Nash Bargaining for Gate-Free Mixture-of-Experts

Anonymous Authors[1]

Abstract

Mixture-of-Experts (MoE) architectures traditionally rely on a parameterized gating network to route inputs and achieve conditional computation. In computer vision, explicit routing often suffers from optimization instability and specialization collapse while ensembling alternatives bypass routing at substantial computational cost and exhibit destructive interference under naive logit aggregation. We propose gate-free MoE (gfMoE), an architecture that frames expert collaboration as a Nash Bargaining problem. A shared early-feature backbone provides representational stability, and a novel Nash Cooperative Yielding Loss trains each expert to suppress its own activations whenever its marginal contribution to the coalition prediction is negative, instantiating the individual rationality condition of the Nash Bargaining Solution. On CIFAR-10, gfMoE attains a Unified test accuracy of $89.93\% \pm 0.68$, statistically indistinguishable from a dense ResNet-18 baseline ($89.74\% \pm 0.53$) and a gated MoE counterpart ($90.38\% \pm 1.11$), while reducing the destructive-interference gap of Stochastic Multiple Choice Learning (sMCL) ensembles from 25.12 to 2.54 percentage points. We additionally report results on MNIST, CIFAR-100, and Imagenette, and ablate the contribution of the Nash regularizer and the number of experts.

1. Introduction

The scaling of deep neural networks is frequently constrained by the quadratic increase in computational requirements. Mixture of Experts (MoE) architectures address this by decoupling parameter count from active floating-point operations (FLOPs). Foundational work by Shazeer et al. and subsequent developments like the Switch Transformer (Fedus et al., 2022) demonstrated that parameterized gating networks can successfully route tokens to specialized sub-networks in autoregressive language modeling.

However, applying MoE to computer vision tasks introduces distinct challenges. Explicit routing networks in vision models often collapse into uniform utilization or fail to partition complex continuous manifolds organically. An alternative approach is Stochastic Multiple Choice Learning (sMCL) proposed by Lee et al. which trains an ensemble of independent networks using a Winner-Take-Gradient (WTG) mechanism. While sMCL forces specialization, it requires computing N parallel networks from pixel to prediction. Furthermore, sMCL is designed for multi-hypothesis generation. If an sMCL ensemble is forced to produce a single unified prediction via logit averaging, the out-of-distribution experts produce uncalibrated activations that cause destructive interference.

We introduce a decentralized gate-free MoE (gfMoE) architecture designed to aggregate specialized knowledge without a routing network. We stabilize representation efficiency by utilizing a shared early backbone. To solve the destructive interference of un-gated logit aggregation, we apply cooperative game theory. We frame the logit aggregation step as a Nash Bargaining problem where experts evaluate their own marginal contribution and learn to yield the floor (i.e. output zero vectors) when they cannot confidently assist the global coalition.

2. Methods

2.1. Architectural Formulation

We transition from an ensemble of independent models to a unified MoE framework. The architecture consists of a shared ResNet18 feature extraction backbone and N independent expert heads (last block and a fully connected classification layer). This structure ensures that early layers capture universal low-level features (e.g. edges and color gradients) while forcing the expert heads to diverge on high-level semantic abstractions.

2.2. Nash Cooperative Yielding Loss

In a traditional logit-averaged ensemble, the coalition prediction is the mean of the N expert logits, $\bar{Z} =$

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

$\frac{1}{N}\sum_{i=1}^{N} Z_i$. Without a gate to silence experts whose activations are misaligned with the target, low-confidence or out-of-distribution experts contaminate $\bar{Z}$, thus producing destructive interference (Lee et al., 2016).

