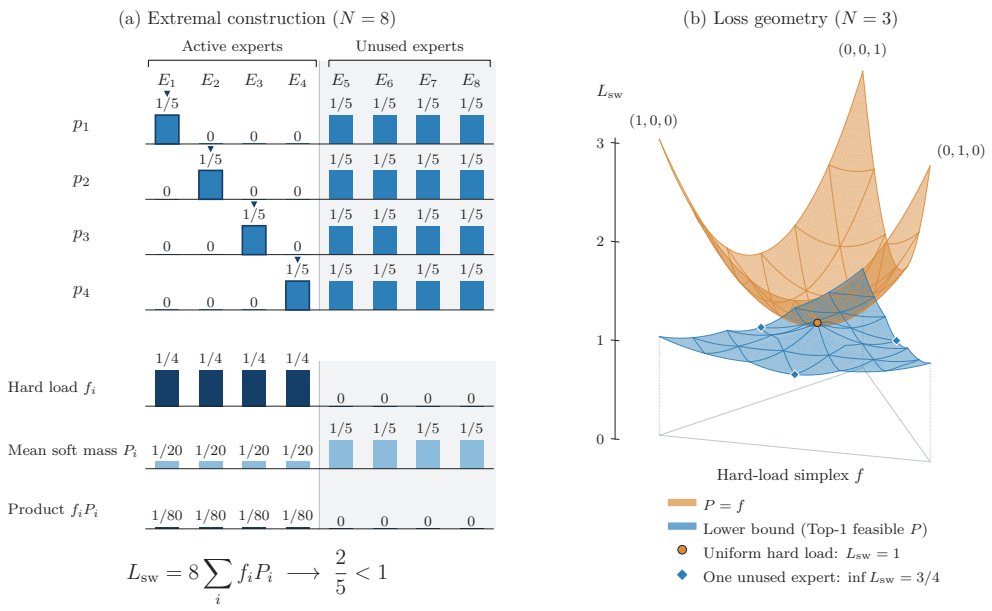

Figure 1: **The Switch loss can favor unused experts over uniform routing.** (a) An eight-expert, four-token construction with half the experts unused. Each token's soft mass approaches an equal split among its selected expert and the four unused experts, giving $L_{\text{sw}} \to 0.4$. (b) Loss surfaces over the three-expert hard-load simplex. Under $P = f$, the loss $L_{\text{sw}} = 3\|f\|_2^2$ (orange) attains its minimum of 1 at uniform routing. Allowing soft probabilities to vary subject to top-1 routing gives the exact lower bound at each fixed hard load f (blue). This surface has global minimum $3/4$ at the three edge midpoints, where one expert is unused and the other two share the load equally.

We disprove the claim that the Switch loss attains its minimum value of 1 under uniform routing. For example, with $N = 3$, take two tokens whose gates approach $(1/2, 1/2, 0)$ and $(1/2, 0, 1/2)$, keeping experts 2 and 3 as the respective unique winners. Then

$$f = (0, 1/2, 1/2), \qquad P \longrightarrow (1/2, 1/4, 1/4), \qquad L_{\text{sw}} \longrightarrow 3/4 < 1.$$

Expert 1 has the largest mean soft mass but receives no tokens (Figure 1(b)). Because $f_1 = 0$, its soft mass contributes nothing to the loss. The example suggests a way to lower the loss at a fixed hard load: shift soft mass from more-used to less-used experts, where it is multiplied by a smaller f_i. In fact, writing $u = (1/N, \ldots, 1/N)$, we have

$$L_{\text{sw}} - 1 = N\langle f - u, P - u\rangle.$$

A loss below 1 corresponds exactly to negative hard–soft alignment: across experts, higher soft mass tends to accompany lower hard load.

We use this intuition to solve the minimization problem, first at a fixed hard load and then over all hard-load vectors. For $N \geq 3$, the sharp global infimum is

$$\inf L_{\text{sw}} = \frac{N}{\lfloor (N+1)^2/4 \rfloor} = \frac{4 + o(1)}{N}.$$

To approach this bound, we distribute tokens evenly among roughly half the experts and leave the others unused. Each token's gate approaches a uniform distribution over its assigned expert and all unused experts, while the assigned expert remains the unique winner (Figure 1(a)). We prove that all extremal hard loads have this form. Our stability result further shows that the same structure governs near-minimal losses: their hard loads must approach this half-dead family.

Even away from the extremal configurations, a decrease in the Switch loss need not mean that work has been distributed more evenly among experts. We show that changing soft probabilities can lower the loss while every token keeps its expert assignment. We also construct cases where routed outputs remain unchanged under top-k routing with post-selection renormalization.

Beyond these constructions, we observe losses below 1 in layers of released Switch Transformer, Qwen, and DeepSeek models. In Switch Transformer and Qwen, such values coexist with unused experts that retain above-uniform soft mass. One justification for a minimum of 1 under uniform routing adopts the approximation $f \approx P$ (Su & Liu, 2026). This replaces the Switch loss by $N\|P\|_2^2$, whose minimum is 1 at the uniform distribution, as illustrated by the $P = f$ surface in Figure 1(b). Yet controlled causal-MoE training produces losses below 1 alongside unused experts, and Switch regularization can turn hard–soft correlation negative. Switch optimization can therefore undermine the correspondence used to justify that minimum.

Contributions.

1. **Sharp infimum and extremal geometry.** We disprove the widely repeated claim that uniform routing minimizes the Switch loss at 1. We solve the fixed-load problem and prove the sharp global infimum $N/\lfloor (N+1)^2/4 \rfloor = (4 + o(1))/N$ for $N \geq 3$. We characterize all extremal hard loads, prove that near-optimal loads approach the half-dead family, and quantify finite-batch corrections (Sections 3–4, Appendix D).
2. **Loss-decreasing directions and top-k extensions.** We identify strict loss-decreasing directions that preserve hard assignments and, under full-softmax top-k routing with post-selection renormalization, routed outputs. We also construct top-2 routers with roughly half the experts unused and limiting loss $(8 + o(1))/N$ (Section 5 and Appendix F).
3. **Observations in pretrained models and controlled training.** We observe losses below 1 with unused experts in pretrained MoEs and controlled training. Training comparisons show that Switch regularization can induce negative hard–soft correlation (Section 6, Appendices H–I).

2 SETUP

Write $[N] = \{1, \dots, N\}$. Let $(\mathcal{X}, \mu)$ be a probability space of token inputs and let $p : \mathcal{X} \to \Delta_N$ be a measurable router, where $\Delta_N = \{q \in \mathbb{R}_+^N : \sum_i q_i = 1\}$. Let $a(x) \in \arg\max_i p_i(x)$ denote the hard assignment, with ties resolved by any fixed measurable rule. Define

$$f_i = \mu\{x : a(x) = i\}, \qquad P_i = \int p_i(x)\, d\mu(x), \qquad L_{\mathrm{sw}}(p) = N \sum_{i=1}^{N} f_i P_i. \tag{2}$$

The vectors f and P are the hard and soft routing marginals, respectively.

The lower bounds derived in Sections 3–4 hold on every probability space and for every router. To show that they are sharp, we assume a nonatomic token space and an unrestricted measurable router, which allow a partition with any prescribed masses f_i. With finite logits, closed-cell extremizers that contain zero probabilities or winner ties can be approached, but need not be attained, by strict softmax gates. On a fixed finite dataset, hard loads also lie on a lattice; Appendix D gives the corresponding corrections. Unless stated otherwise, we study the raw pre-capacity top-1 statistic in the Switch definition; post-capacity dispatch and top-k routing have different feasible geometry.

3 THE FIXED-LOAD PROBLEM

We first fix the hard-load vector f and optimize over routers that realize f. For $i \in [N]$, define the closed top-1 cell

$$C_i = \{q \in \Delta_N : q_i \geq q_j \text{ for every } j\}$$

If a token is routed to expert i, its gate belongs to C_i. Thus $m_i(f) = \min_{q \in C_i} \langle f, q \rangle$ is the smallest per-token value of $\langle f, q \rangle$ compatible with that assignment.

Theorem 3.1 (Fixed-load formula)**.** *For every $f \in \Delta_N$,*

$$m_i(f) = \min_{S \subseteq [N] : i \in S} \frac{1}{|S|} \sum_{j \in S} f_j. \tag{3}$$

Every router with hard load f satisfies

$$\frac{L_{\mathrm{sw}}(p)}{N} \geq F_N(f) := \sum_{i=1}^{N} f_i m_i(f). \tag{4}$$

If $(\mathcal{X}, \mu)$ is nonatomic and the router class is unrestricted, the right-hand side is the infimum over routers with hard load f. The same infimum is approached by finite-logit softmax routers with strict argmax almost surely.

The vertices of C_i are precisely $u_S = |S|^{-1}\mathbf{1}_S$ with $i \in S$. Since a linear functional attains its minimum at a vertex, this proves Eq. (3). Writing $A_i = \{x : a(x) = i\}$ for the hard-routing regions gives

$$\langle f, P\rangle = \sum_i \int_{A_i} \langle f, p(x)\rangle \, d\mu(x) \geq \sum_i f_i m_i(f).$$

A nonatomic partition of masses f_i, with a nearly optimal constant gate on each part, approaches equality. Appendix A gives a constructive vertex decomposition and strict-softmax perturbation. The minimizing vertices expose the mechanism: on tokens routed to i, soft mass is distributed uniformly over i and a subset of low-load experts.

