CONCENTRATION-SCALED GRASSMANNIAN ROUTING FOR CONTROLLABLE MIXTURE-OF-EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A sparse mixture-of-experts model is useful only if its router can specialize experts without starving them and can adapt inference cost after training. Standard top-k routers expose the first problem during training but provide only an indirect post-hoc compute interface. We propose Grassmannian MoE (GrMoE), which represents each expert by an orthonormal subspace and scores a normalized token by concentration-scaled projection energy. The resulting affinity is basis invariant and bounded, while a global concentration scale α controls the sharpness of the routing distribution from one checkpoint. We distinguish these geometry-specific properties from generic consequences of softmax scaling and give a conditional load-balance bound under explicit mixture-balance and effective-margin assumptions. At 350M parameters, GrMoE improves perplexity and severe-concentration frequency over standard top-k routing; after matching descriptor parameters, scoring MACs, diversity regularization, expert compute, and realized fan-out, it retains a smaller 0.14-PPL residual over the strongest Euclidean control. A fixed checkpoint reaches two active experts across $\tau \in \{0.03, 0.05, 0.07\}$ with only 0.03 PPL spread, and downstream, $N = 64/128$, second-family, expert-parallel, and memory studies preserve the same scoped picture. We therefore present GrMoE as a geometric routing primitive for stable one-checkpoint compute control, not as a universal or frontier-scale routing claim.

1 INTRODUCTION

A sparse MoE deployment often changes its compute budget after training, yet the router is usually trained around a fixed dispatch rule. Moving a conventional model from top-2 toward top-1 can change utilization abruptly, while temperature scaling alone does not specify how many experts a thresholded executor will realize. The practical problem is therefore twofold: the router must avoid persistent expert starvation during training, and the deployed checkpoint should expose a predictable quality–compute operating curve.

Existing routing mechanisms address different parts of this problem. Sparsely gated top-k routing is simple and scalable (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022), but its balance is sensitive to the learned logits and auxiliary objectives. Expert Choice and auxiliary-loss-free routing improve utilization (Zhou et al., 2022; Wang et al., 2024); Hash routing removes learned allocation (Roller et al., 2021); Soft MoE replaces hard token choice with soft token–expert interactions (Puigcerver et al., 2024); and fine-grained routed/shared-expert designs change the expert organization itself (Dai et al., 2024). These methods are useful comparators, but none makes a bounded higher-rank expert descriptor the basis of a post-training concentration dial.

GrMoE starts from a different representation of expertise. An expert that responds to several correlated directions is more naturally described by a subspace than by one routing vector. Each expert therefore owns an orthonormal frame $U_e \in \mathbb{R}^{d \times k_r}$, and a normalized token x receives affinity $a_e(x) = \|U_e^\top x\|_2^2$. A learned positive concentration κ_e and an inference-time scale α form logits $\alpha\kappa_e a_e(x)$. The Matrix Bingham family motivates the concentration semantics (Chikuse, 2003; Hoff, 2009; Kume & Wood, 2005), but the implemented router is an ordinary discriminative softmax and never evaluates a Bingham normalizer.

The revision is explicit about what geometry does and does not establish. Basis invariance, bounded projection energy, and multi-directional affinity follow from the subspace construction. Entropy

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.

monotonicity and top-k mass concentration under positive logit scaling are generic softmax facts. The load-balance result is conditional on balanced mixture mass and a positive effective margin; a held-out audit checks whether that premise occurs in trained representations rather than treating it as automatic.

The empirical claim is similarly narrow. At 350M, ten-seed results show lower severe-concentration frequency and better validation perplexity than standard top-k routing. A capacity- and FLOP-matched three-seed study reduces the quality gap substantially but leaves a consistent -0.14 PPL difference relative to the strongest low-rank Euclidean control. More importantly for deployment, one GrMoE checkpoint reaches a two-active-expert target across nearby thresholds with nearly unchanged perplexity. Downstream evaluation, two decoder families, $N = 64$ and $N = 128$ routing regimes, two-GPU expert parallelism, and peak-memory accounting test the main alternative explanations. The 1.3B and 3B experiments remain trend evidence because they have three seeds and a narrower evaluation suite.

Our contributions are therefore: a basis-invariant concentration-scaled subspace router; a scoped theory separating geometric properties from generic softmax concentration; matched controls that isolate a small residual beyond router capacity and diversity; and a measured one-checkpoint compute-control protocol whose calibration, tail fan-out, systems cost, and failure boundaries are reported explicitly. Complete proofs, seed-level controls, dense sensitivity tables, secondary scale results, and implementation audits are in the appendix.

2 Grassmannian Routing

2.1 Subspace affinities and concentration

Let $x_{\mathrm{raw}} \in \mathbb{R}^d$ denote the representation entering an MoE router. The router uses the explicit normalization $x = x_{\mathrm{raw}}/(\|x_{\mathrm{raw}}\|_2 + \epsilon)$; LayerNorm or RMSNorm alone is not assumed to produce unit Euclidean norm. Expert $e \in \{1, \dots, N\}$ is represented by $U_e \in \mathbb{R}^{d\times k_r}$ with $U_e^\top U_e = I$. Its affinity is

$$a_e(x) = \|U_e^\top x\|_2^2 = x^\top U_e U_e^\top x. \tag{1}$$

This score depends only on the subspace spanned by U_e, not on the selected basis. Computing all expert affinities costs $O(Ndk_r)$ rather than $O(Nd)$ for a rank-1 linear router.

Each expert has a positive concentration

$$\kappa_e = \mathrm{softplus}(\eta_e) + \kappa_{\min}, \qquad g_e^{(\alpha)}(x) = \frac{\exp\{\alpha\kappa_e a_e(x)\}}{\sum_{j=1}^N \exp\{\alpha\kappa_j a_j(x)\}}. \tag{2}$$

Training uses $\alpha = 1$. At inference, $\alpha = 0$ is uniform and increasing α concentrates probability on high-affinity experts without changing the learned subspaces. For token-dependent ambiguity, the amortized variant uses $h_\phi(x) = N\,\mathrm{softmax}(\mathrm{MLP}_\phi(x))$ and replaces κ_e by $h_\phi(x)_e\kappa_e$, so $\sum_e h_\phi(x)_e = N$ and the amortizer redistributes rather than globally rescales concentration.

2.2 Sparse dispatch and training

For threshold τ, the sparse set is $S_\tau^{(\alpha)}(x) = \{e : g_e^{(\alpha)}(x) \geq \tau\}$. If it is empty, the expert with the largest routing probability is activated. Probabilities are renormalized over the actually dispatched set. A fixed $K_{\max}$ may cap that thresholded set for static buffers; this systems cap is not part of the theory, and the reported $K_{\max} = 4$ operating point clips only 0.7% of tokens at the two-active-expert calibration.

Training minimizes language-model loss plus a redundancy penalty,

$$\mathcal{L} = \mathcal{L}_{\mathrm{LM}} + \beta \sum_{e<j} \max\big(0, \|U_e^\top U_j\|_F^2 - \rho_0 k_r\big)\,. \tag{3}$$

The default is $\beta = 0.01$ and $\rho_0 = 0.3$. At 350M, $d = 768$ and $k_r = 48$, so independent random subspaces have expected overlap $k_r^2/d = 3.0$, well below the 14.4 penalty threshold; the regularizer prevents near-duplicate spans rather than forcing orthogonality between experts.

The sparse execution contract is common across compared routers. Training uses capacity factor 1.25; overflow is rerouted to the next-highest eligible non-full expert, gate weights are renormalized over the realized set, and a token uses the residual path if every eligible expert is full. Non-dispatched experts receive zero token-path gradient for that token. The maximum observed residual-only incidence is 0.08%; no input token is discarded from the network.

2.3 Maintaining the Stiefel constraint

Frames are initialized from the Haar distribution by QR decomposition of Gaussian matrices. After every AdamW update, each frame receives a thin-QR retraction with deterministic diagonal-sign correction; AdamW moments remain in ambient coordinates, so this is not presented as exact Riemannian Adam (Absil et al., 2008; Becigneul & Ganea, 2019). Across training, the median and maximum $\|U_e^\top U_e - I\|_F$ errors are 1.9×10^{-5} and 7.6×10^{-5}. The retraction adds 0.42% to measured training-step time and has no inference-time cost.

2.4 Bingham motivation and implementation cost

The Matrix Bingham perspective supplies concentration semantics but is not part of the deployed likelihood computation. GrMoE never evaluates a Bingham normalizer; training and inference use the discriminative softmax in Eq. 2, so all guarantees in Section 3 apply directly to the implemented gate rather than to exact Bayesian inference. For N experts, scoring costs $O(Ndk_r)$ and the full redundancy penalty has $\binom{N}{2}$ terms. We evaluate all 28 pairs for $N = 8$ and sample $M = 4N$ unordered pairs with the appropriate rescaling for $N \geq 16$; matched short runs at 1.3B/3B change CV by at most 0.002 relative to full-pair regularization. The additional router state is small relative to the model, but the $N = 128$ narrow-expert stress test shows that scoring can become material when expert computation shrinks, so the method is not claimed to have scale-free routing overhead.

3 Theory

The first two properties are specific to the subspace descriptor; the next two are generic consequences of scaling a fixed logit vector. This separation matters because α is not itself a Grassmannian theorem: geometry supplies a bounded, basis-invariant score to which the generic concentration mechanism is applied.

Proposition 1 (Subspace invariance and bounded affinity)**.** *For any orthogonal $O \in O(k_r)$, replacing U_e by $U_e O$ leaves $a_e(x)$ unchanged. If $\|x\|_2 = 1$ and $U_e^\top U_e = I$, then $0 \leq a_e(x) \leq 1$.*

With the measured orthogonality error $\varepsilon = 7.6 \times 10^{-5}$, the empirical relaxation is $a_e(x) \leq 1 + \varepsilon$ for unit inputs. Thus concentration acts on a norm-controlled affinity rather than an unconstrained dot-product magnitude.

Theorem 1 (Generic softmax concentration)**.** *Let $s_e(x) = \kappa_e a_e(x)$ and $g^{(\alpha)} = \mathrm{softmax}(\alpha s)$. Then*

$$\frac{\partial H(\alpha, x)}{\partial \alpha} = -\alpha \mathrm{Var}_{e \sim g^{(\alpha)}}[s_e(x)] \leq 0. \tag{4}$$

For $\alpha > 0$, the inequality is strict whenever the logits are not all equal.

This identity holds for any scaled softmax logits. The corresponding mass statement is also generic once the logits are fixed.

Corollary 1 (Top-k mass control)**.** *If $s_{(1)} \geq \cdots \geq s_{(N)}$ and $\delta_k = s_{(k)} - s_{(k+1)} \geq 0$, then $G_k^{(\alpha)} = \sum_{i=1}^k g_{(i)}^{(\alpha)}$ satisfies*

$$G_k^{(\alpha)} \geq 1 - (N - k)e^{-\alpha \delta_k}. \tag{5}$$

The value of GrMoE is therefore not that softmax scaling is new, but that the scaled pre-concentration affinities have the invariance and boundedness of Proposition 1.