We resolve this by treating the N experts as players in a cooperative coalition that jointly minimize the cross-entropy of the averaged logits:

$$\mathcal{L}_{\text{global}} = \text{CE}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i, y\right) \tag{1}$$

To establish a disagreement point for each player, we evaluate the coalition loss when expert i unilaterally yields the floor — operationalized by replacing Z_i with the zero vector (a maximum-entropy logit, equivalent to a uniform softmax):

$$\mathcal{L}_{\text{yielded},i} = \text{CE}\left(\frac{1}{N}\sum_{j\neq i} Z_j, y\right) \tag{2}$$

The marginal contribution of expert i is the improvement it confers on the coalition relative to its disagreement point:

$$\text{MC}_i = \mathcal{L}_{\text{yielded},i} - \mathcal{L}_{\text{global}} \tag{3}$$

If $\text{MC}_i > 0$, expert i improves the coalition prediction. If $\text{MC}_i < 0$, the expert actively harms the coalition by introducing destructive interference. We apply a targeted L_2 shrinkage penalty P exclusively to harmful experts, scaled by the magnitude of the damage:

$$\text{P}_i = \begin{cases} 0 & \text{if } \text{MC}_i \geq 0, \\ |\text{MC}_i| \cdot \|Z_i\|_2^2 & \text{if } \text{MC}_i < 0, \end{cases} \tag{4}$$

where Z_i is the raw logit vector of expert i. The full training objective is the Nash Cooperative Yielding Loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \lambda \sum_{i=1}^{N} \text{P}_i \tag{5}$$

where $\lambda \geq 0$ controls the strength of the disagreement-point regularization. The marginal contributions MC_i are computed in the forward pass with N extra cheap leave-one-out averages and are detached from the autograd graph when used as scalar weights, so gradients flow only through $\mathcal{L}_{\text{global}}$ and the $\|Z_i\|_2^2$ shrinkage term.

This formulation operationalizes the individual rationality condition of the Nash Bargaining Solution (Nash, 1950) where each expert is incentivized to project confident activations only when doing so yields a Pareto improvement over its disagreement point. Otherwise, the expert is encouraged to yield, hence allowing the rest of the coalition to dominate the prediction without a learned routing signal.

3. Experiments

We evaluate gfMoE on four image classification benchmarks: MNIST, CIFAR10, CIFAR100, and Imagenette. For MNIST, we employ a lightweight 4-stage residual network we refer to as ResNet4. For all other datasets we use a pretrained ResNet18 model. Across all MoE variants, the shared backbone comprises stages 1–3 of the underlying topology, and each expert head comprises stage 4 plus a fully connected classifier. For the CIFAR10 ablations in Section 3.3, we train on a uniformly subsampled 50% subset of the CIFAR10 training set (which refer as "50% CIFAR10" in the rest of this paper) to reduce wall-clock cost while preserving relative comparisons.

We trained all models for 100 epochs on a mini-batch size of 300 using SGD with momentum of 0.9, weight decay of 1×10^{-4}, with a learning rate of 1×10^{-1}(Polyak, 1964). Then we used OneCycleLR to anneal the learning rate (Smith & Topin, 2019).

We perform no hyperparameter tuning since achieving state-of-the-art accuracy is beyond the scope of this study. Our objective is to test whether the individual-rationality condition of the Nash Bargaining Solution is sufficient for an expert to participate only when its marginal contribution is non-negative, thereby eliminating the need for a gating network. Each configuration is repeated for R independent runs (reported in each table caption: R=10 for MNIST, R=5 for CIFAR10/100, R=3 for Imagenette) with different random seeds; we report the mean and standard deviation of test accuracy across runs.

Metrics We report three quantities for every model:

1. **Unified Accuracy** is the test accuracy of the deployable prediction obtained by averaging the logits of all experts and taking the $\arg\max$. This is the metric a practitioner would observe at inference time.

2. **Oracle Accuracy** is the test accuracy obtained when for each test example, an oracle is allowed to select the single expert whose individual prediction matches the label (if any). Oracle accuracy upper-bounds the achievable accuracy of any post-hoc routing scheme over the trained experts and quantifies the capacity of the ensemble.

3. **Interference Gap** is the difference between the oracle accuracy OracleACC and unified accuracy UnifiedACC. It measures how much accuracy is lost

in the aggregation step due to destructive interference between experts.