4 Sharp global geometry

Section 3 reduces the problem to minimizing $F_N(f)$ over hard-load vectors. Let

$$M_N = \left\lfloor \frac{(N+1)^2}{4} \right\rfloor = \max_{1 \leq k \leq N} k(N-k+1). \tag{5}$$

The following theorem gives the sharp global value and characterizes all hard-load vectors that minimize the fixed-load envelope.

Theorem 4.1 (Sharp global infimum and all extremal loads)**.** *For unrestricted top-1 routers on a nonatomic token space,*

$$\inf_p L_{\mathrm{sw}}(p) = \begin{cases} 7/8, & N = 2, \\ N/M_N, & N \geq 3. \end{cases} \tag{6}$$

For $N = 2$, the minimizers of F_2 are permutations of $(1/4, 3/4)$. For $N \geq 3$, the minimizers of F_N are, up to permutation,

$$v^{(k)} = (\underbrace{1/k, \ldots, 1/k}_{k}, \underbrace{0, \ldots, 0}_{N-k}), \qquad k(N-k+1) = M_N. \tag{7}$$

Thus if $N = 2h-1$ is odd, $k = h$; if $N = 2h$ is even, $k = h$ or $h+1$.

For $N = 2$, write the sorted load as $(x, 1-x)$ with $0 \leq x \leq 1/2$. Then

$$2F_2(x, 1-x) = 2x^2 + 1 - x = \frac{7}{8} + 2\left(x - \frac{1}{4}\right)^2.$$

This proves the $N = 2$ case. We henceforth assume $N \geq 3$.

4.1 The extremal construction

For the upper bound in Theorem 4.1, fix k with $k(N-k+1) = M_N$, and let A and D be the k active and $N-k$ dead experts. Partition tokens equally among $i \in A$. On the region routed to i, use a gate approaching the uniform distribution on $D \cup \{i\}$ while keeping i the unique winner. In this limit,

$$P_i \longrightarrow \frac{1}{k(N-k+1)} \; (i \in A), \qquad P_j \longrightarrow \frac{1}{N-k+1} \; (j \in D), \tag{8}$$

and therefore

$$L_{\mathrm{sw}} \longrightarrow \frac{N}{k(N-k+1)} = \frac{N}{M_N}.$$

Finite-logit routers approach but do not attain this limit, establishing the upper bound in Theorem 4.1. Each dead expert receives k times as much mean soft mass as each active expert. For $N = 64$, the sharp value is $64/1056 \approx 0.0606$, versus one for every perfectly balanced hard load.

This construction also has a linear-router realization on k feature groups. Relabel $A = \{1, \dots, k\}$, represent group i by $e_i \in \mathbb{R}^k$, and choose a logit matrix whose ith column is $\delta > 0$ at expert i, zero at unused experts, and $-B$ at the other active experts. Taking $\delta \downarrow 0$ and $B \to \infty$ gives Eq. (8).

On a fixed T-token batch, the same construction uses integer assignment counts. For $N \geq 3$ and $T \geq \lceil N/2 \rceil$, the batch infimum is at most $N/(4T^2)$ above N/M_N, and equals it whenever $\lceil N/2 \rceil$ divides T. Appendix D proves this bound and treats expected IID minibatch losses separately.

4.2 Optimality via a ranked quadratic inequality

The construction gives the upper bound N/M_N. To prove that no other hard-load vector can do better, it remains to show

$$F_N(f) \geq \frac{1}{M_N} \qquad \text{for every } f \in \Delta_N.$$

Since F_N is invariant under permutations, sort the coordinates as $0 \leq x_1 \leq \cdots \leq x_N$ and set $s_0 = 0$ and $s_t = \sum_{j=1}^{t} x_j$ for $1 \leq t \leq N$. For an expert at rank r, Theorem 3.1 gives

$$m_r(x) = \min_{1 \leq k \leq r} \frac{x_r + s_{k-1}}{k}. \tag{9}$$

For a set of size $k \leq r$, the smallest average is obtained by pairing rank r with the $k-1$ smallest coordinates. If $k > r$, the best set is the first k coordinates, whose average is at least s_r/r. Thus larger sets cannot improve on the candidate $k = r$.

Lemma 4.2 (Ranked quadratic inequality)**.** *For $N \geq 3$, every sorted $x \in \Delta_N$ satisfies*

$$\sum_{r=1}^{N} x_r \min_{1 \leq k \leq r} \frac{x_r + s_{k-1}}{k} \geq \frac{1}{M_N}. \tag{10}$$

Equality holds exactly at the sorted step vectors in Eq. (7)*.*

Proof sketch. For each r, choose $k_r \in \{1, \dots, r\}$ attaining the minimum in Eq. (9), and set $\boldsymbol{k} = (k_1, \dots, k_N)$. Define

$$Q_{\boldsymbol{k}}(x) = \sum_{r=1}^{N} \frac{x_r(x_r + s_{k_r - 1})}{k_r}.$$

Thus $F_N(x) = Q_{\boldsymbol{k}}(x)$; we seek a lower bound independent of $\boldsymbol{k}$.

Set $d_a = x_a - x_{a-1} \geq 0$, $x_0 = 0$, and $w_a = N - a + 1$. Each increment d_a raises the suffix $a, \dots, N$ uniformly. If it is the only nonzero increment, normalization gives a step load with w_a active experts. More generally, x is a convex combination of these step loads with weights $w_a d_a$, so $w^\top d = 1$. Substitution gives $Q_{\boldsymbol{k}} = d^\top A d$ with A symmetric, and the desired bound becomes $M_N Q_{\boldsymbol{k}} \geq (w^\top d)^2$.

Since $d \geq 0$, entrywise lower bounds on A also bound the quadratic form from below. For $a \leq b$, the rth summand contributes $g_{ab}(k_r)$ to A_{ab} when $r \geq b$, where

$$g_{ab}(k) = \frac{1}{k} + \frac{(k-a)_+ + (k-b)_+}{2k}.$$

Here $(t)_+ = \max\{t, 0\}$. Checking the three ranges $k \leq a$, $a < k \leq b$, and $b < k$ gives

$$g_{ab}(k) \geq c_{ab} := \begin{cases} 1, & a = b = 1, \\ (b+1)/(2b), & a = 1 < b, \\ 1/a, & 2 \leq a \leq b. \end{cases} \tag{11}$$

There are w_b indices $r \geq b$, so $A_{ab} \geq B_{ab} := w_b c_{ab}$ for $a \leq b$, with B symmetric. Comparing this lower bound with the target gives

$$M_N Q_{\boldsymbol{k}} - (w^\top d)^2 \geq d^\top C d, \qquad C := M_N B - w w^\top. \tag{12}$$

For $N \geq 5$, C is entrywise nonnegative. Indeed, for $2 \leq a \leq b$, $C_{ab} = w_b(M_N/a - w_a) \geq 0$ because $M_N \geq aw_a$; also $C_{11} = N(M_N - N) \geq 0$, while $C_{1b} = w_b[M_N(b+1)/(2b) - N] \geq 0$ follows from $b \leq N$ and $M_N \geq 2N^2/(N+1)$. Hence Eq. (12) proves $Q_{\boldsymbol{k}} \geq 1/M_N$. Appendix B supplies the three-case calculation and explicit sum-of-squares certificates for $N = 3, 4$.

For $N \geq 5$, equality requires $C_{aa}d_a^2 = 0$ for every a. Since $C_{aa} > 0$ whenever $a(N-a+1) < M_N$, we must have $d_a = 0$ at these indices. The $N = 3, 4$ certificates impose the same restriction. This leaves one jump when N is odd and two adjacent possible jumps, h and $h+1$, when $N = 2h$ is even. The two jumps cannot both be positive at equality: for $a \geq 2$, the diagonal bound $g_{aa}(k) \geq 1/a$ is an equality only at $k = a$. The $r = h+1$ summand contributes to both diagonal coefficients. If both jumps were positive, equality would therefore require $k_{h+1} = h$ and $k_{h+1} = h+1$, a contradiction. Only a single jump remains, yielding precisely the step vectors in Eq. (7). Appendix B also computes the exact excess along the even-dimensional two-increment slice.

By Eq. (9), the lemma gives $F_N(x) \geq 1/M_N$, and Theorem 3.1 therefore yields $L_{\mathrm{sw}} \geq N/M_N$. Together with the construction, this proves Theorem 4.1 for $N \geq 3$.

4.3 Near-optimal values force near-step loads

Finite-logit routers can approach the sharp infimum without attaining it. Since Theorem 3.1 gives $L_{\mathrm{sw}}/N \geq F_N(f)$, a near-optimal router has a near-minimal envelope value. For its sorted hard load x, retain the increments d_a and weights w_a from the preceding proof, and let $A^\star = \{a : aw_a = M_N\}$ denote the optimal jump positions.

Proposition 4.3 (Increment stability)**.** *For $N \geq 5$, a sorted load with increments $d_a = x_a - x_{a-1}$ satisfies*

$$M_N F_N(x) - 1 \geq \sum_{a=1}^{N} \frac{w_a}{a}[M_N - aw_a]d_a^2. \tag{13}$$

Consequently, if $F_N(x) \leq M_N^{-1} + \epsilon$, then

$$\sum_{a \notin A^\star} d_a^2 \leq \frac{M_N \epsilon}{\eta_N}, \qquad \eta_N = \min_{a \notin A^\star} \frac{w_a}{a}[M_N - aw_a] > 0. \tag{14}$$

Retaining only the diagonal $C_{aa} = w_a[M_N - aw_a]/a$ of the entrywise-nonnegative certificate in Eq. (12) gives Eq. (13); the definition of η_N then gives Eq. (14).