Theorem 2 (Conditional load-balance bound)**.** *Let inputs follow a uniform mixture $x \sim N^{-1} \sum_i \mathcal{D}_i$. For $x \sim \mathcal{D}_i$, suppose $\|U_i^\top x\|^2 \geq \gamma$ and $\|U_j^\top x\|^2 \leq \rho\gamma$ for $j \neq i$. With*

Table 1: Held-out audit of the effective-margin premise. Clusters are formed without router outputs and matched to experts only on a separate calibration split.

Diagnostic	Value
Positive effective margin	96.8%
Median / 5th-percentile margin	0.42 / 0.018
Independent cluster-mass CV	0.083
Cluster–expert bootstrap stability	89.4%

$\kappa_{\min} = \min_e \kappa_e$, $\kappa_{\max} = \max_e \kappa_e$, *and* $\Delta = \gamma(\kappa_{\min} - \rho\kappa_{\max}) > 0$, *the coefficient of variation of expected expert load satisfies*

$$\mathrm{CV}(\bar{g}) \le (N-1)e^{-\alpha\Delta}, \qquad \bar{g}_e = \mathbb{E}_x[g_e^{(\alpha)}(x)]. \tag{6}$$

Corollary 2 (Amortized conditional load balance)**.** *If* $h_\phi(x)_i\kappa_i a_i(x) - h_\phi(x)_j\kappa_j a_j(x) \ge \Delta_h > 0$ *for every non-target* j *on mixture component* i*, then* $\mathrm{CV}(\bar{g}) \le (N-1)e^{-\alpha\Delta_h}$.

The load-balance argument has a simple proof structure. On component i, the assumed projection and concentration margins lower-bound the target logit gap by Δ, which makes every non-target probability at most exponentially small in $\alpha\Delta$. Uniform mixture mass then places each expected expert load inside an $O(e^{-\alpha\Delta})$ envelope around $1/N$; the amortized result repeats the same step with the effective token-dependent gap Δ_h. The theorem is therefore conditional on the representation supplying the margin rather than a claim that optimization must create it.

A held-out audit checks that premise without using router outputs to define the groups. Pre-router representations are clustered into N groups, clusters are matched to experts by Hungarian assignment on a calibration split, and the effective margin is evaluated on a disjoint 100k-token split. Positive margins occur for 96.8% of tokens, with median 0.42, fifth percentile 0.018, cluster-mass CV 0.083, and 89.4% cluster-to-expert bootstrap match stability (Appendix E). These diagnostics make the premise empirically plausible in the evaluated model but do not imply that the conservative exponential bound tightly predicts observed language-model load. Complete proofs are in Appendix I.

4 Experiments

4.1 Setup

All models are trained from scratch. The primary model has 350M total and approximately 150M active parameters, 12 layers, $d = 768$, eight experts in every second block, and $k_r = 48$. It is trained for 15B tokens; the 1.3B and 3B pilots use 30B tokens and three seeds each. The training mixture is 60% FineWeb-Edu, 25% The Stack v2, and 15% OpenWebMath, drawn by fixed token quotas after source-level shuffling and without replacement until each quota is filled, with manifest seed 2026. All runs use the GPT-2 tokenizer, sequence length 1024, global batch 256, AdamW with peak learning rate 3×10^{-4} and cosine decay. Baselines include Softmax Top-k (Shazeer et al., 2017), Switch (Fedus et al., 2022), Expert Choice (Zhou et al., 2022), Hash routing (Roller et al., 2021), Soft MoE (Puigcerver et al., 2024), DeepSeekMoE-style routing (Dai et al., 2024), auxiliary-loss-free routing (Wang et al., 2024), and geometry controls described below. Unless stated otherwise, $\pm$ is standard deviation across independent model seeds.

For utilization, let p_e be the fraction of realized expert assignments received by expert e and $N_{\text{eff}} = (\sum_e p_e^2)^{-1}$. We call a run *persistently severely concentrated* only when three conditions hold together during the final 20% of training: $N_{\text{eff}}/N \le 0.50$, at least 25% of experts each receive less than $0.1/N$ of total load, and both conditions persist for at least 10 consecutive evaluations. Moderate nonuniformity that does not satisfy this rule is treated as specialization rather than collapse.

4.2 Primary routing results

Table 2 gives the ten-seed 350M comparison. GrMoE with amortized concentration reaches 17.8 ± 0.1 PPL and load CV 0.075 ± 0.02, compared with 18.6 ± 0.4 and 0.245 ± 0.08 for Softmax Top-2. No GrMoE seed meets the severe-concentration rule, whereas four of ten Softmax Top-2 seeds do.

Table 2: 350M MoE results over 10 seeds. "Severe" uses the persistent three-part concentration rule in Section 4.

Method	PPL↓	CV↓	Severe↓	H
Softmax Top-2	$18.6_{\pm 0.4}$	$.245_{\pm .08}$	4/10	1.10
Switch	$18.9_{\pm 0.3}$	$.155_{\pm .05}$	2/10	0.88
Expert Choice	$18.8_{\pm 0.2}$	$.068_{\pm .01}$	0/10	1.71
Hash	$20.2_{\pm 0.1}$	$.038_{\pm .01}$	0/10	2.05
Soft MoE	$18.2_{\pm 0.3}$	$.115_{\pm .04}$	1/10	1.41
DeepSeekMoE-style	$18.0_{\pm 0.2}$	$.095_{\pm .03}$	0/10	1.35
AuxFree	$18.2_{\pm 0.2}$	$.088_{\pm .02}$	0/10	1.05
GrMoE	$18.0_{\pm 0.2}$	$.089_{\pm .02}$	0/10	1.34
GrMoE + amort.	$\mathbf{17.8_{\pm 0.1}}$	$\mathbf{.075_{\pm .02}}$	0/10	1.28

Table 3: Mechanism ablations at 350M. Full regularization and routing-rank sweeps are in Appendix N.

Variant	PPL↓	CV↓	Severe↓
Global κ only	$18.0_{\pm 0.2}$	$.089_{\pm .02}$	0/10
Bias-only token correction	$17.9_{\pm 0.2}$	$.086_{\pm .02}$	0/10
Amortized concentration	$\mathbf{17.8_{\pm 0.1}}$	$\mathbf{.075_{\pm .02}}$	0/10
No subspace regularizer	$18.1_{\pm 0.3}$	$.168_{\pm .05}$	2/10
$k_r = 96$	$18.3_{\pm 0.3}$	$.105_{\pm .03}$	0/10

Expert Choice attains slightly lower CV than non-amortized GrMoE, but not the same perplexity or post-hoc compute interface.

The severe-concentration label is not a proxy for ordinary specialization. Within Softmax Top-2, the four severe seeds finish 1.01 PPL worse than the six nonsevere seeds, and event-aligned PPL rises after the first severe event. By contrast, a method-adjusted comparison among nonsevere runs gives only $+0.01$ PPL for moderate specialization versus near-uniform use, with a 95% interval spanning zero (Appendix E).

4.3 Mechanism ablations

The main components are not interchangeable sources of extra router capacity. Table 3 shows that multiplicative amortized concentration improves both PPL and load CV beyond a bias-only token correction, while removing the subspace-overlap regularizer approximately doubles CV and reintroduces severe concentration. The routing-rank sweep is also non-monotone: $k_r = 48$ attains 17.8 ± 0.1 PPL and 0.075 ± 0.02 CV, whereas doubling the rank to 96 worsens these to 18.3 ± 0.3 and 0.105 ± 0.03. Thus the gains are not explained by increasing descriptor dimension alone; the full rank and regularization sweeps are in Appendix N.

4.4 Stability and specialization diagnostics

The severe-event analysis also connects utilization to quality continuously rather than relying only on a binary count. Across saved checkpoints, minimum N_{eff}/N and subsequent validation PPL have Spearman correlation -0.61 with seed-clustered bootstrap 95% CI $[-0.72, -0.47]$. In an event-aligned analysis, the two evaluations after the first severe event are $+0.38$ PPL above the pre-event level and $+0.34$ above matched stable windows. Conversely, among nonsevere runs, moderate specialization ($0.50 < N_{\text{eff}}/N \leq 0.80$) differs from near-uniform use by only $+0.01$ PPL with 95% interval $[-0.07, 0.09]$. This distinction is why the paper reserves "severe concentration" for persistent starvation rather than equating all nonuniform routing with failure.

The learned geometry is likewise heterogeneous rather than collapsing to a common temperature. At 350M, the maximum normalized subspace overlap is 0.28, learned concentrations span 0.6–3.8, and routing entropy falls from 1.62 in early layers to 1.20 in late layers; at 1.3B the corresponding values are 0.31, 0.4–4.2, 1.65, and 1.15. The correlation between expert concentration and token-type entropy is -0.81 at 350M and -0.87 at 1.3B, consistent with high-κ experts behaving as narrower specialists. Initialization is not load-bearing: Haar, PCA, and identity-block starts finish at 17.8, 17.7, and 17.9 PPL, respectively, with no observed severe event. Full token-level diagnostics and routing-rank sweeps appear in Appendix M and Appendix N.

Table 4: Capacity/MAC-matched 350M control at two realized active experts, three paired seeds. Full router latency, thresholds, fan-out, and seed-level values are in Appendix A.

Router	Params/MACs	Diversity	PPL↓	CV↓
Rank-1 Euclidean	.006M/.006M	no	$18.42_{\pm.16}$	$.172_{\pm.028}$
Multi-key, 48 keys	.295M/.295M	no	$18.08_{\pm.12}$	$.106_{\pm.016}$
Multi-key	.295M/.295M	yes	$18.01_{\pm.10}$	$.094_{\pm.014}$
Low-rank	.295M/.295M	no	$18.03_{\pm.10}$	$.099_{\pm.013}$
Low-rank	.295M/.295M	yes	$17.96_{\pm.05}$	$.088_{\pm.011}$
Small MLP	.295M/.295M	no	$18.18_{\pm.15}$	$.125_{\pm.020}$
Small MLP	.295M/.295M	yes	$18.11_{\pm.14}$	$.112_{\pm.018}$
GrMoE + amort.	.295M/.295M	yes	$\mathbf{17.82}_{\pm.06}$	$\mathbf{.076}_{\pm.009}$

Table 5: Calibration of one GrMoE checkpoint to two active experts under nearby frozen thresholds.