A model that produces well-specialized experts will exhibit a high Oracle Accuracy; a model that successfully aggregates them will exhibit a small Interference Gap. The central design goal of gfMoE is to obtain both simultaneously without a learned gating network.

3.1. Baseline Models

We compare gfMoE against three baselines:

1. Single. A single network of the underlying topology (ResNet4 for MNIST, ResNet18 elsewhere) trained with standard cross-entropy. By construction, Unified ACC = Oracle ACC and Interference Gap = 0.

2. sMCL (Lee et al., 2016). Stochastic Multiple Choice Learning trains N independent networks of the same topology under the Winner-Take-Gradient rule, in which the example's loss is back-propagated only through the expert with the lowest per-example loss. sMCL is designed for multi-hypothesis prediction; we report Unified ACC for completeness as a worst-case logit-averaging deployment.

3. GatedMoE. A standard top-k gated MoE built from the same shared backbone and N expert heads as gfMoE with a softmax gating function over a linear projection of the post-backbone features, top-k routing with $k = 1$, and a load-balancing auxiliary loss with coefficient. The gate replaces the Nash Cooperative Yielding Loss but the architecture is otherwise identical.

For all MoE variants the default configuration uses $N = 5$. The gfMoE-specific hyperparameter λ defaults to 1.0.

3.2. Main Results

Tables 1–4 report Unified Accuracy, Unified standard deviation, Interference Gap, Oracle Accuracy, and Oracle standard deviation for each model on each dataset. Tables 5–6 report t-tests of gfMoE against each baseline.

MNIST. gfMoE attains the highest Unified Accuracy of any model ($98.74\% \pm 0.46$, Table 1), exceeding both Single ($98.15\% \pm 1.04$) and GatedMoE ($97.82\% \pm 1.32$) while preserving an Oracle Accuracy ($99.43\% \pm 0.25$) within 0.11 points of sMCL ($99.54\% \pm 0.09$). The Interference Gap of 0.69 is the smallest non-trivial gap of any multi-expert method on this dataset. The improvement over GatedMoE trends positive but does not reach significance at the 0.05 level ($t = 1.926$, $p = 0.0863$, Table 5); the improvement over sMCL is highly significant ($t = 4.225$, $p = 0.0022$).

CIFAR10. GatedMoE achieves a marginally higher Unified Accuracy than gfMoE ($90.38\% \pm 1.11$ vs. $89.93\% \pm 0.68$, Table 2); the difference is well within one standard deviation. The salient observation is that sMCL collapses on CIFAR10, with a Unified Accuracy of $69.02\% \pm 3.38$ and an Interference Gap of 25.12 points while preserving an Oracle Accuracy of $94.14\% \pm 0.39$. gfMoE reduces this Interference Gap by an order of magnitude ($25.12 \rightarrow 2.54$) while remaining within 0.45 points of GatedMoE on Unified Accuracy. We omit the t-test for CIFAR10 here as the gap between gfMoE and GatedMoE is well within noise; per-comparison statistics are reported only where they materially affect the conclusions.

CIFAR100. gfMoE achieves the highest Unified Accuracy on CIFAR100 ($41.00\% \pm 1.80$, Table 3) which exceeds both Single ($38.80\% \pm 0.45$) and GatedMoE ($39.11\% \pm 0.84$). gfMoE also attains the highest Oracle Accuracy ($51.57\% \pm 2.10$), which is roughly 12 points above GatedMoE ($39.11\% \pm 0.84$), thereby suggesting that the Nash regularization actively encourages a degree of expert specialization that the gated baseline does not recover. The Interference Gap of 10.57 is non-trivial but is roughly half of sMCL's gap (16.06). The improvement over GatedMoE does not reach significance at $\alpha = 0.05$ ($t = 1.576$, $p = 0.190$, Table 6), which we attribute primarily to the small run count ($R = 5$).