By Eq. (14), increments outside $A^\star$ vanish as $\epsilon \to 0$. For odd N, only one optimal jump is available; $w^\top d = 1$ fixes its limiting height, so x approaches the unique optimal step load. For even $N = 2h$, $A^\star = \{h, h+1\}$, and the bound first places x near the line segment joining the two optimal step loads. On this segment, $F_N > 1/M_N$ except at the endpoints (Eq. (20)), so near-minimizers approach one of them. Appendix C supplies the continuity and compactness argument.

5 Optimization within routing cells

The extremal construction lowers the loss by placing soft mass on unused experts. To see how the Switch update can favor this redistribution, consider a finite batch with $p_t = \mathrm{softmax}(z_t)$ and strict top-1 assignments. Within a routing cell, each token keeps the same assignment, so the hard loads f are constant. Differentiating with respect to the logits gives

$$\frac{\partial L_{\mathrm{sw}}}{\partial z_{t,i}} = \frac{N}{T} p_{t,i}\big(f_i - \langle f, p_t \rangle\big). \tag{15}$$

Thus logit-space gradient descent raises $z_{t,i}$ when f_i is below the soft-weighted load $\langle f, p_t \rangle$, and lowers it when f_i is above. In particular, every zero-load expert has a strictly negative derivative on every token. Raising such an expert's logit by a sufficiently small common amount increases its mean soft mass and strictly decreases the loss without changing any assignment. Within a routing cell, the loss can therefore improve through soft redistribution alone. Across routing boundaries, these cellwise gradients need not come from a globally continuous potential (Appendix E).

At finite logits, Eq. (15) vanishes for every token and expert exactly when $f = u$, since every $p_{t,i}$ is positive. Nonuniform loads therefore have no stationary point in independent logit coordinates inside a routing cell. The extremal loads characterize the scalar's infimum; the dynamics of a trained router also depend on crossing cell boundaries, shared parameters, and the task gradient.

Loss reduction without changing routed output. Consider a full-softmax top-k router that renormalizes the selected experts' weights (Jiang et al., 2024). Let S_t be the selected set for token t, and let $f_i^{(k)} = (kT)^{-1} \sum_t \mathbf{1}\{i \in S_t\}$. If expert j is unused on a finite batch and its logit lies strictly below the selected logits on every token, a sufficiently small increase in its logit changes neither the selected sets nor their renormalized weights. With expert outputs fixed, the routed outputs are therefore unchanged. Yet every other expert's full-softmax probability decreases, so $L_k = N \sum_i f_i^{(k)} P_i$ strictly decreases because $f_j^{(k)} = 0$. Appendix F.2 gives the precise perturbation and a population analogue; Appendix J illustrates the descent in a frozen language model.

6 Experiments

We examine losses below one in released checkpoints and causal-MoE training. For top-k checkpoints, we use $L_k = N \sum_i f_i^{(k)} P_i$ from Section 5, which equals one under uniform routing.

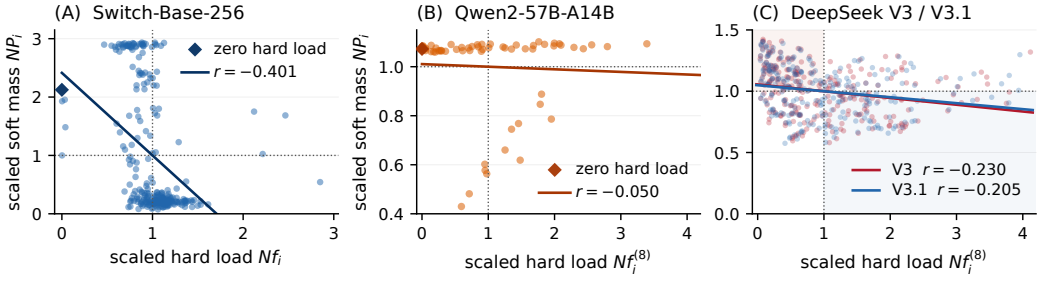

Figure 2: **Hard–soft separation in released MoEs evaluated on C4.** (A) Hard loads and soft masses of all 256 experts in Switch-Base-256 encoder layer 11. Expert 37 (diamond) receives no assignments, while its soft mass is 2.12 times the uniform value. (B) The corresponding quantities for Qwen2-57B-A14B layer 11. Expert 49 (diamond) receives no assignments despite above-uniform soft mass. (C) Hard loads and soft masses in the first sparse layers of DeepSeek-V3/V3.1 (256 experts). Solid lines are least-squares fits over all experts; dotted lines mark uniform values.

6.1 A public pretrained realization in Switch-Base-256

We evaluate the public `google/switch-base-256` checkpoint on C4 validation (Fedus et al., 2022; Raffel et al., 2020), computing f from top-1 choices before capacity truncation and P from the full-softmax probabilities. On a 1.18M-token evaluation block, encoder layer 11 has $L_{\mathrm{sw}} = 0.86936$ and a Pearson correlation of -0.401 between hard load and soft mass across its 256 experts (Figure 2(A)). Expert 37 receives no hard assignment but retains mean soft mass 0.00830, or 2.12 times the uniform reference. Negative correlation also appears in decoder layer 1, where $L_{\mathrm{sw}} = 0.97347$ and $r = -0.050$. Appendix H gives the settings and full results.

6.2 Top-8 realizations in public base MoEs

We next examine whether soft mass tracks expert utilization in public top-8 MoEs. We evaluate Qwen2-57B-A14B and DeepSeek-V3/V3.1 on the same 4,096 C4 validation documents, taking the first 256 tokens of each document with each model's tokenizer.

Qwen2-57B-A14B has 64 routed experts and uses standard full-softmax top-8 routing, with router auxiliary coefficient 0.001 (Yang et al., 2024). Layers 11–16 all have $L_8 < 1$ and negative correlations between hard load and soft mass. In layer 11, $L_8 = 0.99302$ and expert 49 receives no assignments

across 1,048,576 tokens (Figure 2(B)). Its mean soft mass is 0.016759, above the uniform value $1/64 = 0.015625$. Across the six layers, L_8 ranges from 0.96604 to 0.99761.

DeepSeek-V3 and V3.1 use bias-assisted top-8 routing, with P formed from normalized unbiased sigmoid scores (DeepSeek-AI, 2024; 2025). V3 primarily balances load through bias updates and also uses a small sequence-wise auxiliary loss. Here L_8 measures the relation between bias-assisted dispatch and unbiased scores. In layer 3, the first sparse layer, $L_8 = 0.96175$ for V3 and 0.96601 for V3.1. The hard–soft correlation across its 256 experts is -0.230 in V3 and remains negative at -0.205 in V3.1 after 839B additional pretraining tokens (Figure 2(C)). Removing the correction bias raises L_8 to 0.98127 and 0.98536, respectively. Bias-assisted selection therefore contributes to, but does not fully explain, the separation between soft mass and expert use.

Appendix H tabulates the public-checkpoint measurements and gives per-layer results.

6.3 Observations during end-to-end MoE training

We train a four-layer, 150M-parameter causal language model with two MoE layers from scratch on 16.8M C4 tokens to study how Switch regularization affects routing. Of its parameters, 135M belong to the expert networks. Each MoE layer selects one of 256 experts and weights the expert's output by its full-softmax probability. All parameters are optimized jointly with the language-model loss, Switch loss, and router z-loss (Zoph et al., 2022). We compare coefficients $\alpha \in \{0, 0.01, 0.03\}$ and evaluate routing statistics and negative log-likelihood (NLL) on a fixed set of 1,048,576 held-out C4 tokens. Appendix I gives the settings and repetitions across seeds.

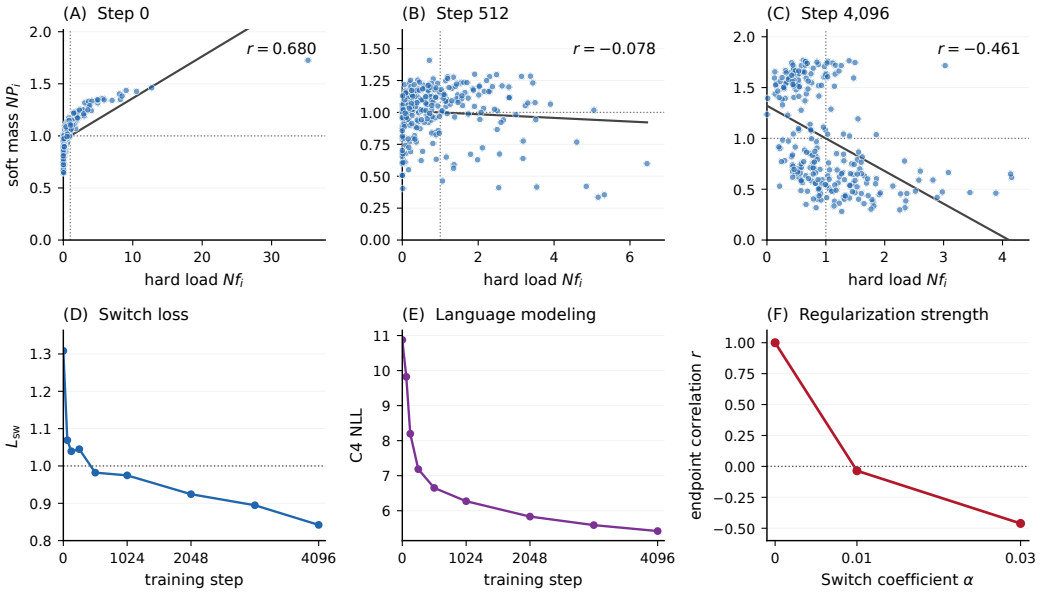

Figure 3: **Routing and language-model loss during C4 training.** (A–C) Hard load and soft mass of the 256 experts in layer 3 at steps 0, 512, and 4,096 of one run with $\alpha = 0.03$. Solid lines are least-squares fits; dotted lines mark the uniform values. (D–E) Layer-3 L_{sw} and NLL on held-out C4 along the same training run. (F) Layer-3 endpoint correlations for different coefficients of the mean Switch loss over the two MoE layers.