τ	α	Active	p95/p99	PPL↓
0.03	2.25	2.00	4/6	17.98
0.05	1.78	2.00	3/4	17.95
0.07	1.49	2.00	3/4	17.96

4.5 Geometry versus router capacity

The submitted rank-1 Euclidean α-router rules out thresholding as the entire explanation, but it does not match GrMoE's descriptor capacity. We therefore compare multi-key, low-rank, and small-MLP routers at 350M while matching descriptor parameters and scoring MACs, using the same amortizer, data, optimizer, expert architecture, method-blind training fan-out controller, cumulative expert compute, and held-out-calibrated test fan-out. Diversity-matched variants receive normalized span or logit penalties with the same initial loss contribution. For $N = 8, d = 768, k_r = 48$, multi-key, low-rank, and GrMoE descriptors each have 294,912 parameters and approximately 0.295M scoring MACs per token; the MLP width is chosen to match that budget within 0.2%. A disjoint calibration batch updates the training threshold every 2,000 steps to target two active experts without using validation quality, and each saved router is calibrated once on the same held-out split before test evaluation. Cumulative expert FLOPs differ by only 0.8% among the matched methods.

Table 4 shows the decision-critical comparison. The strongest Euclidean control is low-rank routing with matched span diversity at 17.96 ± 0.05 PPL and 0.088 ± 0.011 CV; GrMoE reaches 17.82 ± 0.06 and 0.076 ± 0.009. The paired seed differences are $-0.16, -0.13, -0.13$ PPL, giving a mean difference of -0.14 with 95% CI $[-0.18, -0.10]$. The gap is much smaller than the rank-1 comparison, so the supported conclusion is a residual advantage after matching capacity and diversity, not that those factors are irrelevant.

4.6 One-checkpoint compute control

Realized fan-out depends jointly on α and τ, so α should be calibrated rather than treated as a direct expert-count parameter. The deployment procedure first selects τ or $K_{\max}$ from a held-out tail-cardinality constraint, checks aggregate fan-out over a coarse α bracket, and then bisects to the desired mean fan-out. Calibration over 10M held-out tokens costs 2.8 A100 GPU-minutes once per deployment budget.

A single 350M checkpoint is stable across nearby thresholds (Table 5). At a two-active-expert target, $\tau \in \{0.03, 0.05, 0.07\}$ gives PPL $17.98, 17.95, 17.96$. The strongest matched Euclidean control spans 0.14 PPL over the same thresholds, compared with 0.03 for GrMoE; at $\tau = 0.03$ the Euclidean control is actually 0.04 PPL better, so the result is threshold robustness rather than uniform quality dominance. Tokenwise active count need not be monotone: 1.3% of tokens are non-monotone across adjacent grid points even though mean fan-out decreases throughout the tested range.

The original fixed-τ operating curve remains useful for deployment context. At $\tau = 0.05$, increasing α from 1 to 2 changes average active experts from 3.1 to 1.8, PPL from 17.8 to 18.0, and measured single-A100 throughput by $1.24\times$; at $\alpha = 5$, the model approaches one active expert at 18.6 PPL. A $K_{\max} = 4$ cap changes the calibrated two-expert point by only $+0.01$ PPL. We therefore call α a post-hoc compute-control dial, not a guarantee of an optimal operating point.

Table 6: One-checkpoint operating curve at fixed $\tau = 0.05$. Throughput is relative to $\alpha = 1$.

α	Avg. active	PPL↓	Throughput↑
0.5	5.2	18.1	0.82×
1.0	3.1	17.8	1.00×
2.0	1.8	18.0	1.24×
5.0	1.1	18.6	1.41×

Table 7: Full parameter/FLOP-matched 350M router comparison, three paired seeds. Latency is measured at batch 16 and sequence length 1024.

Router	Desc. params	MACs/token	Latency	Div.	Train fan-out	Frozen τ	Test fan-out	PPL↓	CV↓
Rank-1 Euclidean α-router	.006M	.006M	.30 ms	no	$2.01 \pm .03$	.061	$2.00 \pm .02$	$18.42 \pm .16$	$.172 \pm .028$
Multi-key Euclidean, 48 keys	.295M	.295M	.43 ms	no	$2.02 \pm .02$	.041	$2.00 \pm .02$	$18.08 \pm .12$	$.106 \pm .016$
Multi-key + span diversity	.295M	.295M	.43 ms	yes	$2.02 \pm .02$	.044	$2.00 \pm .02$	$18.01 \pm .10$	$.094 \pm .014$
Unconstrained low-rank	.295M	.295M	.39 ms	no	$2.01 \pm .03$	.047	$2.01 \pm .02$	$18.03 \pm .10$	$.099 \pm .013$
Low-rank + span diversity	.295M	.295M	.39 ms	yes	$2.02 \pm .02$	.049	$2.00 \pm .01$	$17.96 \pm .05$	$.088 \pm .011$
Small MLP, width 380	.295M	.295M	.40 ms	no	$2.01 \pm .02$	.034	$2.00 \pm .02$	$18.18 \pm .15$	$.125 \pm .020$
Small MLP + logit diversity	.295M	.295M	.40 ms	yes	$2.01 \pm .02$	.036	$2.00 \pm .02$	$18.11 \pm .14$	$.112 \pm .018$
GrMoE + amortizer	.295M	.295M	.40 ms	yes	$2.02 \pm .02$	.050	$2.00 \pm .01$	$\mathbf{17.82 \pm .06}$	$\mathbf{.076 \pm .009}$

A global α is the primary interface, but a saved-checkpoint early/late two-group calibration tests whether layer-dependent concentration is needed. At the same 2.0 mean fan-out, grouped calibration lowers layerwise fan-out CV from 0.23 to 0.06 while changing PPL from 17.95 to 17.92, with paired 95% CI $[-0.08, 0.02]$ (Appendix F). The near-null quality change favors the simpler global dial while documenting grouped calibration as an optional systems refinement.

A controlled synthetic routing task isolates the intended geometry before moving to language data. With eight ground-truth subspaces in $d = 128$, GrMoE obtains 91.7% assignment accuracy in the easy regime versus 88.5% for the strongest listed non-GrMoE baseline, while the hard regime gives 78.3% versus 68.1% for Softmax Top-1. Softmax Top-1 severely concentrates in 6/10 hard-setting seeds and GrMoE in none. The synthetic result is mechanism evidence only; it does not substitute for the language-model comparisons (Appendix K).

4.7 Deployment diagnostics and failure boundaries

The one-dimensional deployment interface is useful only after separating aggregate fan-out control from tokenwise behavior. On the full saved-checkpoint grid $\alpha \in \{0.5, 1, 1.5, 2, 3, 5\}$ and $\tau \in \{0.03, 0.05, 0.07\}$, mean fan-out decreases monotonically for every tested threshold, but 1.3% of tokens have a non-monotone active-count sequence between adjacent grid points. This is expected because scaling a softmax guarantees entropy contraction, not monotonicity of every nonmaximum probability. The deployment recipe therefore calibrates aggregate fan-out on held-out tokens and separately checks p95/p99 cardinality rather than treating α as an exact per-token expert count. At the two-active operating point, $K_{\max} = 4$ clips only 0.7% of tokens and changes PPL by $+0.01$, giving a static dispatch bound when fixed buffers are required.

Threshold robustness is also comparative rather than automatic. Repeating the same held-out two-active calibration on the strongest matched Euclidean low-rank router yields PPL 17.94 ± 0.05, 17.97 ± 0.05, and 18.08 ± 0.07 for $\tau = 0.03, 0.05, 0.07$, a 0.14-PPL range versus 0.03 for GrMoE. The Euclidean control is 0.04 PPL better at $\tau = 0.03$, so the claim is not uniform quality dominance; it is lower sensitivity to the nearby threshold choices at fixed realized compute. Calibration itself is inexpensive relative to training: 10M held-out tokens require 2.8 A100 GPU-minutes for GrMoE and 2.6 GPU-minutes for the three saved Euclidean checkpoints. These checks make the post-hoc claim auditable while exposing its boundary: α controls a deployment curve, not an exact tokenwise active set.

4.8 Evaluation breadth and systems checks

The perplexity gain is not obtained by sacrificing the evaluated downstream tasks. On frozen 350M checkpoints, the mean over zero-shot HellaSwag, PIQA, ARC-Easy, WinoGrande and five-shot MMLU is 45.18 for Softmax Top-2, 45.86 for AuxFree, 46.62 for GrMoE at $\alpha = 1$, and 46.34 at matched two-active compute (Table 9). MMLU is close to four-choice chance at this scale and is not used to support a reasoning claim. From one checkpoint, increasing α from 1 to 2 changes

Table 8: Paired comparison against the strongest Euclidean control. Negative differences favor GrMoE.

Seed	LR+div PPL	GrMoE PPL	ΔPPL	LR+div CV	GrMoE CV	ΔCV
17	17.91	17.75	−0.16	.098	.084	−0.014
23	17.97	17.84	−0.13	.090	.078	−0.012
41	18.00	17.87	−0.13	.076	.066	−0.010
Mean ± SD	17.96 ± .05	17.82 ± .06	−0.14	.088 ± .011	.076 ± .009	−0.012

Table 9: Frozen 350M downstream accuracy (%), three paired seeds. MMLU is five-shot; the other tasks are zero-shot.

Method	Hella	PIQA	ARC-E	Wino	MMLU	Mean
Softmax Top-2	41.2 ± .4	62.3 ± .3	44.8 ± .5	51.9 ± .6	25.7 ± .2	45.18
AuxFree	42.3 ± .3	62.9 ± .3	45.7 ± .4	52.5 ± .5	25.9 ± .2	45.86
GrMoE, $\alpha = 1$	43.2 ± .3	63.5 ± .3	46.8 ± .4	53.4 ± .4	26.2 ± .2	46.62
GrMoE, 2.0 active	42.9 ± .3	63.3 ± .3	46.4 ± .4	53.1 ± .4	26.0 ± .2	46.34

the downstream mean only from 46.67 to 46.24; the largest individual drop is -0.9 on HellaSwag (Appendix C).

The 1.3B checkpoints provide a stronger knowledge-oriented check. At matched two-active compute, GrMoE averages 42.74 over SciQ, BoolQ, OpenBookQA, ARC-Challenge, and five-shot TriviaQA, compared with 41.60 for Softmax Top-2 and 42.20 for AuxFree; the seed-paired differences are $+1.14$ points versus Softmax, 95% CI $[0.78, 1.50]$, and $+0.54$ versus AuxFree, 95% CI $[0.26, 0.82]$. After the same 1,500-step UltraChat-200k SFT recipe, GrMoE also improves all four IFEval aggregations over Softmax Top-2: prompt strict/loose is 32.9/40.3 versus 31.5/38.8, and instruction strict/loose is 40.4/48.2 versus 38.9/46.7. These three-seed evaluations support preservation of the tested downstream quality, not a general reasoning or instruction-following advantage.