Imagenette . gfMoE and GatedMoE perform comparably on Imagenette ($75.97\% \pm 0.67$ vs. $75.08\% \pm 0.38$, Table 4) both of which are substantially above sMCL's deployable performance ($53.01\% \pm 9.07$, dominated by an Interference Gap of 30.08). We note that the Single baseline on Imagenette ($37.03\% \pm 6.88$) is substantially weaker than commonly reported ResNet18 results on this dataset. We attribute this to a set of under-trained models, and we caution that this inflates the apparent gain of every multi-expert method on this benchmark. We retain the comparison for completeness but treat Imagenette as a stress test of relative behavior between MoE variants rather than as an absolute accuracy benchmark.

3.3. Ablation Studies

We study two design dimensions on the 50% CIFAR10 training subset: the number of experts N and the disagreement weight λ.

Effect of the number of experts. Figure 1 reports Unified and Oracle accuracy of gfMoE as a function of $N \in 2, 3, 4, 5$ at fixed $\lambda = 1.0$. Performance is non-monotonic in N and peaks sharply at $N = 2$, which attains the best Unified Accuracy (87.51) and the smallest Interference Gap (2.62). Each additional expert beyond $N = 2$ degrades Unified Accuracy by approximately 2.6 points and inflates the

Table 1. Performance Results for MNIST (10 runs, ResNet4)

Model	Unified ACC	Unified SD	Interference Gap	Oracle ACC	Oracle SD
Single	98.15	1.04	0.00	98.15	1.04
sMCL	95.45	2.17	4.09	99.54	0.09
GatedMoE	97.82	1.32	0.00	97.82	1.32
gfMoE	**98.74**	0.46	0.69	99.43	0.25

Table 2. Performance Results for CIFAR10 (5 runs, ResNet18)

Model	Unified ACC	Unified SD	Interference Gap	Oracle ACC	Oracle SD
Single	89.74	0.53	0.00	89.74	0.53
sMCL	69.02	3.38	25.12	94.14	0.39
GatedMoE	**90.38**	1.11	0.00	90.38	1.11
gfMoE	89.93	0.68	2.54	92.47	0.48

Interference Gap by approximately 1.3 points. Oracle Accuracy also declines from 90.13 at $N = 2$ to 87.29 at $N = 5$, indicating over-fragmentation: the marginal-contribution penalty appears to push experts into excessively narrow niches whose individual capacity is too low to contribute confidently to the coalition.

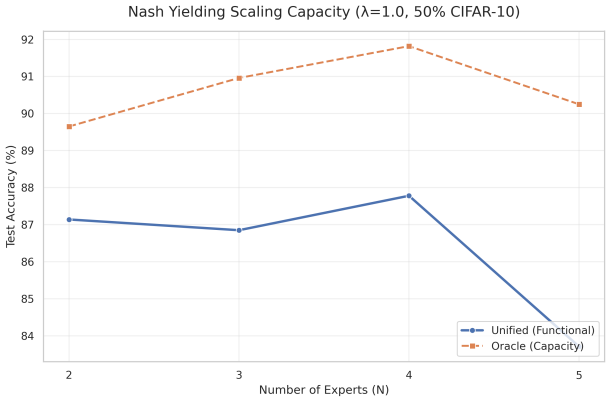

Figure 1. Scaling analysis of Nash Bargaining Solution (NBS) performance. Scaling of Unified and Oracle test accuracies as a function of the number of experts $N \in 2, 3, 4, 5$ on 50% CIFAR10 with a fixed disagreement weight $\lambda = 1.0$.

Effect of the disagreement weight λ. Figure 2 reports the sensitivity of a 3-expert gfMoE to $\lambda \in \{0, 0.1, 0.5, 1, 2, 5\}$ on 50% CIFAR10. The $\lambda = 0$ baseline (no Nash term, i.e. plain logit-averaging on the shared-backbone architecture) attains the highest Oracle Accuracy (92.93) but the worst Interference Gap (7.07) and a middling Unified Accuracy (85.86). For $\lambda \in [0.1, 2]$, Unified Accuracy is approximately flat (84.42–85.22) while the Interference Gap decreases monotonically, indicating that the Nash regularization trades a small, bounded amount of specialization capacity for substantially better deployable performance. At $\lambda = 5$, Unified Accuracy drops to 83.79, indicating over-regularization. We adopt $\lambda \in [1, 2]$ as the default operating range.