At $\alpha = 0.01$, both MoE layers reach Switch losses below one during training, ending at 0.993 in layer 1 and 0.984 in layer 3, while held-out C4 NLL decreases. In layer 1, all 256 experts receive assignments at initialization, but three receive none on the independent evaluation set at the final checkpoint. As in the public checkpoints, a loss below one can coexist with unused experts.

We then examine the relation between hard load and soft mass in layer 3. Without the Switch term, their correlation exceeds 0.999, but dispatch is concentrated on very few experts. With Switch regularization, the endpoint correlation is -0.035 at $\alpha = 0.01$ and becomes more negative at

$\alpha = 0.03$, reaching -0.461 (Figure 3(F)). Hard-load balance improves over the same coefficient comparison (Appendix I). At $\alpha = 0.03$, Figure 3(A–E) shows the correlation turning from positive to negative as the Switch loss falls below one and held-out NLL decreases.

At both regularized endpoints, we independently permute each token's expert labels and recompute its top-1 assignment, preserving all its probability values. All 10,000 randomizations of each endpoint yield positive correlations, with means of 0.531 and 0.452, respectively. Thus Switch training can induce soft-score distributions whose means are anti-aligned with the hard loads they generate. Appendix I.3 reports the control across seeds and proves that its expected loss is at least one.

Summary. Losses below 1, unused experts with above-uniform soft mass, and hard–soft separation occur beyond the extremal constructions. The measured top-1 losses remain far above the sharp infimum. Approaching the infimum requires the selected expert's softmax weight to fall to roughly $2/N$, which can be resisted by the prediction objective as this weight scales the expert's output.

7 Related work

Sparse MoEs scale capacity through conditional expert activation (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Muennighoff et al., 2025). Alongside importance and load penalties, routing research has explored stability (Zoph et al., 2022; Dai et al., 2022), expert-side assignment (Zhou et al., 2022), and balancing through expert bias updates (Wang et al., 2024; Han & Zhong, 2025).

The claim that uniform routing globally minimizes the Switch loss at 1 appears in Switch Transformers and CLIP-MoE; ReMoE attributes this lower bound to softmax normalization (Fedus et al., 2022; Wang et al., 2025; Zhang et al., 2025a). We disprove the claim and derive the sharp infimum and all extremal hard loads. Recent variational theory supports the same interpretation by approximating $f \approx P$ and relating the resulting squared norm to collision entropy (Su & Liu, 2026). Our experiments show that optimizing the Switch loss can itself drive f and P apart.

The ϕ-balancing framework directly balances the mean soft distribution through convex potentials (Chen et al., 2026), avoiding the cross term. Yet soft balance alone does not ensure uniform dispatch: nearly tied gates can send every token to the same expert. Mutual-information and dual-entropy objectives also encourage confident individual routing alongside marginal utilization (Shen et al., 2023; Su & Liu, 2026; Zhang et al., 2025b; Mustafa et al., 2022). Appendix G gives sufficient conditions under which soft balance and routing confidence yield quantitative hard-load guarantees.

8 Conclusion

The Switch loss is not minimized by uniform expert use. For $N \geq 3$, its sharp infimum is approached by routing tokens evenly to roughly half the experts while concentrating soft mass on the unused ones. We characterize all extremal hard loads and show that near-optimal loads must approach this family.

Public MoE checkpoints and controlled training exhibit losses below 1 alongside unused experts. We can also lower the auxiliary loss while keeping expert assignments and routed outputs fixed under top-k routing with post-selection renormalization. Thus, the loss alone cannot establish that dispatch is balanced, and lowering it need not improve balance.

The sharp infimum tells us what the auxiliary loss permits, not which router a language model will learn. Prediction loss, shared parameters, and changing assignments all shape training; their effect on the routing patterns that can be reached remains open.

AI Use Statement

We used generative AI tools to assist with literature searches, checking mathematical arguments, revising the manuscript's wording and structure, and debugging code. Throughout the project, we also used these tools to discuss and obtain feedback on proof strategies, implementation choices, experimental design, and presentation. The authors developed the final arguments and methods, made

all research decisions, and reviewed the mathematical reasoning, references, experimental code, and reported results. The authors take responsibility for the final manuscript.

Reproducibility Statement

Proofs are provided in the main text and appendices. Appendices H–J describe the public-checkpoint evaluations, controlled training, and output-invariant intervention. The supplementary package includes code, configurations, and reported measurements for reproducing the experiments and figures.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Lizhang Chen, Jonathan Li, Qi Wang, Runlong Liao, Shuozhe Li, Chen Liang, Ni Lao, and Qiang Liu. ϕ-balancing for mixture-of-experts training. In *International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.15403.

Damai Dai, Li Dong, Shuming Ma, Bo Zheng, Zhifang Sui, Baobao Chang, and Furu Wei. StableMoE: Stable routing strategy for mixture of experts. In *Annual Meeting of the Association for Computational Linguistics*, pp. 7085–7095, 2022. URL https://aclanthology.org/2022.acl-long.489/.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. DeepSeek-V3.1 release. DeepSeek API documentation, 2025. URL https://api-docs.deepseek.com/news/news250821/.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569, 2022. URL https://proceedings.mlr.press/v162/du22c.html.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Trevor Gale, Deepak Narayanan, Cliff Young, and Matei Zaharia. MegaBlocks: Efficient sparse training with mixture-of-experts. In *Proceedings of Machine Learning and Systems*, volume 5, pp. 288–304, 2023. URL https://proceedings.mlsys.org/paper_files/paper/2023/hash/5a54f79333768effe7e8927bcccffe40-Abstract-mlsys2023.html.

X. Y. Han and Yuan Zhong. A theoretical framework for auxiliary-loss-free load balancing of sparse mixture-of-experts in large-scale AI models. *arXiv preprint arXiv:2512.03915*, 2025. URL https://arxiv.org/abs/2512.03915.

Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. FasterMoE: Modeling and optimizing training of large-scale dynamic pre-trained models. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pp. 120–134, 2022. doi: 10.1145/3503221.3508418. URL https://doi.org/10.1145/3503221.3508418.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. URL https://arxiv.org/abs/2006.16668.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9564–9576, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3e67e84abf900bb2c7cbd5759bfce62d-Abstract.html.

Zihan Qiu, Zeyu Huang, Bo Zheng, Kaiyue Wen, Zekun Wang, Rui Men, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Demons in the detail: On implementing load balancing loss for training specialized mixture-of-expert models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 5005–5018, 2025. URL https://aclanthology.org/2025.acl-long.249/.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Yikang Shen, Zheyu Zhang, Tianyou Cao, Shawn Tan, Zhenfang Chen, and Chuang Gan. ModuleFormer: Modularity emerges from mixture-of-experts. *arXiv preprint arXiv:2306.04640*, 2023. URL https://arxiv.org/abs/2306.04640.

Ye Su and Yong Liu. Variational inference, entropy, and orthogonality: A unified theory of mixture-of-experts. *arXiv preprint arXiv:2601.03577*, 2026. URL https://arxiv.org/abs/2601.03577.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024. URL https://arxiv.org/abs/2408.15664.

Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94dc604e115237a7f4a758b3146cd976-Abstract-Conference.html.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. URL https://arxiv.org/abs/2407.10671.

Longrong Yang, Dong Shen, Chaoxiang Cai, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Solving token gradient conflict in mixture-of-experts for large vision-language model. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/532ce4fcf853023c4cf2ac38cbc5d002-Abstract-Conference.html.

Minghao Yang, Ren Togo, Guang Li, Takahiro Ogawa, and Miki Haseyama. L2R: Low-rank and lipschitz-controlled routing for mixture-of-experts. *arXiv preprint arXiv:2601.21349*, 2026. URL https://arxiv.org/abs/2601.21349.

Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. CLIP-MoE: Towards building mixture of experts for CLIP with diversified multiplet upcycling. In *Conference on Empirical Methods in Natural Language Processing*, pp. 5406–5419, 2025a. URL https://aclanthology.org/2025.emnlp-main.275/.

Tianyu Zhang, Suyuchen Wang, Chao Wang, Juan Rodriguez, Ahmed Masry, Xiangru Jian, Yoshua Bengio, and Perouz Taslakian. SCOPE: Selective cross-modal orchestration of visual perception experts. *arXiv preprint arXiv:2510.12974*, 2025b. URL https://arxiv.org/abs/2510.12974.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7103–7114, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. URL https://arxiv.org/abs/2202.08906.

