UNNORMALIZED RELAXATIONS FOR ROBUST MODULAR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Modular learning scales Large Language Models by routing inputs through p frozen, domain-specific experts rather than retraining a monolithic model. The robust gating framework of Cortes et al. (2026) achieves minimax optimal performance under arbitrary shifts in the mixture proportions of the known source domains, but its strict global normalization ($\mathcal{G}_1$: $Z_g = 1$) couples the gate to every expert's predictions, creating a costly synchronization bottleneck at training time. We prove that this constraint is not only computationally unnecessary but *yields less favorable provable guarantees*. We introduce two *unnormalized* gating families and develop a complete theoretical, algorithmic, and empirical framework around them. The first bounded family $\mathcal{G}_{\infty,1}$ replaces the data-coupled mass constraint with an expert-independent structural ceiling ($\sum_k \max_x g(x,k) \leq 1$), depending only on the gating function g. We prove that $\mathcal{G}_{\infty,1}$ achieves a provably tighter minimax robust guarantee in specialized regimes via a novel *Normalization Bonus* $\Delta(Z_{g^*}) \geq 0$, with a matching lower bound; an ordering of empirical Rademacher complexities under mild conditions ($\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ for $p \geq 3$); *Exact Rejection Sampling* for unbiased inference with a structural efficiency bound; and a *Distillation Inheritance* theorem proving that causal student routers preserve the teacher's minimax bounds up to a controlled $\delta + C\sqrt{2\delta}$ gap, with no causal projection penalty and an assumption-free total variation variant. The second locally-normalized family $\mathcal{G}_0$ enables fully pointwise updates, with robust minimax guarantees via a *Selection Gain* and parallel distillation guarantees. Experiments on 1B+ parameter LLMs and 20-expert code generation ensembles confirm the theory: unnormalized gates reduce wall-clock time by >50% while strictly improving NLL over both monolithic and $\mathcal{G}_1$ baselines.

1 INTRODUCTION

As Large Language Models scale to unprecedented sizes (Brown et al., 2020; Radford et al., 2019; Bommasani et al., 2021), monolithic training on heterogeneous data becomes increasingly impractical: it is compute-intensive (Strubell et al., 2019; Schwartz et al., 2020), vulnerable to distribution shift (Koh et al., 2021), and prone to catastrophic forgetting (Kirkpatrick et al., 2017; Parisi et al., 2019). *Modular* learning sidesteps these issues by combining frozen, domain-specific experts through a learned gating function (Cortes et al., 2026). Weight-merging methods offer a cheaper alternative (Wortsman et al., 2022; Matena & Raffel, 2022; Ilharco et al., 2022) but provably suffer from interference on disjoint domains (Yu et al., 2020) and lack formal guarantees; see Appendix A. Building on Multiple-Source Adaptation theory (Mansour et al., 2008; Hoffman et al., 2018; Cortes et al., 2021b), Cortes et al. (2026) formalized robust modularity via an adversarial minimax game: given base experts $\widehat{\pi}_k$ modeling each task, a gated mixture $\pi_g(x) = \sum_k g(x,k)\widehat{\pi}_k(x)$ minimizes the KL divergence against the worst-case task mixture. Under the globally normalized family $\mathcal{G}_1 = \{g : g(x,\cdot) \in \Delta([1,p])\ \forall x,\ Z_g = 1\}$, the resulting gate exploits task diversity as a *Separability Bonus* ($-\mathsf{D}_{\mathrm{JSD}}$), measured in terms of Jensen-Shannon Divergence, provably outperforming monolithic baselines via the *Symmetric Divergence Effect*: task diversity ($\mathsf{D}_{\mathrm{JSD}}$) loosens the monolithic per-task lower bound while simultaneously tightening the modular risk ceiling.

However, the global normalization constraint $Z_g = 1$ of the partition function couples the gate to the data-dependent predictions of every expert across the entire input support. This creates a syn-

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.

Table 1: Comparison of robust modularity under different gating relaxations.

Feature	Global ($\mathcal{G}_1$)	Bounded ($\mathcal{G}_{\infty,1}$)	Local ($\mathcal{G}_0$)
Constraint	Strict (Simplex, $Z_g = 1$)	Data-Independent ($\sum_k \max_x g \le 1$)	Pointwise ($\sum_k g = 1$)
Robust Guarantee	Strong (Overlap Gain H_σ)	**Tighter in specialized regimes** ($\Delta(Z^*) > 0$) (Normalization Bonus $\Delta(Z)$)	Loosest (Selection Gain $\mathcal{A}_{\text{selection}}$)
Generalization ($\mathfrak{R}_m$)	Moderate	**Tightest** ($\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ for $p \ge 3$)	No Cross-Input Coupling (Lookup Table)
Update Efficiency	Low (Global Sync)	High (Async Mini-batch)	**Maximum** (Fully Decoupled)
Sampling	Exact, acceptance $1/M_h$ ($1/p$ for one-hot gates)	**Exact**, acceptance $\ge Z_g$ guaranteed ($M_g \le 1$)	Exact, acceptance Z_g/p ($\approx 1/p$ at $Z_g \approx 1$)
Distillation Gap	$+\delta + C\sqrt{2\delta}$	$+\delta + C\sqrt{2\delta}$	$+\delta + C\sqrt{2\delta}$

chronization bottleneck: enforcing the probability budget during distributed stochastic optimization requires cross-node, cross-batch reduction at every step (Dean et al., 2012; Lian et al., 2017). At scale, communication overhead dominates training time.

A natural conjecture is that relaxing this constraint trades statistical optimality for computational efficiency. *We prove the opposite: unnormalized gating is not only computationally cheaper but yields provably tighter robustness bounds in specialized regimes, tighter generalization bounds under a mild expert-separation condition, and enables efficient exact inference.* This paper introduces and comprehensively analyzes two unnormalized gating families, $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$, establishing a hierarchy of theoretical, algorithmic, and empirical results summarized in Table 1.

Our contributions are organized around three axes: theory, algorithms, and experiments. For the bounded family $\mathcal{G}_{\infty,1}$, we establish:

(i) Minimax Robustness with a Normalization Bonus. The $\mathcal{G}_{\infty,1}$ constraint replaces the data-coupled global mass budget with an expert-independent structural ceiling ($\sum_k \max_x g(x,k) \le 1$), determined solely by the p gate values. We prove (Theorem 5) that the resulting robust gate retains the $\mathsf{D}_{\mathrm{JSD}}$ Separability Bonus of $\mathcal{G}_1$ and earns an additional *Normalization Bonus* ($-\Delta(Z_{g^*}) \le 0$), yielding a strictly tighter provable KL bound in specialized regimes. A matching lower bound (Theorem 15) confirms near-optimality (Section 3).

(ii) Complexity Ordering. Under a mild expert-separation condition (Assumption 7), we prove that the expert-independent constraint restricts the Rademacher complexity of the hypothesis space: $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ for $p \ge 3$ (up to a vanishing term for $p = 2$, Theorem 8). The generalization gap (Theorem 19) is governed by the empirical Rademacher complexity of the gating class, with a confidence term of only $\mathcal{O}(\sqrt{\log p/m})$ in the number of experts.

(iii) Exact Rejection Sampling. Since $\mathcal{G}_{\infty,1}$ provides a data-independent geometric envelope, it admits *Exact Rejection Sampling* for unbiased inference with a structural lower bound on the acceptance rate (Proposition 9), whereas $\mathcal{G}_1$ leaves the envelope mass uncontrolled: it reaches p for the one-hot gates it learns.

(iv) Distillation Inheritance. Structural Distillation of the non-causal $\mathcal{G}_{\infty,1}$ teacher into a causal student preserves the teacher's robust KL bound up to $\delta + C\sqrt{2\delta}$ (Theorem 10), with no causal projection gap, and an assumption-free Total Variation variant.

For the locally-normalized family $\mathcal{G}_0$, we prove robust minimax existence with a *Selection Gain*, characterize the *Volume Penalty*, and provide parallel algorithmic and distillation guarantees (Section 4). Empirically, we evaluate both families at a substantially larger scale than the main experiments of prior work: 1B+ parameter experts and 20-expert code ensembles demonstrate >50% wall-clock speedup, strict NLL improvement, and scaling consistent with the $\mathcal{O}(\sqrt{\log p/T})$ convergence guarantee and the $\mathcal{O}(\sqrt{\log p/m})$ confidence term of the generalization bound (Section 5).

Novelty relative to (Cortes et al., 2026). While we inherit the problem formulation and notation of Cortes et al. (2026), every result in this paper is new. The two gating families ($\mathcal{G}_{\infty,1}$, $\mathcal{G}_0$), their distinct theoretical profiles (Normalization Bonus, Selection Gain, Volume Penalty), the Rademacher complexity comparison, the exact rejection sampling scheme with structural efficiency bounds, the formal distillation inheritance theorems, and the large-scale empirical validation are all original contributions of this work.

2 PRELIMINARIES

We adopt the notation and setup of (Cortes et al., 2026), summarized here for reference. Let $\widehat{\mathsf{p}}_1, \ldots, \widehat{\mathsf{p}}_p$ be p source distributions over the finite support $\mathcal{X}_0 = \bigcup_k \mathrm{supp}(\widehat{\mathsf{p}}_k)$, with $p \geq 2$ pre-trained frozen experts $\widehat{\pi}_k$, probability distributions over $\mathcal{X}_0$, with finite individual expert errors $\epsilon_k = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \,\|\, \widehat{\pi}_k) < \infty$. An input-dependent gating function $g(x,k) \geq 0$ yields the mixture $\pi_g(x) = \sum_{k=1}^p g(x,k)\widehat{\pi}_k(x)$ with partition function $Z_g = \sum_{x\in\mathcal{X}_0} \pi_g(x)$. The goal is to approximate an arbitrary test mixture $\widehat{\mathsf{p}}_\lambda = \sum_{k=1}^p \lambda_k \widehat{\mathsf{p}}_k$ for any $\lambda \in \Delta([1,p])$. For any $\lambda \in \Delta([1,p])$, task diversity is measured by the *Jensen-Shannon Divergence* $\mathsf{D}^\lambda_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\}) = \sum_{k=1}^p \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \,\|\, \widehat{\mathsf{p}}_\lambda) \in [0, H(\lambda)]$, where $H(\lambda)$ is the entropy, $H(\lambda) = -\sum_{k=1}^p \lambda_k \log \lambda_k$.

Unnormalized Relaxations and Divergences. To enable *decoupled, asynchronous updates*, this paper studies robust gating functions within hypothesis spaces where the mixture mass is allowed to fluctuate ($Z_g \neq 1$). Evaluating unnormalized mixture models requires generalized loss functions:

1. Unnormalized Relative Entropy ($\widetilde{\mathsf{D}}_{\mathrm{KL}}$). For the model family $\mathcal{G}_{\infty,1}$ where the total mass is bounded ($Z_g \leq 1$), we use the *Unnormalized Relative Entropy*, a natural Bregman divergence extension of the standard KL divergence (Csiszár, 1975). For a target distribution p (where $\sum_x \mathsf{p}(x) = 1$) and an unnormalized model π_g, it is defined as:

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\mathsf{p} \,\|\, \pi_g) = \mathop{\mathbb{E}}_{x\sim\mathsf{p}}\left[\log \frac{\mathsf{p}(x)}{\pi_g(x)}\right] + \sum_{x\in\mathcal{X}_0} (\pi_g(x) - \mathsf{p}(x)) = \mathop{\mathbb{E}}_{x\sim\mathsf{p}}\left[\log \frac{\mathsf{p}(x)}{\pi_g(x)}\right] + Z_g - 1. \tag{1}$$

For a fixed target p, minimizing $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ is mathematically equivalent to minimizing the *penalized negative log-likelihood*: $\min_g \mathbb{E}_{x\sim\mathsf{p}}[-\log \pi_g(x)] + Z_g$. This metric penalizes the model for misallocating mass without requiring a hard, synchronous equality constraint on Z_g (we evaluate the true NLL of these unnormalized models using a new unbiased exact rejection sampling estimator in Appendix E).

2. Negative Log-Likelihood ($\mathcal{L}_{\log}$). For the model family $\mathcal{G}_0$ requiring only a pointwise normalization, we rely on the standard Negative Log-Likelihood loss, which ignores global mass entirely:

$$\mathcal{L}_{\log}(\mathsf{p}, \pi_g) = -\mathop{\mathbb{E}}_{x\sim\mathsf{p}}[\log \pi_g(x)]. \tag{2}$$

While $\mathcal{L}_{\log}$ allows fully local and decoupled updates, controlling the partition function must be handled carefully via specific gating constraints or regularization, which we formalize for $\mathcal{G}_0$ below.

3 EXPERT-INDEPENDENT BOUNDED GATING ($\mathcal{G}_{\infty,1}$)

We first study the $\mathcal{G}_{\infty,1}$ family, which interpolates between the strict global normalization of $\mathcal{G}_1$ and the purely local constraints of $\mathcal{G}_0$.

Definition 1 (Expert-Independent Bounded Family $\mathcal{G}_{\infty,1}$)**.** *The family of $\mathcal{G}_{\infty,1}$ gates is defined as the set of non-negative gating functions satisfying an expert-independent norm constraint:*

$$\mathcal{G}_{\infty,1} = \left\{ g \geq 0 \colon \sum_{k=1}^p \max_{x\in\mathcal{X}_0} g(x,k) \leq 1 \right\}. \tag{3}$$

Unlike the $\mathcal{G}_1$ constraint, which couples the gate to the data-dependent predictions $\widehat{\pi}_k(x)$ of every expert, the $\mathcal{G}_{\infty,1}$ constraint is *expert-independent*: it depends solely on the gate values $g(x,k)$ and is fully decoupled from the expert models. This decoupling enables a projection operator computable without cross-expert synchronization. (In practice, the exact evaluation of $\max_{x\in\mathcal{X}_0}$ over combinatorial sequence spaces is replaced by batch-wise approximations or architectural bounds on the gate outputs; see Section 3.3 and Appendix D.1.)

Lemma 2 (Properties of $\mathcal{G}_{\infty,1}$)**.** *The family $\mathcal{G}_{\infty,1}$ is non-empty, compact, and convex. Furthermore, for any $g \in \mathcal{G}_{\infty,1}$, the total mass of the resulting mixture model π_g is bounded by unity:*

$$Z_g = \sum_{x\in\mathcal{X}_0} \pi_g(x) \leq 1. \tag{4}$$

The proof is given in Appendix B.1. Unlike the $\mathcal{G}_1$ constraint, which enforces global mass unity ($Z_g = 1$) through a data-dependent coupling of gate and expert predictions, the $\mathcal{G}_{\infty,1}$ constraint decomposes into per-expert capacity budgets $w_k = \max_x g(x,k)$ that are determined independently of the data distribution and of other experts' predictions, enabling fully decoupled optimization.

3.1 Fixed Mixture and Robust Solutions

We prove the existence of a robust solution and characterize its structure, proceeding from the fixed-mixture case to the full minimax game. In the appendix, we further characterize the optimal gate as a capacity-constrained relative entropy projection (Lemma 13, Appendix B.2) and derive the optimal capacity vector in the specialization limit (Proposition 14, Appendix B.3).

A. Fixed Mixture Solutions. Since $Z_g \le 1$ for all $g \in \mathcal{G}_{\infty,1}$, we formulate the fixed-mixture problem using $\widetilde{\mathsf{D}}_{\mathrm{KL}}$: $\min_{g\in\mathcal{G}_{\infty,1}} \mathbb{E}_{x\sim\widehat{\mathsf{p}}_\lambda}[-\log \pi_g(x)] + Z_g$. Evaluating the objective at the input-independent gate $g(x,k) = w_k$ with $\sum_k w_k \le 1$ (which lies in $\mathcal{G}_{\infty,1}$ since $\max_x g(x,k) = w_k$), combined with the joint convexity of $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ in (g,λ) (which must account for the additional mass penalty $Z_g - 1$), yields:

Proposition 3 (Fixed Mixture Guarantee for $\mathcal{G}_{\infty,1}$)**.** *For any fixed $\lambda \in \Delta$, $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) \le \sum_{k=1}^{p} \lambda_k \epsilon_k \le \max_k \epsilon_k$.*

B. Robust Solutions. We now prove the existence of a robust gating function $g^* \in \mathcal{G}_{\infty,1}$ with a favorable guarantee for *any* distribution $\widehat{\mathsf{p}}_\lambda$, $\lambda \in \Delta([1,p])$. We define the value of this robust minimax game as: $V_\infty = \min_{g\in\mathcal{G}_{\infty,1}} \max_{\lambda\in\Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g)$. We first establish the following identity.

Lemma 4 (Unnormalized KL of a Mixture)**.** *For any mixture weight $\lambda = (\lambda_1, \ldots, \lambda_p) \in \Delta([1,p])$, the following identity holds for the unnormalized KL divergence:*

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \sum_{k=1}^{p} \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\}).$$

The following is the main result of this section.

Theorem 5 (Robust Existence and Guarantees for $\mathcal{G}_{\infty,1}$)**.** *The robust minimax game for the $\mathcal{G}_{\infty,1}$ family admits a solution $g^* \in \mathcal{G}_{\infty,1}$. For any mixture $\lambda \in \Delta$, the resulting model satisfies:*

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) \le \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\}). \tag{5}$$

Here, e^{ϵ_k} denotes the exponential of the expert error ϵ_k, distinct from the mixture weight $\lambda_k \in \Delta$.

C. Geometric Interpretation.

The bound retains the $\mathcal{G}_1$ framework's *Separability Bonus* ($\mathsf{D}_{\mathrm{JSD}}$) while circumventing global normalization, see Figure 1. Evaluating the normalized density $\mathsf{p}_{g^*} = \pi_{g^*}/Z_{g^*}$ via the exact decomposition $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\mathsf{p} \parallel q) = \mathsf{D}_{\mathrm{KL}}(\mathsf{p} \parallel q/Z_q) + \Delta(Z_q)$ with penalty $\Delta(Z) = Z - 1 - \log Z \ge 0$ yields the implied guarantee:

$$\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_{g^*}) \le \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\widehat{\mathsf{p}}_k) - \Delta(Z_{g^*}). \tag{6}$$

Since $\mathcal{G}_{\infty,1}$ allows sub-normalization ($Z_{g^*} \le 1$), the penalty $\Delta(Z_{g^*})$ acts as a strictly subtractive *Normalization Bonus*, tightening the provable bound below the classical capacity cost. We note that $\mathcal{G}_{\infty,1}$ is at least as expressive as $\mathcal{G}_1$ (any $h \in \mathcal{G}_1$ rescales to $h/\sum_k \max_x h(x,k) \in \mathcal{G}_{\infty,1}$ with the same normalized density), so its achievable normalized risk is never worse, and its further advantages are (i) a tighter *provable* bound in specialized regimes (Eq. 6 vs. the known $\mathcal{G}_1$ guarantee), (ii) lower gating complexity (Theorem 8), and (iii) fully decoupled optimization. A matching lower bound (Theorem 15, Appendix B.5) shows that dense normalized systems ($Z_g = M = 1$) cannot avoid the full capacity cost.

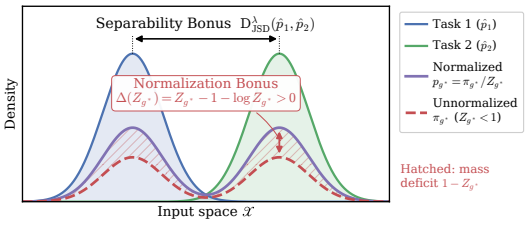


Figure 1: Geometric interpretation of the $\mathcal{G}_{\infty,1}$ guarantee: task diversity yields the Separability Bonus ($\mathsf{D}_{\mathrm{JSD}}$), and sub-normalization ($Z_{g^*} < 1$, hatched) yields the Normalization Bonus ($\Delta(Z_{g^*})$), both subtracted in Eq. 6.

This geometry governs two limits (Appendix B.6). In the *Specialization Limit* (disjoint supports), maximal diversity ($\mathsf{D}^{\lambda}_{\mathrm{JSD}} = H(\lambda)$) perfectly offsets the capacity cost, rendering the gate an optimal

switch. In the *Redundancy Limit* ($\mathsf{D}_{\mathrm{JSD}}^{\lambda} = 0$, identical experts with error ϵ), the bound reduces to $\epsilon + \log p$ (a single-expert gate attains ϵ, Appendix B.6), averting monolithic gradient conflicts. In both regimes, the $\mathcal{G}_{\infty,1}$ gate uses the task geometry ($\mathsf{D}_{\mathrm{JSD}}^{\lambda}$) to compensate for sub-normalization, effectively self-correcting without explicit enforcement of $Z_g = 1$.

D. Comparison with Monolithic Baselines. A monolithic model $\widehat{\pi}_\lambda$ from a class with best-in-class per-task errors ϵ_k satisfies the floor (Cortes et al., 2026, Theorem 6) $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \widehat{\pi}_\lambda) \geq \sum_k \lambda_k \epsilon_k - \mathsf{D}_{\mathrm{JSD}}^{\lambda}(\widehat{\mathsf{p}}_k)$ (Appendix B.7). The $\mathsf{D}_{\mathrm{JSD}}$ term thus enters the monolithic floor and the modular ceiling with the same coefficient -1, the *Symmetric Divergence Effect* (Theorem 16, Appendix B.8): both mixture risks are per-task risks discounted by the same mixing credit, so the comparison between the two architectures rests on the per-task errors of the monolithic class against those of the specialists.

3.2 Algorithm and Convergence

The robust solution g^* is computed by solving the linearized convex-concave game $\widetilde{L}(\lambda, g) = \sum_k \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$ using no-regret dynamics. The λ-player uses Exponentiated Gradient; the g-player uses Projected Gradient Descent with gradient $v_t(x,k) = \widehat{\pi}_k(x)\Big(1 - \frac{\widehat{\mathsf{p}}_{\lambda_{t+1}}(x)}{\pi_{g_t}(x)}\Big)$. Unlike $\mathcal{G}_1$, the $\mathcal{G}_{\infty,1}$ *projection* $\Pi_{\mathcal{G}_{\infty,1}}$ is expert-free: it depends solely on the gate values and is decoupled from the expert model values $\widehat{\pi}_k(x)$. To formalize this, we provide the complete non-parametric learning procedure (Algorithm 2 in Appendix B.9) and an exact Euclidean projection algorithm onto $\mathcal{G}_{\infty,1}$, based on coordinate-wise clipping and bisection on a single Lagrange multiplier, that runs in $\mathcal{O}(p \cdot |\mathcal{X}_0| \log |\mathcal{X}_0| + p \cdot |\mathcal{X}_0| \log(1/\epsilon))$ time for numerical tolerance $\epsilon > 0$ (Algorithm 5 in Appendix D.1, with a proof of correctness; the convergence guarantee below is preserved by the approximate projection for tolerance $\epsilon = \mathcal{O}(1/T)$, see Remark 17).

Algorithm 1 Robust Parametric Gate for $\mathcal{G}_{\infty,1}$ (uKL-EG + SGD)

1: **Input:** Models $\widehat{\pi}_1, \ldots, \widehat{\pi}_p$; datasets $D_1, \ldots, D_p$; learning rates $\eta_\lambda, \eta_\theta$; iterations T; penalty coefficient α.
2: **Initialize:** $\lambda_0(k) = 1/p$; router parameters θ_0 (shared weights).
3: **for** $t = 0, 1, \ldots, T-1$ **do**
4: **Sample** mini-batch $\mathcal{B} \sim \cup D_k$.
5: **(Compute Gains)** For $k \in [1, p]$, compute $\ell_t(k) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_{g_\theta})$ over $\mathcal{B}$.
6: **(λ-update)** Update mixture weights using Exponentiated Gradient: $\lambda_{t+1}(k) \propto \lambda_t(k) \exp(\eta_\lambda \ell_t(k))$.
7: **(θ-update)** Update gate parameters θ via gradient descent on the composite objective:
$\mathcal{L}_{\text{total}} = \underbrace{\mathbb{E}_{x \sim \widehat{\mathsf{p}}_{\lambda_{t+1}}}[-\log \pi_{g_\theta}(x)] + Z_{g_\theta}}_{\text{Penalized Log-Likelihood}} + \underbrace{\alpha \left[\mathrm{ReLU}\Big(\sum_{k=1}^{p} \max_{x \in \mathcal{B}} g_\theta(x,k) - 1\Big)\right]^2}_{\text{Group-Norm Penalty}}$.
8: **end for**
9: **Output:** The trained gate parameters θ_T.

Theorem 6 (Convergence Guarantee for $\mathcal{G}_{\infty,1}$)**.** *Let $V_\infty^\eta = \min_{g \in \mathcal{G}_{\infty,1}^\eta} \max_{\lambda \in \Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g)$ be the robust value of the game over the non-degenerate family $\mathcal{G}_{\infty,1}^\eta$ (Appendix B.11). Under Assumption 18, with step sizes $\eta_\lambda, \eta_g \propto 1/\sqrt{T}$, the time-averaged gate $\overline{g}_T$ produced by Algorithm 2 (with projection onto $\mathcal{G}_{\infty,1}^\eta$) converges to the robust solution with rate:*

$$\max_{\lambda \in \Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{\overline{g}_T}) - V_\infty^\eta \leq \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right). \tag{7}$$

In practice, the soft penalty in Algorithm 1 requires only $\mathcal{O}(|\mathcal{B}| \times p)$ operations on the lightweight gating logits, bypassing expert evaluation and cross-device synchronization entirely. The partition function Z_{g_θ} in the penalized log-likelihood is estimated at every step from sequences sampled from the frozen experts via the identity $Z_{g_\theta} = \sum_k \mathbb{E}_{x \sim \widehat{\pi}_k}[g_\theta(x,k)]$, and the gradient flows through this estimate, so the mini-batch gradient of the linear mass term is unbiased (Appendix G.1). The soft group-norm penalty on $\sum_k \max_{x \in \mathcal{B}} g_\theta(x,k)$ enforces the structural capacity constraint $\sum_k w_k \leq 1$. An alternative primal-dual formulation that dynamically learns α via dual ascent is given in Appendix B.10.

3.3 Generalization and Sample Efficiency

Adapting the framework of Cortes et al. (2026) to our unnormalized setting, we establish formal generalization bounds under mild positivity conditions (Assumption 18, Theorem 19, Appendix B.11; simplified in Corollary 20 for uniform sample sizes). The generalization gap is isolated into the empirical Rademacher complexity of the gating class, with a confidence term of only $\mathcal{O}(\sqrt{\log p/m})$ in the number of experts. The bound applies verbatim to any subclass of $\mathcal{G}_{\infty,1}$, in particular to the lightweight parametric gating network used in practice, whose complexity governs the gap.

The following theorem compares the two constraint sets under a mild expert-separation condition: $\mathcal{G}_{\infty,1}$ caps each expert's total weight across inputs, whereas $\mathcal{G}_1$ can fit any assignment on the sample.

Assumption 7 (Sample Size vs. Expert Separation)**.** *There exist at least two experts j and l whose total-variation distance bounds the sample size multiplied by the maximum point-mass probability of any expert: $m \max_{x \in \mathcal{X}_0, k} \widehat{\pi}_k(x) \le \|\widehat{\pi}_j - \widehat{\pi}_l\|_{\mathrm{TV}}$.*

The left side (exponentially small for LLMs) bounds the mass any mixture of experts can assign to m distinct sequences, while the right side is close to 1 for specialized experts (Appendix B.11).