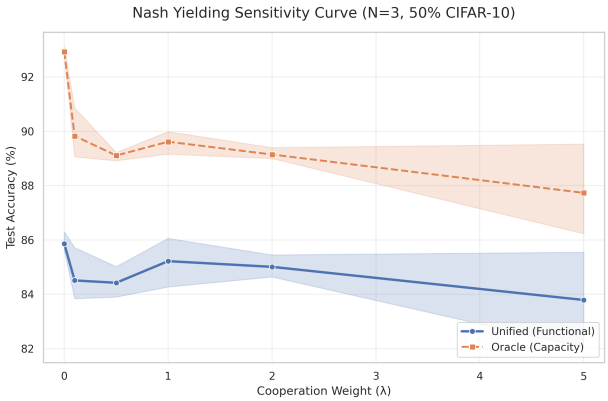

Figure 2. Sensitivity of Nash objective to disagreement weight. Impact of the weight parameter λ on test accuracy for a 3-expert ensemble on 50% CIFAR10. The plot illustrates the trade-off between task performance and the disagreement point surplus, with shaded regions denoting confidence intervals across multiple experimental runs.

4. Related works

4.1. Mixture-of-Experts Foundations.

The Mixture of Experts (MoE) framework was introduced by Jacobs et al. who proposed a supervised learning procedure for systems composed of many separate networks, each learning to handle a subset of training cases, demonstrating that the architecture divides a vowel discrimination task into appropriate sub-tasks solvable by simple expert networks. Jordan & Jacobs extended this to the Hierarchical Mixtures of Experts (HME) model, a tree-structured variant that leverages the Expectation-Maximization (EM)

Table 3. Performance Results for CIFAR100 (5 runs, ResNet18)

Model	Unified ACC	Unified SD	Interference Gap	Oracle ACC	Oracle SD
Single	38.8	0.45	0	38.8	0.45
sMCL	32.16	0.40	16.06	48.22	0.63
GatedMoE	39.11	0.84	0	39.11	0.84
gfMoE	**41.00**	1.80	10.57	51.57	2.10

Table 4. Performance Results for Imagenette (3 runs, ResNet18)

Model	Unified ACC	Unified SD	Interference Gap	Oracle ACC	Oracle SD
Single	37.03	6.88	0	37.03	6.88
sMCL	53.01	9.07	30.08	83.09	1.27
GatedMoE	75.08	0.38	0	75.08	0.38
gfMoE	**75.97**	0.67	4.47	80.44	0.82

algorithm for training and can summarize data at multiple scales of resolution.

The modern revival of MoE at scale was catalyzed by Shazeer et al. who introduced the Sparsely-Gated MoE layer, achieving greater than 1000x improvements in model capacity with only minor losses in computational efficiency on modern GPU clusters. Their key innovation is a top-k routing mechanism paired with load-balancing auxiliary losses which has established the dominant paradigm for conditional computation in deep learning. Fedus et al. further simplified this with the Switch Transformer, which uses top-1 routing to scale to trillion-parameter models with simple and efficient sparsity.

4.2. Expert Specialization.

A central tenet of MoE architectures is that individual experts develop distinct competencies over disjoint regions of the input space. Such specialization allows experts to develop expertise in specific knowledge domains; however, input data should ideally be distributed evenly among experts, as uneven distribution can lead to model collapse, reducing the efficiency of the MoE framework by under-utilizing experts. This creates a fundamental tension between specialization and load balance, objectives that are often in conflict under standard training regimes.

The MoE architecture is described as consisting of expert modules and a gate where experts compute functions useful in different regions of the input space. Improving and measuring the quality of such specialization has attracted growing attention. Recent work on ReMoE (Wang et al., 2024) found that continuous routing promotes stronger specialization: ReMoE exhibits dynamic expert allocation, assigning more computational resources to challenging tokens and achieving domain specialization more effectively than traditional MoE.