Routing difficulty is tested separately from model scale. In the corrected $N = 64$ study, GrMoE calibrated to four experts reaches 17.98 ± 0.11 PPL and CV 0.084 ± 0.014, versus 18.71 ± 0.25 and 0.258 ± 0.045 for Softmax Top-4; calibrating to 2.5 experts gives 18.14 ± 0.12 PPL. The $N = 128$ stress test extends the candidate pool further: GrMoE at eight active experts reaches 22.97 ± 0.12 PPL and 0.098 ± 0.014 CV versus 23.84 ± 0.29 and 0.312 ± 0.041 for Softmax Top-8. Because narrow experts make router scoring 8.3% of counted forward MACs there, this is routing-quality evidence rather than an efficiency result (Appendix D). A matched 350M LLaMA-style decoder gives 17.31 ± 0.10 PPL and 0.078 ± 0.013 CV for GrMoE versus 17.62 ± 0.16 and 0.093 ± 0.018 for AuxFree, extending the result to a second evaluated decoder family (Appendix F).

The throughput result comes from sparse expert execution rather than cheaper routing. On one A100 at batch 16 and sequence length 1024, $\alpha = 1$ and $\alpha = 2$ take 12.50 and 10.08 ms, while an $\alpha = 2$ dense-expert control takes 12.45 ms. On two A100-80GB GPUs with EP=2, matched two-active GrMoE reaches $56{,}900 \pm 460$ aggregate tokens/s versus $54{,}600 \pm 420$ for Softmax Top-2, with nearly unchanged all-to-all share; peak allocated memory changes from 29.84 to 29.95 GiB in training and 7.42 to 7.43 GiB in inference. These measurements establish single-node inference compatibility, not multi-node or distributed-training scalability (Appendix G).

Finally, the 1.3B and 3B pilots preserve the language-model trend: GrMoE + amortization reaches 13.3 ± 0.1 and 11.1 ± 0.1 PPL, respectively, with no severe event in the three runs at either scale. Their smaller seed count, fewer baselines, and incomplete sensitivity suite are why they remain trend evidence rather than a scaling claim (Appendix L).

5 Related Work

Sparse MoE routing trades dense model capacity for conditional expert execution. Sparsely gated MoE, GShard, and Switch establish token-choice top-k routing at scale (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022). Expert Choice changes the allocation direction by letting experts select tokens (Zhou et al., 2022); Hash routing removes a learned router (Roller et al., 2021); Soft MoE uses soft token-slot mixtures (Puigcerver et al., 2024); auxiliary-loss-free routing changes how balance is enforced (Wang et al., 2024); and DeepSeekMoE-style designs refine the organization of routed and shared experts (Dai et al., 2024). GrMoE changes the descriptor used to score

Table 10: 1.3B quality beyond perplexity, three seeds. "Know." averages SciQ, BoolQ, OpenBookQA, ARC-Challenge, and five-shot TriviaQA; IFEval uses the identical UltraChat-200k SFT recipe.

Method	Know.	Prompt S.	Prompt L.	Instr. S.	Instr. L.
Softmax Top-2	41.60	$31.5 \pm .8$	$38.8 \pm .9$	$38.9 \pm .7$	$46.7 \pm .8$
AuxFree	42.20	$32.1 \pm .7$	$39.5 \pm .8$	$39.6 \pm .7$	$47.4 \pm .7$
GrMoE, $\alpha = 1$	**43.12**	$\mathbf{32.9 \pm .6}$	$\mathbf{40.3 \pm .7}$	$\mathbf{40.4 \pm .6}$	$\mathbf{48.2 \pm .7}$

Table 11: Fine-grained $N = 64$ routing, three seeds. Rerouted is capacity-overflow incidence; residual-only is the final fallback after all eligible experts are full.

Method	Active	PPL↓	CV↓	$N_{\text{eff}}/N \uparrow$	Severe	Rerouted	Resid.-only
Softmax Top-4	4.00	$18.71 \pm .25$	$.258 \pm .045$	$.938 \pm .020$	1/3	1.30%	.08%
AuxFree Top-4	4.00	$18.29 \pm .17$	$.101 \pm .018$	$.990 \pm .004$	0/3	.25%	.02%
GrMoE, 4.0	$4.01 \pm .03$	$17.98 \pm .11$	$.084 \pm .014$	$.993 \pm .002$	0/3	.18%	.01%
GrMoE, 2.5	$2.50 \pm .02$	$18.14 \pm .12$	$.088 \pm .015$	$.992 \pm .003$	0/3	.08%	$<.01\%$

an expert: a token is compared with an orthonormal subspace rather than one Euclidean key, after which ordinary sparse dispatch is applied.

The geometric component is related to neural representations and optimization on non-Euclidean or matrix manifolds (Liu et al., 2018; Ganea et al., 2018; Absil et al., 2008; Becigneul & Ganea, 2019). The Grassmannian identifies frames that span the same subspace, making projection energy invariant to basis rotations. Matrix Bingham models provide concentration semantics for distributions over subspaces (Chikuse, 2003; Hoff, 2009; Kume & Wood, 2005); GrMoE borrows that semantics but does not perform normalized Bingham inference. The closest controls in this paper are therefore not only rank-1 routers but also parameter- and MAC-matched multi-key, unconstrained low-rank, and small-MLP routers. Their comparison narrows the empirical claim from a broad geometry advantage to a small residual quality difference plus a more stable measured compute-control interface.

Post-hoc compute control also intersects with the systems problem of variable fan-out. A temperature can sharpen any softmax, but realized sparse compute is set jointly by probability scale, a dispatch rule, per-expert capacity, and the serving buffer. GrMoE does not remove these constraints: the measured protocol calibrates α after fixing τ or $K_{\max}$, reports p95/p99 cardinality, and separates router time from expert execution. The distinction is important because a smoother probability distribution is not itself an inference-speed result. Our systems evidence is therefore paired with dense-expert, batch-size, and EP=2 controls rather than inferred from average active count alone.

6 Discussion, Limitations, and Conclusion

GrMoE replaces rank-1 expert keys with orthonormal subspaces and uses concentration-scaled projection energy as a one-checkpoint compute-control interface. Capacity- and diversity-matched controls reduce the original rank-1 gap substantially but leave a paired 0.14-PPL residual at 350M, while nearby thresholds recover the same two-expert budget within 0.03 PPL. Mechanism ablations show that neither added rank nor token-dependent flexibility alone reproduces the full result, and downstream, fine-grained, second-family, and EP=2 checks preserve the scoped empirical pattern. The evidence remains bounded: several breadth studies use three seeds, the 1.3B/3B results are trend evidence, and the $N = 128$ narrow-expert setting makes router cost material. GrMoE is not a zero-shot retrofit for pretrained MoEs, and the systems study does not establish multi-node or distributed-training scalability. Thin-QR retraction adds 0.42% measured step overhead, while peak memory rises by 0.4% in training and 0.1% in inference. The supported conclusion is therefore a geometric routing primitive with stable sparse execution and post-hoc compute control in the tested regimes, with fixed-k large-pool routing, broader families, multi-node serving, and explicit router adaptation left for further validation.

Use of AI Assistance

Generative AI tools were used only to assist with language editing, clarity, and formatting of the manuscript. All research ideas, theoretical derivations, experimental design, implementation, anal-

ysis, and conclusions were developed and verified by the authors. AI-generated suggestions were reviewed and edited by the authors, who take full responsibility for the content of the paper.

ETHICS STATEMENT

This work studies routing and compute control in mixture-of-experts language models and does not involve human subjects or the collection of private or personally identifiable data. The experiments use existing large-scale language-model training and evaluation datasets described in the paper. As with other methods that improve the efficiency or capability of language models, the proposed approach may inherit broader risks associated with their deployment. We therefore limit our claims to the evaluated settings and explicitly report the method's computational costs, deployment assumptions, and observed failure boundaries.

REPRODUCIBILITY STATEMENT

We provide the architectural, optimization, data-mixture, routing, capacity, and calibration details required to reproduce the reported experiments. The paper specifies model configurations, training budgets, random-seed counts, routing and overflow rules, inference-time calibration, matched baselines, and evaluation protocols. Complete proofs, seed-level results, sensitivity analyses, routing diagnostics, and additional implementation details are included in the appendices. We also distinguish primary multi-seed results from smaller-scale exploratory experiments and report the limitations of the evaluated scaling and systems settings.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Gary Becigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019.

Yasuko Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2003.

Damai Dai, Chengqi Deng, Chenggang Zhao, R.x. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y.k. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Peter D Hoff. Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2): 438–456, 2009.

Alfred Kume and Andrew TA Wood. Saddlepoint approximations for the bingham and fisher–bingham normalising constants. *Biometrika*, 92(2):465–476, 2005.

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. In *International Conference on Learning Representations*, 2021.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024.

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason E Weston. Hash layers for large sparse models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

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.

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.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

A Matched Router Controls

The capacity-matched study uses three paired seeds because its purpose is causal attribution rather than re-estimating the ten-seed primary result. All compared routers share the 350M backbone, experts, data order, optimizer, concentration amortizer, capacity factor, overflow rule, and token budget. A method-blind controller updates the training threshold every 2,000 steps from a disjoint calibration batch to target two active experts without using validation quality. At inference, each router's threshold is calibrated once on the same held-out split to two average active experts and then frozen. This per-router calibration avoids favoring any method through an arbitrary common numerical threshold when logit scales differ.

For $N = 8$, $d = 768$, and $k_r = 48$, multi-key, low-rank, and GrMoE descriptors contain $Ndk_r = 294{,}912$ parameters and require approximately 0.295M scoring MACs per token. The small MLP uses width 380 and matches this budget within 0.2%. The shared amortizer contributes 0.075M parameters, so full matched routers contain 0.370M router parameters; the rank-1 reference contains 0.081M. Cumulative expert FLOPs differ by only 0.8% among the matched methods.

The paired GrMoE difference is -0.14 PPL, 95% CI $[-0.18, -0.10]$, and -0.012 CV, 95% CI $[-0.017, -0.007]$. Diversity matching narrows the gap, which is why the paper attributes only this residual to the constrained orthonormal parameterization in the tested setting.

A.1 Submitted 1.3B rank-1 control

The earlier 1.3B control remains useful for a different question: whether dynamic thresholding alone explains the effect. It is not capacity-matched and is therefore kept separate from Table 7.

Table 12: Rank-1 Euclidean dynamic-dispatch control on the 1.3B model, three seeds.

Method	Descriptor	Control	Avg. active	PPL↓	Severe↓
Softmax Top-2	rank-1 Euclidean	fixed top-2	2.0	$14.3_{\pm 0.4}$	3/3
Softmax, temp.-scaled	rank-1 Euclidean	temperature	2.0	$14.5_{\pm 0.4}$	3/3
Euclidean α-router	rank-1 Euclidean	α + threshold	2.0 avg.	$14.1_{\pm 0.3}$	2/3
GrMoE + amort., $\alpha = 1.9$	rank-k_r subspace	α + threshold	2.0 avg.	$13.4_{\pm 0.1}$	0/3
GrMoE + amort., $\alpha = 1.0$	rank-k_r subspace	α + threshold	3.2 avg.	$13.3_{\pm 0.1}$	0/3

Table 13: Systems checks. The single-A100 speedup disappears when expert FFNs remain dense; EP=2 is single-node only.