A ROUTING-CELL GEOMETRY

Full proof of Theorem 3.1. Fix i and $q \in C_i$. Order the other coordinates as $q_{j_1} \geq \cdots \geq q_{j_{N-1}}$ and set $q_{j_N} = 0$. Then

$$q = (q_i - q_{j_1})e_i + \sum_{k=1}^{N-1} (k+1)(q_{j_k} - q_{j_{k+1}})u_{\{i,j_1,\ldots,j_k\}}. \tag{16}$$

All coefficients are nonnegative, and their sum telescopes to $q_i + \sum_{j \neq i} q_j = 1$. Hence $C_i = \mathrm{conv}\{u_S : S \ni i\}$, and minimizing $f^\top q$ proves Eq. (3).

For any router with hard regions A_i, conditional expectation and the cell constraint give Eq. (4). On a nonatomic space, choose disjoint measurable A_i of masses f_i and use a minimizing closed-cell vertex on each. To approximate it with a finite-logit strict router, let D be its support excluding i, put probability $q + \eta$ on i, probability q on D, and probability η^2 on every other coordinate, with q chosen to normalize. For small $\eta > 0$, i is the unique winner, every probability is positive, and the vertex is approached.

B Proof of the ranked quadratic inequality

Fix arbitrary choices $k_r \in \{1, \dots, r\}$, write $\boldsymbol{k} = (k_1, \dots, k_N)$, and set

$$Q_{\boldsymbol{k}}(x) = \sum_{r=1}^{N} \frac{x_r(x_r + s_{k_r - 1})}{k_r}. \tag{17}$$

It suffices to prove $Q_{\boldsymbol{k}} \geq 1/M_N$ for every choice. Let $d_a = x_a - x_{a-1} \geq 0$, $x_0 = 0$, and $w_a = N - a + 1$. Then $\sum_r x_r = w^\top d = 1$ and $Q_{\boldsymbol{k}} = d^\top A d$ for a symmetric coefficient matrix A. For $a \leq b$, each term with $r \geq b$ contributes

$$g_{ab}(k_r) = \frac{1}{k_r} + \frac{(k_r - a)_+ + (k_r - b)_+}{2k_r}. \tag{18}$$

The three ranges have the explicit forms

$$g_{ab}(k) = \begin{cases} 1/k, & k \leq a, \\ 1/2 + (2 - a)/(2k), & a < k \leq b, \\ 1 - (a + b - 2)/(2k), & b < k. \end{cases}$$

For $a \geq 2$, the first branch decreases to $1/a$, and the other two are nondecreasing. For $a = 1 < b$, the middle branch decreases to $(b + 1)/(2b)$ and the last branch increases from that value. If $a = b = 1$, the value is identically one. Hence

$$g_{ab}(k) \geq c_{ab} := \begin{cases} 1, & a = b = 1, \\ (b + 1)/(2b), & a = 1 < b, \\ 1/a, & 2 \leq a \leq b. \end{cases} \tag{19}$$

There are w_b indices $r \geq b$, so $A_{ab} \geq B_{ab} := w_b c_{ab}$ for $a \leq b$, with B symmetric. Since $d \geq 0$, it remains to prove copositivity of $C = M_N B - w w^\top$. For $N \geq 5$, C is entrywise nonnegative. Indeed, for $2 \leq a \leq b$,

$$C_{ab} = w_b(M_N/a - w_a) \geq 0$$

because $M_N \geq a w_a$; also $C_{11} = N(M_N - N) \geq 0$. For $1 < b$,

$$C_{1b} = w_b \left[\frac{M_N(b + 1)}{2b} - N \right] \geq 0.$$

The bracket decreases in b, and its value at $b = N$ is nonnegative because $M_N \geq 2N^2/(N + 1)$ for $N \geq 5$.

For the remaining dimensions,

$$C_3 = \begin{pmatrix} 3 & 0 & -1/3 \\ 0 & 0 & 0 \\ -1/3 & 0 & 1/3 \end{pmatrix}, \qquad d^\top C_3 d = \frac{1}{3}(d_3 - d_1)^2 + \frac{8}{3} d_1^2 \geq 0,$$

and

$$C_4 = \begin{pmatrix} 8 & 3/2 & 0 & -1/4 \\ 3/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1/4 & 0 & 0 & 1/2 \end{pmatrix}, \quad d^\top C_4 d = 3 d_1 d_2 + \frac{1}{2}(d_4 - d_1/2)^2 + \frac{63}{8} d_1^2 \geq 0.$$

Therefore $M_N Q_{\boldsymbol{k}} - (w^\top d)^2 \geq d^\top C d \geq 0$, proving Eq. (10). Equality forces increment support inside $\{a : a(N - a + 1) = M_N\}$. For odd N this leaves one step. For even $N = 2h$, it leaves the adjacent increments $h, h + 1$. As shown in the main proof, equality in the two diagonal coefficient bounds prevents these increments from both being positive.

Exact excess for even N. For even $N = 2h \geq 4$, parameterize the two-increment slice as

$$x(t) = \left(0^{h-1}, \frac{t}{h+t}, \left(\frac{1}{h+t}\right)^h\right), \qquad 0 \leq t \leq 1.$$

The coordinate at rank h has $m_h = x_h/h$. Each of the h larger coordinates has

$$m_r = \min\left\{\frac{1}{h(h+t)}, \frac{1+t}{(h+1)(h+t)}\right\}, \qquad r > h.$$

The minimizing support size changes from $h+1$ to h at $t = 1/h$. Substituting these values into $F_N = \sum_r x_r m_r$ gives

$$F_N(x(t)) - \frac{1}{h(h+1)} = \begin{cases} \dfrac{t(h+t-2)}{(h+1)(h+t)^2}, & 0 \leq t \leq 1/h, \\ \dfrac{(1-t)^2}{(h+1)(h+t)^2}, & 1/h \leq t \leq 1. \end{cases} \tag{20}$$

The only zeros are $t = 0$ and $t = 1$, the two loads in Theorem 4.1.

C Near-extremal stability

Keeping the diagonal slack in Eq. (12) yields Eq. (13). For odd $N = 2h-1 \geq 5$, $\eta_N = (h-1)/(h+1)$ and the unique sorted extremal load is $f^\star = (0^{h-1}, (1/h)^h)$. Telescoping and Cauchy–Schwarz give the conservative explicit consequence

$$\mathrm{TV}(x, f^\star) \leq N\sqrt{N-1}\sqrt{\frac{M_N \epsilon}{\eta_N}} \quad \text{when } F_N(x) \leq M_N^{-1} + \epsilon. \tag{21}$$

For even $N \geq 6$, Eq. (14) places near-minimizers close to the two-increment slice in Appendix B. Its exact excess in Eq. (20) is positive except at the two endpoints. Compactness of the slice and continuity of F_N then force near-minimizers to approach one of those endpoints. For $N = 3, 4$, the same qualitative conclusion follows directly from compactness of the sorted simplex and the equality classification in Lemma 4.2.

D Finite batches

D.1 Fixed-batch infima

On a fixed T-token batch with freely chosen logits, $f_i = n_i/T$ is restricted to a lattice, but Theorem 3.1 still gives the infimum at each grid point. For $1 \leq k \leq \min\{T, N-1\}$, use k active experts and $N-k$ dead experts. Writing $T = qk + r$, $0 \leq r < k$, the active counts are most even when r of them equal $q+1$ and the rest equal q. The support construction from Section 4.1 then has limiting loss

$$U_{N,T}(k) = \frac{N}{N-k+1}\sum_{i \in A}\left(\frac{n_i}{T}\right)^2 = \frac{N}{N-k+1}\left[\frac{1}{k} + \frac{r(k-r)}{kT^2}\right]. \tag{22}$$

Corollary D.1 (Finite-batch gap)**.** *Let $L^\star_{N,T}$ be the infimum over finite-logit routers on a fixed T-token batch. Let $N \geq 3$ and choose the smaller $k^\star \in \arg\max_k k(N-k+1)$. If $T \geq k^\star$, then*

$$0 \leq L^\star_{N,T} - \frac{N}{M_N} \leq \frac{N}{4T^2}. \tag{23}$$

If $k^\star$ divides T, equality with the population sharp value is exact.

Proof. In the balanced-count construction, gates for tokens assigned to $i \in A$ approach the uniform distribution on expert i and the $N-k$ dead experts. Hence

$$\frac{U_{N,T}(k)}{N} = \frac{1}{N-k+1}\sum_{i \in A}(n_i/T)^2.$$

The sum of squares is minimized by counts q and $q+1$, proving Eq. (22). Choosing an extremal $k^\star$ and using $r(k-r) \leq k^2/4$ together with $k^\star \leq N - k^\star + 1$ proves Eq. (23). If $k^\star$ divides T, the remainder term vanishes. □

For $N = 2$, let a_T be an integer nearest to $T/4$. Restricting the completed square in Section 4 to the lattice gives $L^\star_{2,T} = 7/8 + 2(a_T/T - 1/4)^2$.

D.2 Expected IID minibatch loss

This fixed-batch infimum differs from the expected loss of a fixed population router on IID mini-batches: $\widehat{f}$ and $\widehat{P}$ use the same tokens, producing a same-token term of weight $1/T$. The distinction is relevant to micro-batch versus global-batch balancing (Chen et al., 2026; Qiu et al., 2025).