A parallel line of work asks whether specialization requires explicit architectural supervision. EMoE (Cheng et al., 2026) argues that current MoE training techniques face a trade-off, achieving either balanced loads or highly specialized experts but not both simultaneously, and proposes routing decisions via projection onto an geometric partitioning of the feature space that naturally assigns coherent input clusters to different experts without a learned gating network or auxiliary objective.

4.3. Specialization Without a Gating Network or Router

A growing body of work challenges the assumption that a centralized learned router is a necessary component of MoE. These approaches distribute routing logic into the experts themselves or derive it from existing representations, avoiding the inductive biases and instabilities introduced by an external gating module.

Routing-Free MoE (Liu et al., 2026) eliminates any hard-coded centralized designs including external routers, Softmax, TopK, and load balancing, instead encapsulating all activation functionalities within individual experts directly optimized through continuous gradient flow, enabling each expert to determine its own activation entirely. This architecture represents a strong prior that routing is a property that should be intrinsic to experts rather than delegated to a bottleneck module. Similarly, Self-Routing (Mohamud et al., 2026) proposes a parameter-free routing mechanism that uses a designated subspace of the token hidden state directly as expert logits which eliminates the router projection entirely, and finds that this yields more balanced expert utilization with approximately 17% higher average normalized routing entropy and no explicit load-balancing loss.

A related differentiable reformulation is offered by ReMoE (Wang et al., 2024) which replaces the conventional discontinuous TopK + Softmax routing with a ReLU function as

the router instead, thereby enabling fully differentiable training with methods to regulate sparsity while balancing load among experts. Although ReMoE retains a router module, it dissolves the hard discretization that severs gradient flow to non-selected experts, moving toward a regime in which each expert's activation is governed by a smooth, continuous gate. These works collectively suggest that the binary "selected / not selected" paradigm enforced by top-k routing is not fundamental to MoE function, and that self-suppression of experts based on their relevance to a given input can emerge from training rather than be mandated by architecture.

4.4. Nash Bargaining and Cooperative Game Theory in Machine Learning

The Nash Bargaining Solution (NBS) originates in the seminal work of Nash et al. who demonstrated a unique solution to a two-person bargaining problem satisfying the axioms of invariance, weak Pareto efficiency, symmetry, and independence of irrelevant alternatives. The unique solution that satisfies these axioms maximizes the product of the players' utility gains over their disagreement point — the Nash product — providing a principled, axiomatic criterion for distributing joint gains among cooperative agents.

The first application of the NBS to gradient-based multi-objective machine learning was introduced by Navon et al. who propose viewing the gradient combination step in multi-task learning as a bargaining game, where tasks negotiate to reach an agreement on a joint direction of parameter update, and derive an efficient algorithm, Nash-MTL, with theoretical convergence guarantees in both convex and non-convex settings. Critically, Nash-MTL is invariant to changes in loss scale and produces solutions that are well balanced across the Pareto front, outperforming heuristic gradient aggregation methods on multiple benchmarks spanning computer vision, quantum chemistry, and reinforcement learning. The appeal of the NBS in this context stems from its axiomatic fairness: unlike linear scalarization, which is sensitive to loss magnitude and dominated by tasks with large gradients, the Nash product formulation ensures that no participating agent's gain is sacrificed beyond what is mutually necessary.

Cooperative game-theoretic notions of marginal contribution have also been explored in the context of neural network design. Work on cooperative game theory for neural network pruning uses a power index akin to the Shapley value, where the marginal contribution of a network module to a coalition reflects the increase in utility the coalition gains when that module joins it, and demonstrates that this approach outperforms existing methods in the tradeoff between parameter count and model accuracy. This line of work establishes the conceptual precedent that individual computational units in a network can be assigned scalar contributions to a collective objective, and that these contributions can be used to regulate the unit's participation — a notion directly relevant to expert self-gating.