Theorem 8 (Rademacher Complexity Comparison)**.** *Under Assumption 7, let S be a sample of m distinct points, then the following holds: if $p = 2$, then $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + 1/(2\sqrt{2m})$; if $p \ge 3$ and $m \ge 3$, then $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) < \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$.*

3.4 Inference: Exact and Efficient Rejection Sampling

An important advantage of $\mathcal{G}_{\infty,1}$ is its compatibility with *Exact Rejection Sampling*. Generating unbiased samples from unnormalized models typically requires evaluating the intractable global mass $Z_{\overline{g}_T}$ over $\mathcal{X}_0$. The data-independent $\mathcal{G}_{\infty,1}$ constraint circumvents this by providing an explicitly computable geometric envelope. Let $w_k^* = \max_{x'} \overline{g}_T(x', k)$ and $M_T = \sum_k w_k^* \le 1$. The capacity-weighted proposal $\mathsf{q}_{w^*}(x) = \frac{1}{M_T} \sum_k w_k^* \widehat{\pi}_k(x)$ bounds the unnormalized density: $\pi_{\overline{g}_T}(x) \le M_T \mathsf{q}_{w^*}(x)$. We sample hierarchically ($k \propto w_k^*$, then $x \sim \widehat{\pi}_k$) and accept with probability $\alpha(x) = \sum_k \overline{g}_T(x,k)\widehat{\pi}_k(x) / \sum_k w_k^* \widehat{\pi}_k(x)$. Since $Z_{\overline{g}_T}$ exactly cancels, accepted samples follow $\mathsf{p}_{\overline{g}_T} = \pi_{\overline{g}_T}/Z_{\overline{g}_T}$ without approximation bias.

Proposition 9 (Structural Efficiency of Exact Sampling)**.** *For any time-averaged gate $\overline{g}_T \in \mathcal{G}^{\eta}_{\infty,1}$, the expected acceptance rate is $\mathbb{P}(\mathrm{accept}) = Z_{\overline{g}_T}/M_T \ge Z_{\overline{g}_T}$. Furthermore, the mass is bounded away from zero: $Z_{\overline{g}_T} \ge \exp(-(C_T + 1)) > 0$, where $C_T = \max_\lambda \widetilde{L}(\lambda, \overline{g}_T) \le \log[\sum_k e^{\epsilon_k}] + \epsilon_T$ is the bounded worst-case risk of the algorithm.*

At the robust solution the acceptance rate is one minus the total variation between the worst-case source mixture and the capacity-weighted expert mixture (Appendix B.12); empirically the *Sparsity Ratio* is $Z_{\overline{g}_T}/M_T \approx 1$ (Figure 8), giving efficient, exact inference.

Comparison with $\mathcal{G}_0$ and $\mathcal{G}_1$. Since $g \le 1$, both $\mathcal{G}_0$ and $\mathcal{G}_1$ admit exact rejection sampling via the uniform mixture envelope (Cortes et al., 2026, App. F.4.2), with mass p and acceptance Z_g/p: $\approx 1/p$ for our $\mathcal{G}_0$ gate ($Z_{g^*} \approx 1$), 5% at $p = 20$ for $\mathcal{G}_1$. The envelope above gives any gate h acceptance Z_h/M_h, but only $\mathcal{G}_{\infty,1}$ enforces $M_h \le 1$; one-hot $\mathcal{G}_1$ gates have $M_h = p$ (Appendix B.12).

Deployment Modes. Computing the true global ceiling $w_k^* = \max_{x \in \mathcal{X}_0} g_\theta(x, k)$ over combinatorial sequence spaces is intractable. We distinguish two modes: (a) *calibration-set maxima* with a safety margin yield approximately exact sampling: on held-out data, only 0.02% of gate values (7 of 36,000) exceeded the raw calibration maxima, and none exceeded them once a 5% safety margin was added; (b) an *architecturally bounded gate* $g_\theta(x,k) = W_k \cdot \sigma(h_\theta(x,k))$ with $\sum_k W_k \le 1$ enforces $w_k^* \le W_k$ by construction for all inputs, giving provably exact and unbiased sampling. Only mode (b) is referred to as provably exact.

3.5 Structural Distillation for $\mathcal{G}_{\infty,1}$

The non-causal gate $g^*(x,k)$ prevents standard autoregressive generation. Following the *Structural Distillation* framework of Cortes et al. (2026), we train a lightweight *Causal Router* $\gamma_\phi(k \mid x_{<t})$ to approximate the normalized teacher $p_{g^*} = \pi_{g^*}/Z_{g^*}$ while preserving the frozen experts. The

student model factors autoregressively as $\pi_\gamma(x) = \prod_t [\sum_k \gamma_\phi(x_{<t}, k)\, \widehat{\pi}_k(x_t \mid x_{<t})]$, with $Z_\gamma = 1$ by construction. The distillation objective is the forward KL divergence from the normalized teacher to the student:

$$\min_\phi \mathcal{J}(\phi) = \mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) = - \mathop{\mathbb{E}}_{x \sim p_{g^*}} \left[\sum_{t=1}^{|x|} \log \left(\sum_{k=1}^{p} \gamma_\phi(x_{<t}, k)\, \widehat{\pi}_k(x_t \mid x_{<t}) \right) \right] + \text{const.} \tag{8}$$

Sequences from p_{g^*} are obtained via the exact rejection sampler (Section 3.4). This is standard next-token prediction MLE on teacher-generated sequences, enabling the cached-logit training of Cortes et al. (2026) (Algorithms 3–6, Appendix B.13.2).

We prove (Proposition 21, Appendix B.13.3) that the structural mismatch between the Mixture-of-Products teacher and the Product-of-Mixtures student is bounded by the joint convexity of KL, vanishing when the gate does not depend on future tokens. The distillation divergence decomposes exactly into token-level conditional KLs (Theorem 22, Appendix B.13.4).

Theorem 10 (Inheritance Guarantee)**.** *If* $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) \le \delta$*, then for any* $\lambda \in \Delta$*:* ***(a)*** $\|\widehat{\mathsf{p}}_\lambda - \pi_\gamma\|_1 \le \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 + \sqrt{2\delta}$*;* ***(b)*** *under Assumption 18,* $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_\gamma) \le \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| p_{g^*}) + \delta + C\sqrt{2\delta}$*, where* $C = \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 / \eta'$ *and* $\eta' = \min_{x \in \mathcal{X}_0} \min\{p_{g^*}(x), \pi_\gamma(x)\} > 0$.

The direct $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$ objective yields a *single-term* bound with no causal projection gap ($\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| p_{g^*}) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_{g^*}) - \Delta(Z_{g^*})$). A complementary per-token TV guarantee, free of the constant C and the boundedness assumption, is presented in Theorem 23 (Appendix B.13.6).

4 Locally-Normalized Gating ($\mathcal{G}_0$)

The $\mathcal{G}_{\infty,1}$ family eliminates data-dependent synchronization by replacing $Z_g = 1$ with $\sum_k \max_x g(x,k) \le 1$, which still couples experts at the mini-batch level. We now ask: *what happens if we push decentralization to its absolute limit?* The $\mathcal{G}_0$ family enforces normalization strictly at the local level, so that each input x is propagated independently of all others.

Definition 11 (Locally-Normalized Family $\mathcal{G}_0$)**.** $\mathcal{G}_0 = \left\{ g \ge 0 : \sum_{k=1}^p g(x,k) = 1,\ \forall x \in \mathcal{X}_0 \right\} = \prod_{x \in \mathcal{X}_0} \Delta([1,p])$.

Motivation and Objective. This family is non-empty, compact, and convex (Lemma 25, Appendix C.1). Unlike $\mathcal{G}_{\infty,1}$, Z_g is unconstrained ($Z_g \le \sum_x \max_k \widehat{\pi}_k(x)$, potentially $\gg 1$). To avoid estimating an intractable partition function, we optimize the NLL $L(\lambda, g) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_g)$ rather than $\widetilde{\mathsf{D}}_{\mathrm{KL}}$, yielding fully decoupled pointwise updates with zero cross-node communication.

Robust Solutions and Selection Gain. For a fixed target, the $\mathcal{G}_0$ gate achieves an *Ambiguity Gain* ($\mathcal{A}^\lambda \ge 0$) from Jensen's inequality (Proposition 26, Appendix C.2), strictly improving upon the weighted average of individual expert log-likelihoods. Building on this, we establish robust minimax existence.

Theorem 12 (Robust Existence and Guarantees for $\mathcal{G}_0$)**.** *The minimax game admits a solution* $g^* \in \mathcal{G}_0$ *that performs hard pointwise expert selection:* $g^*(x,k) = \mathbf{1}[k = \operatorname{argmax}_j \widehat{\pi}_j(x)]$ *(ties broken arbitrarily). For any test mixture* $\lambda \in \Delta$*, with* λ^* *a worst-case mixture at* g^* *and* $\mathcal{A}^{\lambda^*}_{\text{selection}} = \sum_k \lambda^*_k \mathbb{E}_{x \sim \widehat{\mathsf{p}}_k} \left[\log \frac{\max_j \widehat{\pi}_j(x)}{\widehat{\pi}_k(x)} \right] \ge 0$ *the pointwise selection gain:*

$$\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g^*}) \le \max_{k \in [1,p]} (\epsilon_k + H(\widehat{\mathsf{p}}_k)) - \mathcal{A}^{\lambda^*}_{\text{selection}}.$$

Selection Gain vs. Separability Bonus. This bound reveals a core trade-off: $\mathcal{G}_{\infty,1}$ benefits from the *Separability Bonus* ($-\mathsf{D}_{\mathrm{JSD}}$), tightening as tasks diverge, while $\mathcal{G}_0$ benefits from the *Selection Gain* ($\mathcal{A}_{\text{selection}}$), tightening when expert predictions overlap (Figure 5(b), Appendix C.4). Intuitively, $\mathcal{G}_{\infty,1}$ rewards distributional separation; $\mathcal{G}_0$ rewards pointwise variation in expert quality.

Volume Penalty and Regularization. On the normalized KL scale, a *Volume Penalty* $\log Z_{g^*} \ge 0$ appears. We augment the objective with a soft regularizer $\beta(Z_g - 1)^2$ that drives the mass toward unity (Theorem 28, Appendix C.6). The regularizer gradient requires a periodic scalar aggregation of Z_g (a single all-reduce), which is qualitatively lighter than the full gradient synchronization of $\mathcal{G}_1$;

the pointwise gate updates remain fully decoupled. The game is solved via a parallelizable no-regret procedure converging at $\mathcal{O}(\sqrt{\log p/T})$ (Algorithm 4, Theorem 27, Appendix C.5).

Inference via Structural Distillation. Since $Z_g \not\leq 1$, the tight structural envelope of $\mathcal{G}_{\infty,1}$ is unavailable; the uniform-mixture envelope still yields an exact sampler, but with acceptance rate $Z_g/p \approx 1/p$ (Section 3.4). We therefore deploy the $\mathcal{G}_0$ teacher via Structural Distillation into a causal student $\gamma_\phi(k \mid x_{<t})$, minimizing $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$. The inheritance guarantee mirrors Theorem 10: both the TV and KL bounds hold with no causal projection gap (Theorem 29, Appendix C.8).

Comparative Analysis. As summarized in Table 1, the two relaxations occupy distinct Pareto positions. $\mathcal{G}_{\infty,1}$ exploits the Normalization Bonus $\Delta(Z^*)$ (vs. $\mathcal{G}_1$'s Overlap Gain H_σ, Cortes et al. 2026), achieves tighter generalization ($\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$, Theorem 8), and enables efficient exact rejection sampling. $\mathcal{G}_0$ trades these for maximum throughput via fully decoupled updates, deployable autoregressively through Structural Distillation.

5 Experiments

We evaluate $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ against the globally normalized $\mathcal{G}_1$ baseline, monolithic models, and standard MoE models following the multi-domain setup of Cortes et al. (2026).

Setup. We evaluate across four capacity regimes: *Lightweight* (~6.5M per expert, 3 domains), *1B+* (~1.6B per expert, 3 domains), *20-Task Scaling* (~130M total, 20 code languages), and *Pre-Trained Frozen* (GPT-2 + CodeParrot-small, ~124M). Experts are causal Transformer decoders; gating networks are bidirectional Transformer encoders (~290K–2.5M parameters). The *Monolithic* baseline (referred to as the *Retrained* model by Cortes et al. (2026)) and the standard MoE baseline are capacity-matched to the combined modular system: their parameter count matches the sum of the experts plus the gating network. For all experiments, domains span Wikipedia (Wikimedia Foundation, 2024), FineWeb-Edu (Penedo et al., 2024), and code (Kocetkov et al., 2022; Chaudhary, 2023). All results are averaged over 5 runs with fixed test sets, varying seeds or training splits (see Appendix G.1). All reported results enforce the $\mathcal{G}_{\infty,1}$ constraint via the parametric soft-penalty implementation (Algorithm 1); the non-parametric exact projection (Appendix D.1) is used solely in the convergence analysis (Theorem 6). For full implementation details, see Appendix G.1.

Table 2: Results for Lightweight (~6.5M) and 1B+ parameter regimes. Individual experts report in-domain NLL; all other rows report NLL on the union test mixture. Unnormalized gates outperform all baselines across seeds and data splits.

Model	**Lightweight**		**1B+**	
	Diff Seed	**Diff Data**	**Diff Seed**	**Diff Data**
Wiki Expert	4.938 ± 0.008	4.934 ± 0.002	4.756 ± 0.005	4.743 ± 0.023
Code Expert	5.638 ± 0.022	5.109 ± 0.025	5.215 ± 0.007	4.938 ± 0.036
FineWeb Expert	5.471 ± 0.002	5.467 ± 0.003	5.390 ± 0.005	5.387 ± 0.002
Monolithic Model	5.311 ± 0.017	5.155 ± 0.017	4.987 ± 0.006	4.886 ± 0.021
Standard MoE Model	5.271 ± 0.014	5.117 ± 0.014	4.982 ± 0.005	4.879 ± 0.012
Gate Model $\mathcal{G}_1$ (Baseline)	5.143 ± 0.005	4.983 ± 0.009	4.895 ± 0.010	4.769 ± 0.017
Gate Model $\mathcal{G}_{\infty,1}$ (Ours)	$\mathbf{5.123 \pm 0.004}$	$\mathbf{4.962 \pm 0.009}$	$\mathbf{4.835 \pm 0.005}$	$\mathbf{4.749 \pm 0.015}$
Gate Model $\mathcal{G}_0$ (Ours)	5.129 ± 0.008	4.967 ± 0.008	4.859 ± 0.002	4.758 ± 0.015

Experimental Results. Our primary result is that the unnormalized relaxations consistently improve over the $\mathcal{G}_1$ baseline. At the 1B+ regime (Table 2), $\mathcal{G}_{\infty,1}$ outperforms $\mathcal{G}_1$ in both setups. In the 20-Task Scaling regime (Table 3), where the $\mathcal{O}(\log p)$ capacity term becomes pronounced, the unnormalized gates outperform $\mathcal{G}_1$ on 19 of the 20 languages (Table 8, Appendix G.4). These advantages persist at the *Lightweight* scale (~6.5M), where both $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ outperform $\mathcal{G}_1$. Post-training mass estimation confirms the geometric predictions: $\mathcal{G}_{\infty,1}$ is sub-normalized ($\log Z_{g^*} \approx -0.197$, $\Delta(Z_{g^*}) \approx 0.018$), while $\mathcal{G}_0$ incurs a slight Volume Penalty ($\log Z_{g^*} \approx 0.049$).

Furthermore, in the *Pre-Trained Frozen Experts* regime (Table 3, Appendix G.5), our unnormalized $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ gates successfully combine off-the-shelf foundation models (GPT-2 and CodeParrot-small) without any expert fine-tuning, outperforming the $\mathcal{G}_1$ baseline, standard MoE, and monolithic model. All reported NLL values represent the mean and standard deviation over five independent runs

Table 3: Aggregate Test NLL for 20-Task and Frozen Experts. $\mathcal{G}_{\infty,1}$ achieves the lowest NLL.

Model	20-Task	Frozen Experts
Monolithic Model	2.467 ± 0.006	2.736 ± 0.008
Standard MoE Model	2.400 ± 0.006	2.586 ± 0.007
Gate Model $\mathcal{G}_1$ (Baseline)	2.365 ± 0.004	2.466 ± 0.005
Gate Model $\mathcal{G}_0$ (Ours)	2.356 ± 0.003	2.446 ± 0.004
Gate Model $\mathcal{G}_{\infty,1}$ (Ours)	$\mathbf{2.353 \pm 0.002}$	$\mathbf{2.396 \pm 0.003}$

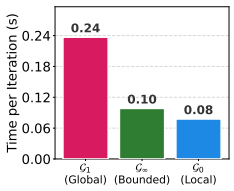


Figure 2: Wall-clock time for the pre-trained frozen experts task.

with different seeds. By keeping the base models entirely frozen, the unnormalized gates function as a universal routing mechanism that dynamically shifts probability mass to the most competent expert without requiring task-specific parameter updates, effortlessly combining structurally distinct foundation models. *Note:* Tables 2 and 3 report the non-causal teacher ($\mathcal{G}_{\infty,1}$ / $\mathcal{G}_0$) NLL; the causal $\mathcal{G}_{\infty,1}$ student, the strict autoregressive comparison, attains 2.401 on the frozen-experts union (Appendix G.5.3), preserving the ordering in Table 3.

Computational Efficiency. Wall-clock timing for the pre-trained frozen experts task (Figure 2), averaged over the final 4,980 of 5,000 iterations on an A100 GPU (first 20 warmup steps discarded), confirms that $\mathcal{G}_{\infty,1}$ reduces iteration time by $> 50\%$ over $\mathcal{G}_1$ (0.10s vs. 0.24s), with $\mathcal{G}_0$ achieving just 0.08s, validating that relaxing global normalization eliminates the dominant synchronization overhead. On a single device, this overhead manifests as sequential expert evaluation and global partition function computation. In distributed settings, it additionally requires cross-node communication for global normalization. Total compute across all regimes is approximately 200 A100-hours.

Scaling Laws and Inference Validation. To isolate the $\mathcal{O}(\sqrt{\log p})$ scaling, we measured the Excess Gating Risk, (Gate NLL minus the Oracle Expert NLL), with $m_k = 500$ held constant across $p \in \{2, 5, 10, 15, 20\}$. Figure 3 (Left, Middle) shows the excess risk scaling with $\sqrt{\log p}$ and the regret levels separating according to p, consistent with the $\log p$ dependence in Theorems 8 and 6.

Inference and Distillation. The empirical acceptance rate $Z_{\bar{g}_T}/M_T$ stably converges to a highly efficient regime (Figure 8, Appendix G.5.2), validating Proposition 9. The observed $Z_{\bar{g}_T} \approx M_T$ keeps exact sampling efficient. Figure 3 (Right) confirms the Distillation Inheritance guarantee (Theorem 10): the distillation KL $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$ converges to a small residual and the student–teacher NLL gap stays within $\delta + C\sqrt{2\delta}$ across all domains (Figure 9, Appendix G.5.3). Ablation on α (Appendix G.5.4) validates $\alpha \geq 10$. See Appendix G for extended results.

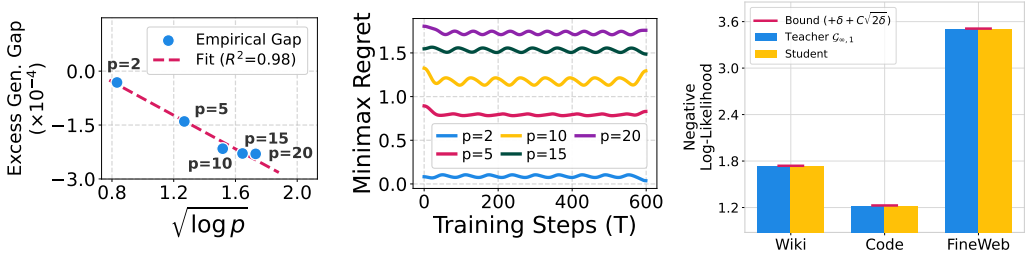


Figure 3: (Left) Excess gating risk (Gate NLL minus Oracle Expert NLL) scales with $\mathcal{O}(\sqrt{\log p})$. (Middle) Minimax regret levels separate according to the $\log p$ dependence of Theorem 6. (Right) Distillation NLL Gaps across Domains.

6 Conclusion

We introduced two unnormalized relaxations of the robust gating framework $\mathcal{G}_1$ and proved that relaxing global normalization does not merely save compute, it yields provably tighter robustness bounds in specialized regimes, tighter generalization bounds, and enables efficient exact inference. The bounded $\mathcal{G}_{\infty,1}$ family earns a *Normalization Bonus*, a favorable complexity ordering, and efficient exact sampling; the locally-normalized $\mathcal{G}_0$ achieves maximum throughput via fully decoupled updates, deployable autoregressively through Structural Distillation. Together, they delineate a Pareto front between theoretical rigor ($\mathcal{G}_{\infty,1}$) and computational throughput ($\mathcal{G}_0$), enabling robust modular models at scales where global synchronization was previously prohibitive. More broadly, our results demonstrate that structural gating constraints can simultaneously improve computational efficiency, statistical generalization, and inference exactness. See Appendix F for limitations.

AI USE STATEMENT

In the preparation of this manuscript, generative AI tools were used to assist with text editing (grammar, clarity, and style refinement) as well as code debugging and formatting and illustrative figure generation. No generative AI tools were used for core research ideation, theoretical formulation, or analyzing experimental outcomes. All theoretical proofs, experimental designs, code execution, and data analysis were rigorously verified by the authors, who remain fully responsible for the accuracy and integrity of all content in this paper.

ETHICS STATEMENT

This work is primarily theoretical and algorithmic, and does not involve human subjects. All datasets and pre-trained models used in our experiments are publicly available and used under their respective open-source licenses (listed in Appendix G.1). A detailed discussion of potential negative societal impacts, including dual-use risks from lowering the computational barrier to large-scale generative modeling and the possible amplification of biases present in individual expert training corpora, together with mitigations, as well as privacy considerations and environmental impact, is provided in Appendix F.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results, together with the underlying assumptions, are provided in the appendices (Appendices B–C.8), and the full non-parametric algorithms, the exact projection procedure, and the distillation loops are given in Algorithms 2–6. All experimental details needed to reproduce our empirical results, including model architectures and sizes, datasets and their exact splits, tokenization, training hyperparameters, penalty coefficients, and evaluation protocols (5 runs with fixed test sets and varied seeds/splits), are documented in Appendix G.1, and the estimation protocol used to report normalized NLL for unnormalized models is described in Appendix E.

REFERENCES

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. LLM augmented LLMs: Expanding capabilities through composition. In *International Conference on Learning Representations*, 2024.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Kshipra Bhawalkar, Jeff Dean, Christopher Liaw, Aranyak Mehta, and Neel Patel. Equilibria and learning in modular marketplaces. *arXiv preprint arXiv:2502.20346*, 2025.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

James P. Boyle and Richard L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, volume 37 of *Lecture Notes in Statistics*, pp. 28–47. Springer, 1986.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pp. 1877–1901, 2020.

Sahil Chaudhary. Code alpaca: An instruction-following LLaMA model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*, pp. 67–82, 2016a.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2514–2522, 2016b.

Corinna Cortes, Mehryar Mohri, and Ananda Theertha Suresh. Relative deviation margin bounds. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021a.

Corinna Cortes, Mehryar Mohri, Ananda Theertha Suresh, and Ningshan Zhang. A discriminative technique for multiple-source adaptation. In *International Conference on Machine Learning*, pp. 2132–2143, 2021b.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, 92(2):277–315, 2024.

Corinna Cortes, Mehryar Mohri, and Yutao Zhong. A theoretical framework for modular learning of robust generative models. *arXiv preprint arXiv:2602.17554*, 2026.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

Imre Csiszár and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

Christoph Dann, Yishay Mansour, Teodor V. Marinov, and Mehryar Mohri. Principled model routing for unknown mixtures of source domains. In *Advances in Neural Information Processing Systems*, 2025.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

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

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

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

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301, 2017.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pp. 8246–8256, 2018.

Bodun Hu, Shuozhe Li, Saurabh Agarwal, Myungjin Lee, Akshay Jajoo, Jiamin Li, Le Xu, Geon-Woo Kim, Donghyun Kim, Hong Xu, Amy Zhang, and Aditya Akella. StitchLLM: Serving LLMs, one block at a time. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 26887–26903, 2025.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Michael G Jacobides, Stefano Brusoni, and Francois Candelon. The evolutionary dynamics of the artificial intelligence ecosystem. *Strategy Science*, 6(4):412–435, 2021.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Junlin Julian Jiang and Xin Li. Look ahead text understanding and LLM stitching. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 751–760, 2024.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The stack: 3 TB of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In Gükhan Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan (eds.), *Predicting Structured Data*. MIT Press, 2006.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Yonghui Wu. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations (ICLR)*, 2021.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pp. 276–284, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, 2008.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, pp. 4753–4761, 2024a.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable H-consistent and Bayes-consistent loss functions for learning to defer. In *Advances in Neural Information Processing Systems*, 2024b.

Michael Matena and Colin Raffel. Merging models with Fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, 2022.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings*, pp. 3–17. Springer, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, second edition, 2018.

Mehryar Mohri, Judy Hoffman, and Ningshan Zhang. Multiple-source adaptation theory and algorithms. *Annals of Mathematics and Artificial Intelligence*, 89(3):237–270, 2021.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087. PMLR, 2020.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, pp. 30811–30849, 2024.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Ponti. Modular deep learning. *Transactions on Machine Learning Research*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

Donald B Rubin. Using the SIR algorithm to simulate posterior distributions. *Bayesian Statistics*, 3: 395–402, 1988.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

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.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, 2019.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 11415–11434. PMLR, 2023.

Leandro von Werra. Training CodeParrot from scratch. Hugging Face Blog, https://huggingface.co/blog/codeparrot, 2021.

Wikimedia Foundation. Wikimedia downloads, 2024. URL https://dumps.wikimedia.org.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pp. 23965–23998, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

CONTENTS OF APPENDIX

A Extended Discussion of Related Work **17**

A.1 Mixture of Experts and Modular Deep Learning . . . 17
A.2 Robustness and Multiple-Source Adaptation . . . 17
A.3 Model Composition, Merging, and Stitching . . . 17
A.4 Learning to Defer and Routing Policies . . . 17
A.5 Unnormalized Models and Distillation . . . 18
A.6 AI Ecosystems and Modular Economics . . . 18

B Expert-Independent Bounded Gating ($\mathcal{G}_{\infty,1}$) **18**

B.1 Definition and Properties . . . 18
B.2 Fixed Mixture Solutions . . . 19
B.3 Optimal $\mathcal{G}_{\infty,1}$ solution for a Fixed-Mixture . . . 19
B.4 Robust Solutions . . . 21
B.5 Fundamental Lower Bound for $\mathcal{G}_{\infty,1}$ . . . 22
B.6 Geometric Interpretation of the $\mathcal{G}_{\infty,1}$ Bound . . . 24
B.7 Comparison with Monolithic Baselines . . . 25
B.8 Symmetric Divergence Effect . . . 25
B.9 Algorithm and Convergence . . . 26
B.10 Alternative Formulation: A Primal-Dual Approach . . . 28
B.11 Generalization and Sample Efficiency . . . 29
B.12 Inference: Exact and Efficient Rejection Sampling . . . 33
B.13 Structural Distillation for $\mathcal{G}_{\infty,1}$ . . . 34
B.13.1 Background: Structural Distillation . . . 34
B.13.2 Model Definitions . . . 34
B.13.3 The Posterior Probability Causal Router . . . 35
B.13.4 Distillation Objective and Chain Rule Decomposition . . . 36
B.13.5 Distillation Performance Inheritance . . . 36
B.13.6 Average Per-Token Distillation Guarantee . . . 37

C Locally-Normalized Gating ($\mathcal{G}_0$) **38**

C.1 Motivation, Definition and Properties . . . 38
C.2 Fixed Mixture Solutions . . . 38
C.3 Robust Solutions . . . 39
C.4 Geometric Interpretation of the $\mathcal{G}_0$ Bound . . . 40
C.5 Algorithm and Convergence . . . 41
C.6 Global Regularization and Mass Stability . . . 43
C.7 Generalization and Sample Efficiency . . . 44

C.8 Inference: Approximate Sampling-Importance-Resampling and Structural Distillation 44

D Implementation Details **45**

D.1 Exact Euclidean Projection onto $\mathcal{G}_{\infty,1}$ 45

D.2 Structural Distillation Procedure 47

E Evaluation of Normalized NLL for Unnormalized Families **47**

E.1 Method 1: Monte Carlo and Importance Sampling 47

E.2 Method 2: Acceptance-Rate Mass Estimation (Rejection Sampling) 48

E.3 Properties of the Acceptance-Rate Estimator 48

F Comprehensive Limitations and Societal Impact **49**

F.1 Technical Limitations and Failure Modes 49

F.2 Environmental Impact: Modularity for Greener AI 50

F.3 Privacy and Data Sovereignty 50

F.4 Potential Negative Societal Impacts 50

G Experimental Results **50**

G.1 Experimental Setup 51

G.2 Real-World Robustness Results (Lightweight Models) 52

G.3 Scaling to the 1B+ Parameter Regime 53

G.4 Scaling Modularity: 20-Task Code Generation 54

G.5 Scaling with Pre-Trained Frozen Experts 54

G.5.1 Computational Efficiency: Wall-Clock Time 56

G.5.2 Generative Exact Inference and Acceptance Rates 56

G.5.3 Causal Structural Distillation Dynamics 56

G.5.4 Ablation Studies: Capacity Ceiling Constraint (α) 57

A Extended Discussion of Related Work

Our proposed unnormalized gating framework sits at the intersection of several research areas. We provide a concise overview below and refer the reader to the comprehensive survey in Cortes et al. (2026) for an extended discussion of Mixture-of-Experts, Multiple-Source Adaptation, model composition, and learning-to-defer literature.