Let $\widehat{f}, \widehat{P}$ be computed from T IID tokens using one population router. A diagonal/off-diagonal decomposition gives

$$\mathbb{E}\sum_i \widehat{f}_i \widehat{P}_i = \left(1 - \frac{1}{T}\right)\sum_i f_i P_i + \frac{1}{T}\mathbb{E}p_{a(X)}(X). \tag{24}$$

Proposition D.2 (Fixed-load IID minibatch formula)**.** *Under the unrestricted nonatomic setup, for prescribed population hard load f,*

$$\inf \mathbb{E}L^{(T)}_{\mathrm{sw}} = N\sum_i f_i \min_{S\ni i}\left\{\left(1 - \frac{1}{T}\right)\frac{\sum_{j\in S} f_j}{|S|} + \frac{1}{T|S|}\right\}. \tag{25}$$

At $T = 1$ the global infimum is one; as $T \to \infty$, Eq. (25) decreases to Theorem 3.1. For the k-active step load it becomes

$$\min\left\{1, \frac{N}{N-k+1}\left(\frac{1-1/T}{k} + \frac{1}{T}\right)\right\}. \tag{26}$$

Proof. Expand $\widehat{f}_i\widehat{P}_i = T^{-2}\sum_{s,t}\mathbf{1}\{a(X_s) = i\}p_i(X_t)$. The $s \neq t$ terms factor as f_iP_i, while $s = t$ gives $\mathbb{E}\mathbf{1}\{a(X) = i\}p_i(X)$, proving Eq. (24). Conditional on A_i, the pointwise linear functional is $q \mapsto (1 - 1/T)f^\top q + q_i/T$. Its value at u_S is the expression in braces in Eq. (25); Theorem 3.1's construction completes the proof. For a k-active step load, a candidate support containing the winner, all dead experts, and s other active experts is monotone in s; endpoints $s = 0, k-1$ yield Eq. (26). □

E The detached update has no global potential

The detached update is a gradient within each routing cell, as described in Section 5. These local gradients cannot in general be joined into the gradient of a globally continuous objective.

Theorem E.1 (Minimal non-potential witness)**.** *For $N = 2$ experts and $T = 2$ tokens, no globally continuous function has a gradient agreeing with the detached-frequency Switch field on every routing cell; for $T = 1$, such a potential exists.*

Let z_t be the binary logit difference and $p_t = \sigma(z_t)$. Then $f_1 = (\mathbf{1}\{z_1 > 0\} + \mathbf{1}\{z_2 > 0\})/2$ and $\gamma = 2f_1 - 1 \in \{-1, 0, 1\}$. Up to a common positive factor, the detached gradient is

$$G(z_1, z_2) = \gamma\big(p_1(1-p_1), p_2(1-p_2)\big). \tag{27}$$

On the both-negative cell, any cellwise potential must be $-p_1 - p_2 + C_-$; on the adjacent mixed cell it must be a constant C_{mix}. Continuity along $z_1 = 0, z_2 < 0$ would require $-1/2 - p_2 + C_- = C_{\mathrm{mix}}$ for every $p_2 \in (0, 1/2)$, a contradiction. When $T = 1$, the field is proportional to $d|p - 1/2|/dz$, which does have a continuous potential.

F Top-k routing

For top-2 routing, we normalize hard loads as $f^{(2)}_i = \frac{1}{2}\mathbb{E}[\mathbf{1}\{i \in S(X)\}]$ and the objective as $L_2 = N\sum_i f^{(2)}_i P_i$, so that uniform routing has reference value one.

Proposition F.1 (A top-2 half-dead construction)**.** *Fix $N \geq 3$ and an integer $m \in \{2, \ldots, N-1\}$. There are finite-logit top-2 routers with m hard-active and $d = N - m$ hard-dead experts whose loss approaches $L_2^{\mathrm{constr}} = 2N/[m(N - m + 2)]$. Their hard-load total variation is d/N and the dead experts carry limiting soft mass $d/(d + 2)$. Choosing m nearest $(N + 2)/2$ gives*

$$\inf L_2 \leq (8 + o(1))/N, \tag{28}$$

while both the dead fraction and hard-load TV approach $1/2$.

F.1 Proof of Proposition F.1

Use one equiprobable routing domain for every unordered pair of the m hard-active experts. On the domain associated with pair $\{i, j\}$, give i and j logit $\varepsilon > 0$, every hard-dead expert logit zero, and every other active expert logit $-M$. Taking $M \to \infty$ and then $\varepsilon \downarrow 0$ preserves the declared top-2 set. Each active expert appears in $m-1$ of the $\binom{m}{2}$ pairs, hence $f_i^{(2)} = 1/m$ on the active set and zero elsewhere. Conditional soft mass is asymptotically uniform over the selected pair and the d dead experts, so $P_i = 2/[m(d+2)]$ for active i and $P_j = 1/(d+2)$ for dead j. Substitution gives $L_2 = 2N/[m(d+2)]$. The TV and dead-mass claims follow directly. Maximizing the concave quadratic $m(N-m+2)$ places m at an integer nearest $(N+2)/2$. □

For $N \in \{8, 16, 32, 64\}$ and $m = (N+2)/2$, the finite-logit construction with $M = 30$ and $\varepsilon = 10^{-5}$ preserves all selected pairs and gives losses within 10^{-4} of their limiting values.

F.2 Output-invariant descent

Finite-batch perturbation. Let j be unused on a finite batch, with

$$\Delta = \min_t \left(\min_{i \in S_t} z_{t,i} - z_{t,j} \right) > 0.$$

For $0 < s < 1$, set

$$z_{t,i}^{(s)} = z_{t,i} + s\Delta \mathbf{1}\{i = j\}. \tag{29}$$

Then j remains outside every S_t, while all selected logits and their post-selection weights stay fixed. With expert outputs fixed, the routed outputs are unchanged. For every $i \neq j$, the full-softmax probability satisfies $p_{t,i}^{(s)} < p_{t,i}$. Because $f_j^{(k)} = 0$, it follows that $L_k^{(s)} < L_k$.

On a population token space, we can reduce the loss without changing routed outputs by lowering all active logits together. Consider finite logits and normalized top-k hard loads $f^{(k)}$, with ties resolved by the same fixed ordering of experts. Let $A = \operatorname{supp}(f^{(k)})$ and $D = [N] \setminus A \neq \varnothing$, and assume $\Delta(x) = \min_{i \in S(x)} z_i(x) - \max_{j \in D} z_j(x) > 0$ almost surely. Since $S(x) \subseteq A$ almost surely, the set $G = \{x : S(x) \subseteq A,\ \Delta(x) > 0\}$ has full measure. For $0 < t < 1$, set

$$z_i^{(t)}(x) = \begin{cases} z_i(x) - t\Delta(x), & x \in G \text{ and } i \in A, \\ z_i(x), & \text{otherwise.} \end{cases}$$

All active–active logit differences are preserved. On G, selected logits remain above every logit in D; outside G, no logit changes. Thus the selected set and the hard loads are unchanged. The selected weights $\widetilde{p}_i(x) = e^{z_i(x)} / \sum_{j \in S(x)} e^{z_j(x)}$ are also unchanged, since on G their common factor cancels. With expert outputs fixed, the routed layer therefore computes the same function.

For $x \in G$, write $Z_A = \sum_{i \in A} e^{z_i}$, $Z_D = \sum_{j \in D} e^{z_j}$, and $q = e^{-t\Delta} \in (0, 1)$. The full-softmax probabilities satisfy

$$p_i^{(t)} = \frac{q(Z_A + Z_D)}{qZ_A + Z_D}\, p_i < p_i \qquad (i \in A).$$

Every active expert's mean soft mass decreases, while $f_i^{(k)} = 0$ on D. Consequently, $L_k^{(t)} = N \sum_i f_i^{(k)} P_i^{(t)} < L_k$ even though the hard loads and model output are unchanged.

G From soft balance to hard balance

A nearly uniform mean soft distribution need not give uniform hard routing. For example, gates that are constant across tokens can approach $u = \mathbf{1}/N$ while keeping expert 1 as the unique winner. Then $P \to u$, but $f = e_1$ and $\mathrm{TV}(f, u) = 1 - 1/N$.

Routing confidence gives a sufficient bound on the hard–soft discrepancy. Define

$$c = \mathbb{E}[1 - \max_i p_i(X)].$$

Pointwise, $\mathrm{TV}(e_{a(x)}, p(x)) = 1 - \max_i p_i(x)$. Taking expectations and using convexity of total variation gives

$$\mathrm{TV}(f, P) \le c, \qquad \mathrm{TV}(f, u) \le \mathrm{TV}(P, u) + c. \tag{30}$$

Thus soft balance together with a small confidence deficit guarantees hard balance. One scalar that controls both quantities is

$$R_\lambda = N\|P\|_2^2 + \lambda c, \qquad \lambda > 0. \tag{31}$$

For a finite collection of softmax gates, R_λ is a continuous, almost-everywhere differentiable function of the logits. Write $\alpha = N\|P - u\|_2^2$ and $\delta = R_\lambda - 1 = \alpha + \lambda c$. The bounds below account for every allocation of δ between these two terms. For $s \ge 0$, let

$$B_s(\delta) := \max_{0 \le z \le \delta} \left\{ s\sqrt{z} + \frac{\delta - z}{\lambda} \right\} = \begin{cases} s\sqrt{\delta}, & \delta \le s^2\lambda^2/4, \\ \delta/\lambda + s^2\lambda/4, & \delta \ge s^2\lambda^2/4. \end{cases} \tag{32}$$

Theorem G.1 (Hard-load certificates)**.** *Let f be the hard argmax load. Then*

$$\mathrm{TV}(f, u) \le B_{s_N}(\delta), \qquad s_N = \frac{\sqrt{\lfloor N^2/4 \rfloor}}{N}, \tag{33}$$

$$\max_i |f_i - 1/N| \le B_{\rho_N}(\delta), \qquad \rho_N = \frac{\sqrt{N-1}}{N}. \tag{34}$$

If $\tau = B_{s_N}(\delta)$, then $|\operatorname{supp}(f)| \ge \lceil N(1-\tau) \rceil$; $B_{\rho_N}(\delta) < 1/N$ certifies no dead expert. Moreover $R_\lambda \ge 1$, with equality in the closed router class exactly for one-hot gates and $f = P = u$.