Positioning. The works above motivate our proposal to reframe expert activation in MoE as a cooperative bargaining problem. Standard MoE architectures delegate activation to a centralized router (Shazeer et al., 2017; Fedus et al., 2022); recent router-free approaches (Liu et al., 2026; Mohamud et al., 2026) instead allow each expert to self-activate, but do not provide a principled, axiomatic criterion for when an expert should suppress itself. We adopt the individual rationality axiom of the Nash Bargaining Solution (Nash et al., 1950; Navon et al., 2022): an expert participates in the coalition only when its marginal contribution MC_i is non-negative relative to its disagreement point. The shared joint-prediction loss $\mathcal{L}_{\text{global}}$ provides the Pareto-improving objective that the coalition collectively optimizes, and the Nash Cooperative Yielding Loss (Eq. 5) operationalizes this criterion as a differentiable training signal. To our knowledge, this is the first MoE formulation in which expert self-suppression is governed by a cooperative-game-theoretic axiom rather than by a learned gate, a load-balancing heuristic, or an architectural constraint on the activation function.

5. Conclusion

gfMoE is a gate-free Mixture-of-Experts model that replaces a learned router with a Nash Bargaining–based rule: experts are encouraged to suppress themselves when their contribution hurts the overall prediction. It achieves competitive accuracy with standard gated MoE across four image-classification benchmarks and greatly reduces destructive interference in sMCL ensembles, though its gains over Gated-MoE are not statistically significant. The main contribution is therefore conceptual which shows that cooperative-game-theoretic individual rationality can govern expert participation without a gate rather than proving superior empirical performance. We also note important limitations: gfMoE is not sparse because all experts are still evaluated, performance works best at two experts, and results are only shown on small-to-medium convolutional benchmarks. Future work includes richer coalition baselines, true sparse compute by skipping harmful experts, and applying the method to transformers.

Impact Statement

This paper presents work with the goal of helping to advance the field of Machine Learning. There may be potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Cheng, A., Duan, S., Li, S., Yin, C., Cheng, M., Nazarian, S., Thompson, P., and Bogdan, P. Emoe: Eigenbasis-guided routing for mixture-of-experts. *arXiv preprint arXiv:2601.12137*, 2026.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994.

Lee, S., Purushwalkam Shiva Prakash, S., Cogswell, M., Ranjan, V., Crandall, D., and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems*, 29, 2016.

Liu, Y., Han, J., Yan, S., Tresp, V., and Ma, Y. Routing-free mixture-of-experts. *arXiv preprint arXiv:2604.00801*, 2026.

Mohamud, J. H., Wagner, D., and Ravanelli, M. Self-routing: Parameter-free expert routing from hidden states. *arXiv preprint arXiv:2604.00421*, 2026.

Nash, J. F. et al. The bargaining problem. *Econometrica*, 18(2):155–162, 1950.

Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., and Fetaya, E. Multi-task learning as a bargaining game. *arXiv preprint arXiv:2202.01017*, 2022.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pp. 369–386. SPIE, 2019.

Wang, Z., Zhu, J., and Chen, J. Remoe: Fully differentiable mixture-of-experts with relu routing. *arXiv preprint arXiv:2412.14711*, 2024.

We present the statistical tests on MNIST and CIFAR100 to compare the performance difference among the models.

Table 5. Statistical Comparison for MNIST (10 runs, ResNet4)

Comparison	t-value	p-value	Result	Significant
gfMoE vs Single	1.546	0.1566	Not Significant	FALSE
gfMoE vs sMCL	4.225	0.0022	Significant	TRUE
gfMoE vs GatedMoE	1.926	0.0863	Not Significant	FALSE

Table 6. Statistical Comparison for CIFAR100 (5 runs, ResNet18)

Comparison	t-value	p-value	Result	Significant
gfMoE vs Single	2.178	0.0949	Not Significant	FALSE
gfMoE vs sMCL	8.363	0.0011	Significant	TRUE
gfMoE vs GatedMoE	1.576	0.1901	Not Significant	FALSE