A.1 Mixture of Experts and Modular Deep Learning

Early MoE architectures (Jacobs et al., 1991; Jordan & Jacobs, 1994; Shazeer et al., 2017; Fedus et al., 2022; Lepikhin et al., 2021; Du et al., 2022) scale parameter counts without proportionally increasing inference compute but rely on heuristic load-balancing losses. Our modular framework operates on frozen experts (in the sense of the modular deep learning taxonomy of Pfeiffer et al. 2023), yielding strict generalization bounds on the lightweight gate alone (Section B.11).

Architectural distinction. Standard sparse-MoE routes each token to $\text{top-}k$ jointly-trained feed-forward sub-networks within a single transformer (intra-model routing), using auxiliary load-balancing losses that lack worst-case guarantees. Our framework routes entire sequences to frozen, pre-trained domain experts (inter-model routing), optimizing a minimax objective over sequence-level log-likelihoods with provable robustness bounds. This distinction also means our gating function need not be sparse: the optimal $\mathcal{G}_{\infty,1}$ gate can simultaneously assign positive mass to every expert for a given input, unlike top-k routing.

A.2 Robustness and Multiple-Source Adaptation

Our minimax objective is rooted in Multiple-Source Domain Adaptation (MSA) (Mansour et al., 2008; Hoffman et al., 2018; Mohri et al., 2021; Cortes et al., 2021b). While Dann et al. (2025) applied minimax routing to value-based settings, extending it to generative models requires the KL divergence and normalization constraints introduced by Cortes et al. (2026). Our $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ relaxations advance this paradigm by demonstrating that strict normalization can be safely dropped while maintaining minimax optimality and solving the distributed synchronization bottleneck. Our approach also intersects with DRO (Ben-Tal et al., 2013; Sagawa et al., 2020; Duchi & Namkoong, 2021), though we target the geometry of sequence generation rather than general ERM.

A.3 Model Composition, Merging, and Stitching

A popular alternative for leveraging multiple pre-trained models is weight merging. Techniques like Model Soups (Wortsman et al., 2022), Fisher-Weighted Averaging (Matena & Raffel, 2022), and Task Arithmetic (Ilharco et al., 2022) linearly combine the weights of fine-tuned models to create a single static generalist. More recent methods such as TIES-Merging (Yadav et al., 2023) and DARE (Yu et al., 2024) explicitly address parameter redundancy and sign interference to improve merging quality. Nevertheless, even optimized merging produces a static model that cannot adapt to per-input domain shifts. When domains are highly disjoint, weight interference remains a fundamental limitation (Yu et al., 2020). Recent works like CALM (Bansal et al., 2024) augment LLMs via cross-attention, while others explore block-level stitching (Jiang & Li, 2024; Hu et al., 2025). Unlike these approaches, which either result in a static monolithic model or lack statistical performance guarantees, our framework uses dynamic, input-dependent gating. The *Symmetric Divergence Effect* makes the comparison explicit: task diversity enters the monolithic floor and our modular ceiling identically, so the advantage of dynamic gating rests on the specialists' per-task errors rather than on merging.

A.4 Learning to Defer and Routing Policies

The learning-to-defer (L2D) literature (Cortes et al., 2016a; 2024; Mozannar & Sontag, 2020; Verma et al., 2023; Mao et al., 2024a;b) focuses on hard classification deferral under average-case accuracy. Our framework differs by inducing a soft generative mixture, optimizing sequence-level log-likelihoods under worst-case adversarial mixtures.

A.5 Unnormalized Models and Distillation

Evaluating sub-normalized gating families requires navigating unnormalized relative entropy and information geometry (Csiszár, 1975; Csiszár & Matus, 2003; van Erven & Harremoës, 2014; Amari, 2016). Training unnormalized Energy-Based Models (EBMs) traditionally requires heavy MCMC approximations or Noise Contrastive Estimation (LeCun et al., 2006; Gutmann & Hyvärinen, 2010). A major contribution of our $\mathcal{G}_{\infty,1}$ family is showing that by establishing a data-independent structural bound, we bypass these approximations entirely and achieve Exact Rejection Sampling (Section B.12) (Rubin, 1988). This connects our work to the broader literature on tractable unnormalized models, including score matching (Hyvärinen, 2005) and Stein discrepancies (Liu et al., 2016; Gorham & Mackey, 2017), though our approach is the first to provide rejection-sampling guarantees for modular mixtures. Furthermore, we extend sequence-level knowledge distillation (Hinton et al., 2015; Kim & Rush, 2016) to the unnormalized setting, providing the first proof (Section B.13.5) that a causal student inherits the minimax robustness bounds of an unnormalized, non-causal teacher.

A.6 AI Ecosystems and Modular Economics

As Foundation Models proliferate (Bommasani et al., 2021), there is growing interest in the evolutionary dynamics (Jacobides et al., 2021) and economic equilibria (Bhawalkar et al., 2025) of modular AI marketplaces. Our framework provides the necessary statistical scaffolding for such ecosystems. By allowing fully decentralized training via $\mathcal{G}_0$ and eliminating the need to synchronize data across expert boundaries, our approach supports privacy-preserving, plug-and-play modularity where domain owners can independently train experts and dynamically combine them with provable robustness guarantees.

B Expert-Independent Bounded Gating ($\mathcal{G}_{\infty,1}$)

We first explore the $\mathcal{G}_{\infty,1}$ family, which provides a robust middle ground between the strict global normalization of $\mathcal{G}_1$ and the purely local constraints of $\mathcal{G}_0$.

B.1 Definition and Properties

We provide the proofs for the properties of the bounded family $\mathcal{G}_{\infty,1}$ defined in Section 3.

Lemma 2 (Properties of $\mathcal{G}_{\infty,1}$)**.** *The family $\mathcal{G}_{\infty,1}$ is non-empty, compact, and convex. Furthermore, for any $g \in \mathcal{G}_{\infty,1}$, the total mass of the resulting mixture model π_g is bounded by unity:*

$$Z_g = \sum_{x \in \mathcal{X}_0} \pi_g(x) \leq 1. \tag{4}$$

Proof. The set is defined by the constraint $\|g\|_{\infty,1} \leq 1$. Since the maximum function is convex and the sum of convex functions is convex, the constraint defines a convex sublevel set. Compactness follows from the set being closed and bounded in finite dimensions. To prove the mass bound, we observe that for any $g \in \mathcal{G}_{\infty,1}$:

$$Z_g = \sum_{x,k} g(x,k)\widehat{\pi}_k(x) = \sum_k \sum_x g(x,k)\widehat{\pi}_k(x) \leq \sum_k \max_{x'} g(x',k) \sum_x \widehat{\pi}_k(x) = \sum_k \max_{x'} g(x',k) \leq 1. \tag{9}$$

The last inequality follows from the definition of the $\mathcal{G}_{\infty,1}$ family and the fact that each expert $\widehat{\pi}_k$ is a valid probability distribution. □

This expert-independent constraint essentially limits the total capacity allocated to each expert across the entire input space $\mathcal{X}_0$. Unlike the $\mathcal{G}_1$ family, which requires experts to sum to unity at every individual point x, the $\mathcal{G}_{\infty,1}$ family allows for structural sparsity. An expert may be weighted for a specific subset of sequences while remaining inactive for others, provided the sum of all weights across the ensemble does not exceed one. This property makes the $\mathcal{G}_{\infty,1}$ family robust to expert misalignment or the presence of "dead" experts that do not contribute to specific domains.

B.2 Fixed Mixture Solutions

We now provide the full proof of Proposition 3.

Proposition 3 (Fixed Mixture Guarantee for $\mathcal{G}_{\infty,1}$)**.** *For any fixed $\lambda \in \Delta$, $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) \le \sum_{k=1}^p \lambda_k \epsilon_k \le \max_k \epsilon_k$.*

Proof. Existence of g^* follows from the Weierstrass theorem on the compact convex set $\mathcal{G}_{\infty,1}$. The bound is obtained by evaluating at the constant gate $g_\lambda(x,k) = \lambda_k \in \mathcal{G}_{\infty,1}$, for which $Z_{g_\lambda} = 1$ and $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ reduces to D_{KL}. The joint convexity of D_{KL} then gives $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g_\lambda}) \le \sum_k \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k) = \sum_k \lambda_k \epsilon_k$, exactly as in the corresponding result for $\mathcal{G}_1$ (Cortes et al., 2026, Prop. 1). □

B.3 Optimal $\mathcal{G}_{\infty,1}$ solution for a Fixed-Mixture

Lemma 13 (Structure of the Optimal $\mathcal{G}_{\infty,1}$ Fixed-Mixture Solution)**.** *Let $\widehat{\mathsf{p}}_\lambda = \sum_k \lambda_k \widehat{\mathsf{p}}_k$ be a fixed data mixture. An optimal gating function $g^* \in \operatorname{argmin}_{g \in \mathcal{G}_{\infty,1}} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g)$ can be taken to be a scaled version of a capacity vector $w^* \in \Delta([1,p])$. Specifically:*

$$g^*(x,k) = w_k^* \cdot \rho^*(x), \quad \textit{where} \quad \rho^*(x) = \min\left(1, \frac{\widehat{\mathsf{p}}_\lambda(x)}{\sum_{j=1}^p w_j^* \widehat{\pi}_j(x)}\right), \tag{10}$$

with the convention $\rho^*(x) = 0$ when $\sum_j w_j^* \widehat{\pi}_j(x) = 0$. The vector w^* represents the optimal allocation of the unit capacity bound $\|g\|_{\infty,1} \le 1$ across experts and is a minimizer of the capacity-constrained relative entropy*:*

$$w^* \in \operatorname*{argmin}_{w \in \Delta} \left\{ \sum_{x \in \mathcal{X}_0} \left[\widehat{\mathsf{p}}_\lambda(x) \log \frac{\widehat{\mathsf{p}}_\lambda(x)}{\min\{\widehat{\mathsf{p}}_\lambda(x),\, e_w(x)\}} + \min\{\widehat{\mathsf{p}}_\lambda(x),\, e_w(x)\} \right] \right\}, \tag{11}$$

where $e_w(x) = \sum_k w_k \widehat{\pi}_k(x)$ is the capacity envelope *induced by w. The objective in Equation 11 equals $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g_w^*}) + 1$ for the gate $g_w^*(x,k) = w_k \min\{1, \widehat{\mathsf{p}}_\lambda(x)/e_w(x)\}$, and only the sub-envelope points $\{x : e_w(x) < \widehat{\mathsf{p}}_\lambda(x)\}$ contribute non-trivially to it.*

Proof. By the definition of $\mathcal{G}_{\infty,1}$, we have $\sum_k \max_x g(x,k) \le 1$. Let $w_k = \max_x g(x,k)$. The constraint becomes $\sum_k w_k \le 1$ and $0 \le g(x,k) \le w_k$ for all x. Writing $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \sum_x \left[\widehat{\mathsf{p}}_\lambda(x) \log \frac{\widehat{\mathsf{p}}_\lambda(x)}{\pi_g(x)} + \pi_g(x) \right] - 1$, the objective decomposes into a sum of pointwise terms $\phi_x(\pi) = \widehat{\mathsf{p}}_\lambda(x) \log \frac{\widehat{\mathsf{p}}_\lambda(x)}{\pi} + \pi$, each of which is convex in π and minimized at $\pi = \widehat{\mathsf{p}}_\lambda(x)$.

For any fixed capacity w, the pointwise mixture $\pi_g(x) = \sum_k g(x,k) \widehat{\pi}_k(x)$ ranges over $[0, e_w(x)]$, where $e_w(x) = \sum_k w_k \widehat{\pi}_k(x)$ is the capacity envelope. Hence the optimal pointwise mixture is

$$\pi_{g^*}(x) = \min(\widehat{\mathsf{p}}_\lambda(x), e_w(x)),$$

which is achieved by the gating function $g^*(x,k) = w_k \cdot \rho^*(x)$, where $\rho^*(x)$ is the ratio of the target density to the available ensemble capacity, capped at unity. Substituting back yields the reduced objective of Equation 11, up to the additive constant -1.

Each summand in Equation 11 is non-increasing in $e_w(x)$ (its derivative is $1 - \widehat{\mathsf{p}}_\lambda(x)/e_w(x) \le 0$ on the sub-envelope region and 0 elsewhere), so the capacity constraint $\sum_k w_k \le 1$ is active and $w^* \in \Delta$. Since each summand is a convex function of $e_w(x)$, which is linear in w, the reduced objective is convex on the compact set Δ, so a minimizer w^* exists (it is unique whenever the reduced objective is strictly convex in w). □

Remark. Only the sub-envelope region $\{x : e_w(x) < \widehat{\mathsf{p}}_\lambda(x)\}$ contributes non-trivially to the reduced objective in Equation 11; on the remaining points the summand equals the constant $\widehat{\mathsf{p}}_\lambda(x)$. In the specialization limit (disjoint supports), the envelope reduces to $e_w(x) = w_{k(x)} \widehat{\pi}_{k(x)}(x)$, where $k(x)$ is the unique expert whose support contains x, so the reduced objective separates into a sum of per-expert terms, each depending on w only through the single scalar $w_{k(x)}$. On the sub-envelope region the summand further simplifies to $\widehat{\mathsf{p}}_\lambda(x) \log \frac{1}{e_w(x)}$ up to an additive constant.

Remark (Capacity saturation). For any $g \in \mathcal{G}_{\infty,1}$ with $M_g = \sum_k \max_x g(x,k) < 1$ and any $c \in (1, 1/M_g]$, the scaled gate cg lies in $\mathcal{G}_{\infty,1}$ and, since $\widehat{\mathsf{p}}_k$ has unit mass, $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_{cg}) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) - \log c + (c-1)Z_g$ for every k. As $Z_g \le M_g < 1$, the right-hand side is strictly smaller than $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$ for c slightly above 1. Hence every finite-risk minimizer of the unrestricted robust game saturates the capacity, $M_{g^*} = 1$, whereas its mass Z_{g^*} may remain below 1: sub-normalization is a property of the mass, not of the envelope.

Proposition 14 (Optimal $\mathcal{G}_{\infty,1}$ Capacity)**.** *In the specialization limit where expert supports are disjoint, the robust minimax game for the bounded family $\mathcal{G}_{\infty,1}$ restricted to* input-independent *gates $g(x,k) = w_k$ with $w \in \mathbb{R}^p_+$, $\sum_k w_k \le 1$, has the exact value $V_\infty^{\mathrm{const}} = \log \sum_{k=1}^p e^{\epsilon_k}$, attained by the capacity vector*

$$w_k^* = \frac{\exp(\epsilon_k)}{\sum_{j=1}^p \exp(\epsilon_j)}.$$

Consequently, since input-independent gates form a subset of $\mathcal{G}_{\infty,1}$, the minimax value of the full game satisfies $V_\infty = \min_{g \in \mathcal{G}_{\infty,1}} \max_{\lambda \in \Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_g) \le \log \sum_{k=1}^p e^{\epsilon_k}$.

The proof establishes the saddle point via the *linearized payoff* $\widetilde{L}(\lambda, w) = \sum_k \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \| \pi_g)$, which satisfies $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_g) = \widetilde{L}(\lambda, w) - \mathsf{D}_{\mathrm{JSD}}^\lambda$ (Lemma 4). In the specialization limit (disjoint supports), $\mathsf{D}_{\mathrm{JSD}}^\lambda = H(\lambda)$. Crucially, $\widetilde{L}$ is *linear* in λ and convex in w, and finite on the compact set $\{w : w_k \ge w_{\min},\ \sum_k w_k \le 1\}$ for any $w_{\min} > 0$, so Sion's Minimax Theorem guarantees strong duality for the surrogate game on that set; since the closed-form solution below has every $w_k^* > 0$, the restriction is immaterial. The true game objective $L(\lambda, w) = \widetilde{L} - H(\lambda)$ contains the negative entropy $-H(\lambda) = \sum_k \lambda_k \log \lambda_k$, which is strictly convex in λ; thus, the *minimax* of the true game equals the minimax of the linearized game (the inner maximum over λ is attained at a vertex e_{k^*}, where $H(e_{k^*}) = 0$). Thus the value $V_\infty^{\mathrm{const}} = \log \sum e^{\epsilon_k}$ is exact for both games over input-independent gates. Input-dependent gates $g(x,k) \le w_k$ can only lower the minimax value, which is why the full-family value V_∞ is bounded above, but not in general equal to, $\log \sum_k e^{\epsilon_k}$. This upper bound is precisely what is used in the robust guarantee of Theorem 5.

Proof. Consider the regime where the supports of the target task distributions $\widehat{\mathsf{p}}_k$ and the expert models $\widehat{\pi}_k$ are mutually disjoint across tasks. For any mixture $\lambda \in \Delta$ and input-independent gate $g(x,k) = w_k$, the mixture density $\pi_g(x)$ equals $w_k \widehat{\pi}_k(x)$ for all $x \in \mathrm{supp}(\widehat{\mathsf{p}}_k)$.

The unnormalized relative entropy objective $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \mathbb{E}_{x \sim \widehat{\mathsf{p}}_\lambda}\left[\log \frac{\widehat{\mathsf{p}}_\lambda(x)}{\pi_g(x)}\right] + (Z_g - 1)$ decomposes as follows:

$$\begin{aligned} L(\lambda, w) &= \sum_{k=1}^p \lambda_k \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_k}\left[\log\left(\frac{\lambda_k \widehat{\mathsf{p}}_k(x)}{w_k \widehat{\pi}_k(x)}\right)\right] + \sum_{k=1}^p w_k - 1 \\ &= \sum_{k=1}^p \lambda_k \left[\log\left(\frac{\lambda_k}{w_k}\right) + \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k)\right] + \sum_{k=1}^p w_k - 1. \end{aligned}$$

Since $\epsilon_k = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k)$ by definition, and using $\log(\lambda_k / w_k) = \log \lambda_k - \log w_k$, the game value is:

$$V_\infty^{\mathrm{const}} = \min_{w \in \mathbb{R}_+^p} \max_{\lambda \in \Delta} \left[\sum_{k=1}^p \lambda_k (\epsilon_k - \log w_k + \log \lambda_k) + \sum_{k=1}^p w_k - 1\right].$$

As noted in the preamble, the inner objective contains the negative entropy $-H(\lambda) = \sum_k \lambda_k \log \lambda_k$, which is strictly convex. Hence, the maximum over $\lambda \in \Delta$ is attained at a vertex $\lambda = e_{k^*}$ (where $H(e_{k^*}) = 0$), yielding $\max_k (\epsilon_k - \log w_k)$. This reduces the minimax to:

$$V_\infty^{\mathrm{const}} = \min_{w \in \mathbb{R}_+^p} \left[\max_{k \in [p]} (\epsilon_k - \log w_k) + \sum_{k=1}^p w_k - 1\right].$$

Setting $t = \max_k (\epsilon_k - \log w_k)$, we first observe that if any $w_k \to 0$, then $\epsilon_k - \log w_k \to +\infty$, so $t \to +\infty$. Since we are minimizing over w and any uniform vector $w_k = 1/p > 0$ yields a finite objective, the global minimum must occur in the interior with $w_k > 0$ for all k. At the optimum, all penalties must be exactly equalized: $\epsilon_k - \log w_k = t$ for all k (if there were some k where $\epsilon_k - \log w_k < t$, we could slightly decrease the strictly positive w_k, strictly lowering $\sum_j w_j$ without increasing t,

which contradicts optimality). This equalization condition implies $w_k = e^{\epsilon_k - t}$. Substituting into the objective yields the univariate function:

$$f(t) = t + e^{-t} \sum_{j=1}^{p} e^{\epsilon_j} - 1.$$

Setting $f'(t) = 1 - e^{-t} \sum_j e^{\epsilon_j} = 0$ gives $e^{t^*} = \sum_j e^{\epsilon_j}$, i.e., $t^* = \log\left(\sum_{j=1}^{p} e^{\epsilon_j}\right)$. Moreover, $f''(t) > 0$, so t^* is the unique global minimum. Back-substituting:

$$w_k^* = \frac{e^{\epsilon_k}}{e^{t^*}} = \frac{\exp(\epsilon_k)}{\sum_{j=1}^{p} \exp(\epsilon_j)}, \qquad \sum_{k=1}^{p} w_k^* = 1, \qquad V_\infty^{\text{const}} = \log\left(\sum_{k=1}^{p} e^{\epsilon_k}\right).$$

The optimal capacity vector is the softmax of the expert errors, and the minimax value V_∞^{const} is the log-sum-exp of the individual errors $\epsilon_1, \ldots, \epsilon_p$. □

B.4 Robust Solutions

This section provides the complete proofs for the robust existence guarantees presented in Section 3.1.

Lemma 4 (Unnormalized KL of a Mixture)**.** *For any mixture weight $\lambda = (\lambda_1, \ldots, \lambda_p) \in \Delta([1, p])$, the following identity holds for the unnormalized KL divergence:*

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \sum_{k=1}^{p} \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) - \mathsf{D}_{\mathrm{JSD}}^{\lambda}(\{\widehat{\mathsf{p}}_k\}).$$

Proof. The proof parallels the D_{KL} mixture identity in (Cortes et al., 2026, Lemma 2). Expanding both sides using $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\mathsf{p} \parallel q) = \mathsf{D}_{\mathrm{KL}}(\mathsf{p} \parallel q) + (Z_q - 1)$, the mass terms $(Z_g - 1)$ cancel by linearity of λ. The residual is exactly $-\sum_k \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\mathsf{p}}_\lambda) = -\mathsf{D}_{\mathrm{JSD}}^{\lambda}(\{\widehat{\mathsf{p}}_k\})$. □

Theorem 5 (Robust Existence and Guarantees for $\mathcal{G}_{\infty,1}$)**.** *The robust minimax game for the $\mathcal{G}_{\infty,1}$ family admits a solution $g^* \in \mathcal{G}_{\infty,1}$. For any mixture $\lambda \in \Delta$, the resulting model satisfies:*

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) \le \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \mathsf{D}_{\mathrm{JSD}}^{\lambda}(\{\widehat{\mathsf{p}}_k\}). \tag{5}$$

Proof. The proof follows the linearization strategy of Cortes et al. (2026, Thm. 1). Define the linearized payoff $\widetilde{L}(\lambda, g) = \sum_k \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$, which satisfies $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \widetilde{L}(\lambda, g) - \mathsf{D}_{\mathrm{JSD}}^{\lambda}$ by Lemma 4. Since $\widetilde{L}$ is linear in λ, its maximum over Δ is attained at a vertex, so the worst-case risk of a gate is $F(g) = \max_{\lambda \in \Delta} \widetilde{L}(\lambda, g) = \max_{k \in [1,p]} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$, which may take the value $+\infty$ (for instance at the all-zero gate). Each map $g \mapsto \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$ is convex and lower semicontinuous on $\mathcal{G}_{\infty,1}$ with values in $(-\infty, +\infty]$ (it is a finite sum of terms $-\widehat{\mathsf{p}}_k(x) \log \pi_g(x) + \pi_g(x)$, each of which is convex and lower semicontinuous in g), hence so is F. Since $\mathcal{G}_{\infty,1}$ is non-empty and compact, F attains its minimum at some $g^* \in \mathcal{G}_{\infty,1}$, which is the robust solution. (When the game is restricted to the non-degenerate family $\mathcal{G}_{\infty,1}^{\eta}$ of Appendix B.11, the payoff is finite and since $\mathcal{G}_{\infty,1}^{\eta}$ is compact and convex, Sion's theorem additionally yields a saddle point (λ^*, g^*).)

The bound follows by evaluating F at the softmax witness $g(x, k) = \sigma_k = e^{\epsilon_k} / (\sum_j e^{\epsilon_j})$, which is feasible in $\mathcal{G}_{\infty,1}$ since $\sum_k \sigma_k = 1$. The identical computation as in Cortes et al. (2026, Thm. 1) yields $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) \le \epsilon_k - \log \sigma_k = \log[\sum_j e^{\epsilon_j}]$ for every k, whence $F(g^*) \le F(g) \le \log[\sum_k e^{\epsilon_k}]$. Finally, for any $\lambda \in \Delta$, Lemma 4 gives $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) = \widetilde{L}(\lambda, g^*) - \mathsf{D}_{\mathrm{JSD}}^{\lambda} \le F(g^*) - \mathsf{D}_{\mathrm{JSD}}^{\lambda} \le \log[\sum_k e^{\epsilon_k}] - \mathsf{D}_{\mathrm{JSD}}^{\lambda}$, completing the proof. □

Unlike the $\mathcal{G}_1$ family, which incurs a logarithmic capacity cost tied to explicitly constrained global normalization, the bounded relaxation exploits the unnormalized geometry to explicitly leverage task separability to ensure that the ensemble matches the fidelity of the best constituent expert regardless of task diversity or the number of experts p.

Theoretical Implications and Comparison with the $\mathcal{G}_1$ Family. To clarify the theoretical power of the $\mathcal{G}_{\infty,1}$ relaxation, we analyze what Theorem 5 implies for standard probability distributions and how it compares to the results established for the $\mathcal{G}_1$ family.

Implied KL Bound and the Normalization Bonus. The robust existence result provides an upper bound on the unnormalized relative entropy $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*})$. In generative modeling, however, the primary interest often lies in the fidelity of the normalized density $\mathsf{p}_{g^*} = \pi_{g^*}/Z_{g^*}$. By recalling the fundamental decomposition:

$$\widetilde{\mathsf{D}}_{\mathrm{KL}}(\mathsf{p} \parallel q) = \mathsf{D}_{\mathrm{KL}}(\mathsf{p} \parallel q/Z_q) + \Delta(Z_q), \tag{12}$$

where $\Delta(Z) = Z - 1 - \log Z$ is the non-negative *normalization penalty*, we can derive a rigorous guarantee for the rescaled model. Since $Z_{g^*} \le 1$ for any optimal gate under the capacity constraint, the penalty $\Delta(Z_{g^*})$ is always non-negative. This yields the following chain of inequalities:

$$\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_{g^*}) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) - \Delta(Z_{g^*}) \le \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\widehat{\mathsf{p}}_k) - \Delta(Z_{g^*}). \tag{13}$$

Since $\Delta(Z_{g^*})$ is positive for sub-normalized models ($Z_{g^*} < 1$), the provable D_{KL} bound of the normalized model is *tighter* than the classical capacity cost $\log[\sum e^{\epsilon_k}]$ suggests. We refer to the term $-\Delta(Z_{g^*})$ as the *Normalization Bonus*.