Proof. Let $v = P - u$, so $\mathbf{1}^\top v = 0$. With r positive coordinates, Cauchy–Schwarz on this set and its complement gives

$$\mathrm{TV}(P, u)^2 \le \frac{r(N-r)}{N}\|v\|_2^2 \le \frac{\lfloor N^2/4 \rfloor}{N^2}\alpha.$$

For any coordinate, minimizing the squared norm of the remaining coordinates at fixed zero sum gives $v_i^2 \le (N-1)\|v\|_2^2/N = (N-1)\alpha/N^2$. The TV constant is sharp when v is constant on each of two sign sets whose sizes differ by at most one; the coordinate constant is sharp when all other coordinates are equal. Since f and P are probability vectors, $\max_i |f_i - P_i| \le \mathrm{TV}(f, P) \le c$. Splitting $f - u$ through P therefore gives $\mathrm{TV}(f, u) \le s_N\sqrt{\alpha} + c$ and $\max_i |f_i - 1/N| \le \rho_N\sqrt{\alpha} + c$. Maximizing $s\sqrt{\alpha} + c$ subject to $\alpha + \lambda c = \delta$ selects $\alpha = \delta$ when $\sqrt{\delta} \le s\lambda/2$, and $\alpha = s^2\lambda^2/4$ otherwise, yielding Eq. (32).

A distribution supported on k coordinates is at TV distance at least $(N-k)/N$ from u, giving the support claim. The coordinate bound rules out a zero entry when it is below $1/N$. Finally, $R_\lambda \ge 1$ because $\alpha, c \ge 0$. Equality requires $P = u$ and $c = 0$; the latter makes the gate one-hot almost surely and hence $f = P$. Conversely, one-hot gates with $f = P = u$ give $R_\lambda = 1$. □

Strongly convex marginal potentials. More generally, suppose a symmetric ϕ is minimized at u and is μ-strongly convex in ℓ_2. If $\phi(P) - \phi(u) + \lambda c \le \epsilon$, then $c \le \epsilon/\lambda$ and $\|P - u\|_2 \le \sqrt{2\epsilon/\mu}$. The triangle inequality and $\mathrm{TV}(P, u) \le \sqrt{N}\|P - u\|_2/2$ give

$$\mathrm{TV}(f, u) \le \frac{\epsilon}{\lambda} + \frac{\sqrt{N}}{2}\sqrt{\frac{2\epsilon}{\mu}}. \tag{35}$$

H PUBLIC-CHECKPOINT EVALUATION

All public-checkpoint measurements use the English validation split of C4 (Raffel et al., 2020). Qwen and DeepSeek use the same 4,096 documents, taking the first 256 tokens of each with the model's tokenizer. Each checkpoint therefore evaluates $T = 1{,}048{,}576$ tokens and $8T = 8{,}388{,}608$ top-8 dispatch positions. The documents have at least 256 tokens under both tokenizer families. Pearson r is computed across the hard-load and soft-mass pairs of all routed experts within a layer. The reported Qwen and DeepSeek layers also have $L_8 < 1$ on each of the 16 disjoint groups of 256 documents that partition the evaluation set. Exact checkpoint revisions, document selection, preliminary evaluations, and additional settings are recorded in the supplementary materials.

Table 1: **Routing statistics in public MoEs evaluated on held-out C4.** Here $L_1 = L_{\mathrm{sw}}$; Pearson r measures the correlation between hard load and soft mass across experts. The Qwen entry gives the ranges over layers 11–16.

Checkpoint	Layer(s)	k	L_k	Pearson r
Switch-Base-256	Encoder 11	1	0.86936	-0.401
Switch-Base-256	Decoder 1	1	0.97347	-0.050
Qwen2-57B-A14B	11–16	8	0.96604–0.99761	$[-0.282, -0.017]$
DeepSeek-V3	3	8	0.96175	-0.230
DeepSeek-V3.1	3	8	0.96601	-0.205

H.1 SWITCH-BASE-256

For `google/switch-base-256` (Fedus et al., 2022), the preliminary and evaluation blocks contain 288 and 2,300 span-corrupted C4 examples, respectively. Each example contains 512 encoder tokens and 114 decoder tokens, giving 1,177,600 tokens for encoder layer 11 and 262,200 tokens for decoder layer 1 on the evaluation block. These two layers were identified on the preliminary block and measured on the separate evaluation block. Table 2 reports their losses and hard–soft correlations on both blocks. Hard loads count top-1 assignments before capacity truncation, and soft masses are the mean full-softmax probabilities. Expert 37 in encoder layer 11 receives no assignments on either block.

Table 2: Switch-Base-256 routing losses and hard–soft correlations on the preliminary and evaluation blocks. Encoder and decoder token counts within each block come from the same input–target pairs.

Layer	Block	Tokens	L_{sw}	Pearson r
Encoder 11	Preliminary	147,456	0.87075	-0.389
Encoder 11	Evaluation	1,177,600	0.86936	-0.401
Decoder 1	Preliminary	32,832	0.97983	-0.038
Decoder 1	Evaluation	262,200	0.97347	-0.050

H.2 QWEN2-57B-A14B

For Qwen2-57B-A14B (Yang et al., 2024), both hard load and soft mass are computed over the 64 routed experts. Hard load is the top-8 assignment count divided by $8T$, and P is the mean full-softmax probability. Layers 11–16 were identified in a preliminary evaluation and re-evaluated on the 4,096-document set (Table 3). On this set, expert 49 in layer 11 has zero assignments and mean soft mass 0.016759.

Table 3: Qwen2-57B-A14B routing statistics on C4. The preliminary set contains 8,192 tokens in packed sequences of length 256; evaluation uses the first 256 tokens of each of 4,096 documents, totaling 1,048,576 tokens. Z counts experts with zero assignments on the corresponding set.

	Preliminary		Evaluation		
Layer	L_8	Z	L_8	Z	Pearson r
11	0.993398	2	0.993017	1	-0.050
12	0.993227	0	0.994662	0	-0.047
13	0.996808	0	0.996026	0	-0.042
14	0.996916	0	0.997609	0	-0.017
15	0.979826	0	0.978967	0	-0.167
16	0.966079	0	0.966041	0	-0.282

H.3 DEEPSEEK-V3 TO V3.1

We evaluate the first sparse router, layer 3, in the official DeepSeek-V3-Base and DeepSeek-V3.1-Base checkpoints (DeepSeek-AI, 2024; 2025). Evaluation runs the checkpoint prefix up to this gate.

The router adds a correction bias to sigmoid scores, restricts routing to four of eight expert groups, and selects eight experts. Hard loads record these selections among the 256 routed experts, normalized by $8T$. Soft mass P averages the per-token distributions obtained by normalizing the unbiased sigmoid scores over these experts. Pearson r is computed across the 256 pairs $(f_i^{(8)}, P_i)$.

For the zero-bias control, we set the correction bias to zero and recompute grouped top-8 assignments on the same inputs, leaving the soft-mass calculation unchanged. Table 4 reports both evaluations. Removing the correction bias raises L_8, but both values remain below one.

Table 4: Layer-3 routing statistics in DeepSeek-V3 and V3.1 on 1,048,576 held-out C4 tokens per checkpoint. Pearson r uses the original bias-assisted assignments.

Checkpoint	L_8	Pearson r	Zero-bias L_8
DeepSeek-V3	0.96175	-0.230	0.98127
DeepSeek-V3.1	0.96601	-0.205	0.98536

I CAUSAL-MOE TRAINING AND COEFFICIENT COMPARISONS

I.1 MODEL, DATA, AND OPTIMIZATION

The causal MoE has 149,781,760 trainable parameters, including 134,610,944 in the expert feed-forward networks. It has four Transformer layers, width 256, and four attention heads. Layers 1 and 3 (zero-based) each contain 256 experts of hidden width 512; the dense feed-forward hidden width is 1,024. The vocabulary has 50,257 entries and the context length is 256. Each token selects $a = \arg\max_i p_i$, and the MoE output is $p_a E_a(x)$, using the full-softmax probability. The language-model loss thus supplies gradients to the router through the selected gate. All tokens are dispatched without capacity truncation.

Training uses the first English training shard of C4, and evaluation uses validation shard 7 (Raffel et al., 2020). We tokenize documents with the GPT-2 tokenizer, append EOS, and pack the stream into 257-ID blocks with stride 256, giving 256 next-token predictions per block. Context resets at each block. The training set contains 65,536 blocks and the evaluation set 4,096 blocks, or 1,048,576 prediction tokens. The two sets share no exact document texts. Dataset and tokenizer revisions are provided in the supplementary materials.