Setting	Reference	GrMoE operating point	Change
A100 latency, batch 16	12.50 ms ($\alpha = 1$)	10.08 ms ($\alpha = 2$)	1.24×
Dense-expert control	12.50 ms	12.45 ms	1.00×
EP=2 tokens/s	$54{,}600 \pm 420$	$56{,}900 \pm 460$	+4.2%
Train / inference peak	29.84 / 7.42 GiB	29.95 / 7.43 GiB	+0.4 / +0.1%

Table 14: PPL / average active experts from one frozen 350M GrMoE checkpoint.

τ	$\alpha = .5$	$\alpha = 1$	$\alpha = 1.5$	$\alpha = 2$	$\alpha = 3$	$\alpha = 5$
.03	18.05 / 6.0	17.78 / 4.0	17.82 / 2.9	17.92 / 2.2	18.18 / 1.55	18.55 / 1.16
.05	18.10 / 5.2	17.80 / 3.1	17.86 / 2.3	18.00 / 1.8	18.27 / 1.34	18.60 / 1.10
.07	18.16 / 4.5	17.86 / 2.7	17.95 / 1.98	18.10 / 1.55	18.38 / 1.20	18.72 / 1.04

B Concentration and Threshold Sensitivity

Mean fan-out decreases monotonically over the tested α range for every threshold, but this aggregate observation is not a tokenwise theorem. Across adjacent grid points, 1.3% of tokens have a non-monotone active-count sequence because a nonmaximum softmax probability can increase before eventually decreasing under scaling.

The matched Euclidean control spans 0.14 PPL over these thresholds versus 0.03 for GrMoE. At $\tau = .03$ it is 0.04 PPL better than GrMoE, which is why the paper describes threshold stability rather than uniform dominance. The Euclidean calibration requires 2.6 A100 GPU-minutes for the three saved checkpoints.

C Downstream Evaluation

The matched-compute GrMoE point is $+1.16$ mean accuracy points over Softmax Top-2 and $+0.48$ over AuxFree. MMLU is near the four-choice chance level at this scale and is not used to support a reasoning claim.

At matched compute, the seed-paired mean-task difference is $+1.14$ points versus Softmax Top-2, 95% CI $[0.78, 1.50]$, and $+0.54$ versus AuxFree, 95% CI $[0.26, 0.82]$. These are prespecified descriptive contrasts and are not multiplicity-adjusted across tasks or methods.

The base checkpoints are not instruction tuned, so IFEval is evaluated after the same UltraChat-200k supervised fine-tuning recipe for 1,500 optimizer steps, with identical data order, optimizer, schedule, and three paired seeds. GrMoE is trained and evaluated at $\alpha = 1$, $\tau = .050$; decoding is greedy with temperature 0, top-$p = 1$, and at most 1,024 new tokens.

The paired 95% intervals for GrMoE minus Softmax are $[0.38, 2.42]$, $[0.33, 2.67]$, $[0.54, 2.46]$, and $[0.44, 2.56]$ in table order. These exploratory intervals are not multiplicity-adjusted; the study shows no instruction-following loss under this matched SFT recipe rather than a general post-SFT advantage.

D Fine-Grained Expert Pools

The $N = 64$ comparison uses the same backbone and expert widths within that configuration, 352M total parameters, 151M active parameters at four experts, the same data, capacity rule, and token budget. Because the total-to-active parameter ratio changes with N/k, no claim is made that one expert-width rescaling simultaneously matches both quantities to the original $N = 8$ model.

The effective-count values in Table 21 are the corrected aggregation: for each seed p_e is computed from the same realized assignment vector as CV, $N_{\text{eff}} = (\sum_e p_e^2)^{-1}$, and $N_{\text{eff}}/N = 1/(1 + \text{CV}(p)^2)$. The Severe column is pathwise, while CV and N_{eff}/N summarize the terminal held-out window; one severe Softmax seed rebalanced before that terminal window.

Table 15: Identical two-active-expert calibration for the strongest matched Euclidean control.

τ	Calibrated α	Active	p95/p99	PPL↓
.03	2.10	2.00	4/6	$17.94 \pm .05$
.05	1.66	2.00	4/5	$17.97 \pm .05$
.07	1.34	2.00	3/4	$18.08 \pm .07$

Table 16: Original post-hoc operating curve at $\tau = 0.05$ on the 350M checkpoint. Throughput is relative to $\alpha = 1$.

α	Avg. active experts	PPL↓	Throughput↑
0.5	5.2	18.1	.82×
1.0	3.1	17.8	1.00×
2.0	1.8	18.0	1.24×
5.0	1.1	18.6	1.41×
Softmax Top-2	2.0	18.6	1.00×
Softmax Top-1	1.0	19.5	1.35×

To stress a much larger candidate pool, the $N = 128$ experiment uses a 160M model with $d = 512$, 12 layers, MoE in every second block, expert hidden width 128, $k_r = 16$, nominal top-8, capacity factor 1.25, three seeds, and 5B training tokens.

At $N = 128, d = 512, k_r = 16$, GrMoE scoring costs 1.05M MACs per token per MoE layer, or 16× rank-1 scoring. Across six MoE layers this is 6.29M MACs, 8.3% of the explicitly counted 75.8M end-to-end forward MACs including attention at sequence length 1024 and the language-model head. This configuration therefore tests routing quality rather than serving efficiency. A fixed-k comparison at $N \in \{64, 128\}$ and the quality–cost tradeoff from reducing k_r remain open.

E Collapse and Margin Audits

The severe rule is fixed before joining anonymized utilization traces to method identity or validation quality: $N_{\text{eff}}/N \leq 0.50$, at least 25% of experts below $0.1/N$ load, and persistence for ten consecutive evaluations within the final 20% of training. Across saved checkpoints, minimum N_{eff}/N and subsequent PPL have Spearman correlation -0.61, seed-clustered bootstrap 95% CI $[-0.72, -0.47]$.

Pre-router representations are clustered into N groups without router outputs, clusters are matched to experts by a Hungarian assignment on a calibration split, and margins are evaluated on a disjoint split. The audited margin is $m(x) = h_i \kappa_i \|U_i^\top x\|^2 - \max_{j \neq i} h_j \kappa_j \|U_j^\top x\|^2$. The positive fifth percentile supports the premise in this audit, but the theorem additionally idealizes mixture balance and uses a conservative worst-case exponential bound.

F Second Decoder Family and Layerwise Calibration

The LLaMA-style decoder uses RMSNorm, SwiGLU, and RoPE while holding data, token budget, expert count, total/active parameters, and training protocol fixed. The result supports the two evaluated families only.

The -0.03 PPL difference has paired 95% CI $[-0.08, 0.02]$. The global dial remains the primary interface; grouped concentration is an optional systems refinement when balanced per-layer fan-out matters.

G Systems and Memory Accounting

The router cost is unchanged across α; the gain appears only when fewer expert FFNs execute. On the same hardware and sequence length, the $\alpha = 1 \to 2$ speedup is 1.12×, 1.18×, 1.24×, and 1.26× at batch sizes 1, 4, 16, and 32, respectively.

The 1.3B, $N = 16$ benchmark uses global batch 16, sequence length 1024, bf16, experts 1–8 and 9–16 on separate shards, 50 warm-up and 200 synchronized timed iterations, and NCCL

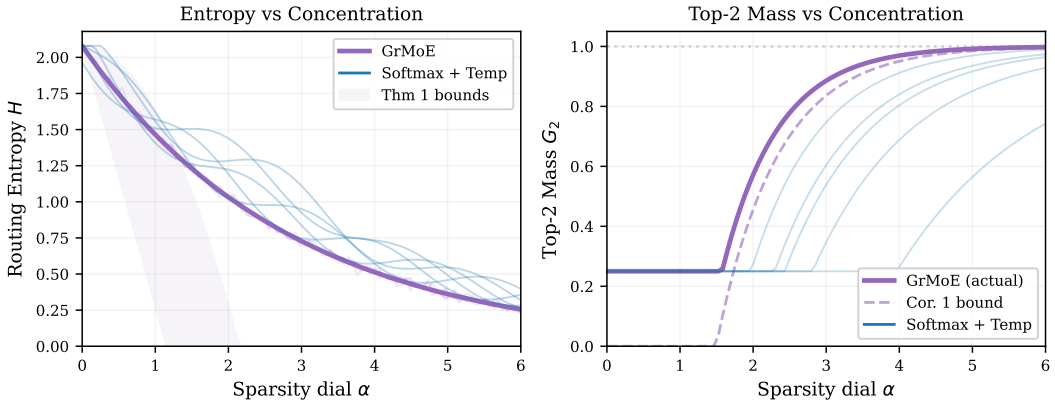

Figure 1: Post-hoc concentration control from one checkpoint. Routing entropy decreases as α increases, while top-2 routing mass concentrates. The entropy identity is generic to scaled softmax logits; the bounded projection-energy construction is GrMoE-specific.

Table 17: Frozen 350M downstream accuracy (%), three paired seeds. HellaSwag, PIQA, ARC-Easy, and WinoGrande are zero-shot; MMLU is five-shot.

Method	HellaSwag	PIQA	ARC-Easy	WinoGrande	MMLU	Mean
Softmax Top-2	$41.2 \pm .4$	$62.3 \pm .3$	$44.8 \pm .5$	$51.9 \pm .6$	$25.7 \pm .2$	45.18
Rank-1 Euclidean α	$41.8 \pm .4$	$62.6 \pm .3$	$45.2 \pm .4$	$52.2 \pm .5$	$25.8 \pm .2$	45.52
AuxFree	$42.3 \pm .3$	$62.9 \pm .3$	$45.7 \pm .4$	$52.5 \pm .5$	$25.9 \pm .2$	45.86
GrMoE, $\alpha = 1$	$43.2 \pm .3$	$63.5 \pm .3$	$46.8 \pm .4$	$53.4 \pm .4$	$26.2 \pm .2$	46.62
GrMoE, matched 2.0 active	$42.9 \pm .3$	$63.3 \pm .3$	$46.4 \pm .4$	$53.1 \pm .4$	$26.0 \pm .2$	46.34

`all_to_all_single` with realized variable split sizes. Staging buffers use $K_{\max} = 4$. At matched 2.0-active compute GrMoE is 4.2% faster than Softmax Top-2; at the $\alpha = 2$ point it is 7.7% faster. The within-GrMoE $\alpha = 1 \to 2$ speedup is $1.28\times$, 95% CI $[1.26, 1.30]\times$. No multi-node or communication-dominated expert-parallel claim is made.