Relating the Upper and Lower Bounds: The Sparsity Mechanism. To formally establish that this improvement is a fundamental structural advantage rather than an artifact of a loose upper bound, we prove a minimax lower bound for the $\mathcal{G}_{\infty,1}$ family in Appendix B.5. The lower bound demonstrates that in the specialization limit, the normalized risk for any valid gate $g \in \mathcal{G}_{\infty,1}$ is bounded below by $\log[\sum_k e^{\epsilon_k}] - \log(M/Z_g)$, where $M = \sum_k \max_x g(x,k) \le 1$ is the total envelope mass. This shows a theoretical relationship:

- **The Lower Bound (The Necessity of Sub-normalization):** The lower bound proves that a dense, normalized system (where $Z_g = M = 1$) is locked into paying the full capacity cost of $\log[\sum e^{\epsilon_k}]$. To pierce this fundamental floor and achieve a tighter bound, a model is required to maintain a Sparsity Ratio $M/Z_g > 1$. (Note: a gate $h \in \mathcal{G}_1$ can have $M_h = \sum_k \max_x h(x,k) > 1 = Z_h$, e.g., a hard expert switch, so the floor applies specifically to dense normalized constructions, not to all $\mathcal{G}_1$ gates.)
- **The Upper Bound (The Realization of the Bonus):** Our upper bound confirms that optimizing the $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective realizes this structural potential. By allowing the mixture mass to dynamically float below its maximum envelope bounds, the optimal gate g^* actively filters out expert errors. The model harvests this sparsity as a rigorous Normalization Bonus ($-\Delta(Z_{g^*})$), pushing the provable bound strictly below the capacity cost, and below the known globally normalized guarantee whenever $\Delta(Z_{g^*})$ exceeds the Overlap Gain H_σ of that guarantee, in particular in specialized regimes where H_σ vanishes.

Together, these bounds demonstrate that $\mathcal{G}_{\infty,1}$ is not merely a computational convenience for asynchronous training; it admits tighter provable bounds by leveraging the structural flexibility of sub-normalization and sparsity. This comparison concerns the bounds. On the scale of the achievable normalized risk, $\mathcal{G}_{\infty,1}$ is at least as expressive as $\mathcal{G}_1$: any $h \in \mathcal{G}_1$ rescales to $h/\sum_k \max_x h(x,k) \in \mathcal{G}_{\infty,1}$ with the same normalized density (the rescaling is well defined since $\sum_k \max_x h(x,k) \ge Z_h = 1$), so the set of normalized densities induced by $\mathcal{G}_{\infty,1}$ contains the set induced by $\mathcal{G}_1$, and the minimax normalized risk of $\mathcal{G}_{\infty,1}$ is never worse than that of $\mathcal{G}_1$. The converse inclusion fails in general, since the rescaling g/Z_g of a gate $g \in \mathcal{G}_{\infty,1}$ need not lie on the pointwise simplex. The further advantages of $\mathcal{G}_{\infty,1}$ are the tighter provable guarantee in specialized regimes, the complexity ordering of Theorem 8, the efficient exact sampler, and decoupled optimization.

B.5 Fundamental Lower Bound for $\mathcal{G}_{\infty,1}$

Theorem 15 (Minimax Lower Bound for $\mathcal{G}_{\infty,1}$)**.** *In the specialization limit (disjoint expert supports), for any gating function $g \in \mathcal{G}_{\infty,1}$ with capacity vector $w_k = \max_x g(x,k)$ and total capacity envelope $M = \sum_k w_k \le 1$, the worst-case unnormalized relative entropy plus diversity is bounded below by:*

$$\max_{\lambda \in \Delta}\left[\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) + \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\})\right] \ge \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \log M + Z_g - 1, \tag{14}$$

Furthermore, whenever $Z_g > 0$, the implied standard D_{KL} divergence of the normalized model $\mathsf{p}_g = \pi_g / Z_g$ satisfies the lower bound:

$$\max_{\lambda \in \Delta} \left[\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_g) + \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\}) \right] \geq \log \left[\sum_{k=1}^{p} e^{\epsilon_k} \right] - \log \left(\frac{M}{Z_g} \right). \tag{15}$$

Proof. Assume the pre-trained experts $\{\widehat{\pi}_k\}$ target mutually disjoint supports. Let $g \in \mathcal{G}_{\infty,1}$ be an arbitrary capacity-constrained gate, with allocated capacity $w_k = \max_x g(x,k)$. By definition, the envelope mass is $M = \sum_{k=1}^{p} w_k \leq 1$.

For any input $x \in \mathrm{supp}(\widehat{\mathsf{p}}_k)$, the mixture density simplifies to $\pi_g(x) = g(x,k)\widehat{\pi}_k(x) \leq w_k \widehat{\pi}_k(x)$. Using the standard decomposition (Lemma 4), the unnormalized relative entropy for a specific task k evaluates to:

$$\begin{aligned}
\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) &= \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g) + Z_g - 1 \\
&\geq \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_k} \left[\log \frac{\widehat{\mathsf{p}}_k(x)}{w_k \widehat{\pi}_k(x)} \right] + Z_g - 1 \\
&= \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k) - \log w_k + Z_g - 1 \\
&= \epsilon_k - \log w_k + Z_g - 1.
\end{aligned}$$

To evaluate the robust minimax risk, we consider the adversary's maximization over the mixture $\lambda \in \Delta$. Under disjoint supports, the Jensen-Shannon Divergence is exactly the Shannon entropy, $\mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\}) = -\sum_{k=1}^{p} \lambda_k \log \lambda_k$. Combining the per-task lower bound above with Lemma 4, the worst-case unnormalized relative entropy plus diversity is bounded below by a maximum of a sum over tasks:

$$\max_{\lambda \in \Delta} \left[\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) + \mathsf{D}^{\lambda}_{\mathrm{JSD}} \right] \geq \max_{\lambda \in \Delta} \left[\sum_{k=1}^{p} \lambda_k (\epsilon_k - \log w_k) \right] + Z_g - 1. \tag{16}$$

Since the right-hand side maximizes a linear function over the simplex Δ, the maximum is attained at the largest individual component, so the worst-case objective is at least

$$\max_{k \in [1,p]} (\epsilon_k - \log w_k) + Z_g - 1. \tag{17}$$

We now rigorously lower-bound this maximum. Let $t = \max_{k \in [1,p]} (\epsilon_k - \log w_k)$. By definition of the maximum, it holds that for every expert k:

$$t \geq \epsilon_k - \log w_k \Rightarrow \log w_k \geq \epsilon_k - t \Rightarrow w_k \geq e^{\epsilon_k} e^{-t}. \tag{18}$$

Summing both sides of this inequality over all experts k, and recalling that the total envelope mass is fixed at $\sum_{k=1}^{p} w_k = M$:

$$M \geq \sum_{k=1}^{p} e^{\epsilon_k} e^{-t} = e^{-t} \sum_{k=1}^{p} e^{\epsilon_k}. \tag{19}$$

Rearranging to solve for t yields the absolute lower bound on the maximum component:

$$e^t \geq \frac{1}{M} \sum_{k=1}^{p} e^{\epsilon_k} \Rightarrow t \geq \log \left[\sum_{k=1}^{p} e^{\epsilon_k} \right] - \log M. \tag{20}$$

Substituting this back into the worst-case risk equation yields the unnormalized lower bound:

$$\max_{\lambda \in \Delta} \left[\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) + \mathsf{D}^{\lambda}_{\mathrm{JSD}} \right] \geq \log \left[\sum_{k=1}^{p} e^{\epsilon_k} \right] - \log M + Z_g - 1. \tag{21}$$

Finally, applying the fundamental identity $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g) = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_g) + \Delta(Z_g)$ where $\Delta(Z_g) = Z_g - 1 - \log Z_g$, we can rearrange the bound to evaluate the implied normalized probability distribution

p_g:

$$\begin{aligned}\max_{\lambda\in\Delta}\left[\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_g) + \mathsf{D}^{\lambda}_{\mathrm{JSD}}\right] &\geq \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \log M + Z_g - 1 - (Z_g - 1 - \log Z_g) \\ &= \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \log M + \log Z_g \\ &= \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \log\left(\frac{M}{Z_g}\right).\end{aligned}$$

By the definition of Z_g and w_k, we have $Z_g = \sum_{k=1}^{p}\sum_{x\in\mathcal{X}_0} g(x,k)\widehat{\pi}_k(x) \leq \sum_{k=1}^{p} w_k \sum_{x\in\mathcal{X}_0}\widehat{\pi}_k(x) = M$. Thus, the ratio $M/Z_g \geq 1$ and the term $-\log(M/Z_g) \leq 0$ serves as a rigorous *sparsity bonus*. This confirms that by dynamically using a sub-normalized mass ($Z_g \leq M \leq 1$), the $\mathcal{G}_{\infty,1}$ architecture achieves a theoretically advantageous robust anchor compared to the global baseline $\log[\sum_k e^{\epsilon_k}]$. □

B.6 Geometric Interpretation of the $\mathcal{G}_{\infty,1}$ Bound

The upper bound $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) \leq \log\left[\sum_{k=1}^{p} e^{\epsilon_k}\right] - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\{\widehat{\mathsf{p}}_k\})$ established in Theorem 5 reveals how the robust gate leverages task geometry while circumventing global synchronization. To interpret these performance guarantees on a standard probabilistic scale, we evaluate the implied D_{KL} divergence of the normalized distribution $\mathsf{p}_{g^*} = \pi_{g^*}/Z_{g^*}$. Recall the fundamental decomposition for unnormalized models:

$$\mathsf{D}_{\mathrm{KL}}\left(\widehat{\mathsf{p}}_\lambda \parallel \frac{\pi_{g^*}}{Z_{g^*}}\right) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{g^*}) - \Delta(Z_{g^*}), \tag{22}$$

where $\Delta(Z) = Z - 1 - \log Z$ is the non-negative *normalization penalty*. Since $Z_g \leq 1$ for all $g \in \mathcal{G}_{\infty,1}$, we refer to the term $-\Delta(Z_{g^*})$ as the *normalization bonus*. This leads to the following analysis across limiting geometric regimes:

The geometric behavior of the $\mathcal{G}_{\infty,1}$ bound parallels the three limiting regimes identified for $\mathcal{G}_1$ by Cortes et al. (2026, §4.1), with a key additional element, the Normalization Bonus, that arises from sub-normalization. Table 4 summarizes how each quantity adapts, and Figure 5(a) illustrates the switch behavior in the specialization limit.

Table 4: Geometric regimes: $\mathcal{G}_1$ (MOD) vs. $\mathcal{G}_{\infty,1}$ (this work).

	Specialization (Disjoint supports)	**Redundancy** (Identical experts, error ϵ)	**General** (Arbitrary)
$\mathsf{D}^{\lambda}_{\mathrm{JSD}}$	$H(\lambda)$	0	$\in [0, H(\lambda)]$
$\mathcal{G}_1$ bound	$\log\sum e^{\epsilon_k} - H(\lambda) - H_\sigma$	ϵ	$\log\sum e^{\epsilon_k} - \mathsf{D}_{\mathrm{JSD}} - H_\sigma$
$\mathcal{G}_{\infty,1}$ bound	$\log\sum e^{\epsilon_k} - H(\lambda)$	$\epsilon + \log p$	$\log\sum e^{\epsilon_k} - \mathsf{D}_{\mathrm{JSD}}$
Implied KL	$\log\sum e^{\epsilon_k} - H(\lambda) - \Delta(Z_{g^*})$	$\epsilon + \log p - \Delta(Z_{g^*})$	$\log\sum e^{\epsilon_k} - \mathsf{D}_{\mathrm{JSD}} - \Delta(Z_{g^*})$

In the Specialization limit, the $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_1$ upper bounds coincide. In the Redundancy limit (identical source distributions $\widehat{\mathsf{p}}_k = \widehat{\mathsf{p}}$, so that $\mathsf{D}^{\lambda}_{\mathrm{JSD}} = 0$, and identical experts $\widehat{\pi}_k = \widehat{\pi}$ with common error ϵ), the $\mathcal{G}_1$ bound of Cortes et al. (2026) evaluates to ϵ because its Overlap Gain H_σ equals $\log p$ for the uniform softmax weights, whereas Theorem 5 yields $\log\sum_k e^{\epsilon} = \epsilon + \log p$. This $\log p$ gap is a looseness of the upper bound rather than of the family: the single-expert gate $g(x,k) = \mathbb{1}_{k=k_0}$ lies in $\mathcal{G}_{\infty,1}$ with $Z_g = 1$ and attains $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_g) = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \widehat{\pi}) = \epsilon$ for every λ, so the robust $\mathcal{G}_{\infty,1}$ gate also achieves worst-case risk at most ϵ in this limit. The distinction is the additional *Normalization Bonus* $-\Delta(Z_{g^*})$, which applies whenever the $\mathcal{G}_{\infty,1}$ gate exploits sub-normalization ($Z_{g^*} < 1$). Since $\Delta(Z) = Z - 1 - \log Z \geq 0$, this bonus strictly tightens the provable D_{KL} bound relative to the $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ bound itself. Relative to the known $\mathcal{G}_1$ guarantee, whose Overlap Gain H_σ plays the analogous role (Table 4), the implied bound is tighter exactly when $\Delta(Z_{g^*}) > H_\sigma$: this holds in specialized regimes, where H_σ vanishes, but not in the redundancy limit, where $H_\sigma = \log p$ and the $\mathcal{G}_1$ guarantee is the tighter one. The $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective also provides implicit mass regularization, penalizing extreme mass deviations while allowing the gate to "breathe" without requiring global synchronization.

B.7 Comparison with Monolithic Baselines

The upper bound derived in Theorem 5 highlights that the $\mathcal{G}_{\infty,1}$ relaxation maintains the guarantees of the foundational modular models and tightens them in specialized regimes. We analyze these benefits by contrasting the interaction between task geometry and architectural priors.

The Monolithic Floor. A monolithic model $\widehat{\pi}_\lambda$ trained on the mixture $\widehat{\mathsf{p}}_\lambda$ is constrained by the capacity of its class. As established in Theorem 6 of the foundational framework, the risk of an optimal retrained model is lower-bounded by the average best-in-class error discounted by the Jensen-Shannon Divergence:

$$\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \widehat{\pi}_\lambda) \geq \sum_{k=1}^{p} \lambda_k \epsilon_k - \mathsf{D}^{\lambda}_{\mathrm{JSD}}(\widehat{\mathsf{p}}_k). \tag{23}$$

Here, as in Cortes et al. (2026, Thm. D.1), ϵ_k denotes the best-in-class error $\min_{\pi\in\Pi} \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\pi)$ of the monolithic class Π. The bound follows from the identity $\sum_k \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\pi) = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda\|\pi) + \mathsf{D}^{\lambda}_{\mathrm{JSD}}$, valid for any model π, by lower-bounding each $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_\lambda)$ by the best-in-class error. It therefore holds with the frozen-expert errors $\epsilon_k = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_k)$ exactly when the retrained model is no better than each specialist on its own domain, $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_\lambda) \geq \epsilon_k$ for all k, which is the regime in which the modular comparison is of interest (if the monolithic model dominated every specialist, the experts would not be worth reusing). The $\mathsf{D}_{\mathrm{JSD}}$ term is a credit rather than a penalty: by the identity above, every mixture-level risk $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda\|\pi)$ equals the average per-task risk $\sum_k \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\pi)$ minus $\mathsf{D}^{\lambda}_{\mathrm{JSD}}$, for the monolithic and the modular model alike. In particular, an unrestricted monolithic class attains $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda\|\widehat{\mathsf{p}}_\lambda) = 0$ for any task geometry, and the floor is informative only through the capacity restriction encoded in the best-in-class errors ϵ_k. What we call interference is the gap between a monolithic model's per-task risks $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_\lambda)$ and its class's best-in-class errors; Theorem 16 does not quantify it, and the experiments of Section 5 measure it.

The Modular Ceiling: Diversity as Separability. The robust $\mathcal{G}_{\infty,1}$ gate exploits task diversity as a *Separability Gain*: the negative $\mathsf{D}_{\mathrm{JSD}}$ term in Theorem 5 cancels the capacity cost $\log \sum_k e^{\epsilon_k}$ in specialized regimes (the Specialization Limit of Appendix B.6), so the ceiling reaches the specialists' own errors. This is a statement about the modular ceiling, not a comparative advantage over the monolithic floor, which receives the same credit; the comparison between the two architectures rests on the per-task errors of the monolithic class against those of the specialists.

Statistical Advantage via Data-Independent Bounds. While both $\mathcal{G}_1$ and $\mathcal{G}_{\infty,1}$ benefit from this geometric decoupling, the unnormalized relaxation provides tighter provable statistical guarantees. The $\mathcal{G}_1$ global normalization constraint is a single scalar equation on the data-dependent expert predictions ($\widehat{\pi}_k$), which leaves the gate free to fit any expert assignment on a sample by compensating elsewhere. Conversely, the $\mathcal{G}_{\infty,1}$ constraint ($\sum_k \max_x g(x,k) \leq 1$) is a purely data-independent structural ceiling that caps the total weight each expert can receive across all inputs. As we will prove formally in Theorem 8, under a mild expert-separation condition this ceiling restricts the gating Rademacher complexity such that $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ for $p \geq 3$. Thus, $\mathcal{G}_{\infty,1}$ holds a distinct statistical advantage: it matches the $\mathcal{G}_1$ minimax robustness anchor with a generalization term that is no larger, all while bypassing the global synchronization bottleneck.

The contrast between the modular and monolithic settings is symmetric: the Jensen-Shannon Divergence $\mathsf{D}_{\mathrm{JSD}}$ appears with identical coefficient -1 in both the monolithic floor and the modular separability ceiling, with the same consequence: a discount of the average per-task risk by the mixing credit. We formalize this as the *Symmetric Divergence Effect* (Theorem 16, Appendix B.8).

B.8 Symmetric Divergence Effect

The Symmetric Divergence Effect was established by Cortes et al. (2026, Thm. 6, §4.3) for $\mathcal{G}_1$: the $\mathsf{D}_{\mathrm{JSD}}$ term appears with coefficient -1 in both the monolithic risk floor and the modular risk ceiling, lowering both by the same mixing credit.

The extension to $\mathcal{G}_{\infty,1}$ is immediate.

Theorem 16 (Symmetric Divergence Effect for $\mathcal{G}_{\infty,1}$)**.** *Let $\epsilon_k = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_k)$ be the frozen-expert errors of Theorem 5, and assume that the retrained monolithic model is no better than each specialist on its own domain, $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k\|\widehat{\pi}_\lambda) \geq \epsilon_k$ for all k (Appendix B.7). Then the monolithic floor and*

modular ceiling satisfy $R_{\mathrm{mono}}(\lambda) \geq \sum_k \lambda_k \epsilon_k - \mathsf{D}_{\mathrm{JSD}}$ *and* $R_{\mathrm{mod}}(\lambda) \leq \log[\sum_k e^{\epsilon_k}] - \mathsf{D}_{\mathrm{JSD}} - \Delta(Z_{g^*})$. *The only change from the* $\mathcal{G}_1$ *statement is the additional Normalization Bonus* $-\Delta(Z_{g^*}) \leq 0$ *in the modular ceiling. (The floor alone also holds with* ϵ_k *the best-in-class error of the monolithic class, as in Cortes et al. (2026, Thm. D.1), but the ceiling requires the frozen-expert errors, so a common* ϵ_k *needs the assumption above.)*

Proof. The floor follows from Cortes et al. (2026, Thm. 6) via the identity of Appendix B.7 and the assumption $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \| \widehat{\pi}_\lambda) \geq \epsilon_k$. The ceiling combines Theorem 5 with the decomposition $\widetilde{\mathsf{D}}_{\mathrm{KL}} = \mathsf{D}_{\mathrm{KL}} + \Delta(Z)$. □

B.9 Algorithm and Convergence

We present the detailed algorithm and the convergence proof for Theorem 6.

Algorithm 2 Robust Non-Parametric Gate for $\mathcal{G}_{\infty,1}$ (uKL-EG + PGD)

Input: Models $\widehat{\pi}_1, \ldots, \widehat{\pi}_p$; datasets $D_1, \ldots, D_p$; learning rates η_λ, η_g; iterations T.
Initialize: $\lambda_0(k) = 1/p$; $g_0(x,k) = 1/p$.
▷ The problem is solved using the linearized payoff $\widetilde{L} = \sum_{k=1}^p \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$.
for $t = 0, 1, \ldots, T-1$ **do**
  (Compute Gains) For $k \in [1,p]$, compute $\ell_t(k) = \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_{g_t})$.
  (λ-update) Update $\lambda_{t+1}(k) \propto \lambda_t(k) \exp(\eta_\lambda \ell_t(k))$.
  (g-update)
  Compute gradient $v_t(x,k) = \widehat{\pi}_k(x)\Big(1 - \frac{\widehat{\mathsf{p}}_{\lambda_{t+1}}(x)}{\pi_{g_t}(x)}\Big)$.
  $g'_{t+1} = g_t - \eta_g v_t$.
  Project onto $\mathcal{G}^\eta_{\infty,1}$: $g_{t+1} = \Pi_{\mathcal{G}^\eta_{\infty,1}}(g'_{t+1})$. ▷ Expert-Free projection.
end for
Output: The time-averaged gate $\overline{g}_T$.

Theorem 6 (Convergence Guarantee for $\mathcal{G}_{\infty,1}$)**.** *Let* $V^\eta_\infty = \min_{g \in \mathcal{G}^\eta_{\infty,1}} \max_{\lambda \in \Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_g)$ *be the robust value of the game over the non-degenerate family* $\mathcal{G}^\eta_{\infty,1}$ *(Appendix B.11). Under Assumption 18, with step sizes* $\eta_\lambda, \eta_g \propto 1/\sqrt{T}$*, the time-averaged gate* $\overline{g}_T$ *produced by Algorithm 2 (with projection onto* $\mathcal{G}^\eta_{\infty,1}$*) converges to the robust solution with rate:*

$$\max_{\lambda \in \Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \pi_{\overline{g}_T}) - V^\eta_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right). \tag{7}$$

Proof. Let $\widetilde{L}(\lambda, g) = \sum_{k=1}^p \lambda_k \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \pi_g)$ be the payoff of the linearized convex-concave surrogate game, which is linear in λ and convex in g. Algorithm 2 simulates a sequence of plays $(\lambda_t, g_t)_{t=1}^T$ where the λ-player uses Exponentiated Gradient (EG) and the g-player uses Projected Gradient Descent (PGD).

We define the regret for each player after T iterations as:

$$R^\lambda_T = \max_{\lambda \in \Delta} \sum_{t=1}^T \widetilde{L}(\lambda, g_t) - \sum_{t=1}^T \widetilde{L}(\lambda_t, g_t), \quad R^g_T = \sum_{t=1}^T \widetilde{L}(\lambda_{t+1}, g_t) - \min_{g \in \mathcal{G}^\eta_{\infty,1}} \sum_{t=1}^T \widetilde{L}(\lambda_{t+1}, g). \tag{24}$$

All iterates lie in $\mathcal{G}^\eta_{\infty,1}$, so under Assumption 18 the gains and gradients are bounded ($\pi_{g_t}(x) \geq \eta\, e^{-M}$; in practice, the parametric neural gate naturally avoids the degenerate $g = 0$ region). Since Δ and $\mathcal{G}^\eta_{\infty,1}$ are compact convex sets, standard no-regret analysis for EG and PGD yields the following bounds with step sizes $\eta_\lambda, \eta_g \propto 1/\sqrt{T}$:

$$R^\lambda_T \leq M_\lambda \sqrt{2T \log p} = \mathcal{O}(\sqrt{T \log p}), \quad R^g_T \leq D_{\mathcal{G}_{\infty,1}} M_g \sqrt{T} = \mathcal{O}(\sqrt{T}). \tag{25}$$

Let $\overline{g}_T = \frac{1}{T}\sum_{t=1}^T g_t$ and $\overline{\lambda}_T = \frac{1}{T}\sum_{t=1}^T \lambda_{t+1}$. Summing the regrets introduces an alternating update difference:

$$\max_{\lambda \in \Delta} \sum_{t=1}^T \widetilde{L}(\lambda, g_t) - \min_{g \in \mathcal{G}^\eta_{\infty,1}} \sum_{t=1}^T \widetilde{L}(\lambda_{t+1}, g) = R^\lambda_T + R^g_T + \sum_{t=1}^T \big[\widetilde{L}(\lambda_t, g_t) - \widetilde{L}(\lambda_{t+1}, g_t)\big]. \tag{26}$$

Since λ_{t+1} is computed via EG from λ_t with step size $\eta_\lambda \propto 1/\sqrt{T}$, the parameter shift $\|\lambda_{t+1} - \lambda_t\|_1$ is bounded by $\mathcal{O}(1/\sqrt{T})$ per step. Summing over T bounds the cumulative difference by $\mathcal{O}(\sqrt{T})$. Thus, the duality gap of the time-averaged strategies satisfies:

$$\max_{\lambda\in\Delta} \widetilde{L}(\lambda, \overline{g}_T) - \min_{g\in\mathcal{G}^\eta_{\infty,1}} \widetilde{L}(\overline{\lambda}_T, g) \le \frac{R^\lambda_T + R^g_T + \mathcal{O}(\sqrt{T})}{T} = \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right). \tag{27}$$

Let $\widetilde{V} = \min_{g\in\mathcal{G}^\eta_{\infty,1}} \max_\lambda \widetilde{L}(\lambda, g)$ be the minimax value of the linearized game over $\mathcal{G}^\eta_{\infty,1}$. By weak duality, $\min_g \widetilde{L}(\overline{\lambda}_T, g) \le \widetilde{V}$. Therefore, the time-averaged gate bounds the linearized payoff:

$$\max_{\lambda\in\Delta} \widetilde{L}(\lambda, \overline{g}_T) - \widetilde{V} \le \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right). \tag{28}$$

We now rigorously relate the linearized game back to the true objective $\widetilde{\mathsf{D}}_{\mathrm{KL}}$. Since the unnormalized relative entropy $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\mathsf{p} \,\|\, q)$ is strictly convex in its first argument p (due to the underlying negative entropy term $-H(\mathsf{p})$), the composite objective $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \,\|\, \pi_g)$ is convex with respect to $\lambda \in \Delta$. Thus, its maximum over the simplex must occur at an extreme point (a vertex e_k). At any vertex, the Jensen-Shannon Divergence vanishes ($\mathsf{D}^{e_k}_{\mathrm{JSD}} = 0$), so by the decomposition in Lemma 4, $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_{e_k} \,\|\, \pi_g) = \widetilde{L}(e_k, g)$.

Furthermore, because the surrogate $\widetilde{L}(\lambda, g)$ is linear in λ, its maximum over Δ also occurs at a vertex. Thus, for *any* fixed gate g, the worst-case risks are perfectly identical:

$$\max_{\lambda\in\Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \,\|\, \pi_g) = \max_{k\in[1,p]} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \,\|\, \pi_g) = \max_{k\in[1,p]} \widetilde{L}(e_k, g) = \max_{\lambda\in\Delta} \widetilde{L}(\lambda, g). \tag{29}$$

Taking the minimum over $g \in \mathcal{G}^\eta_{\infty,1}$ on both sides proves that their robust minimax values are exactly equal: $V^\eta_\infty = \widetilde{V}$. Applying this exact equality for $g = \overline{g}_T$ and substituting into Equation (28) directly yields the final bound:

$$\max_{\lambda\in\Delta} \widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \,\|\, \pi_{\overline{g}_T}) - V^\eta_\infty = \max_{\lambda\in\Delta} \widetilde{L}(\lambda, \overline{g}_T) - \widetilde{V} \le \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right). \tag{30}$$

This completes the proof. □

Explicit constants. With the step sizes $\eta_\lambda = \sqrt{2\log p/T}/M_\lambda$ and $\eta_g = D_{\mathcal{G}_{\infty,1}}/(M_g\sqrt{T})$, the argument above gives the explicit bound $\big(M_\lambda\sqrt{2\log p} + D_{\mathcal{G}_{\infty,1}} M_g + c_\lambda\big)/\sqrt{T}$, where $M_\lambda \le M + \log(1/\eta)$ bounds the gains $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \,\|\, \pi_{g_t})$ over $\mathcal{G}^\eta_{\infty,1}$, $M_g \le (e^M/\eta)\big(\sum_x \|\widehat{\boldsymbol{\pi}}(x)\|_2^2\big)^{1/2}$ bounds the ℓ_2-norm of the gradient v_t (using $|1 - \widehat{\mathsf{p}}_\lambda(x)/\pi_g(x)| \le e^M/\eta$ on $\mathcal{G}^\eta_{\infty,1}$), $D_{\mathcal{G}_{\infty,1}} \le 2\sqrt{|\mathcal{X}_0|}$ is the ℓ_2-diameter of $\mathcal{G}^\eta_{\infty,1}$ (since $\sum_{x,k} g(x,k)^2 \le |\mathcal{X}_0| \sum_k w_k^2 \le |\mathcal{X}_0|$), and c_λ collects the alternating-update term, which is $\mathcal{O}(M_\lambda^2\sqrt{\log p})$. The $\mathcal{O}(\cdot)$ notation of Theorem 6 treats M, η, M_g and $|\mathcal{X}_0|$ as constants.