Each run starts from random initialization and trains for 4,096 updates with a batch of 16 sequences, processing 16,777,216 prediction tokens. We use seeds 8601, 8602, and 8603, denoted seeds 1–3 in the figures and tables, at each fixed coefficient $\alpha \in \{0, 0.01, 0.03\}$. Within a seed, initialization and training-data order are shared across coefficients. AdamW uses learning rate 3×10^{-4} after a 256-update linear warm-up, weight decay 0.1, and gradient clipping at norm one. All parameters are optimized with

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \alpha \, \frac{L_{\mathrm{sw}}^{(1)} + L_{\mathrm{sw}}^{(3)}}{2} + 0.001 \, \frac{L_z^{(1)} + L_z^{(3)}}{2},$$

where L_z is the mean squared log-sum-exp of the router logits. Each training Switch loss uses the 4,096 positions in the current batch, with no gradient accumulation. Training uses BF16 with FP32 routers, input jitter 0.01, and no dropout.

We evaluate at steps 0, 64, 128, 256, 512, 1,024, 2,048, 3,072, and 4,096. Evaluation uses FP32 forward passes with TF32 disabled; softmax probabilities and routing sums are computed in FP64. Reported L_{sw} and correlations use f and P aggregated over the entire evaluation set. These are distinct from the minibatch losses used for training. Zero dispatch means zero assignments on this evaluation set.

I.2 Training dynamics and endpoint comparisons

Section 6.3 and Figure 3 report seed 1. Panels A–E show its training run at $\alpha = 0.03$. Its layer-3 correlation changes from 0.680 at initialization to -0.078 at step 512 and -0.461 at step 4,096. Panel F compares its endpoints at the three coefficients. Table 5 reports the final NLL and routing statistics of both MoE layers across all three seeds.

Table 5: Training endpoints on 1,048,576 held-out C4 tokens. Here r is the hard–soft Pearson correlation, $\mathrm{TV} = \frac{1}{2}\sum_i |f_i - 1/256|$ measures hard-load imbalance, and Z counts experts with zero assignments. NLL is the language-model negative log-likelihood on the same evaluation set.

			Layer 1				Layer 3			
Seed	α	NLL	L_{sw}	r	TV	Z	L_{sw}	r	TV	Z
1	0	5.429	54.509	0.978	0.992	252	160.702	1.000	0.990	247
1	0.01	5.418	0.993	−0.027	0.304	3	0.984	−0.035	0.500	0
1	0.03	5.420	1.012	0.067	0.223	0	0.842	−0.461	0.257	0
2	0	5.433	49.833	0.995	0.996	251	179.930	1.000	0.991	245
2	0.01	5.423	0.882	−0.303	0.362	2	0.901	−0.218	0.462	1
2	0.03	5.418	0.936	−0.249	0.228	0	0.851	−0.456	0.244	0
3	0	5.424	42.362	0.997	0.996	251	120.391	0.999	0.988	235
3	0.01	5.432	0.977	−0.076	0.306	5	0.911	−0.207	0.500	1
3	0.03	5.443	0.967	−0.154	0.211	0	0.889	−0.334	0.248	0

At $\alpha = 0.01$, all three layer-1 endpoints have $L_{\mathrm{sw}} < 1$ and zero-dispatch experts. The mean maximum gate probability in layer 3 is 0.0281–0.0597 across the six regularized endpoints.

In layer 3, increasing α from 0.01 to 0.03 lowers the Switch loss and makes the hard–soft correlation more negative in all three seeds. Over the same comparison, hard-load TV decreases in both layers and every expert receives assignments at the endpoint. The unregularized runs instead concentrate dispatch on very few experts. Held-out NLL decreases along each training trajectory; final NLLs across coefficients lie in the range 5.418–5.443.

I.3 Token-wise randomization control

For each of the six layer-3 endpoints at $\alpha = 0.01$ and 0.03, we independently permute the expert labels of every evaluation token's probability vector. We then recompute top-1 selection and aggregate f^{π} and P^{π} over all $T = 1{,}048{,}576$ tokens. Each token retains its complete probability multiset; only its assignment to expert identities changes. Every endpoint uses 10,000 independent randomizations.

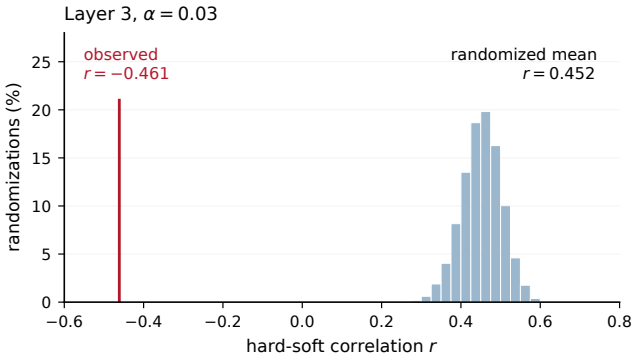

Figure 4: Token-wise randomization of the layer-3 endpoint in Figure 3, C, at $\alpha = 0.03$. The histogram shows 10,000 randomized correlations; the red line marks the observed correlation.

Table 6 gives the six comparisons; Figure 4 shows the endpoint from Figure 3, C. Every randomized correlation is positive, giving a one-sided Monte Carlo p-value of 0.000600 for each endpoint after Bonferroni correction across the six comparisons.

Table 6: Token-wise randomization of layer-3 training endpoints.

Seed	α	Observed r	Randomized mean r
1	0.01	-0.035	0.531
1	0.03	-0.461	0.452
2	0.01	-0.218	0.498
2	0.03	-0.456	0.402
3	0.01	-0.207	0.520
3	0.03	-0.334	0.399

The expected loss under this control can be computed directly. Writing a_t^π for the recomputed assignment, expand

$$L_{\mathrm{sw}}^\pi = \frac{N}{T^2} \sum_{s,t} p_{t,a_s^\pi}^\pi.$$

For $s \neq t$, independence gives $\mathbb{E}_\pi[p_{t,a_s^\pi}^\pi] = 1/N$; for $s = t$, the selected probability is $\max_i p_{t,i}$. Hence

$$\mathbb{E}_\pi[L_{\mathrm{sw}}^\pi] = 1 + \frac{N\bar{m} - 1}{T} \geq 1, \qquad \bar{m} = \frac{1}{T} \sum_t \max_i p_{t,i}. \tag{36}$$

The centered hard–soft inner product is therefore nonnegative in expectation. For these six endpoints, the expected randomized loss is between $1 + 5.9 \times 10^{-6}$ and $1 + 1.4 \times 10^{-5}$, in contrast to the observed $L_{\mathrm{sw}} = 0.842$–0.984.

J Output-invariant intervention in frozen Pythia

Model and expert construction. We use `EleutherAI/pythia-410m-deduped` (Biderman et al., 2023), with all model parameters frozen. The final MLP is replaced by eight experts arranged in four pairs. For each pair, split the original MLP's intermediate neurons into even and odd indices, divide its output bias equally, and multiply each resulting expert output by two. Each pair reconstructs the original MLP under equal weighting, so the routing intervention can be evaluated in the frozen model without training new experts.

Router and data. The router applies a linear map to layer-normalized hidden states. It assigns one logit to each pair and repeats that logit for both members, so top-2 routing selects a complete pair with conditional weights $(1/2, 1/2)$. The first two pair weights are opposite random unit vectors scaled by 0.3, with bias 0.3; the remaining pair weights and biases are zero. Pair j also receives a deterministic offset $-10^{-7}j$. Two pairs are initially active and two unused. There are no gradient updates.

Evaluation uses the first eight nonoverlapping sequences of 64 tokens from the WikiText-2 validation stream (Merity et al., 2017). Nonempty texts are joined by two newlines and tokenized with the checkpoint tokenizer without added special tokens. The 512 routing positions provide 504 next-token predictions, with context reset at each sequence. Five independent random projection directions use seeds 1611, 1722, 1833, 1944, and 2055.

Intervention. Let A and D be the sets of experts with positive and zero hard load, respectively, at $s = 0$ on the evaluation tokens. For token t, define its margin to the unused set by

$$\Delta_t = \min_{i \in S_t} z_{t,i} - \max_{j \in D} z_{t,j} > 0.$$

At strength $s \in \{0, 0.25, 0.5, 0.75, 0.9, 0.99\}$, use

$$z_{t,i}^{(s)} = z_{t,i} - s\Delta_t \mathbf{1}\{i \in A\}.$$

Table 7: Frozen-Pythia intervention endpoints from $s = 0$ to $s = 0.99$ across five router initializations.

Seed	Initial L_2	Final L_2	Reduction	Fixed hard-load TV
1611	1.1863	0.9408	20.69%	0.5000
1722	1.3803	1.0257	25.69%	0.6074
1833	1.1603	0.9278	20.04%	0.5000
1944	1.4138	1.0407	26.39%	0.6152
2055	1.2272	0.9582	21.92%	0.5000

The relative order of active logits is unchanged and every selected logit remains above every unused logit. Selected logits all decrease by the same amount, leaving their conditional weights unchanged. Thus both dispatch and routed outputs are fixed, while full-softmax mass moves from active to unused experts and L_2 decreases. The intervention is applied during the model's forward pass at each strength.

Results. Table 7 reports the endpoints for all five router initializations. At every intermediate strength, the selected sets and conditional weights remain unchanged in the recorded evaluation, with zero measured NLL change. Hard-load TV and the number of unused experts also remain fixed along each path.