The backward pass stores $U_e^\top x$ for each token, expert, and subspace dimension. At 350M with $N = 8$, $k_r = 48$, six MoE layers, batch 16, sequence length 1024, and bf16, the projection tensor occupies 75.5 MB across six layers. Relative to rank-1 logits, the incremental saved projection activation is 73.9 MB. GrMoE with amortization adds approximately 0.364M router parameters per MoE layer; parameters, gradients, and two fp32 AdamW moment buffers add 26.2 MB across six layers under the no-fp32-master-copy configuration, for an analytical incremental minimum of about 100.1 MB before temporary workspaces.

At inference, projection activations are not retained; the full bf16 router-parameter increase including the amortizer is approximately 4.37 MB. Dispatch staging buffers are allocated to the same $K_{\max} = 4$ bound for both methods.

Table 18: Downstream accuracy under the α sweep from one frozen 350M checkpoint.

α	Avg. active	Downstream mean (%)	Largest per-task drop vs. $\alpha = 1$
1.0	3.1	46.67	–
1.5	2.3	46.52	-0.4 on ARC-Easy
2.0	1.8	46.24	-0.9 on HellaSwag

Table 19: Knowledge-oriented accuracy (%) on existing 1.3B checkpoints, three seeds. TriviaQA reports five-shot exact match; the other tasks are zero-shot.

Method	SciQ	BoolQ	OpenBookQA	ARC-C	TriviaQA	Mean
Softmax Top-2	$76.8 \pm .5$	$58.4 \pm .7$	$34.0 \pm .6$	$29.7 \pm .5$	$9.1 \pm .4$	41.60
AuxFree	$77.5 \pm .4$	$59.0 \pm .6$	$34.6 \pm .5$	$30.3 \pm .5$	$9.6 \pm .3$	42.20
GrMoE, $\alpha = 1$	$78.4 \pm .4$	$60.1 \pm .5$	$35.6 \pm .5$	$31.2 \pm .4$	$10.3 \pm .3$	43.12
GrMoE, 2.0 active	$78.0 \pm .4$	$59.7 \pm .5$	$35.2 \pm .5$	$30.8 \pm .4$	$10.0 \pm .3$	42.74

H Routing Algorithm

Algorithm 1 GrMoE sparse forward pass

Input: token x, frames $\{U_e\}_{e=1}^N$, concentrations $\{\kappa_e\}_{e=1}^N$, amortizer h_ϕ, concentration α, threshold τ, optional cap $K_{\max}$
Output: MoE output y

$\bar{x} \leftarrow x/(\|x\|_2 + \epsilon)$
for $e = 1, \dots, N$ **do**
  $a_e \leftarrow \|U_e^\top \bar{x}\|_2^2$
end for
$h \leftarrow h_\phi(x)$; $\ell_e \leftarrow \alpha h_e \kappa_e a_e$; $g \leftarrow \mathrm{softmax}(\ell)$
$S \leftarrow \{e : g_e \geq \tau\}$; if $S = \emptyset$, set $S \leftarrow \{\arg\max_e g_e\}$
if $K_{\max}$ is enabled and $|S| > K_{\max}$ **then**
  retain the $K_{\max}$ elements of S with largest g_e
end if
apply capacity-aware rerouting; if no eligible expert remains, retain the residual path
$\tilde{g}_e \leftarrow g_e / \sum_{j \in S} g_j$ for realized $e \in S$
$y \leftarrow \sum_{e \in S} \tilde{g}_e f_e(x)$
return y

I Proofs

I.1 Proof of Proposition 1

Proof. For any $O \in O(k_r)$, $(U_e O)^\top x = O^\top U_e^\top x$. Therefore

$$\|(U_e O)^\top x\|^2 = x^\top U_e O O^\top U_e^\top x = x^\top U_e U_e^\top x = \|U_e^\top x\|^2. \tag{7}$$

Since $P_e = U_e U_e^\top$ is an orthogonal projector, its eigenvalues are 0 or 1. Hence $0 \leq x^\top P_e x \leq x^\top x$. If $\|x\|_2 = 1$, the affinity lies in $[0, 1]$. □

I.2 Proof of Theorem 1

Proof. Let $Z(\alpha) = \sum_j \exp(\alpha s_j)$ and $g_i = \exp(\alpha s_i)/Z(\alpha)$. The entropy can be written as $H(\alpha) = \log Z(\alpha) - \alpha \sum_i g_i s_i$. Since $\partial_\alpha \log Z = \sum_i g_i s_i$, differentiating gives

$$\frac{\partial H}{\partial \alpha} = \sum_i g_i s_i - \sum_i g_i s_i - \alpha \frac{\partial}{\partial \alpha} \sum_i g_i s_i. \tag{8}$$

The remaining derivative is the variance of s under $g^{(\alpha)}$:

$$\frac{\partial}{\partial \alpha} \sum_i g_i s_i = \sum_i g_i s_i^2 - \left(\sum_i g_i s_i \right)^2 = \mathrm{Var}_{g^{(\alpha)}}[s_i]. \tag{9}$$

Thus $\partial_\alpha H = -\alpha \mathrm{Var}_{g^{(\alpha)}}[s_i] \leq 0$. □

Table 20: Matched-SFT IFEval accuracy (%) on 1.3B checkpoints, three paired seeds.

Method	Prompt strict	Prompt loose	Instr. strict	Instr. loose
Softmax Top-2 + SFT	$31.5 \pm .8$	$38.8 \pm .9$	$38.9 \pm .7$	$46.7 \pm .8$
AuxFree + SFT	$32.1 \pm .7$	$39.5 \pm .8$	$39.6 \pm .7$	$47.4 \pm .7$
GrMoE + SFT	$32.9 \pm .6$	$40.3 \pm .7$	$40.4 \pm .6$	$48.2 \pm .7$

Table 21: Corrected $N = 64$ fine-grained study, three seeds. Rerouted is overflow incidence; residual-only is the final fallback rate after all eligible experts are full.

Method	Active	PPL↓	Downstream↑	CV↓	N_{eff}/N ↑	Severe	Rerouted	Residual-only
Softmax Top-4	4.00	$18.71 \pm .25$	$45.02 \pm .31$	$.258 \pm .045$	$.938 \pm .020$	1/3	1.30%	.08%
AuxFree Top-4	4.00	$18.29 \pm .17$	$45.81 \pm .27$	$.101 \pm .018$	$.990 \pm .004$	0/3	.25%	.02%
GrMoE, calibrated 4.0	$4.01 \pm .03$	$17.98 \pm .11$	$46.51 \pm .24$	$.084 \pm .014$	$.993 \pm .002$	0/3	.18%	.01%
GrMoE, calibrated 2.5	$2.50 \pm .02$	$18.14 \pm .12$	$46.13 \pm .25$	$.088 \pm .015$	$.992 \pm .003$	0/3	.08%	<.01%

I.3 Proof of Corollary 1

Proof. Let $s_{(1)} \geq \cdots \geq s_{(N)}$. The tail-to-head mass ratio satisfies

$$\frac{1 - G_k^{(\alpha)}}{G_k^{(\alpha)}} = \frac{\sum_{j=k+1}^{N} \exp(\alpha s_{(j)})}{\sum_{i=1}^{k} \exp(\alpha s_{(i)})} \leq \frac{(N-k)\exp(\alpha s_{(k+1)})}{\exp(\alpha s_{(k)})}. \tag{10}$$

Thus $(1 - G_k^{(\alpha)})/G_k^{(\alpha)} \leq (N-k)\exp(-\alpha\delta_k)$. Since $G_k^{(\alpha)} \leq 1$, it follows that $1 - G_k^{(\alpha)} \leq (N-k)\exp(-\alpha\delta_k)$. □

I.4 Proof of Theorem 2

Proof. For $x \sim \mathcal{D}_i$, the target concentrated affinity satisfies $s_i \geq \kappa_{\min}\gamma$, while any non-target $j \neq i$ satisfies $s_j \leq \kappa_{\max}\rho\gamma$. Hence $s_i - s_j \geq \Delta$. The target routing probability is bounded as

$$g_i^{(\alpha)}(x) = \frac{1}{1 + \sum_{j \neq i} \exp[-\alpha(s_i - s_j)]} \geq \frac{1}{1 + (N-1)\exp(-\alpha\Delta)}. \tag{11}$$

Let $\epsilon = (N-1)\exp(-\alpha\Delta)$. Then $g_i^{(\alpha)}(x) \geq 1 - \epsilon$. For $x \sim \mathcal{D}_j$ with $j \neq i$, the same margin gives $g_i^{(\alpha)}(x) \leq \exp(-\alpha\Delta) = \epsilon/(N-1)$. Therefore the expected load of expert i obeys $(1-\epsilon)/N \leq \bar{g}_i \leq (1+\epsilon)/N$. Since $N^{-1}\sum_i \bar{g}_i = 1/N$, the across-expert variance is at most ϵ^2/N^2, and

$$\mathrm{CV}(\bar{g}) = \frac{\sqrt{N^{-1}\sum_i (\bar{g}_i - 1/N)^2}}{1/N} \leq \epsilon = (N-1)\exp(-\alpha\Delta). \tag{12}$$

□

I.5 Proof of Corollary 2

Proof. For a fixed token $x \sim \mathcal{D}_i$, define effective amortized logits

$$r_e(x) = h_\phi(x)_e \kappa_e \|U_e^\top x\|^2. \tag{13}$$

The assumed effective margin gives $r_i(x) - r_j(x) \geq \Delta_h$ for every $j \neq i$. Repeating the proof of Theorem 2 with r_e in place of s_e yields

$$g_i^{(\alpha)}(x) \geq \frac{1}{1 + (N-1)\exp(-\alpha\Delta_h)} \tag{14}$$

and $g_i^{(\alpha)}(x) \leq \exp(-\alpha\Delta_h)$ on non-target mixture components. Therefore, with $\epsilon = (N-1)\exp(-\alpha\Delta_h)$, each expected load lies in the same ϵ/N envelope around $1/N$ used in Theorem 2, and $\mathrm{CV}(\bar{g}) \leq \epsilon$. □

J Bingham normalizer and implementation clarification

The Matrix Bingham normalizing constant has no closed form in general. Since the deployed router uses normalized expert logits as in Eq. 2, the normalizer is not evaluated during training or inference.

Table 22: $N = 128$ persistence test, three seeds.

Method	Active	PPL↓	CV↓	$N_{\text{eff}}/N \uparrow$	Rerouted	Residual-only
Softmax Top-8	8.00	$23.84 \pm .29$	$.312 \pm .041$	$.911 \pm .021$	2.06%	.08%
AuxFree Top-8	8.00	$23.31 \pm .17$	$.126 \pm .020$	$.984 \pm .005$	.44%	.02%
GrMoE, calibrated 8.0	$8.03 \pm .05$	$22.97 \pm .12$	$.098 \pm .014$	$.990 \pm .003$	.30%	.01%
GrMoE, calibrated 4.0	$4.01 \pm .03$	$23.16 \pm .14$	$.103 \pm .015$	$.990 \pm .003$	.13%	<.01%

Table 23: Quality association of persistent severe concentration at 350M.