Since $\mathcal{G}^\eta_{\infty,1} \subseteq \mathcal{G}_{\infty,1}$, the benchmark satisfies $V^\eta_\infty \ge V_\infty$, with equality whenever some robust solution of the full game satisfies $\sum_k g^*(x,k) \ge \eta$ on $\mathcal{X}_0$. The restriction to $\mathcal{G}^\eta_{\infty,1}$ is what makes the gains and gradients bounded, and the same restriction underlies the generalization bounds of Appendix B.11.

Remark 17 (Robustness to approximate projections)**.** *The proof above assumes the exact projection $\Pi^{\mathrm{exact}}_{\mathcal{G}^\eta_{\infty,1}}$. The guarantee also holds for a projection operator satisfying the bounded-error condition $\|\Pi_{\mathcal{G}^\eta_{\infty,1}}(g') - \Pi^{\mathrm{exact}}_{\mathcal{G}^\eta_{\infty,1}}(g')\|_2 \le \epsilon_{\mathrm{proj}}$, at a cost that depends on the step size: the standard perturbation analysis of Projected Gradient Descent with inexact projections (expanding $\|g_{t+1} - g\|_2^2$ and absorbing the projection error) adds a term of order $T(2\epsilon_{\mathrm{proj}} D_{\mathcal{G}_{\infty,1}} + \epsilon^2_{\mathrm{proj}})/(2\eta_g)$ to R^g_T, where $D_{\mathcal{G}_{\infty,1}}$ is the diameter of $\mathcal{G}^\eta_{\infty,1}$. With $\eta_g \propto 1/\sqrt{T}$, this contributes $\mathcal{O}(\epsilon_{\mathrm{proj}}\sqrt{T})$ to the averaged bound, so the rate of Theorem 6 is preserved provided $\epsilon_{\mathrm{proj}} = \mathcal{O}(1/T)$. The bisection procedure of Algorithm 5 with tolerance ϵ returns an approximate projection with $\epsilon_{\mathrm{proj}} = \mathcal{O}(\epsilon)$ at a cost logarithmic in $1/\epsilon$, so the tolerance $\epsilon = \mathcal{O}(1/T)$ required to retain the rate costs only $\mathcal{O}(\log T)$ bisection steps per iteration.*

Implicit Mass Regularization. While the $\mathcal{G}_{\infty,1}$ family allows the total mass Z_g to float below unity, it does not collapse to zero. The $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective inherently penalizes extreme under-estimation of mass. Thus, the sub-normalized state observed during optimization is actively used by the model to safely secure the Normalization Bonus $-\Delta(Z_g)$ discussed in Section B.4. The robustness guarantees derived for g^* therefore apply directly to this optimally sub-normalized measure, and its corresponding normalized distribution rigorously respects the tightened implied D_{KL} bounds.

Characterization of the Projection. In the non-parametric regime, where $g(x,k)$ is treated as an independent variable for each x and k, a primary advantage of the $\mathcal{G}_{\infty,1}$ family is the efficiency of the exact projection $\Pi_{\mathcal{G}_{\infty,1}}$. Given an intermediate gate g', the projection onto the $\mathcal{G}_{\infty,1}$ constraint involves solving:

$$\min_{g \in \mathcal{G}_{\infty,1}} \|g - g'\|_2^2$$

$\sum_k \max_x g(x,k) \le 1$ is separable across experts k (up to the global sum), this can be computed exactly. The projection proceeds in two stages: for a fixed capacity vector $w \ge 0$ with $\sum_k w_k \le 1$, the optimal gate is the coordinate-wise clipping $g_w(x,k) = \min(w_k, (g'(x,k))_+)$; the remaining problem over w is solved via bisection on a single Lagrange multiplier, with total complexity $\mathcal{O}(p \cdot |\mathcal{X}_0| \log |\mathcal{X}_0| + p \cdot |\mathcal{X}_0| \log(1/\epsilon))$ for tolerance ϵ (Algorithm 5).

Parametric Implementation and Tractability. While the exact projection is efficient in function space, modern generative routing relies on parametric neural networks $g_\theta(x,k)$ with shared representations across experts. In this parametric setting, finding the exact global maximum over $\mathcal{X}_0$ and projecting shared weights without cross-expert interference is intractable. Therefore, in practical deep learning implementations, we approximate this exact projection. Rather than computing a hard projection over the entire dataset, we enforce the bounded norm constraint via a soft penalty term over empirical mini-batches $\mathcal{B}$, optimizing:

$$\mathcal{L}_{\text{penalty}} = \alpha \left[\mathrm{ReLU}\left(\sum_{k=1}^{p} \max_{x \in \mathcal{B}} g_\theta(x,k) - 1 \right) \right]^2. \tag{31}$$

This approach can be viewed as an empirical relaxation of a $(1,\infty)$-group-norm projection, where the norm is computed over the active mini-batch. This penalty introduces negligible computational overhead. Specifically, computing the batch-wise maximum and subsequent sum requires only $\mathcal{O}(|\mathcal{B}| \times p)$ operations applied directly to the lightweight gating logits. It bypasses the need to evaluate the p heavy expert models $\widehat{\pi}_k(x)$ or synchronize across devices. This relaxation preserves the decoupled nature of the training, avoiding the global cross-batch synchronization required by $\mathcal{G}_1$, while encouraging the shared neural network to adhere to the theoretically optimal capacity bounds.

B.10 Alternative Formulation: A Primal-Dual Approach

While the soft penalty method is a pragmatic and widely-used technique for handling the constraint in parametric settings, an alternative and more theoretically grounded approach exists that eliminates the need to tune the penalty hyperparameter. This is achieved by reformulating the problem within a primal-dual framework, where the Lagrange multiplier for the constraint is not a fixed value but a dynamic variable learned via dual ascent. This transforms the problem into a three-player game between the mixture player (λ), the gating player (g), and a new *constraint player* (α). The objective becomes solving the saddle-point problem on the Lagrangian:

$$\min_{g_\theta} \max_{\lambda \in \Delta, \alpha \ge 0} \quad \mathcal{L}(g_\theta, \lambda, \alpha) = \widetilde{L}(\lambda, g_\theta) + \alpha \left(\sum_{k=1}^{p} \max_{x \in \mathcal{B}} g_\theta(x,k) - 1 \right). \tag{32}$$

The algorithm would then iterate through three updates:

1. **λ-player (EG):** The mixture weights are updated using Exponentiated Gradient as before, responding to the unnormalized relative entropy of the current gate.
2. **g-player (Primal Step):** The gate parameters θ are updated by descending on the Lagrangian's subgradient: $\theta \leftarrow \theta - \eta_g \nabla_\theta \mathcal{L}(g_\theta, \lambda, \alpha)$.
3. **α-player (Dual Step):** The dual variable α is updated using a projected gradient ascent step to enforce the non-negativity constraint $\alpha \ge 0$:

$$\alpha \leftarrow \max\left(0, \alpha + \eta_\alpha \left(\sum_{k=1}^{p} \max_{x \in \mathcal{B}} g_\theta(x,k) - 1 \right)\right). \tag{33}$$

Advantages and Challenges. This primal-dual formulation offers distinct theoretical benefits. It possesses stronger guarantees; under standard conditions, the algorithm is guaranteed to converge to a solution satisfying the Karush-Kuhn-Tucker (KKT) conditions. The learned multiplier α dynamically adjusts to the "hardness" of the task diversity, ensuring that the model exploits its sub-normalized capacity to maximize the *normalization bonus* established in Section B.4 without violating the global norm. The challenges of this approach are related to optimization stability. Both the Lagrangian and the quadratic soft penalty are non-differentiable (relying on subgradient computation for the $\max$ operator). The key differences are practical:

- **Gradient Behavior:** The quadratic soft penalty, $[\mathrm{ReLU}(S(\theta)-1)]^2$, where $S(\theta) = \sum_k \max_x g_\theta(x,k)$, has a gradient that smoothly vanishes as the constraint becomes satisfied ($S(\theta) \to 1^+$). This provides stability near the constraint boundary. In contrast, the gradient of the Lagrangian's penalty term, $\alpha(S(\theta)-1)$, has a magnitude that remains constant with respect to the violation, which can cause oscillations or "chattering" around the boundary unless η_α is carefully decayed.

- **Mini-batch Constraint Estimates:** In both formulations the ceiling $\sum_k \max_x g_\theta(x,k)$ is evaluated on the current mini-batch $\mathcal{B}$, and $\max_{x\in\mathcal{B}} g_\theta(x,k) \le \max_{x\in\mathcal{X}_0} g_\theta(x,k)$, so the mini-batch statistic is a downward-biased estimate of the global ceiling. The penalty (or multiplier) therefore enforces the constraint only up to this estimation gap, which is why the deployed envelope is calibrated on a held-out set with a safety margin (Section 3.4); the exact projection of Algorithm 5 avoids the issue in the non-parametric setting.

- **Optimization Complexity:** The primal-dual method replaces the hyperparameter α with the dual learning rate η_α. While this eliminates manual tuning of the penalty strength, the stability of the resulting three-player game depends on the relative time-scales of the primal and dual updates.

In summary, the choice represents a trade-off between theoretical formalism and practical stability. The primal-dual method offers a principled way to enforce the $\mathcal{G}_{\infty,1}$ constraint, but the quadratic soft penalty is often preferred for its smoother landscape and simpler optimization dynamics.

B.11 Generalization and Sample Efficiency

While the robust guarantees established previously ensure the existence of an optimal gate, we now analyze the generalization guarantees for learning such a gate from a finite sample $S = (x_1, \dots, x_m)$ of size m. Our analysis adapts the structural definitions of the foundational modular framework (Cortes et al., 2026) to our unnormalized setting. We demonstrate that the generalization gap scales purely with the capacity of the lightweight gating class $\mathcal{G}_{\infty,1}$, rather than the massive complexity of the generative experts.

To ensure a Lipschitz-continuous loss landscape, we adopt the standard positivity condition.

Assumption 18 (Bounded Expert Likelihoods)**.** *We assume that the pre-trained experts are bounded away from zero on the support of the data distribution, at both the sequence and the token level. That is, there exist constants $M > 0$ and $M_{\mathrm{tok}} > 0$ such that for all $x \in \mathcal{X}_0$ and all $k \in [1,p]$,*

*(i) (*sequence level*) the negative log-likelihood is bounded: $|\log \widehat{\pi}_k(x)| \le M$;*
*(ii) (*token level*) for every position $t \le |x|$, the next-token conditional is bounded: $|\log \widehat{\pi}_k(x_t \mid x_{<t})| \le M_{\mathrm{tok}}$.*

Thus, for any valid probability mixture $\pi_g(x)$, the probability is lower-bounded by e^{-M}, and, writing $T = \max_{x\in\mathcal{X}_0} |x|$ for the maximal sequence length, any causal product of per-token expert mixtures $\prod_t \sum_k \gamma(x_{<t},k)\widehat{\pi}_k(x_t \mid x_{<t})$ with $\gamma(x_{<t},\cdot) \in \Delta([1,p])$ is lower-bounded by $e^{-TM_{\mathrm{tok}}}$ on $\mathcal{X}_0$. Part (i) is the condition used in the generalization analysis below (with $M \le TM_{\mathrm{tok}}$ by the chain rule), while part (ii) is used in the distillation analysis (Appendix B.13.5) to lower-bound the likelihood of the causal student.

Restriction to non-degenerate gates. The $\mathcal{G}_{\infty,1}$ constraint allows the degenerate all-zero gate $g = 0$, for which $\pi_g(x) = 0$ and the loss diverges. To ensure a uniformly bounded loss function over the hypothesis class, a prerequisite for valid Rademacher complexity bounds, we formally restrict the

gating family to the non-degenerate subset:

$$\mathcal{G}_{\infty,1}^{\eta} = \Big\{ g \in \mathcal{G}_{\infty,1} : \sum_{k=1}^{p} g(x,k) \geq \eta \ \ \forall\, x \in \mathcal{X}_0 \Big\}, \quad 0 < \eta \leq 1,$$

which is non-empty since it contains the uniform gate $g \equiv 1/p$. Under Assumption 18, any $g \in \mathcal{G}_{\infty,1}^{\eta}$ satisfies $\pi_g(x) \geq \eta\, e^{-M} > 0$, so the NLL loss is bounded by $M + \log(1/\eta)$. The set $\mathcal{G}_{\infty,1}^{\eta}$ is compact and convex for any $\eta > 0$. All generalization bounds (Theorems 19 and 20) are stated over $\mathcal{G}_{\infty,1}^{\eta}$, and we write $\mathcal{G}_{\infty,1}$ for brevity throughout when the context is clear.

As previously established in the literature (Cortes et al., 2026), this assumption naturally holds for Large Language Models (LLMs) with finite vocabularies. The standard softmax bottleneck ensures positive token probabilities; unless an expert explicitly masks a token to $-\infty$, the sequence log-likelihood remains finite.

We bound the generalization gap using vector-valued Rademacher complexity also referred to as *factor graph Rademacher complexity* (Cortes et al., 2016b). Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}_0$ to $\mathbb{R}^p$, and denote by f_j the j-th component of $f \in \mathcal{F}$. The empirical Rademacher complexity of $\mathcal{F}$ for a sample S is defined as:

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) = \frac{1}{m}\, \mathbb{E}_{\boldsymbol{\sigma}}\Big[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sum_{j=1}^{p} \sigma_{i,j} f_j(x_i) \Big],$$

where $\sigma_{i,j}$ are independent uniformly distributed Rademacher variables taking values in $\{-1, +1\}$. Its expectation over samples of size m, $\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_S\big[\widehat{\mathfrak{R}}_S(\mathcal{F})\big]$, is the population Rademacher complexity. Note that for $p = 1$, this trivially collapses to the standard real-valued definition.

We denote by p_k the true distribution generating the dataset D_k (or sample S_k), and $\mathfrak{R}_{m_k}^{k}(\mathcal{G}_{\infty,1})$ as the Rademacher complexity of $\mathcal{G}_{\infty,1}$ over p_k for a sample of size m_k.

The Expert Coincidence Norm (C_Π). To capture the geometric relationship between the pre-trained models, we use the maximum expert coincidence norm (Cortes et al., 2026). Letting $\widehat{\boldsymbol{\pi}}(x) = (\widehat{\pi}_1(x), \ldots, \widehat{\pi}_p(x))$ be the vector of expert likelihoods at input x, we define:

$$C_\Pi = \sup_{x \in \mathcal{X}_0} \|\widehat{\boldsymbol{\pi}}(x)\|_2.$$

This data-dependent constant $C_\Pi \in [0, \sqrt{p}]$ acts as a condition number for modularity, measuring the ensemble's diversity:

- **Orthogonal Experts (Best Case):** If experts represent distinct tasks with disjoint supports, at most one expert assigns non-negligible probability per input. Here, $\|\widehat{\boldsymbol{\pi}}(x)\|_2 \approx \widehat{\pi}_k(x) \leq 1$, yielding $C_\Pi \leq 1$.
- **High Entropy (Typical Case):** For LLMs generating over large vocabularies, sequence probabilities are typically minuscule. Even if experts are correlated, the ℓ_2-norm of these tiny probabilities remains small, frequently ensuring $C_\Pi \ll 1$.
- **Redundant Experts (Worst Case):** The maximum $C_\Pi = \sqrt{p}$ is achieved only if all p experts are perfectly redundant, simultaneously assigning a probability of 1 to the identical input.

Let $\mathcal{L}_{\log}(g) = \mathbb{E}_{x \sim \mathsf{p}}[-\log(\pi_g(x))]$ denote the expected Negative Log-Likelihood, and $\widehat{\mathcal{L}_{\log}}(g)$ its empirical counterpart over S. Since the generative structure of our mixture matches the globally normalized formulation, we can cleanly extend their generalization bounds to our unnormalized family.

Theorem 19 (Generalization Bound for Modular Gate Models)**.** *Under Assumption 18, for any $\delta > 0$ and $\eta > 0$, with probability at least $1 - \delta$ over the draw of samples $S_k \sim \mathsf{p}_k^{m_k}$, the following inequality holds simultaneously for all $g \in \mathcal{G}_{\infty,1}^{\eta}$ and $\lambda \in \Delta$:*

$$\mathcal{L}_{\log}(g) \leq \widehat{\mathcal{L}_{\log}}(g) + \sum_{k=1}^{p} \lambda_k \left[\frac{2\sqrt{2}\, C_\Pi\, e^{M}}{\eta}\, \mathfrak{R}_{m_k}^{k}(\mathcal{G}_{\infty,1}) + \Big(M + \log \tfrac{1}{\eta}\Big) \sqrt{\frac{\log \frac{p}{\delta}}{2m_k}} \right].$$

Proof. The bound follows from standard Rademacher complexity bounds (Mohri et al., 2018) and vector-contraction (Cortes et al., 2016b; Maurer, 2016). For each expert k, the loss $\ell_g(x) = -\log \pi_g(x)$ is bounded by $M + \log(1/\eta)$ (since $\pi_g(x) \geq \eta\, e^{-M}$ for $g \in \mathcal{G}^{\eta}_{\infty,1}$) and $(C_\Pi e^M/\eta)$-Lipschitz in g (since $\|\nabla_g \log \pi_g\|_2 = \|\widehat{\boldsymbol{\pi}}(x)\|_2/\pi_g(x) \leq C_\Pi/(\eta\, e^{-M}) = C_\Pi e^M/\eta$). Applying the Rademacher generalization bound per expert and combining via the mixture weights λ_k gives the first term. The confidence term follows from Hoeffding's inequality and a union bound over the p experts. □

Corollary 20 (Equal Sample Sizes)**.** *In the common setting where all experts have equal sample sizes $m_k = m$ and the complexity is uniform across distributions (denoted $\mathfrak{R}_m(\mathcal{G}_{\infty,1})$), the bound simplifies. For any $\delta > 0$, with probability at least $1-\delta$, for all $g \in \mathcal{G}^{\eta}_{\infty,1}$ and $\lambda \in \Delta$:*

$$\mathcal{L}_{\log}(g) \leq \widehat{\mathcal{L}_{\log}}(g) + \frac{2\sqrt{2}\, C_\Pi\, e^M}{\eta}\, \mathfrak{R}_m(\mathcal{G}_{\infty,1}) + \Big(M + \log\tfrac{1}{\eta}\Big)\sqrt{\frac{\log \frac{p}{\delta}}{2m}}.$$

Remark (Unbounded losses). Assumption 18 requires the loss to be uniformly bounded by M, which may be large for long sequences (since M scales with the maximum sequence length). The relative deviation margin bounds of Cortes et al. (2021a) provide an alternative that handles unbounded loss functions: their Corollary 8 bounds the generalization gap by $\gamma \sqrt[\alpha]{\mathcal{L}_\alpha(h)}\sqrt{\Delta_m/m^{2(\alpha-1)/\alpha}} + \rho$, where $\mathcal{L}_\alpha$ is the α-moment of the loss and Δ_m depends on covering numbers rather than Rademacher complexities. When M is prohibitively large, these bounds offer tighter guarantees at the cost of a margin parameter trade-off.

The theorem and corollary isolate the generalization gap entirely into the Rademacher complexity of the gating class. The same argument applies verbatim to any subclass $\mathcal{G} \subseteq \mathcal{G}^{\eta}_{\infty,1}$, in particular to the parametric class of the gating network, with $\widehat{\mathfrak{R}}_S(\mathcal{G})$ in place of $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1})$. By Theorem 8, under Assumption 7 the complexity ordering $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \leq \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ holds for every sample of $m \geq 3$ distinct points when $p \geq 3$, and up to a vanishing $\mathcal{O}(1/\sqrt{m})$ term when $p = 2$. Hence the bound of Theorem 19 is at least as tight as the same bound stated with $\widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ in place of $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1})$ (up to the vanishing term when $p = 2$), confirming that $\mathcal{G}_{\infty,1}$ generalizes at least as well. We now contrast the statistical stability of our unnormalized gate with both monolithic baselines and the original modular framework.

A. Modular vs. Monolithic Complexity. A standard monolithic model requires a hypothesis class Π_{mono} to internally represent the diverse supports of all underlying tasks, inflating its empirical Rademacher complexity $\widehat{\mathfrak{R}}_S(\Pi_{\text{mono}})$. While a monolithic Large Language Model may contain 10^9 to 10^{11} parameters, our gate is parameterized by a lightweight network (e.g., a small Transformer encoder or MLP) with merely 10^5 to 10^6 parameters (~290K–2.5M in our experiments), so the parametric gating class $\mathcal{G}$ satisfies $\mathfrak{R}_m(\mathcal{G}) \ll \mathfrak{R}_m(\Pi_{\text{mono}})$.

By freezing the high-capacity experts, our modular framework isolates the learning problem strictly to the gating class: by Theorem 19, the generalization gap is governed by the complexity of that class, and the number of experts enters the confidence term only through $\sqrt{\log p}$.

B. Bounded ($\mathcal{G}_{\infty,1}$) vs. Globally Normalized ($\mathcal{G}_1$) Gates.

Comparing our unnormalized relaxation to the previous framework reveals a critical architectural advantage. While both modular approaches benefit from the dampening effect of the expert coincidence norm, the $\mathcal{G}_1$ global normalization constraint, $\sum_{x,k} g(x,k)\widehat{\pi}_k(x) = 1$, is a single scalar equation that couples the gate to the expert distributions. Beyond the pointwise simplex, it imposes no restriction on how the gate behaves at individual inputs: any assignment of experts to the sampled points can be realized within $\mathcal{G}_1$, as long as the global mass is balanced elsewhere on the support (Step 2 of the proof below). On any sample satisfying Assumption 7, the feasible hypothesis space therefore contains the full lookup-table behavior of $\mathcal{G}_0$.

Conversely, the $\mathcal{G}_{\infty,1}$ constraint $\sum_{k=1}^{p} \max_x g(x,k) \leq 1$ is purely expert-independent. Applying a strict $(1,\infty)$-group norm to the gating predictions caps the total weight that each expert can receive across *all* inputs: raising the gate for one expert at one input consumes capacity available at every other input. This cross-input coupling is what restricts $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1})$, offering a cleaner, more stable capacity limit while avoiding computational synchronization bottlenecks.

The theoretical advantage of the unnormalized relaxation is most evident when formally comparing the intrinsic complexities of the two gating families. The comparison requires a mild condition relating the sample size to the separation of the experts, which we restate here.

Assumption 7 (Sample Size vs. Expert Separation)**.** *There exist at least two experts j and l whose total-variation distance bounds the sample size multiplied by the maximum point-mass probability of any expert: $m \max_{x \in \mathcal{X}_0, k} \widehat{\pi}_k(x) \le \|\widehat{\pi}_j - \widehat{\pi}_l\|_{\mathrm{TV}}$.*

The left side bounds the absolute maximum probability mass any combination of base models could possibly assign to a sample of m distinct sequences. For sequence models like LLMs, individual sequences carry *exponentially small* probability, so this maximum mass is negligible for any realistic sample size ($m \ll |\mathcal{X}_0|$). Meanwhile, since base models typically specialize with little support overlap, the right side is close to 1. Thus, this condition ensures the base models' separation comfortably exceeds the gate's capacity to memorize the sample. Nothing is required of the other experts, so duplicated experts are allowed. (The identical-experts case is trivial and separate: there $Z_g \equiv 1$ on $\mathcal{G}_1$ and no restriction on m is needed.)

Theorem 8 (Rademacher Complexity Comparison)**.** *Under Assumption 7, let S be a sample of m distinct points, then the following holds: if $p = 2$, then $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_1) + 1/(2\sqrt{2m})$; if $p \ge 3$ and $m \ge 3$, then $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) < \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$.*

Proof. Write $a = \max_{x,k} \widehat{\pi}_k(x)$ and $\Delta = \|\widehat{\pi}_j - \widehat{\pi}_l\|_{\mathrm{TV}}$ for the pair of Assumption 7, so that $ma \le \Delta$. Fix $\sigma \in \{\pm 1\}^{m \times p}$ and let $\Phi_{\mathcal{G}}(\sigma) = \sup_{g \in \mathcal{G}} \sum_{i,k} \sigma_{ik}\, g(x_i, k)$, so that $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) - \widehat{\mathfrak{R}}_S(\mathcal{G}_1) = \frac{1}{m}\,\mathbb{E}_\sigma\big[\Phi_{\mathcal{G}_{\infty,1}}(\sigma) - \Phi_{\mathcal{G}_1}(\sigma)\big]$.

Step 1: $\Phi_{\mathcal{G}_{\infty,1}}(\sigma) \le \max_k N_k^+$, where $N_k^+ = |\{i : \sigma_{ik} = +1\}|$. For $g \in \mathcal{G}_{\infty,1}$ put $w_k = \max_x g(x,k)$, so that $\sum_k w_k \le 1$ and $0 \le g(x_i,k) \le w_k$; then $\sum_{i,k} \sigma_{ik}\, g(x_i,k) \le \sum_k w_k N_k^+ \le \max_k N_k^+$.

Step 2: $\Phi_{\mathcal{G}_1}(\sigma) \ge \sum_i \max_k \sigma_{ik}$. It suffices to show that any prescribed values $r_i \in \Delta([1,p])$ at the sampled points extend to a gate in $\mathcal{G}_1$; the claim then follows by taking $r_i = e_{k(i)}$ with $\sigma_{ik(i)} = \max_k \sigma_{ik}$. Let $A \subseteq \mathcal{X}_0$ attain $\widehat{\pi}_j(A) - \widehat{\pi}_l(A) = \Delta$, and define the pointwise-simplex gates

$$h^-(x,\cdot) = \begin{cases} e_l & x \in A \\ e_j & x \notin A \end{cases}, \qquad h^+(x,\cdot) = \begin{cases} e_j & x \in A \\ e_l & x \notin A \end{cases},$$

whose masses are $Z_{h^-} = \widehat{\pi}_l(A) + \widehat{\pi}_j(A^c) = 1 - \Delta$ and $Z_{h^+} = \widehat{\pi}_j(A) + \widehat{\pi}_l(A^c) = 1 + \Delta$. Now overwrite both gates at the sampled points with the prescribed r_i. At each x_i the mixture value $\sum_k g(x_i,k)\,\widehat{\pi}_k(x_i)$ lies in $[0, a]$ before and after the change, so the total mass moves by at most $ma \le \Delta$: the modified gates $\widetilde{h}^{\pm}$ satisfy $Z_{\widetilde{h}^-} \le 1 \le Z_{\widetilde{h}^+}$. Since Z_g is linear in g, a convex combination $t\,\widetilde{h}^- + (1-t)\,\widetilde{h}^+$ has mass exactly 1 for some $t \in [0,1]$; it lies on the simplex at every point and agrees with r_i on S, hence belongs to $\mathcal{G}_1$.

Step 3: expectation of the difference. By Steps 1–2,

$$\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) - \widehat{\mathfrak{R}}_S(\mathcal{G}_1) \le \frac{1}{m}\,\mathop{\mathbb{E}}_{\sigma}\left[\max_k N_k^+ - \sum_{i=1}^m \max_k \sigma_{ik}\right]. \tag{34}$$

The second term has expectation $m(1 - 2^{1-p})$, since $\max_k \sigma_{ik} = -1$ exactly when all p signs at x_i are negative.