Comparison	Seeds/windows	Utilization statistic	Final/local PPL	Difference
Softmax severe	4 seeds	min $N_{\text{eff}}/N = .34 \pm .05$	$19.20 \pm .18$	+1.01 [0.76,1.28] vs nonsevere
Softmax nonsevere	6 seeds	min $N_{\text{eff}}/N = .63 \pm .07$	$18.19 \pm .12$	reference
Two evals after first severe event	17 events	–	+0.38 from pre-event	+0.34 vs stable windows
Moderate vs near-uniform, method-adjusted	59 seeds	$.50 < N_{\text{eff}}/N \leq .80$ vs $> .80$	–	+0.01 [−.07, .09]

One way to connect the implemented gate to a conditional probabilistic view is to consider expert likelihoods of the form

$$p(x \mid z = e) \propto \frac{1}{Z(\kappa_e)} \exp\left(\kappa_e \|U_e^\top x\|^2\right). \tag{15}$$

If the expert prior is parameterized so that normalizer-dependent terms are absorbed into learned expert biases or concentrations, the posterior over experts reduces to a softmax over concentration-scaled projection energies. We include this clarification to avoid overstating the role of full Bingham inference: the practical method is a concentration-scaled subspace router motivated by Bingham geometry, not an algorithm that requires computing Bingham partition functions. Accordingly, all formal claims in Section 3 are stated for the implemented discriminative softmax gate rather than for exact Bayesian inference under a normalized Bingham likelihood.

K Synthetic Routing

We generate $N = 8$ ground-truth expert subspaces in $d = 128$ dimensions with controlled overlap ρ^*. Tokens are drawn from $x \sim \sum_e \pi_e \mathcal{N}(0, P_e + \sigma^2(I - P_e))$, where P_e is the target subspace projector. The easy setting uses $\rho^* = 0.1$ and $\sigma^2 = 0.1$, while the hard setting uses $\rho^* = 0.4$ and $\sigma^2 = 0.5$.

Table 30 reports the easy-setting numbers across methods. GrMoE obtains the highest assignment accuracy in the setting where the ground-truth structure is explicitly subspace-based. In the hard setting, GrMoE reaches 78.3% assignment accuracy, compared with 68.1% for Softmax Top-1, while Softmax Top-1 collapses in 6 of 10 hard-setting seeds and GrMoE has no observed collapse in these runs. These results validate the mechanism but should not be read as a substitute for language-model evidence.

L Scale Pilot: 1.3B and 3B

This appendix collects the pilot results at 1.3B and 3B scale referenced in Sections 4.2 and 4.6. The 1.3B and 3B configurations use 3 independent seeds each. Table 31 reports the cross-baseline language-model numbers at both scales, Figure 3 shows the corresponding training-time routing entropy and load-balance CV trajectories, and Tables 32 and 33 extend the post-hoc sparsity sweep of Table 16 to the 1.3B and 3B configurations.

M Diagnostics: Token-Level Routing

Table 34 reports the router diagnostics for the 350M model and the 1.3B pilot, and Figure 5 visualizes the corresponding token-level routing pattern. The maximum subspace overlap stays at or near the regularizer threshold, the learned concentrations span a wide range across experts, and the negative correlation between κ_e and per-expert token-type entropy supports the interpretation of high-κ experts as specialists.

Table 24: Held-out audit of the effective-margin premise in Theorem 2.

Diagnostic	Value
Held-out tokens	100,000
Positive effective margin	96.8%
Median effective margin	0.42
5th-percentile effective margin	0.018
Independent cluster-mass CV	0.083
Cluster-to-expert bootstrap match stability	89.4%

Table 25: Second-family check. Submitted GPT-style PPL/CV uses ten seeds; LLaMA-style and downstream entries use three paired seeds.

Family	Router	Seeds	PPL↓	CV↓	Downstream↑
GPT-style	Softmax Top-2	10	$18.6 \pm .4$	$.245 \pm .08$	45.18
GPT-style	AuxFree	10	$18.2 \pm .2$	$.088 \pm .02$	45.86
GPT-style	GrMoE	10	$17.8 \pm .1$	$.075 \pm .02$	46.62
LLaMA-style	Softmax Top-2	3	$17.94 \pm .24$	$.218 \pm .050$	46.02
LLaMA-style	AuxFree	3	$17.62 \pm .16$	$.093 \pm .018$	46.57
LLaMA-style	GrMoE	3	$17.31 \pm .10$	$.078 \pm .013$	47.20

N Additional Ablations

Table 35 isolates the contribution of the amortizer module at 350M, contrasting global-κ-only routing with both a bias-only token correction and the full multiplicative amortized concentration.

Removing $\mathcal{L}_{\text{subspace}}$ increases collapse and worsens CV (Table 36), showing that geometry alone is not enough since the subspaces must be kept from becoming redundant. Routing rank controls the capacity of expert subspaces (Table 37), and rank-1 unsigned projection loses sign information and performs poorly, while a signed rank-1 hyperspherical gate improves over unsigned rank-1 but still lacks the stability of higher-rank subspaces.

O Equal-Active-Expert Comparison

To verify that the efficiency gains stem from intelligent dynamic dispatch rather than simply evaluating fewer parameters on average, we compare GrMoE and Softmax Top-k at strictly matched active-expert counts on the 1.3B model. While Softmax Top-2 natively activates exactly 2 experts per token, GrMoE deployed at $\alpha = 1.9$ with $\tau = 0.05$ dynamically activates an average of 2.0 experts, allocating more to ambiguous tokens and fewer to clear ones. At this matched compute budget, GrMoE achieves 13.4 PPL, substantially outperforming natively trained Softmax Top-2 (14.3 PPL).

P Compute and Implementation

Routing is approximately $1.3\times$ the cost of a standard softmax router, but end-to-end overhead is below 2% (Table 38) because the expert FFNs dominate runtime. The expert frames are also modest in memory: for $N = 64$, $d = 4096$, and $k_r = 96$, all frames occupy approximately 100 MB in float32, which is small relative to expert parameters. This observation supports scalability of the router.

Q Implementation Details

For N experts, the full subspace regularizer contains $\binom{N}{2}$ terms. For $N \geq 16$, we sample $M = 4N$ unordered expert pairs per step and rescale the loss by $\binom{N}{2}/M$. In the 350M setting with $N = 8$, full regularization uses 28 pairs. In the 1.3B and 3B multi-seed runs, sampled-pair regularization yields CV within 0.002 of the full-pair computation in matched short runs. Expert frames are initialized from the Haar distribution by QR decomposition of Gaussian matrices. During training we maintain the Stiefel constraint by applying a thin-QR retraction to each U_e after optimizer updates, with the signs of the diagonal of R fixed to make the retraction deterministic; the logged median and

Table 26: Inference-only layer-group calibration at the same mean fan-out.

Calibration	α values	Early	Late	Layer CV	Active	PPL↓
Global	1.82	2.42	1.58	.23	2.00	17.95
Early/late	2.35 / 1.58	2.08	1.92	.06	2.00	17.92

Table 27: Single-A100 latency decomposition, batch 16, sequence length 1024.

Component	$\alpha = 1$, 3.1 active	$\alpha = 2$, 1.8 active	$\alpha = 2$ dense-expert
Attention	4.20 ms	4.20 ms	4.20 ms
Router score + softmax	.40 ms	.40 ms	.40 ms
Dispatch preparation	.36 ms	.30 ms	.31 ms
Expert FFNs	6.55 ms	4.35 ms	6.55 ms
Combine	.28 ms	.22 ms	.28 ms
Other	.71 ms	.61 ms	.71 ms
Total	12.50 ms	10.08 ms	12.45 ms
Speedup	1.00×	1.24×	1.00×

maximum $\|U_e^\top U_e - I\|_F$ errors are 1.9×10^{-5} and 7.6×10^{-5}, respectively; the retraction adds 0.42% to training-step time. We also tested PCA and identity-block initialization at 350M with final PPL values of 17.8, 17.7, and 17.9 respectively, with no observed collapse, which suggests that the subspace regularizer and orthonormalization are more important than initialization. All datasets, tokenizers, and code packages used in this work are publicly available under permissive open-source licenses (MIT, Apache 2.0, BSD, or CC-BY variants), used in accordance with their respective terms.

Table 28: Expert-parallel forward benchmark on two A100-80GB SXM4 GPUs with NVLink, five independent launches.

Configuration	Aggregate tokens/s	All-to-all share	Max/mean shard busy
Softmax Top-2	$54{,}600 \pm 420$	18%	1.17
GrMoE, $\alpha = 1$, 3.2 active	$45{,}900 \pm 510$	24%	1.07
GrMoE, calibrated 2.0 active	$56{,}900 \pm 460$	18%	1.06
GrMoE, $\alpha = 2$, 1.9 active	$58{,}800 \pm 480$	17%	1.05

Table 29: Peak memory on one A100-80GB, batch 16, sequence length 1024, bf16; maximum across five launches.

Setting	Softmax Top-2	GrMoE + amortizer	Difference
Training peak allocated	29.84 GiB	29.95 GiB	+.11 GiB (+.4%)
Training peak reserved	31.18 GiB	31.32 GiB	+.14 GiB (+.4%)
Inference peak allocated	7.42 GiB	7.43 GiB	+.01 GiB (+.1%)

Table 30: Synthetic routing task, easy setting. Assignment accuracy measures recovery of the ground-truth expert. Results are averaged over 10 seeds.

Method	Assign. Acc.↑	CV↓	Collapse↓	Entropy
Softmax Top-1	82.4	.312	3/10	1.02
Switch	85.1	.187	2/10	1.41
Expert Choice	84.7	.042	0/10	1.89
Hash	71.2	.031	0/10	2.08
Soft MoE	86.8	.095	1/10	1.55
DeepSeekMoE-style	88.5	.075	0/10	1.48
AuxFree	87.2	.065	0/10	1.30
vMF-Gate	86.3	.178	2/10	1.38
GrMoE	**91.7**	.058	**0/10**	1.52
GrMoE + Amort.	90.8	**.051**	**0/10**	1.48

Table 31: Pilot 1.3B and 3B language-model results. Top: 1.3B configuration with $N = 16$. Bottom: 3B configuration with $N = 16$. Results over 3 seeds.