Case $p = 2$. Since $\max(N_1^+, N_2^+) = \frac{1}{2}(N_1^+ + N_2^+) + \frac{1}{2}|N_1^+ - N_2^+|$ and $\mathbb{E}|N_1^+ - N_2^+| \le \sqrt{\mathrm{Var}(N_1^+ - N_2^+)} = \sqrt{m/2}$, the right-hand side of (34) is at most $\frac{1}{2} + \frac{1}{2m}\sqrt{m/2} - \frac{1}{2} = \frac{1}{2\sqrt{2m}}$.

Case $p \ge 3$. The quantity $\frac{1}{m}\,\mathbb{E}_\sigma \max_k N_k^+$ is non-increasing in m: writing $Y_i \in \{0,1\}^p$ for the indicator vectors of positive signs, the average $\frac{1}{m}\sum_i Y_i$ is the average over $r \in [m]$ of the leave-one-out averages $\frac{1}{m-1}\sum_{i \ne r} Y_i$, and $v \mapsto \max_k v_k$ is convex, so $\mathbb{E}\max_k \frac{1}{m}\sum_i Y_{ik} \le \mathbb{E}\max_k \frac{1}{m-1}\sum_{i \ne r} Y_{ik}$. It therefore suffices to check that the right-hand side of (34) is negative at one sample size in each case. For $p = 3$, $m = 3$: $\frac{1}{3}\,\mathbb{E}\max_k N_k^+ = \frac{47}{64} < \frac{3}{4} = 1 - 2^{1-3}$. For $p \ge 4$, $m = 2$: $\frac{1}{2}\,\mathbb{E}\max_k N_k^+ = 1 - \frac{1}{2}\big[(\frac{3}{4})^p + (\frac{1}{4})^p\big]$, which is below $1 - 2^{1-p}$ if and only if $3^p + 1 > 2^{p+2}$, true for all $p \ge 4$. Monotonicity extends both cases to all larger m, which proves the claim for $p \ge 3$ and $m \ge 3$. □

This structural analysis highlights exactly why $\mathcal{G}_{\infty,1}$ provides tighter provable bounds. The global normalization of $\mathcal{G}_1$ is an average constraint weighted by the experts: it fixes the total mass but lets the gate select any expert at every sampled point, compensating on the unseen part of the support. In contrast, $\mathcal{G}_{\infty,1}$ imposes a uniform ceiling on the total weight of each expert across all x, independent of the expert distributions. By coupling the gate values across inputs, $\mathcal{G}_{\infty,1}$ restricts the hypothesis space, leading to a tighter generalization bound while entirely sidestepping the computational hurdles of a global partition function.

B.12 Inference: Exact and Efficient Rejection Sampling

A primary technical advantage of the $\mathcal{G}_{\infty,1}$ relaxation is its compatibility with Exact Rejection Sampling. In practice, we apply this protocol to the time-averaged gate $\overline{g}_T$ produced by Algorithm 2. Since the algorithm enforces a data-independent norm constraint, the model is inherently "inference-ready" without requiring global partition function estimation.

We first define a tight envelope for rejection sampling. For a sequence $x \in \mathcal{X}_0$, let $w_k^* = \max_{x'} \overline{g}_T(x', k)$. The unnormalized density $\pi_{\overline{g}_T}(x)$ is bounded by:

$$\pi_{\overline{g}_T}(x) = \sum_{k=1}^{p} \overline{g}_T(x,k)\widehat{\pi}_k(x) \le \sum_{k=1}^{p} w_k^* \widehat{\pi}_k(x). \tag{35}$$

Since $M_T = \sum_{k=1}^{p} w_k^* \le 1$ by the $\mathcal{G}_{\infty,1}$ constraint, we define the normalized proposal distribution $\mathsf{q}_{w^*}(x) = \frac{1}{M_T}\sum_{k=1}^{p} w_k^* \widehat{\pi}_k(x)$. The unnormalized density then satisfies $\pi_{\overline{g}_T}(x) \le M_T \mathsf{q}_{w^*}(x)$. This envelope is significantly tighter than a universal uniform mixture and ensures that we sample from an expert capacity that covers the robust gate. The capacity-weighted envelope is available to any gate, including $h \in \mathcal{G}_1$, with mass $M_h = \sum_k \max_x h(x,k)$. For a $\mathcal{G}_1$ gate that routes each source to its own expert, however, $\max_x h(x,k) = 1$ for every k, so $M_h = p$ and the envelope reduces to the uniform mixture with acceptance rate $1/p$. Near-unit acceptance requires per-expert maxima that sum to about 1, which is exactly what the $\mathcal{G}_{\infty,1}$ ceiling enforces.

We can perform rejection sampling in an exact fashion without knowing the total mass $Z_{\overline{g}_T}$. We adopt the protocol: (1) Sample $x \sim \mathsf{q}_{w^*}(x)$ by selecting expert k with probability w_k^*/M_T and sampling $x \sim \widehat{\pi}_k$; (2) Accept x with probability $\alpha(x) = \frac{\pi_{\overline{g}_T}(x)}{M_T \mathsf{q}_{w^*}(x)} = \frac{\sum_k \overline{g}_T(x,k)\widehat{\pi}_k(x)}{\sum_k w_k^* \widehat{\pi}_k(x)}$.

By the standard rejection sampling theorem, the accepted samples exactly follow the normalized target distribution $\mathsf{p}_{\overline{g}_T}(x) = \pi_{\overline{g}_T}(x)/Z_{\overline{g}_T}$, because the unknown constant $Z_{\overline{g}_T}$ cancels out in the normalization of the accepted proposals.

We provide the formal proof for the structural efficiency of exact rejection sampling introduced in Section 3.4.

Proposition 9 (Structural Efficiency of Exact Sampling)**.** *For any time-averaged gate $\overline{g}_T \in \mathcal{G}^{\eta}_{\infty,1}$, the expected acceptance rate is $\mathbb{P}(\text{accept}) = Z_{\overline{g}_T}/M_T \ge Z_{\overline{g}_T}$. Furthermore, the mass is bounded away from zero: $Z_{\overline{g}_T} \ge \exp(-(C_T+1)) > 0$, where $C_T = \max_\lambda \widetilde{L}(\lambda, \overline{g}_T) \le \log[\sum_k e^{\epsilon_k}] + \epsilon_T$ is the bounded worst-case risk of the algorithm.*

Proof. The expected acceptance rate for rejection sampling is exactly the ratio of the target unnormalized mass to the envelope mass:

$$\mathbb{P}(\text{accept}) = \mathop{\mathbb{E}}_{x\sim \mathsf{q}_{w^*}}[\alpha(x)] = \sum_{x\in\mathcal{X}_0} \frac{\pi_{\overline{g}_T}(x)}{M_T \mathsf{q}_{w^*}(x)} \mathsf{q}_{w^*}(x) = \frac{Z_{\overline{g}_T}}{M_T}.$$

By the definition of the $\mathcal{G}_{\infty,1}$ family, the constraint $\sum_k \max_x \overline{g}_T(x,k) \le 1$ enforces $M_T \le 1$. Therefore, $\frac{Z_{\overline{g}_T}}{M_T} \ge Z_{\overline{g}_T}$. For the lower bound, observe that the unnormalized objective naturally bounds the normalization penalty: $\widetilde{L}(\lambda, \overline{g}_T) \ge \Delta(Z_{\overline{g}_T})$. Therefore, $\Delta(Z_{\overline{g}_T}) \le C_T$. Since $\Delta(Z) = Z - 1 - \log Z \ge -\log Z - 1$, we have $-\log Z_{\overline{g}_T} \le C_T + 1$. Solving for $Z_{\overline{g}_T}$ yields $Z_{\overline{g}_T} \ge \exp(-(C_T+1))$. □

This establishes a robust, positive floor on the sampling efficiency governed by the optimization convergence. Unlike the globally normalized family where $Z_g = M = 1$, the $\mathcal{G}_{\infty,1}$ family actively

sub-normalizes ($Z_{g^*} \le M^* \le 1$) to minimize risk. Therefore, the exact acceptance rate evaluates to the *Sparsity Ratio* $Z_{\bar{g}_T}/M_T$.

We can characterize this efficiency exactly at the robust solution. Let (λ^*, g^*) be a saddle point of the linearized game (Appendix B.1), so that g^* minimizes $\widetilde{\mathsf{D}}_{\mathrm{KL}}(\widehat{\mathsf{p}}_{\lambda^*} \,\|\, \pi_g)$ over $\mathcal{G}_{\infty,1}$ by Lemma 4, and take g^* of the form given by Lemma 13: $g^*(x,k) = w_k^* \rho^*(x)$ with $w^* \in \Delta([1,p])$ and $\rho^*(x) = \min(1, \widehat{\mathsf{p}}_{\lambda^*}(x)/e_{w^*}(x))$, where $e_{w^*}(x) = \sum_k w_k^* \widehat{\pi}_k(x)$. Then $\pi_{g^*}(x) = \min(\widehat{\mathsf{p}}_{\lambda^*}(x), e_{w^*}(x))$, and since $\widehat{\mathsf{p}}_{\lambda^*}$ and e_{w^*} both have unit mass, ρ^* attains the value 1, so the envelope mass is $M_{g^*} = \sum_k w_k^* = 1$ and the capacity-weighted proposal is $\mathsf{q}_{w^*} = e_{w^*}$. The acceptance rate of the exact sampler is therefore

$$\mathbb{P}(\text{accept}) = \frac{Z_{g^*}}{M_{g^*}} = \sum_{x \in \mathcal{X}_0} \min(\widehat{\mathsf{p}}_{\lambda^*}(x), e_{w^*}(x)) = 1 - \|\widehat{\mathsf{p}}_{\lambda^*} - e_{w^*}\|_{\mathrm{TV}}.$$

Acceptance is thus close to 1 exactly when the capacity-weighted expert mixture approximates the worst-case source mixture. Near-disjoint supports alone do not suffice (an expert may be disjoint from the others yet place mass where its own source has none), but specialized experts with small errors ϵ_k do, which is the regime of Figure 8.

While this means the acceptance rate does not artificially equal 100% when domains overlap, this is a vital mathematical feature: the rejected samples are exactly the mechanism the model uses to filter out the errors made by the base experts, effectively harvesting the Normalization Bonus $-\Delta(Z_{\bar{g}_T})$; Regardless of the exact value of $Z_{\bar{g}_T}$, the accepted generative samples are exact and unbiased, bypassing the severe global synchronization bottlenecks of $\mathcal{G}_1$.

B.13 Structural Distillation for $\mathcal{G}_{\infty,1}$

While the rejection sampling method provides an exact generative pathway, the non-causal nature of the robust gate $g^*(x,k)$ prevents standard token-by-token autoregressive generation. To overcome this, we leverage *Structural Distillation* (Cortes et al., 2026), which maps the sequence-level gating decisions of the non-causal teacher into a lightweight *Causal Router* while preserving the original pre-trained experts.

B.13.1 Background: Structural Distillation

Structural Distillation, introduced by Cortes et al. (2026), addresses the fundamental architectural mismatch between the non-causal teacher and autoregressive deployment. The teacher π_{g^*} is a *Mixture of Products*: a single latent expert k is selected per sequence, yielding $\pi_{g^*}(x) = \sum_k g^*(x,k) \prod_t \widehat{\pi}_k(x_t \mid x_{<t})$. The student π_γ is a *Product of Mixtures*: at each token position, the router selects a fresh mixture over experts, yielding $\pi_\gamma(x) = \prod_t [\sum_k \gamma_\phi(x_{<t}, k) \widehat{\pi}_k(x_t \mid x_{<t})]$. As shown by Cortes et al. (2026), a causal product-of-mixtures student can closely approximate the teacher's conditionals; the structural mismatch is bounded by the joint convexity of KL and vanishes when the gate does not depend on future tokens (Proposition 21).

B.13.2 Model Definitions

We define the architectures involved in the distillation process:

- **Robust Non-Causal Teacher (p_{g^*}):** For any sequence x of length $T = |x|$, the robust teacher is defined as the unnormalized measure $\pi_{g^*}(x) = \sum_{k=1}^p g^*(x,k) \prod_{t=1}^T \widehat{\pi}_k(x_t \mid x_{<t})$, where $g^* \in \mathcal{G}_{\infty,1}$ is the non-causal gate. The *normalized teacher* is $p_{g^*} = \pi_{g^*}/Z_{g^*}$, a proper probability distribution.
- **Causal Student Router (π_γ):** The causal student model factors autoregressively:

$$\pi_\gamma(x) = \prod_{t=1}^T \left[\sum_{k=1}^p \gamma_\phi(x_{<t}, k) \widehat{\pi}_k(x_t \mid x_{<t}) \right], \tag{36}$$

  where $\gamma_\phi(x_{<t}, \cdot) \in \Delta$ is a causal router. Since $\sum_k \gamma_\phi(x_{<t}, k) = 1$ and each expert $\widehat{\pi}_k$ is normalized, the student is globally normalized by construction ($Z_\gamma = 1$).

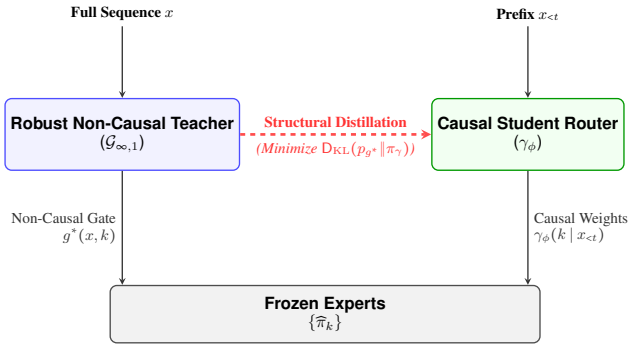


Figure 4: **Structural Distillation from $\mathcal{G}_{\infty,1}$.** The robust non-causal teacher leverages the full sequence context x to compute the gating decisions. The causal student learns to match the teacher's normalized distribution p_{g^*} on a token-by-token basis using only the prefix $x_{<t}$. Both the teacher and student route to the same fixed set of frozen experts $\{\widehat{\pi}_k\}$.

B.13.3 The Posterior Probability Causal Router

The teacher p_{g^*} is a *Mixture of Products* (one latent expert per sequence), while the student π_γ is a *Product of Mixtures* (independent per-token mixing). We characterize the structural gap between these two model classes.

Posterior Probability Router. The joint model $p_{g^*}(x, K{=}k) \propto g^*(x,k)\,\widehat{\pi}_k(x)$ induces a posterior probability for expert k given the prefix:

$$\gamma_k^\dagger(x_{<t}) := P_{p_{g^*}}(K{=}k \mid x_{<t}) = \frac{\sum_{x_{\geq t}} g^*(x,k)\,\widehat{\pi}_k(x)}{\sum_j \sum_{x_{\geq t}} g^*(x,j)\,\widehat{\pi}_j(x)}.$$

Since Z_{g^*} cancels in both numerator and denominator, $\gamma^\dagger$ is well-defined for any $g^* \in \mathcal{G}_{\infty,1}$ with $Z_{g^*} > 0$, and $\sum_k \gamma_k^\dagger(x_{<t}) = 1$.

The Structural Mismatch. By the law of total probability, the teacher's conditional factors as:

$$p_{g^*}(x_t \mid x_{<t}) = \sum_k P(K{=}k \mid x_{<t})\, p_{g^*}(x_t \mid x_{<t}, K{=}k),$$

while the student with router $\gamma^\dagger$ has: $\pi_{\gamma^\dagger}(x_t \mid x_{<t}) = \sum_k \gamma_k^\dagger(x_{<t})\,\widehat{\pi}_k(x_t \mid x_{<t})$. Both are mixtures with the *same weights* $\gamma_k^\dagger = P(K{=}k \mid x_{<t})$, but different component distributions. The component mismatch arises because the gate-reweighted expert conditional $p_{g^*}(x_t \mid x_{<t}, K{=}k) \neq \widehat{\pi}_k(x_t \mid x_{<t})$ in general for non-causal gates (the gate's dependence on future tokens distorts the within-expert conditionals).

Proposition 21 (Structural Mismatch Bound)**.** *For any $g^* \in \mathcal{G}_{\infty,1}$ with $Z_{g^*} > 0$, the posterior probability router $\gamma^\dagger$ achieves a structural mismatch bounded by:*

$$\mathrm{D_{KL}}(p_{g^*} \| \pi_{\gamma^\dagger}) \leq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_{<t} \sim p_{g^*}} \left[\sum_k P(K{=}k \mid x_{<t})\, \mathrm{D_{KL}}(p_{g^*}(\cdot \mid x_{<t}, K{=}k) \,\|\, \widehat{\pi}_k(\cdot \mid x_{<t})) \right].$$

This bound vanishes whenever $p_{g^*}(x_t \mid x_{<t}, K{=}k) = \widehat{\pi}_k(x_t \mid x_{<t})$ a.e., which holds in particular when g^* is constant in x (input-independent gating).

Proof. Since both the teacher conditional $\sum_k \gamma_k^\dagger\, p_{g^*}(\cdot \mid x_{<t}, K{=}k)$ and the student conditional $\sum_k \gamma_k^\dagger\, \widehat{\pi}_k(\cdot \mid x_{<t})$ are mixtures with the same weights, the joint convexity of KL gives:

$$\mathrm{D_{KL}}\big(p_{g^*}(\cdot \mid x_{<t}) \,\|\, \pi_{\gamma^\dagger}(\cdot \mid x_{<t})\big) \leq \sum_k \gamma_k^\dagger(x_{<t})\, \mathrm{D_{KL}}(p_{g^*}(\cdot \mid x_{<t}, K{=}k) \,\|\, \widehat{\pi}_k(\cdot \mid x_{<t})).$$

Applying the standard KL chain rule $\mathrm{D_{KL}}(p_{g^*} \| \pi_{\gamma^\dagger}) = \sum_t \mathbb{E}_{x_{<t}}[\mathrm{D_{KL}}(\mathrm{cond}_t)]$ completes the proof. Each component KL vanishes iff $p_{g^*}(\cdot \mid x_{<t}, K{=}k) = \widehat{\pi}_k(\cdot \mid x_{<t})$, which holds when $g^*(x,k)$ does not depend on $(x_t, x_{>t})$ given $x_{<t}$. □

B.13.4 Distillation Objective and Chain Rule Decomposition

Following Cortes et al. (2026), the distillation objective is the forward KL divergence from the normalized teacher to the causal student:

$$\min_{\phi} \mathcal{J}(\phi) = \mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) = -\mathop{\mathbb{E}}_{x\sim p_{g^*}}[\log \pi_\gamma(x)] + \text{const.} \tag{37}$$

Sequences from p_{g^*} are generated via the *Exact Rejection Sampling* scheme of Section 3.4 (Proposition 9), which produces unbiased samples without approximation. Since π_γ factors autoregressively, the objective decomposes into token-level next-token prediction:

$$\mathcal{J}(\phi) = -\mathop{\mathbb{E}}_{x\sim p_{g^*}}\left[\sum_{t=1}^{T} \log\left(\sum_{k=1}^{p} \gamma_\phi(x_{<t}, k)\, \widehat{\pi}_k(x_t \mid x_{<t})\right)\right] + \text{const.}$$

Theorem 22 (Chain Rule Decomposition)**.** *For any causal router γ_ϕ, the distillation divergence decomposes exactly:*

$$\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x_{<t}\sim p_{g^*}}[\mathsf{D}_{\mathrm{KL}}(p_{g^*}(\cdot \mid x_{<t}) \parallel \pi_\gamma(\cdot \mid x_{<t}))].$$

Proof. This is the standard chain rule for KL divergence applied to autoregressive sequence models (Cortes et al., 2026). Both p_{g^*} and π_γ admit conditional factorizations (the teacher by definition of conditional probability; the student by construction), so the chain rule applies directly. □

Algorithm 3 Structural Distillation via Cached Logits

Input: Teacher gate g^*, frozen experts $\{\widehat{\pi}_k\}$, router γ_ϕ
Phase 1: Teacher Generation & Caching
Generate M sequences $\{x^{(i)}\}$ from p_{g^*} via Exact Rejection Sampling (Section 3.4)
for each sequence $x^{(i)}$ and time step t **do**
    Cache expert probability vectors $\mathbf{P}_t = [\widehat{\pi}_1(x_t \mid x_{<t}), \ldots, \widehat{\pi}_p(x_t \mid x_{<t})]$
end for
Phase 2: Router Training
repeat
    Sample history $x_{<t}$ and cached probabilities $\mathbf{P}_t$ from cache
    Compute student weights $\mathbf{w}_t = \gamma_\phi(x_{<t}, \cdot)$
    Update ϕ by minimizing $\mathcal{L} = -\log(\mathbf{w}_t \cdot \mathbf{P}_t)$
until Convergence

B.13.5 Distillation Performance Inheritance

Theorem 10 (Inheritance Guarantee)**.** *If $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) \le \delta$, then for any $\lambda \in \Delta$: **(a)** $\|\widehat{\mathsf{p}}_\lambda - \pi_\gamma\|_1 \le \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 + \sqrt{2\delta}$; **(b)** under Assumption 18, $\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_\gamma) \le \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| p_{g^*}) + \delta + C\sqrt{2\delta}$, where $C = \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1/\eta'$ and $\eta' = \min_{x\in\mathcal{X}_0} \min\{p_{g^*}(x), \pi_\gamma(x)\} > 0$.*

Proof. The KL difference identity gives:

$$\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| \pi_\gamma) - \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \| p_{g^*}) = \mathop{\mathbb{E}}_{x\sim\widehat{\mathsf{p}}_\lambda}\left[\log \frac{p_{g^*}(x)}{\pi_\gamma(x)}\right] =: \Delta_\lambda.$$

Part (a): Total variation bound. By Pinsker's inequality applied to $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) \le \delta$: $\|p_{g^*} - \pi_\gamma\|_1 \le \sqrt{2\delta}$. The triangle inequality for total variation gives:

$$\|\widehat{\mathsf{p}}_\lambda - \pi_\gamma\|_1 \le \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 + \|p_{g^*} - \pi_\gamma\|_1 \le \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 + \sqrt{2\delta}.$$

Part (b): KL bound. Decompose the expectation via a change of measure:

$$\Delta_\lambda = \underbrace{\mathop{\mathbb{E}}_{x\sim p_{g^*}}\left[\log \frac{p_{g^*}(x)}{\pi_\gamma(x)}\right]}_{=\mathsf{D}_{\mathrm{KL}}(p_{g^*}\|\pi_\gamma)\le\delta} + \underbrace{\sum_x \big(\widehat{\mathsf{p}}_\lambda(x) - p_{g^*}(x)\big) \log \frac{p_{g^*}(x)}{\pi_\gamma(x)}}_{=:R_\lambda}.$$

By definition of $\eta' = \min_{x \in \mathcal{X}_0} \min\{p_{g^*}(x), \pi_\gamma(x)\}$, both $p_{g^*}(x)$ and $\pi_\gamma(x)$ are bounded below by η' on the support. Under Assumption 18, η' is strictly positive whenever the teacher gate is non-degenerate ($g^* \in \mathcal{G}^{\eta}_{\infty,1}$): part (i) gives $p_{g^*}(x) = \pi_{g^*}(x)/Z_{g^*} \geq \eta e^{-M}$ since $Z_{g^*} \leq 1$, and part (ii) gives $\pi_\gamma(x) \geq e^{-TM_{\text{tok}}}$ for the causal student, so that $\eta' \geq \min\{\eta e^{-M}, e^{-TM_{\text{tok}}}\}$ (see the Remark in Appendix C.8 for the corresponding bound for $\mathcal{G}_0$). By the Mean Value Theorem applied to $f(t) = \log t$:

$$\left|\log \frac{p_{g^*}(x)}{\pi_\gamma(x)}\right| = \left|\log p_{g^*}(x) - \log \pi_\gamma(x)\right| \leq \frac{|p_{g^*}(x) - \pi_\gamma(x)|}{\min\big(p_{g^*}(x),\, \pi_\gamma(x)\big)} \leq \frac{|p_{g^*}(x) - \pi_\gamma(x)|}{\eta'}.$$

By Hölder's inequality (ℓ_1–ℓ_∞) and Pinsker's inequality: $|R_\lambda| \leq \frac{1}{\eta'} \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 \cdot \|p_{g^*} - \pi_\gamma\|_\infty \leq \frac{1}{\eta'} \|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1 \sqrt{2\delta}$. Hence:

$$\Delta_\lambda \leq \delta + C\sqrt{2\delta}, \qquad C = \frac{\|\widehat{\mathsf{p}}_\lambda - p_{g^*}\|_1}{\eta'}. \qquad \square$$

The direct $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$ objective yields a *single-term* bound with no causal projection gap. The constant C captures the distribution shift between the test mixture and the teacher, scaled by the minimum likelihood η'; it is independent of the distillation quality δ.

B.13.6 Average Per-Token Distillation Guarantee

At deployment, the student generates text token by token: the operationally relevant metric is how well the student's next-token distribution matches the teacher's at each prefix. Theorem 10 controls the sequence-level divergence, but its Part (b) involves the distribution-shift constant C and the boundedness assumption. The following result bounds the *average per-token conditional TV* directly, with no assumptions.

Define the average per-token conditional KL:

$$\overline{\delta} = \mathop{\mathbb{E}}_{x \sim p_{g^*}}\left[\frac{1}{|x|}\sum_{t=1}^{|x|} \mathsf{D}_{\mathrm{KL}}(p_{g^*}(\cdot \mid x_{<t}) \,\|\, \pi_\gamma(\cdot \mid x_{<t}))\right].$$

Theorem 23 (Average Per-Token Distillation Guarantee)**.** *The average per-token conditional TV is bounded by:*

$$\mathop{\mathbb{E}}_{x \sim p_{g^*}}\left[\frac{1}{|x|}\sum_{t=1}^{|x|} \left\|p_{g^*}(\cdot \mid x_{<t}) - \pi_\gamma(\cdot \mid x_{<t})\right\|_1\right] \leq \sqrt{2\overline{\delta}}.$$

Proof. Apply Pinsker's inequality directly to each conditional pair:

$$\|p_{g^*}(\cdot \mid x_{<t}) - \pi_\gamma(\cdot \mid x_{<t})\|_1 \leq \sqrt{2\,\mathsf{D}_{\mathrm{KL}}(p_{g^*}(\cdot \mid x_{<t}) \,\|\, \pi_\gamma(\cdot \mid x_{<t}))}.$$

Taking the average over t and the expectation over $x \sim p_{g^*}$, then applying Jensen's inequality ($\mathbb{E}[\sqrt{Y}] \leq \sqrt{\mathbb{E}[Y]}$):

$$\mathop{\mathbb{E}}_{x \sim p_{g^*}}\left[\frac{1}{|x|}\sum_t \|p_{g^*}(\cdot \mid x_{<t}) - \pi_\gamma(\cdot \mid x_{<t})\|_1\right] \leq \sqrt{\mathop{\mathbb{E}}_{x \sim p_{g^*}}\left[\frac{1}{|x|}\sum_t 2\,\mathsf{D}_{\mathrm{KL}}(p_{g^*}(\cdot \mid x_{<t}) \,\|\, \pi_\gamma(\cdot \mid x_{<t}))\right]} = \sqrt{2\overline{\delta}},$$

where the equality is the definition of $\overline{\delta}$. $\square$

The bound is *dimension-free*: it depends only on the average per-token training loss $\overline{\delta}$, with no dependence on sequence length T, vocabulary size, or the number of experts p. As the student router converges ($\overline{\delta} \to 0$), the teacher's and student's next-token predictions become indistinguishable at every prefix.

C Locally-Normalized Gating ($\mathcal{G}_0$)

For scenarios requiring maximum computational efficiency, we consider enforcing normalization only locally. The $\mathcal{G}_0$ family represents the limit of computational decentralization within the modular framework. While $\mathcal{G}_{\infty,1}$ enables expert-independent projections, it still requires a global sum-constraint across all experts, which can introduce synchronization overhead in massive-scale distributed settings. In contrast, the locally-normalized $\mathcal{G}_0$ family eliminates the need for any global coordination during the gate update.

C.1 Motivation, Definition and Properties

By enforcing this normalization purely at the local level for each input point x, this relaxation allows the training process to treat each sample in a batch as an independent optimization problem. This provides the most streamlined path for hardware acceleration and asynchronous distributed training.

Definition 24 (Locally-Normalized Family $\mathcal{G}_0$)**.** *The family of locally-normalized gates $\mathcal{G}_0$ is defined as the set of all gating functions that form a valid probability simplex at every individual input point x:*

$$\mathcal{G}_0 = \left\{ g \geq 0 : \sum_{k=1}^{p} g(x,k) = 1, \quad \forall x \in \mathcal{X}_0 \right\} = \prod_{x \in \mathcal{X}_0} \Delta([1,p]). \tag{38}$$

Lemma 25 (Properties of $\mathcal{G}_0$)**.** *The family $\mathcal{G}_0$ is non-empty, compact, and convex.*

Proof. The set $\mathcal{G}_0$ is a Cartesian product of simplices $\Delta([1,p])$. Since the standard simplex is a compact and convex subset of $\mathbb{R}^p$, non-emptiness and convexity follow immediately. The compactness of the product set follows from Tychonoff's Theorem, which establishes that any product of compact spaces is compact under the product topology. □

It is critical to note that while $\mathcal{G}_0$ is compact, it does not naturally satisfy the sub-normalization property ($Z_g \leq 1$) found in $\mathcal{G}_{\infty,1}$. For any $g \in \mathcal{G}_0$, the total mass Z_g depends on the correlation between the local gates and the expert likelihoods:

$$Z_g = \sum_{x \in \mathcal{X}_0} \sum_{k=1}^{p} g(x,k) \widehat{\pi}_k(x) \leq \sum_{x \in \mathcal{X}_0} \max_{k \in [1,p]} \widehat{\pi}_k(x).$$

Since the experts $\widehat{\pi}_k$ are probability distributions over $\mathcal{X}_0$ ($\sum_x \widehat{\pi}_k(x) = 1$), the global mass is bounded by $Z_g \leq \sum_k \sum_x \widehat{\pi}_k(x) = p$. This $\mathcal{O}(p)$ volume is exactly why the tight rejection sampling envelope derived for the bounded family fails here: the uniform-mixture envelope $\mathsf{q}(x) = \frac{1}{p} \sum_k \widehat{\pi}_k(x)$ with envelope mass p still yields an *exact* sampler, with acceptance rate Z_g/p. This rate can be close to 1 when the mass is large (for the hard selector of Theorem 12 with disjoint expert supports, $Z_{g^*} = p$), but it is about $1/p$ for the mass-regularized gate of Section C.6, which drives Z_g toward unity and is the variant used in our experiments, motivating alternative inference strategies like structural distillation.

C.2 Fixed Mixture Solutions

Since the total mass Z_g is unconstrained and variable within the $\mathcal{G}_0$ family, optimizing the unnormalized relative entropy ($\widetilde{\mathsf{D}}_{\mathrm{KL}}$) directly would reintroduce the need to estimate the global partition function, destroying the computational benefits of local normalization. Therefore, to achieve fully decoupled, pointwise updates, we adopt the standard Negative Log-Likelihood (NLL) as our primary objective:

$$L_{\mathcal{L}_{\log}}(\lambda, g) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_g).$$

To characterize the optimal solution, we introduce the *Ambiguity gain* (or Jensen's Gap), which measures the reduction in NLL achieved by pooling expert predictions:

$$\mathcal{A}^\lambda(\{\widehat{\pi}_k\}) = \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_\lambda} \left[\log\left(\sum_{k=1}^{p} \lambda_k \widehat{\pi}_k(x) \right) - \sum_{k=1}^{p} \lambda_k \log \widehat{\pi}_k(x) \right].$$

By Jensen's Inequality for the concave logarithm, $\mathcal{A}^\lambda \geq 0$. It vanishes if and only if the expert predictions are identical (i.e., $\widehat{\pi}_k(x) = \widehat{\pi}_j(x)$ for all k, j and $\widehat{\mathsf{p}}_\lambda$-a.e. x). Note that for disjoint-support experts, $\mathcal{A}^\lambda$ diverges to $+\infty$ (since $\log \widehat{\pi}_k(x) = -\infty$ for inactive experts). In contrast, the pointwise Selection Gain $\mathcal{A}^{\lambda^\star}_{\text{selection}}$ (Section C.4) correctly vanishes for disjoint experts.

Proposition 26 (Fixed Mixture Bound and Ambiguity Gain)**.** *For any fixed mixture $\lambda \in \Delta$, assuming the cross-entropies $\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \widehat{\pi}_k)$ are finite (as under Assumption 18), the constant gate $g_\lambda(x, k) = \lambda_k \in \mathcal{G}_0$ satisfies the following NLL identity:*

$$\begin{aligned} \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g_\lambda}) &= \sum_{k=1}^{p} \lambda_k \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \widehat{\pi}_k) - \mathcal{A}^\lambda(\{\widehat{\pi}_k\}) \\ &\leq \sum_{k=1}^{p} \lambda_k \epsilon_k + H(\widehat{\mathsf{p}}_\lambda), \end{aligned} \tag{39}$$

where $\mathcal{A}^\lambda(\{\widehat{\pi}_k\}) \geq 0$ is the Ambiguity Gain *representing the Jensen gap. Since $\min_{g \in \mathcal{G}_0} \mathcal{L}_{\log} \leq \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g_\lambda})$, the convexity bound applies to the optimal gate as well.*

Proof. The $\mathcal{L}_{\log}$ objective is convex in g and continuous on the compact set $\mathcal{G}_0$, ensuring the existence of a minimizer. Consider the constant mixture gate $g_\lambda(x, k) = \lambda_k$. Evaluating the $\mathcal{L}_{\log}$ at g_λ:

$$\begin{aligned} \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g_\lambda}) &= \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_\lambda}\left[-\log\left(\sum_{k=1}^{p} \lambda_k \widehat{\pi}_k(x) \right) \right] \\ &= \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_\lambda}\left[-\sum_{k=1}^{p} \lambda_k \log \widehat{\pi}_k(x) \right] - \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_\lambda}\left[\log\left(\sum_{k=1}^{p} \lambda_k \widehat{\pi}_k(x) \right) - \sum_{k=1}^{p} \lambda_k \log \widehat{\pi}_k(x) \right] \\ &= \sum_{k=1}^{p} \lambda_k \mathop{\mathbb{E}}_{x \sim \widehat{\mathsf{p}}_\lambda}\left[-\log \widehat{\pi}_k(x) \right] - \mathcal{A}^\lambda(\{\widehat{\pi}_k\}) \\ &= \sum_{k=1}^{p} \lambda_k \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \widehat{\pi}_k) - \mathcal{A}^\lambda(\{\widehat{\pi}_k\}). \end{aligned}$$

This establishes the identity.

The upper bound follows independently by the joint convexity of D_{KL}:

$$\begin{aligned} \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g_\lambda}) &= \mathsf{D}_{\mathrm{KL}}\left(\sum_{k=1}^{p} \lambda_k \widehat{\mathsf{p}}_k \parallel \sum_{k=1}^{p} \lambda_k \widehat{\pi}_k \right) + H(\widehat{\mathsf{p}}_\lambda) \\ &\leq \sum_{k=1}^{p} \lambda_k \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k) + H(\widehat{\mathsf{p}}_\lambda) \\ &\leq \sum_{k=1}^{p} \lambda_k \epsilon_k + H(\widehat{\mathsf{p}}_\lambda). \end{aligned}$$

□

The cross-entropy terms $\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \widehat{\pi}_k) = \mathbb{E}_{\widehat{\mathsf{p}}_\lambda}[-\log \widehat{\pi}_k]$ evaluate each expert under the mixture $\widehat{\mathsf{p}}_\lambda$ rather than under its own domain $\widehat{\mathsf{p}}_k$; they are finite under Assumption 18 but are distinct from the in-domain errors ϵ_k.

C.3 Robust Solutions

We define the value of the robust minimax game for the locally-normalized family as:

$$V_0 = \min_{g \in \mathcal{G}_0} \max_{\lambda \in \Delta} \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_g).$$

We provide the full proof of Theorem 12 establishing the robust guarantees for the locally-normalized family.

Theorem 12 (Robust Existence and Guarantees for $\mathcal{G}_0$)**.** *The minimax game admits a solution $g^* \in \mathcal{G}_0$ that performs hard pointwise expert selection: $g^*(x,k) = \mathbf{1}[k = \operatorname{argmax}_j \widehat{\pi}_j(x)]$ (ties broken arbitrarily). For any test mixture $\lambda \in \Delta$, with λ^* a worst-case mixture at g^* and $\mathcal{A}^{\lambda^*}_{\text{selection}} = \sum_k \lambda^*_k \mathbb{E}_{x\sim\widehat{\mathsf{p}}_k}\left[\log \frac{\max_j \widehat{\pi}_j(x)}{\widehat{\pi}_k(x)}\right] \geq 0$ the pointwise selection gain:*

$$\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g^*}) \leq \max_{k\in[1,p]} (\epsilon_k + H(\widehat{\mathsf{p}}_k)) - \mathcal{A}^{\lambda^*}_{\text{selection}}.$$

Proof. Let $L(\lambda, g) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_g) = \sum_{k=1}^p \lambda_k \mathcal{L}_{\log}(\widehat{\mathsf{p}}_k, \pi_g)$, with values in $(-\infty, +\infty]$ since a gate may assign zero mass to a point charged by some source. Since the gating family $\mathcal{G}_0$ imposes only pointwise constraints, consider the hard selector

$$g^*(x,k) = 1 \Leftrightarrow k = \operatorname*{argmax}_{j\in[1,p]} \widehat{\pi}_j(x) \quad \text{(ties broken arbitrarily)},$$

whose mixture density is $\pi_{g^*}(x) = \max_j \widehat{\pi}_j(x)$. For any $g \in \mathcal{G}_0$ and any x, $\pi_g(x) = \sum_k g(x,k)\widehat{\pi}_k(x)$ is a convex combination of the $\widehat{\pi}_k(x)$, so $\pi_g(x) \leq \max_j \widehat{\pi}_j(x) = \pi_{g^*}(x)$. Since $-\log$ is decreasing, $L(\lambda, g^*) \leq L(\lambda, g)$ for every $\lambda \in \Delta$ and every $g \in \mathcal{G}_0$ simultaneously. Hence g^* minimizes $\max_\lambda L(\lambda, g)$ over $\mathcal{G}_0$, that is, g^* is a robust solution, and $V_0 = \max_{\lambda\in\Delta} L(\lambda, g^*)$. This maximum is finite and attained: for $x \in \operatorname{supp}(\widehat{\mathsf{p}}_k)$ the finite error ϵ_k gives $\widehat{\pi}_k(x) > 0$, hence $\pi_{g^*}(x) > 0$, so $\lambda \mapsto L(\lambda, g^*)$ is a finite linear function on the simplex. Let λ^* be a maximizer, the worst-case mixture at g^*. We now evaluate the minimax value V_0:

$$V_0 = \sum_{k=1}^p \lambda^*_k \mathop{\mathbb{E}}_{x\sim\widehat{\mathsf{p}}_k}\left[-\log \max_j \widehat{\pi}_j(x)\right].$$

Since $\max_j \widehat{\pi}_j(x) \geq \widehat{\pi}_k(x)$ for all k, we have $-\log \max_j \widehat{\pi}_j(x) \leq -\log \widehat{\pi}_k(x)$. We can rewrite the expectation by adding and subtracting the individual expert log-likelihoods:

$$\begin{aligned} V_0 &= \sum_{k=1}^p \lambda^*_k \mathop{\mathbb{E}}_{x\sim\widehat{\mathsf{p}}_k}\left[-\log \widehat{\pi}_k(x)\right] - \sum_{k=1}^p \lambda^*_k \mathop{\mathbb{E}}_{x\sim\widehat{\mathsf{p}}_k}\left[\log \max_j \widehat{\pi}_j(x) - \log \widehat{\pi}_k(x)\right] \\ &= \sum_{k=1}^p \lambda^*_k (\mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_k \parallel \widehat{\pi}_k) + H(\widehat{\mathsf{p}}_k)) - \mathcal{A}^{\lambda^*}_{\text{selection}} \\ &\leq \max_{k\in[1,p]} (\epsilon_k + H(\widehat{\mathsf{p}}_k)) - \mathcal{A}^{\lambda^*}_{\text{selection}}. \end{aligned}$$

This gain represents the expected improvement in log-likelihood achieved by selecting the best expert pointwise over evaluating each expert only on its own generated samples. Since g^* is the robust solution, this bound holds for any mixture λ. □

C.4 Geometric Interpretation of the $\mathcal{G}_0$ Bound

The upper bound $\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g^*}) \leq \max_{k\in[1,p]}(\epsilon_k + H(\widehat{\mathsf{p}}_k)) - \mathcal{A}^{\lambda^*}_{\text{selection}}$ reveals how the locally-normalized gate leverages ensemble selection without global coordination. To interpret these guarantees on a standard probabilistic scale, we evaluate the bound across three limiting geometric regimes.

Case 1: The Specialization Limit (Disjoint Experts). Consider the regime where task supports are mutually disjoint ($\operatorname{supp}(\widehat{\mathsf{p}}_k) \cap \operatorname{supp}(\widehat{\mathsf{p}}_j) = \varnothing$).

- **Selection Gain:** For disjoint supports, at most one expert assigns non-zero probability at any given input x. Thus, $\max_j \widehat{\pi}_j(x) = \widehat{\pi}_k(x)$ for all $x \in \operatorname{supp}(\widehat{\mathsf{p}}_k)$, and the selection gain vanishes ($\mathcal{A}^{\lambda^*}_{\text{selection}} \approx 0$).
- **Result:** The bound reduces to the worst-case individual expert performance: $\mathcal{L}_{\log} \leq \max_k(\epsilon_k + H(\widehat{\mathsf{p}}_k))$.
- **Interpretation:** Without a global mass budget, the $\mathcal{G}_0$ gate cannot "save" capacity by specializing. It defaults to the baseline performance of the expert best suited for the local support.

Case 2: The Selection Limit (Overlapping Experts). Consider the regime where multiple experts provide high-likelihood predictions for the same inputs.

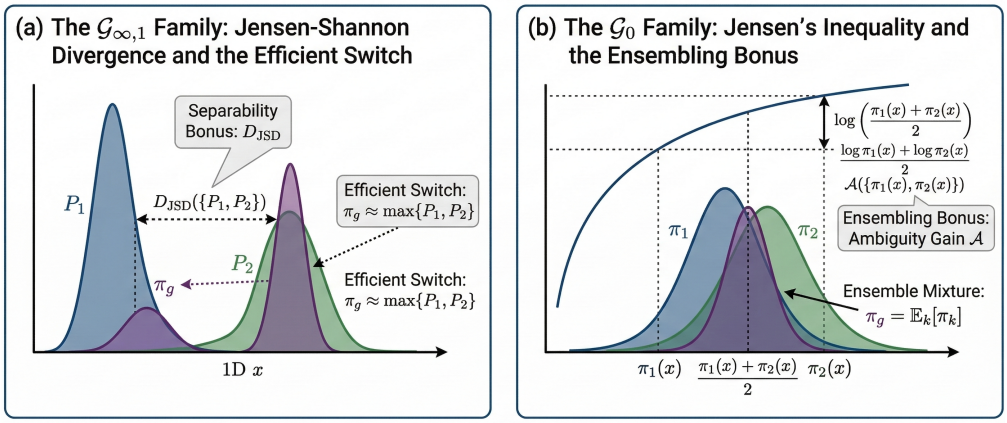


Figure 5: Geometric Interpretation of Robustness Gains. (a) The $\mathcal{G}_{\infty,1}$ family leverages task diversity; the Separability Bonus (D_{JSD}) and the Normalization Bonus ($\Delta(Z)$) enable the model to act as an efficient switch in specialized regimes. (b) The $\mathcal{G}_0$ family excels under expert redundancy, where the concavity of the logarithm provides an ensembling bonus (Ambiguity Gain $\mathcal{A}$) at the cost of global mass stability.

- **Selection Gain:** Since g^* performs a hard selection based on the maximum likelihood pointwise, the gain is positive, $\mathcal{A}^{\lambda^*}_{\text{selection}} > 0$. This gain scales with the ensemble's ability to provide a more certain prediction than any individual expert considered in isolation.
- **Interpretation:** The $\mathcal{G}_0$ family thrives on predictive variety. It rewards the system for using the sharpest expert at every coordinate x, directly compressing the Negative Log-Likelihood beyond the average expert floor.

Case 3: Implied KL and the Volume Penalty. While $\mathcal{G}_0$ gates are normalized locally at each point x, the resulting mixture model π_{g^*} is not a globally normalized distribution. Its total mass Z_{g^*} depends on the alignment of expert peaks across the support. To compare this with normalized families, we consider the rescaled model $\mathsf{p}_{g^*}(x) = \pi_{g^*}(x)/Z_{g^*}$. The relationship between $\mathcal{L}_{\log}$ and D_{KL} is:

$$\mathsf{D}_{\text{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_{g^*}) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g^*}) + \log Z_{g^*} - H(\widehat{\mathsf{p}}_\lambda).$$

Substituting the robust bound from Theorem 12 yields the following performance guarantee:

$$\mathsf{D}_{\text{KL}}(\widehat{\mathsf{p}}_\lambda \parallel \mathsf{p}_{g^*}) \le \max_{k\in[1,p]} (\epsilon_k + H(\widehat{\mathsf{p}}_k)) - H(\widehat{\mathsf{p}}_\lambda) - \mathcal{A}^{\lambda^*}_{\text{selection}} + \log Z_{g^*}.$$

- **The Volume Penalty:** The term $\log Z_{g^*}$ represents a *Volume Penalty* incurred by allowing the model mass to exceed unity. This mass is bounded by $\sum_{x\in\mathcal{X}_0} \max_k \widehat{\pi}_k(x)$. Unlike the $\mathcal{G}_{\infty,1}$ family, where $\log Z \le 0$ acts as a bonus, $\mathcal{G}_0$ incurs a positive penalty that scales with the number of modes in the data.
- **The Phase Transition of Modularity:** This mathematically formalizes a fundamental trade-off: we trade the strict theoretical purity of $\mathcal{G}_{\infty,1}$ (no volume penalty, global diversity gain) for the absolute computational efficiency of $\mathcal{G}_0$ (fully decoupled updates, local selection gain). On the standard probabilistic scale, $\mathcal{G}_0$ remains fundamentally "looser" due to the inflation of unallocated mass across the support.

C.5 Algorithm and Convergence

The robust solution g^* is found by solving the convex-concave game $L(\lambda, g) = \sum_k \lambda_k \mathcal{L}_{\log}(\widehat{\mathsf{p}}_k, \pi_g)$. The L-loss formulation for $\mathcal{G}_0$ offers the most streamlined dynamics in our framework due to the local nature of both the gradients and the projection operator.

The g-player uses a pointwise gradient update. For each sample x, the gradient with respect to $g(x,k)$ is:

$$v_t(x,k) = \nabla_{g(x,k)} L(\lambda_{t+1}, g_t) = -\frac{\widehat{\pi}_k(x)\widehat{\mathsf{p}}_{\lambda_{t+1}}(x)}{\pi_{g_t}(x)}.$$

This update is purely local, it only requires the predictions of the experts and the current gate value at point x. The projection Π_Δ then maps the updated gate back to the simplex independently for each x, enabling full parallelization across batches without any inter-sample communication.

Algorithm 4 Robust Gate for $\mathcal{G}_0$ (NLL-EG + OGD)

Input: Models $\widehat{\pi}_1, \dots, \widehat{\pi}_p$; datasets $D_1, \dots, D_p$; learning rates η_λ, η_g.
Initialize: $\lambda_0(k) = 1/p$; $g_0(x,k) = 1/p$.
for $t = 0, 1, \dots, T-1$ **do**
  (Compute Gains) $\ell_t(k) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_k, \pi_{g_t})$.
  (λ-update) $\lambda_{t+1}(k) \propto \lambda_t(k) \exp(\eta_\lambda \ell_t(k))$.
  (g-update)
  Compute gradient $v_t(x,k) = -\frac{\widehat{\pi}_k(x)\widehat{\mathsf{p}}_{\lambda_{t+1}}(x)}{\pi_{g_t}(x)}$ for each $x \in \cup D_k$.
  $g'_{t+1} = g_t - \eta_g v_t$.
  Project onto local simplex: $g_{t+1}(x,\cdot) = \Pi_\Delta(g'_{t+1}(x,\cdot))$.
end for
Output: The time-averaged gate $\overline{g}_T$.

Theorem 27 (Convergence Guarantee for $\mathcal{G}_0$)**.** *Let V_0 be the value of the L-game over $\mathcal{G}_0$. Under Assumption 18, with step sizes $\eta_\lambda, \eta_g \propto 1/\sqrt{T}$, the time-averaged gate $\overline{g}_T$ produced by Algorithm 4 converges to the robust solution with rate:*

$$\max_{\lambda \in \Delta} L(\lambda, \overline{g}_T) - V_0 \le O\left(\sqrt{\frac{\log p}{T}}\right).$$

Proof. Let $L(\lambda, g) = \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_g)$. By the results in Section C.3, in particular, Theorem 12, L is linear in λ and convex in g over the compact convex set $\Delta \times \mathcal{G}_0$. Algorithm 4 uses Exponentiated Gradient for the λ-player and Online Gradient Descent for the g-player. Since the g-player update uses the advanced weights λ_{t+1}, we define the regrets as:

$$R_T^\lambda = \max_{\lambda \in \Delta} \sum_{t=1}^T L(\lambda, g_t) - \sum_{t=1}^T L(\lambda_t, g_t), \quad R_T^g = \sum_{t=1}^T L(\lambda_{t+1}, g_t) - \min_{g \in \mathcal{G}_0} \sum_{t=1}^T L(\lambda_{t+1}, g).$$

Under Assumption 18, every $g \in \mathcal{G}_0$ satisfies $\pi_g(x) \ge \min_k \widehat{\pi}_k(x) \ge e^{-M}$ since $g(x,\cdot) \in \Delta([1,p])$, so the payoffs are bounded by M and the gradients $v_t(x,k)$ are bounded by $e^M \widehat{\mathsf{p}}_{\lambda_{t+1}}(x)$, as required by the regret bounds below. With step sizes $\eta_\lambda \propto \sqrt{\log p / T}$ and $\eta_g \propto 1/\sqrt{T}$, standard no-regret theory yields $R_T^\lambda \le \mathcal{O}(\sqrt{T \log p})$ and $R_T^g \le \mathcal{O}(\sqrt{T})$. Summing the regrets introduces an alternating update difference:

$$\max_\lambda \sum_t L(\lambda, g_t) - \min_g \sum_t L(\lambda_{t+1}, g) = R_T^\lambda + R_T^g + \sum_{t=1}^T [L(\lambda_t, g_t) - L(\lambda_{t+1}, g_t)].$$

Since λ_{t+1} is computed via EG from λ_t with step size $\eta_\lambda \propto 1/\sqrt{T}$, the parameter shift $\|\lambda_{t+1} - \lambda_t\|_1$ is bounded by $\mathcal{O}(1/\sqrt{T})$ per step, so the cumulative difference is $\mathcal{O}(\sqrt{T})$. Thus the duality gap of the time-averaged strategies $(\overline{\lambda}_T, \overline{g}_T)$ satisfies:

$$\max_{\lambda \in \Delta} L(\lambda, \overline{g}_T) - \min_{g \in \mathcal{G}_0} L(\overline{\lambda}_T, g) \le \frac{R_T^\lambda + R_T^g + \mathcal{O}(\sqrt{T})}{T} = \mathcal{O}\left(\sqrt{\frac{\log p}{T}}\right).$$

To relate this to the robust value $V_0 = \min_g \max_\lambda L(\lambda, g)$, we observe that by the definition of the minimax value:

$$V_0 = \max_{\lambda \in \Delta} \min_{g \in \mathcal{G}_0} L(\lambda, g) \ge \min_{g \in \mathcal{G}_0} L(\overline{\lambda}_T, g).$$

Substituting this into the duality gap inequality, we obtain:

$$\max_{\lambda \in \Delta} L(\lambda, \overline{g}_T) - V_0 \le \max_{\lambda \in \Delta} L(\lambda, \overline{g}_T) - \min_{g \in \mathcal{G}_0} L(\overline{\lambda}_T, g) \le O\left(\sqrt{\frac{\log p}{T}}\right).$$

This confirms that the time-averaged gate $\overline{g}_T$ converges to the robust value V_0 at the stated rate. □

C.6 Global Regularization and Mass Stability

While the $\mathcal{G}_0$ family is optimized using local updates to maximize computational throughput, it lacks the global mass constraints of the $\mathcal{G}_1$ or $\mathcal{G}_{\infty,1}$ families. To bridge this gap and approximate the normalization goals of the original framework, we consider the Regularized $\mathcal{G}_0$ Game. We augment the objective with a soft quadratic penalty on the global partition function Z_g:

$$V_\beta = \min_{g\in\mathcal{G}_0}\max_{\lambda\in\Delta}\{L(\lambda,g) + \beta(Z_g - 1)^2\}.$$

The penalty term explicitly drives the global mass Z_g toward 1 as the regularization strength β increases. This formulation allows the model to remain computationally efficient during training while ensuring the final mixture density is nearly normalized.

Theorem 28 (Robustness and Normalization Stability for $\mathcal{G}_0$)**.** *Under Assumption 18, the regularized game admits a saddle point $(g^*_\beta, \lambda^*_\beta)$. For any mixture $\lambda \in \Delta$, the resulting model $\pi_{g^*_\beta}$ satisfies:*

1. ***Robustness:*** *For any mixture $\lambda \in \Delta$,*

$$\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_{g^*_\beta}) \le \max_{\lambda'\in\Delta}\left[\sum_{k=1}^{p}\lambda'_k\epsilon_k + H(\widehat{\mathsf{p}}_{\lambda'})\right].$$

2. ***Mass Stability:*** *The global mass error satisfies the dimension-free bound:*

$$|Z_{g^*_\beta} - 1| \le \frac{1}{\sqrt{2\beta}}.$$

Proof. Let $L_\beta(\lambda, g) = L(\lambda, g) + \beta(Z_g - 1)^2$. Under Assumption 18, every $g \in \mathcal{G}_0$ satisfies $\pi_g(x) \ge \min_k \widehat{\pi}_k(x) \ge e^{-M}$, so L_β is finite and continuous on the compact convex set $\Delta \times \mathcal{G}_0$, linear in λ and convex in g. By Sion's Minimax Theorem, the saddle point $(g^*_\beta, \lambda^*_\beta)$ exists and the minimax value equals the maximin value: $V_\beta = \min_g \max_\lambda L_\beta = \max_\lambda \min_g L_\beta$. We bound V_β via the maximin formulation. For each fixed λ, evaluate the inner minimum at the *matching* constant gate $g_\lambda(x,k) = \lambda_k$. This witness has $Z_{g_\lambda} = 1$, so the penalty vanishes. Its NLL evaluates to:

$$\min_g L_\beta(\lambda, g) \le L(\lambda, g_\lambda) = \mathsf{D}_{\mathrm{KL}}(\widehat{\mathsf{p}}_\lambda \,\|\, \textstyle\sum_k \lambda_k\widehat{\pi}_k) + H(\widehat{\mathsf{p}}_\lambda) \le \sum_k \lambda_k\epsilon_k + H(\widehat{\mathsf{p}}_\lambda),$$

where the last inequality uses joint convexity of the KL divergence. Maximizing over λ yields the robustness bound. Since $\beta(Z-1)^2 \ge 0$, $\mathcal{L}_{\log} \le V_\beta$ directly.