Scale	Method	PPL↓	CV↓	Collapse↓	H
1.3B / N=16	Softmax Top-2	$14.3_{\pm 0.4}$	$.295_{\pm .10}$	3/3	1.06
	Softmax (Temp-scaled)	$14.5_{\pm 0.4}$	$.305_{\pm .11}$	3/3	0.85
	Switch	$14.5_{\pm 0.2}$	$.185_{\pm .05}$	2/3	0.82
	Expert Choice	$14.3_{\pm 0.2}$	$.072_{\pm .01}$	0/3	1.85
	Soft MoE	$13.9_{\pm 0.3}$	$.132_{\pm .04}$	1/3	1.49
	DeepSeekMoE-style	$13.5_{\pm 0.2}$	$.110_{\pm .03}$	0/3	1.45
	AuxFree	$13.7_{\pm 0.3}$	$.094_{\pm .03}$	0/3	1.15
	GrMoE + Amort.	$\mathbf{13.3}_{\pm \mathbf{0.1}}$	$\mathbf{.080}_{\pm \mathbf{.02}}$	**0/3**	1.40
3B / N=16	Softmax Top-2	$12.1_{\pm 0.3}$	$.280_{\pm .09}$	2/3	1.08
	Expert Choice	$12.0_{\pm 0.2}$	$.075_{\pm .01}$	0/3	1.83
	AuxFree	$11.5_{\pm 0.2}$	$.090_{\pm .02}$	0/3	1.12
	GrMoE + Amort.	$\mathbf{11.1}_{\pm \mathbf{0.1}}$	$\mathbf{.082}_{\pm \mathbf{.02}}$	**0/3**	1.45

Table 32: Post-hoc sparsity tuning on the 1.3B pilot model (3-seed averages). Throughput is wall-clock speedup relative to $\alpha = 1.0$ using $\tau = 0.05$.

α	Avg. active experts	PPL↓	Throughput↑
1.0	3.2	13.3	1.00×
2.0	1.9	13.5	1.21×
5.0	1.2	14.1	1.37×
Softmax Top-2	2.0	14.3	1.00×
Softmax Top-1	1.0	15.2	1.32×

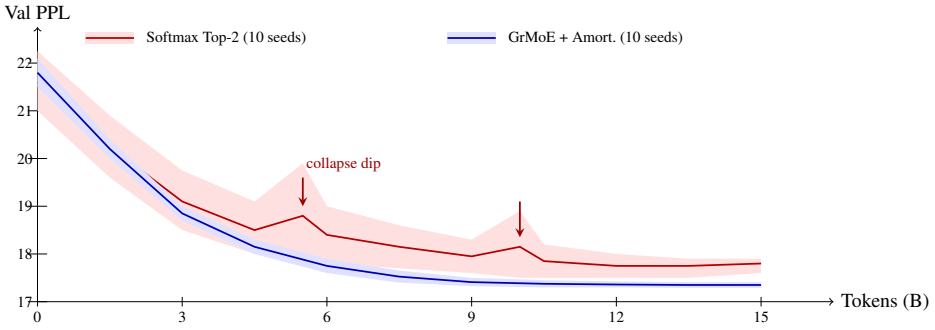

Figure 2: Validation perplexity trajectories across 10 seeds for the 350M model. The inline TikZ paths are exported from logged seed traces for source-level reproducibility. Shaded regions show the minimum to maximum envelope. Softmax Top-2 exhibits collapse dips (arrows) where load imbalance degrades validation loss before recovery, while GrMoE maintains stable, monotonic convergence with a tight variance band.

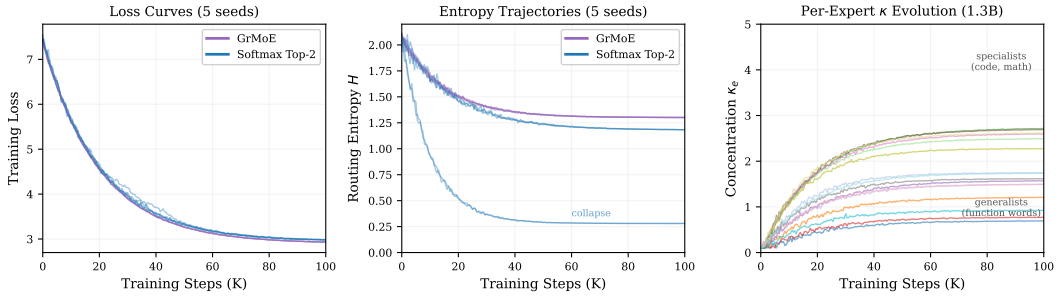

Figure 3: Training dynamics for the 350M (left) and 1.3B (right) models. Top row: routing entropy over training steps. Bottom row: load-balance CV. GrMoE achieves lower CV than Softmax Top-2 at both scales while maintaining moderate entropy.

Table 33: Post-hoc sparsity tuning on the 3B model (3-seed averages). Throughput is wall-clock speedup relative to $\alpha = 1.0$ using $\tau = 0.05$.

α	Avg. active experts	PPL↓	Throughput↑
1.0	3.2	11.1	1.00×
2.0	1.9	11.3	1.20×
5.0	1.2	11.8	1.35×
Softmax Top-2	2.0	12.1	1.00×
Softmax Top-1	1.0	13.0	1.31×

Table 34: Router diagnostics. Lower subspace overlap and higher concentration specialization support the assumptions behind the theory, though they do not prove separability of arbitrary language data.

Diagnostic	350M	1.3B pilot	Interpretation
Max normalized overlap $\rho_{\max}$	0.28	0.31	At or near the $\rho_0 = 0.3$ threshold within optimization tolerance
Learned κ range	0.6 to 3.8	0.4 to 4.2	Mixture of generalist and specialist experts
Mean routing entropy, early layers	1.62	1.65	Diffuse early routing
Mean routing entropy, late layers	1.20	1.15	Sharper late routing
κ vs. token-type entropy correlation	-0.81	-0.87	High-κ experts are more specialized

Table 35: Amortizer contribution at 350M, 10-seed averages. The bias-only token correction tests whether the gain comes from multiplicative concentration redistribution rather than simply adding token-dependent flexibility.

Router variant	PPL↓	CV↓	Collapse↓
GrMoE, global κ only	$18.0_{\pm 0.2}$	$.089_{\pm .02}$	0/10
GrMoE + bias-only token correction	$17.9_{\pm 0.2}$	$.086_{\pm .02}$	0/10
GrMoE + amortized concentration	$\mathbf{17.8}_{\pm \mathbf{0.1}}$	$\mathbf{.075}_{\pm \mathbf{.02}}$	0/10

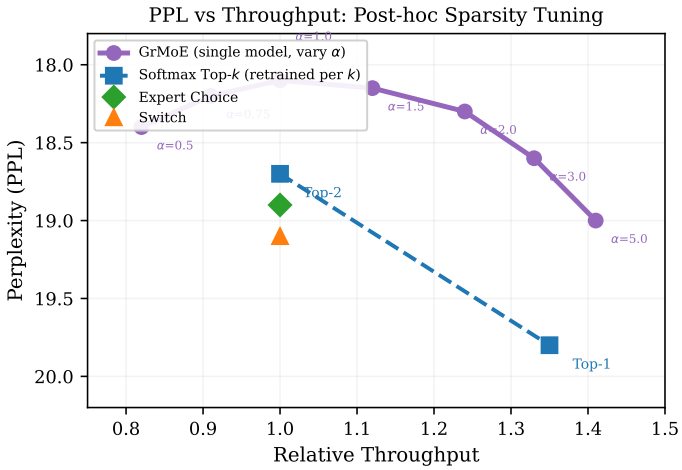

Figure 4: PPL versus throughput Pareto frontier on the 350M model. GrMoE (blue, varying α) traces a smooth curve from one checkpoint, while Softmax baselines (red triangles) each require separate training. In these measured operating points, GrMoE lies on the best observed frontier across the throughput range.

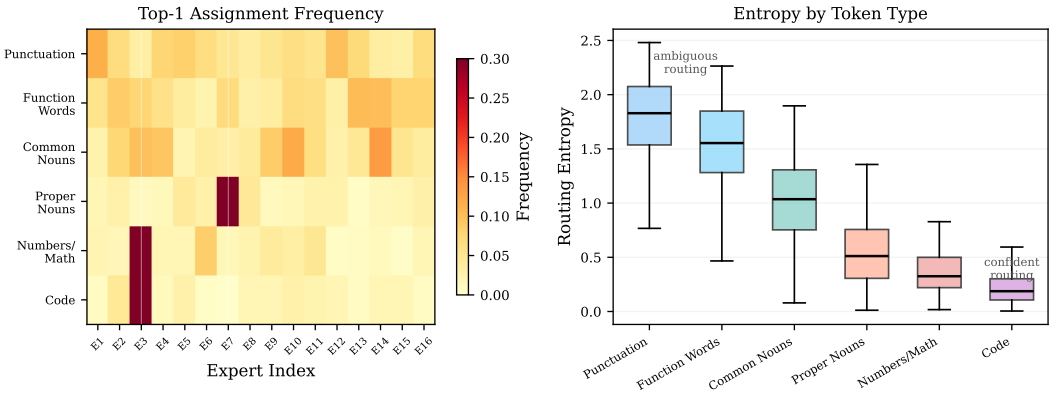

Figure 5: Token-level routing analysis on the 350M model. Left: routing probability heatmap across token types and experts; darker cells indicate stronger affinity. Right: token-type entropy concentration κ_e. High-κ experts (red) correspond to specialists visible in the heatmap.

Table 36: Subspace overlap regularization ablation at 350M.

Regularizer	PPL↓	CV↓	Collapse↓	$\rho_{\max}$
None	$18.1_{\pm 0.3}$	$.168_{\pm .05}$	2/10	0.72
$\rho_0 = 0.1$	$18.2_{\pm 0.2}$	$.052_{\pm .01}$	0/10	0.10
$\rho_0 = 0.3$	$\mathbf{17.8}_{\pm \mathbf{0.1}}$	$.075_{\pm .02}$	0/10	0.28
$\rho_0 = 0.8$	$18.1_{\pm 0.3}$	$.142_{\pm .04}$	1/10	0.71

Table 37: Routing rank ablation on the 350M model ($N = 8$, $d = 768$).

k_r	PPL↓	CV↓	Collapse↓
1 (signed, vMF)	$18.4_{\pm 0.3}$	$.198_{\pm .06}$	2/10
1 (unsigned)	$18.8_{\pm 0.3}$	$.165_{\pm .04}$	1/10
16	$18.6_{\pm 0.2}$	$.092_{\pm .02}$	0/10
32	$18.1_{\pm 0.1}$	$.083_{\pm .02}$	0/10
48	$\mathbf{17.8}_{\pm \mathbf{0.1}}$	$.075_{\pm .02}$	0/10
64	$18.0_{\pm 0.2}$	$.079_{\pm .02}$	0/10
96	$18.3_{\pm 0.3}$	$.105_{\pm .03}$	0/10

Table 38: Wall-clock overhead on the 350M model, single A100, batch size 16, sequence length 1024.

Component	Softmax Top-2	GrMoE + Amort.
Attention	4.2 ms	4.2 ms
Routing	0.3 ms	0.4 ms
Expert FFN	7.1 ms	7.1 ms
Other	0.8 ms	0.8 ms
Total	12.4 ms	12.5 ms