To bound the mass error, we apply the minimum principle to the convex objective L_β at g^*_β. For any $g \in \mathcal{G}_0$:

$$\left\langle \nabla_g L(\lambda^*_\beta, g^*_\beta) + 2\beta(Z_{g^*_\beta} - 1)\nabla_g Z_{g^*_\beta}, g - g^*_\beta \right\rangle \ge 0.$$

Rearranging and noting that $\left\langle \nabla_g Z_{g^*_\beta}, g^*_\beta - g \right\rangle = Z_{g^*_\beta} - Z_g$, we obtain:

$$2\beta(Z_{g^*_\beta} - 1)(Z_{g^*_\beta} - Z_g) \le \left\langle \nabla_g L(\lambda^*_\beta, g^*_\beta), g - g^*_\beta \right\rangle.$$

To prove sufficient reach, let $m_{\min} = \sum_x \min_k \widehat{\pi}_k(x)$ and $m_{\max} = \sum_x \max_k \widehat{\pi}_k(x)$. Since experts are probability distributions, $m_{\min} \le 1 \le m_{\max}$. If $Z_{g^*_\beta} > 1$, choosing g as the mass-minimizing gate yields $Z_g = m_{\min} \le 1$, so $(Z_{g^*_\beta} - Z_g) \ge (Z_{g^*_\beta} - 1)$. If $Z_{g^*_\beta} < 1$, choosing g as the mass-maximizing gate yields $Z_g = m_{\max} \ge 1$, so $(Z_g - Z_{g^*_\beta}) \ge (1 - Z_{g^*_\beta})$. In both cases, $2\beta(Z_{g^*_\beta} - 1)^2 \le \left\langle \nabla_g L, g - g^*_\beta \right\rangle$.

Evaluating the inner product exactly:

$$\left\langle \nabla_g L(\lambda^*_\beta, g^*_\beta), g - g^*_\beta \right\rangle = \sum_x \widehat{\mathsf{p}}_{\lambda^*_\beta}(x)\left(1 - \frac{\pi_g(x)}{\pi_{g^*_\beta}(x)}\right),$$

using $\nabla_{g(x,k)} L(\lambda^*_\beta, g^*_\beta) = -\widehat{\pi}_k(x)\,\widehat{\mathsf{p}}_{\lambda^*_\beta}(x)/\pi_{g^*_\beta}(x)$. Since $\pi_g(x)/\pi_{g^*_\beta}(x) \ge 0$ for all x, each summand is at most 1, and $\widehat{\mathsf{p}}_{\lambda^*_\beta}$ is a probability distribution summing to 1. Hence $\left\langle \nabla_g L, g - g^*_\beta \right\rangle \le 1$, giving $2\beta|Z_{g^*_\beta} - 1|^2 \le 1$. □

Scalability. The bound $|Z_{g^\star_\beta} - 1| \le 1/\sqrt{2\beta}$ is *dimension-free*: it depends only on the regularization strength β and is independent of both the support size $|\mathcal{X}_0|$ and the gradient magnitude. Our empirical results in Section 5 confirm that the mass Z_g stabilizes quickly at moderate β, consistent with this $\mathcal{O}(1/\sqrt{\beta})$ rate.

The Regularization Trade-off ($\beta \to \infty$). Taking $\beta \to \infty$ forces $|Z_g - 1| \to 0$, but an overly aggressive penalty can override the local data fit, pushing the gate toward configurations that minimize global mass error at the expense of the log-likelihood objective. Thus, β must be carefully tuned to balance global mass stability against the quality of the local log-likelihood fit.

C.7 Generalization and Sample Efficiency

While the robust guarantees established previously ensure that an optimal gate exists in the locally-normalized family, we must evaluate the statistical cost of learning this gate from finite data. Unlike the bounded family, the $\mathcal{G}_0$ family operates as a product of independent simplices across the entire input space $\mathcal{X}_0$, granting it the highest possible capacity within the modular framework.

Hypothesis Spaces. The generalization gap for modular mixtures is governed by the Rademacher complexity $\mathfrak{R}_m$ of the gating class. The two unnormalized families constrain the gate in structurally different ways:

- $\mathcal{G}_{\infty,1}$ (Expert-Independent Global Ceiling): This family enforces the expert-independent norm $\sum_k \max_x g(x,k) \le 1$. The ceiling couples gating decisions across inputs (raising g at one input consumes capacity available at every other input) and bounds every gate value by 1.
- $\mathcal{G}_0$ (Pointwise Simplex Constraint): In contrast, $\mathcal{G}_0$ allows independent gating decisions at every point $x \in \mathcal{X}_0$. As proven in Section C.3, the robust solution $g^\star$ is a pointwise maximizer, and the hypothesis space is the product of $|\mathcal{X}_0|$ independent simplices. Without parametric constraints (like those imposed by a shared neural network), a $\mathcal{G}_0$ gate is a lookup table that can fit any assignment on the training sample, so it carries no inductive bias toward unseen inputs, even though each gate value is bounded by 1.

Complexity Comparison. The proven statement is the comparison of Theorem 8 between the bounded relaxation and the globally normalized family: under Assumption 7, $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1}) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_1)$ for $p \ge 3$ (and up to a vanishing $\mathcal{O}(1/\sqrt{m})$ term for $p = 2$). The mechanism is that the single global constraint of $\mathcal{G}_1$ lets the gate realize any expert assignment on the sample by compensating off-sample, so on a sample $\mathcal{G}_1$ can behave as the lookup table $\mathcal{G}_0$, whereas the $\mathcal{G}_{\infty,1}$ ceiling caps the total weight each expert receives across all inputs. Since $\mathcal{G}_1 \subseteq \mathcal{G}_0$ and each $g(x_i,\cdot) \in \Delta([1,p])$, both families satisfy $\widehat{\mathfrak{R}}_S(\mathcal{G}_1) \le \widehat{\mathfrak{R}}_S(\mathcal{G}_0) \le 1$ for any sample S. However, $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ are not nested (a $\mathcal{G}_0$ gate may violate the global ceiling, and a $\mathcal{G}_{\infty,1}$ gate need not lie on the pointwise simplex), so the only ordering between $\widehat{\mathfrak{R}}_S(\mathcal{G}_{\infty,1})$ and $\widehat{\mathfrak{R}}_S(\mathcal{G}_0)$ available is the one implied by Theorem 8. Moreover, a bounded Rademacher complexity for $\mathcal{G}_0$ is not by itself a generalization guarantee: without parametric constraints, $\mathcal{G}_0$ is a lookup table with $|\mathcal{X}_0|$ independent simplices, and its practical utility relies on neural parameterization or structural distillation to impose inductive bias.

Implications for Sample Efficiency. This analysis clarifies the trade-off inherent in local normalization. The $\mathcal{G}_{\infty,1}$ family combines bounded pointwise values with cross-input coupling via the global ceiling, and its generalization gap is controlled by Theorem 19 together with the comparison of Theorem 8. Learning a robust $\mathcal{G}_0$ gate requires either a neural parameterization to impose structure or, as established in the next section, a rigorous structural distillation into a causal student to maintain the robust inheritance.

C.8 Inference: Approximate Sampling-Importance-Resampling and Structural Distillation

Since the $\mathcal{G}_0$ family does not satisfy the $Z_g \le 1$ normalization bound, it cannot benefit from the tight, efficient rejection sampling envelope available to the $\mathcal{G}_{\infty,1}$ family. An exact rejection sampler does exist via the uniform mixture envelope $\mathsf{q}(x) = \frac{1}{p}\sum_k \widehat{\pi}_k(x)$ (with envelope mass p and acceptance rate Z_g/p, about $1/p$ for the mass-regularized gate of Section C.6 used in our experiments). Given

this inefficiency, we provide two primary practical pathways for inference, with a strong preference for structural distillation in autoregressive settings.

Approximate SIR. For a regularized gate where the mass is approximately unity ($Z_g \approx 1$), it is possible to perform *Sampling-Importance-Resampling* (SIR), a Monte Carlo technique that draws candidates from a proposal distribution and filters them based on their importance weights $w \propto \pi_g(x)/q(x)$ to approximate the normalized density. This requires estimating the partition function $\widehat{Z}_g$ via Monte Carlo sampling from a proposal distribution and rescaling the mixture density for importance weighting. However, because $\mathcal{G}_0$ lacks a strict global upper bound on its mass, the variance of this estimator can be extremely high. Thus, SIR may suffer from severe sampling bias and poor sample efficiency, potentially degrading generative quality.

Structural Distillation. The most effective deployment strategy for the $\mathcal{G}_0$ family is to treat the robust gate g^* as a non-causal *teacher* for structural distillation. We train a causal student router $\gamma_\phi(k \mid x_{<t})$ to minimize the forward KL divergence from the normalized teacher p_{g^*} to the student π_γ:

$$\min_\phi \mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) = - \mathop{\mathbb{E}}_{x \sim p_{g^*}} \left[\sum_{t=1}^{|x|} \log \left(\sum_{k=1}^{p} \gamma_\phi(x_{<t}, k) \widehat{\pi}_k(x_t \mid x_{<t}) \right) \right] + \text{const.}$$

Theorem 29 (Distillation Performance Inheritance for $\mathcal{G}_0$)**.** *Let π_γ be the causal mixture model induced by the student router γ_ϕ. If the distillation divergence satisfies $\mathsf{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma) \le \delta$, then for any mixture $\lambda \in \Delta$:*

(a) ***Total Variation bound:*** *$\| \widehat{\mathsf{p}}_\lambda - \pi_\gamma \|_1 \le \| \widehat{\mathsf{p}}_\lambda - p_{g^*} \|_1 + \sqrt{2\delta}$.*

(b) ***NLL bound:*** *Under Assumption 18, $\mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, \pi_\gamma) \le \mathcal{L}_{\log}(\widehat{\mathsf{p}}_\lambda, p_{g^*}) + \delta + C\sqrt{2\delta}$, where $C = \| \widehat{\mathsf{p}}_\lambda - p_{g^*} \|_1 / \eta'$ and $\eta' = \min_{x \in \mathfrak{X}_0} \min\{p_{g^*}(x), \pi_\gamma(x)\} > 0$, as in Theorem 10.*

Proof. By the exact same proof as Theorem 10:

(a) By Pinsker's inequality and the triangle inequality, the total variation distance is bounded.
(b) By change of measure and log-linearization (Appendix B.13.5), the NLL gap is bounded by $\delta + C\sqrt{2\delta}$.

□

Inheritance of Selection Gain. Combining the theorem with the robust guarantees, the causal student directly inherits the *pointwise selection gain* ($\mathcal{A}^{\lambda^*}_{\text{selection}}$) achieved by the robust teacher, up to the distillation error $\delta + C\sqrt{2\delta}$ and the volume penalty. When using the regularized game (Section C.6), the Volume Penalty $\log Z_{g^*}$ is actively constrained. This establishes structural distillation as a principled path for real-time, autoregressive generation without requiring an intermediate causal projection gap.

Remark (Scaling of C with p). The constant $\eta' = \min_{x \in \mathfrak{X}_0} \min\{p_{g^*}(x), \pi_\gamma(x)\}$ has two components. The student component is the same for both families: under the token-level part of Assumption 18, $\pi_\gamma(x) \ge e^{-TM_{\text{tok}}}$, independently of p. The teacher component depends on the family. For $\mathcal{G}_{\infty,1}$ with a non-degenerate gate $g^* \in \mathcal{G}^\eta_{\infty,1}$, the bounded mass $Z_{g^*} \le 1$ ensures $p_{g^*}(x) \ge \eta e^{-M}$, so this component is independent of p. For $\mathcal{G}_0$, the pointwise simplex constraint gives $\pi_{g^*}(x) \ge e^{-M}$, but the unconstrained mass can reach $Z_{g^*} \le p$, so that $p_{g^*}(x) \ge e^{-M}/p$ in the worst case and the teacher component of C can be $\mathcal{O}(p)$. The regularized game (Theorem 28) mitigates this by constraining $Z_{g^*_\beta} \approx 1$, recovering the same p-independent teacher component as $\mathcal{G}_{\infty,1}$.

D Implementation Details

D.1 Exact Euclidean Projection onto $\mathcal{G}_{\infty,1}$

The $\mathcal{G}_{\infty,1}$ constraint requires $\sum_{k=1}^p \max_{x \in \mathfrak{X}_0} g(x,k) \le 1$ and $g \ge 0$. Given an intermediate gate g', the Euclidean projection is the solution to:

$$\Pi_{\mathcal{G}_{\infty,1}}(g') = \mathop{\operatorname{argmin}}_{g \ge 0,\ \sum_{k=1}^p \max_x g(x,k) \le 1} \| g - g' \|_2^2.$$

The projection decouples into two stages. First, for a fixed capacity vector $w \geq 0$ with $\sum_k w_k \leq 1$, the optimal gate is the coordinate-wise clipping $g_w(x,k) = \min(w_k, (g'(x,k))_+)$, where $(a)_+ = \max(a, 0)$. Second, the remaining problem over w is convex and admits a unique Lagrange multiplier $\mu^* \geq 0$ enforcing $\sum_k w_k(\mu^*) = 1$, where $w_k(\mu)$ is the optimal capacity at multiplier μ. The multiplier is found by bisection. Bisection converges to machine precision, so the projection is exact to any desired tolerance $\epsilon > 0$.

Algorithm 5 Exact Euclidean Projection onto $\mathcal{G}_{\infty,1}$

1: **Input:** Intermediate gate values $g'(x,k)$ for all $x \in \mathcal{X}_0, k \in [p]$; tolerance $\epsilon > 0$.
2: **Step 1: Clip negatives.** For each x, k: $g^+(x,k) \leftarrow \max(g'(x,k), 0)$.
3: **Step 2: Check feasibility.** Compute $w'_k = \max_x g^+(x,k)$.
4: If $\sum_k w'_k \leq 1$: **return** $g = g^+$.
5: **Step 3: Bisection on μ.** Find $\mu^* \geq 0$ such that $\sum_k w_k(\mu^*) = 1$:
6: For each candidate μ, compute $w_k(\mu) = \widetilde{w}_k$, where $\widetilde{w}_k$ solves $\min_{w_k \geq 0} \frac{1}{2}\sum_x (\min(w_k, g^+(x,k)) - g'(x,k))^2 + \mu w_k$. ▷ 1D subproblem per expert; $\widetilde{w}_k \leq \max_x g^+(x,k)$ automatically
7: Bisect on μ until $|\sum_k w_k(\mu) - 1| < \epsilon$.
8: **Step 4: Clip to capacity.** For each x, k: $g(x,k) \leftarrow \min(w_k(\mu^*), g^+(x,k))$.
9: **Output:** Projected gate g.

Proposition 30 (Correctness of Algorithm 5)**.** *For any input g', Algorithm 5 computes the unique Euclidean projection $\Pi_{\mathcal{G}_{\infty,1}}(g')$, exactly up to the bisection tolerance ϵ on the capacity constraint.*

Proof. The feasible set $\mathcal{G}_{\infty,1}$ is non-empty, closed, and convex (Lemma 2), so the projection exists and is unique. Introduce auxiliary capacity variables $w \in \mathbb{R}^p$ and rewrite the projection problem equivalently as

$$\min_{w \geq 0,\, \sum_k w_k \leq 1} \; \min_{0 \leq g(\cdot,k) \leq w_k} \frac{1}{2} \sum_{x,k} (g(x,k) - g'(x,k))^2,$$

since any feasible g satisfies $0 \leq g(x,k) \leq \max_{x'} g(x',k) =: w_k$ with $\sum_k w_k \leq 1$, and conversely. For fixed w, the inner problem is separable across coordinates: the minimizer of $(g - g'(x,k))^2$ over the interval $[0, w_k]$ is the clipped value $g_w(x,k) = \min(w_k, (g'(x,k))_+)$, exactly Steps 1 and 4. Substituting yields the reduced objective $\phi(w) = \frac{1}{2}\sum_k \sum_x (\min(w_k, (g'(x,k))_+) - g'(x,k))^2$, which is separable across k and convex in each w_k: for coordinates with $g'(x,k) \leq 0$ the term is constant in w_k, while for $g'(x,k) > 0$ it equals $\frac{1}{2}((g'(x,k) - w_k)_+)^2$, a convex non-increasing function. If the unconstrained minimizer $w'_k = \max_x (g'(x,k))_+$ is feasible ($\sum_k w'_k \leq 1$), it is optimal (Step 2). Otherwise, since ϕ is convex and Slater's condition holds, the KKT conditions apply with a single multiplier $\mu \geq 0$ for the active constraint $\sum_k w_k = 1$: the optimal $w_k(\mu)$ solves the 1-D convex subproblem $\min_{w_k \geq 0} \phi_k(w_k) + \mu w_k$ (Step 3), whose minimizer is unique for $\mu > 0$ and never exceeds $\max_x (g'(x,k))_+$ (beyond that point ϕ_k is constant while μw_k increases), and $w_k(\mu)$ is non-increasing and continuous in μ, so $\sum_k w_k(\mu)$ is monotone and the unique μ^* with $\sum_k w_k(\mu^*) = 1$ is found by bisection to any tolerance ϵ. The resulting $(w(\mu^*), g_{w(\mu^*)})$ satisfies the KKT conditions of the joint convex problem and is therefore the Euclidean projection. □

Complexity. Each bisection iteration evaluates $\sum_k w_k(\mu)$ in $\mathcal{O}(p \cdot |\mathcal{X}_0|)$ time (sorting within each expert gives the per-expert subproblem in $\mathcal{O}(|\mathcal{X}_0| \log |\mathcal{X}_0|)$). The total complexity is $\mathcal{O}(p \cdot |\mathcal{X}_0| \cdot \log |\mathcal{X}_0| + p \cdot |\mathcal{X}_0| \cdot \log(1/\epsilon))$, with explicit dependence on the numerical tolerance ϵ. Unlike the $\mathcal{G}_1$ projection, this procedure is expert-free: it never evaluates the expert likelihoods $\widehat{\pi}_k(x)$.

Projection onto the non-degenerate family $\mathcal{G}^\eta_{\infty,1}$. The convergence analysis of Theorem 6 projects onto $\mathcal{G}^\eta_{\infty,1} = \mathcal{G}_{\infty,1} \cap \bigcap_{x \in \mathcal{X}_0} H_x$, where $H_x = \{g : \sum_{k=1}^p g(x,k) \geq \eta\}$ is a half-space. Algorithm 5 alone projects onto $\mathcal{G}_{\infty,1}$ and may return a gate that violates a floor constraint (for instance, a row of g' with all entries negative is mapped to zero). The Euclidean projection onto the intersection is obtained by Dykstra's alternating projection method (Boyle & Dykstra, 1986) between $\Pi_{\mathcal{G}_{\infty,1}}$ (Algorithm 5) and the half-space projections, each of which is closed-form: Π_{H_x} adds $(\eta - \sum_k g(x,k))_+ / p$ to every entry of row x. Since all sets are closed and convex, the iterates converge to the exact projection $\Pi_{\mathcal{G}^\eta_{\infty,1}}(g')$, and running the scheme to tolerance ϵ yields an approximate projection with $\epsilon_{\text{proj}} = \mathcal{O}(\epsilon)$, which is the setting covered by Remark 17. Each Dykstra iteration costs one call to Algorithm 5 plus $\mathcal{O}(p \cdot |\mathcal{X}_0|)$ for the half-spaces.

Remark (Separation from experiments). All deep learning experiments (Section 5, Appendix G) use the parametric soft-penalty implementation (Algorithm 1). Algorithm 5, combined with the half-space projections above, is the non-parametric exact projection used solely in the convergence proof of Theorem 6.

D.2 Structural Distillation Procedure

The distillation training loop generates sequences from the normalized teacher p_{g^*} and trains the causal router by next-token prediction.

Algorithm 6 Structural Distillation Training Loop

Input: Robust teacher gate g^*, student router γ_ϕ, frozen experts $\{\widehat{\pi}_k\}_{k=1}^p$.
for each training iteration **do**
  Sample sequence $x \sim p_{g^*}$ via Exact Rejection Sampling (Section 3.4).
  for each token $t \in [1, |x|]$ **do**
    Compute student weights $\mathbf{w}_t = \gamma_\phi(x_{<t}, \cdot)$.
    $\mathcal{L}_t = -\log(\mathbf{w}_t \cdot \mathbf{P}_t)$, where $\mathbf{P}_t = [\widehat{\pi}_1(x_t \mid x_{<t}), \ldots, \widehat{\pi}_p(x_t \mid x_{<t})]$.
  end for
  Update ϕ via gradient descent on $\mathcal{L} = \sum_t \mathcal{L}_t$.
end for
Output: Causal student router γ_ϕ.

When using the cached-logit variant (Algorithm 3), the expert evaluations are performed once in an offline phase, so each training iteration involves only a forward pass through the lightweight router and an $\mathcal{O}(p)$ inner product.

E Evaluation of Normalized NLL for Unnormalized Families

To ensure a rigorous and fair comparison between the unnormalized $\mathcal{G}_{\infty,1}$ family and the globally normalized $\mathcal{G}_1$ baseline, we must evaluate the standard Negative Log-Likelihood (NLL) on held-out test data.

Since the optimal $\mathcal{G}_{\infty,1}$ gate actively sub-normalizes to yield the Normalization Bonus ($Z_{g^*} \le M^* \le 1$), the raw unnormalized likelihood $-\log \pi_g(x)$ will be artificially penalized relative to a strictly normalized model. Thus, it must be explicitly corrected to report the true generative performance of the underlying normalized distribution:

$$\mathrm{NLL}_{\mathrm{normalized}} = \mathop{\mathbb{E}}_{x \sim p_{\mathrm{test}}} [-\log \pi_g(x)] + \log Z_g.$$

All reported values are per token: the sequence-level quantity $-\log \pi_g(x) + \log Z_g$ is summed over the test set and divided by the total number of tokens, so the correction enters as $\log Z_g / |x|$ per token of a sequence of length $|x|$, and the same token-weighted average is used for the causal baselines. Since evaluating Z_g exactly requires an intractable summation over the entire support space $\mathcal{X}_0$ for sequence models, we must estimate it. We outline two primary methods for this estimation: a standard Importance Sampling approach and an efficient, near-zero variance Rejection Sampling approach derived from our theoretical bounds.

E.1 Method 1: Monte Carlo and Importance Sampling

The most straightforward approach to estimating Z_g is via Monte Carlo integration. A naive uniform sampling over $\mathcal{X}_0$ suffers from catastrophic variance in high-dimensional spaces. Instead, we can use Importance Sampling using a proposal distribution $q(x)$ that has full support over the data space.

A natural choice for the proposal is the uniform mixture of the base experts, $q(x) = \frac{1}{p} \sum_{k=1}^p \widehat{\pi}_k(x)$. We draw N samples from $q(x)$ and estimate the mass via the importance weights:

$$\widehat{Z}_g^{\mathrm{IS}} = \frac{1}{N} \sum_{i=1}^N \frac{\pi_g(x_i)}{q(x_i)}.$$

While statistically unbiased, this estimator's variance depends heavily on the divergence between the proposal $q(x)$ and the unnormalized target $\pi_g(x)$. If the robust gate $g(x,k)$ is specialized (i.e., highly sparse or peaked), the discrepancy between the uniform expert mixture and the dynamically gated mixture can still lead to high-variance estimates of Z_g, requiring a number of samples N to achieve stable NLL reporting.

E.2 Method 2: Acceptance-Rate Mass Estimation (Rejection Sampling)

We can obtain a vastly superior, low-variance estimate of Z_g almost for free by leveraging the Exact Rejection Sampling protocol established in Section B.12.

Recall that the $\mathcal{G}_{\infty,1}$ family allows a tightly bounded proposal envelope $M^* q_{w^*}(x)$, where $w_k^* = \max_{x \in \mathfrak{X}_0} g(x,k)$ and $M^* = \sum_k w_k^* \le 1$. As proven in Proposition 9, the exact theoretical acceptance rate of this rejection sampling protocol is exactly the ratio of the target mass to the envelope mass:

$$\mathbb{P}(\text{accept}) = \mathop{\mathbb{E}}_{x \sim q_{w^*}}[\alpha(x)] = \frac{Z_g}{M^*}.$$

This theoretical acceptance rate can be estimated via a standard Monte Carlo average. This yields an efficient estimation protocol:

1. **Generate Proposals:** Draw N samples $\{x_i\}_{i=1}^N$ from the envelope proposal $q_{w^*}(x)$.
2. **Compute Empirical Acceptance:** Apply the acceptance criterion $\alpha(x) = \frac{\pi_g(x)}{M^* q_{w^*}(x)}$ to compute the empirical acceptance rate, which is the Monte Carlo estimate for $\mathbb{P}(\text{accept})$:
$$\widehat{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\text{accepted}_i) \approx \mathbb{P}(\text{accept}).$$
3. **Estimate Mass:** By substituting the empirical estimate into the identity above and solving for Z_g, we arrive at the estimator:
$$\widehat{\alpha} \approx \frac{Z_g}{M^*} \Rightarrow Z_g \approx M^* \times \widehat{\alpha}.$$
Thus, our final estimator is $\widehat{Z}_g^{\text{RS}} = M^* \times \widehat{\alpha}$.

E.3 Properties of the Acceptance-Rate Estimator

Variance Reduction. This method acts as a specialized form of importance sampling where the proposal envelope strictly upper-bounds the target density, guaranteeing that the importance weights $\alpha(x)$ are bounded within $[0,1]$. The low variance of this estimator stems from the tightness of the optimal envelope q_{w^*}, which concentrates proposal mass on the target's support. When the acceptance rate $\mathbb{P}(\text{accept}) \approx 1$, the estimator has near-zero variance regardless of the dimensionality of the sequence space.

On Using M^* as a Direct Estimator. One might wonder if M^* itself could serve as a direct, simpler estimate for Z_g, given that the ratio $\widehat{\alpha} = Z_g/M^*$ is mathematically enforced by the protocol. However, this is inadvisable for a rigorous evaluation. Using M^* as an estimator is equivalent to assuming that $\mathbb{P}(\text{accept}) = 1$ exactly, which incorrectly implies $Z_g = M^*$. The term $M^*(1-\widehat{\alpha})$ represents the estimated *mass gap*, the structurally unallocated capacity within the bounded envelope. Discarding $\widehat{\alpha}$ and using only M^* would ignore this physical gap and introduce a systematic upward bias into the estimate of Z_g, artificially degrading the reported NLL (since $\log M^* > \log Z_g$). Therefore, the unbiased estimator $\widehat{Z}_g^{\text{RS}}$ remains the correct choice. Since the NLL involves $\log Z_g$ rather than Z_g, the plug-in $\log \widehat{Z}_g^{\text{RS}}$ is biased. Unconditionally, Jensen's inequality gives $\mathbb{E}[\log \widehat{Z}_g^{\text{RS}}] \le \log Z_g$, but the zero-acceptance event, of probability $(1-\mathbb{P}(\text{accept}))^N$ for N proposals, makes this expectation $-\infty$; conditioning on at least one acceptance removes the event but also the exact unbiasedness of $\widehat{Z}_g^{\text{RS}}$ (its conditional mean is $Z_g/(1-(1-\mathbb{P}(\text{accept}))^N) > Z_g$). The relevant statement is asymptotic: by the delta method,

$$\mathbb{E}[\log \widehat{Z}_g^{\text{RS}}] - \log Z_g = -\frac{1-\mathbb{P}(\text{accept})}{2N\,\mathbb{P}(\text{accept})} + \mathcal{O}(N^{-2}),$$

with relative standard error $\sqrt{(1-\mathbb{P}(\text{accept}))/(N\,\mathbb{P}(\text{accept}))}$. We use $N = 10{,}000$ proposals; at the observed $\mathbb{P}(\text{accept}) \approx 0.82$, the bias is about -1.1×10^{-5} nats and the standard error of $\log \widehat{Z}_g^{\text{RS}}$ about 0.005 nats, below the run-to-run standard deviations reported in the tables. Averaging the acceptance probabilities $\alpha(x_i)$ of the proposals instead of the accept indicators yields an estimator with the same expectation and lower variance (a Rao–Blackwellized variant) that is positive whenever some $\alpha(x_i) > 0$, and coincides with importance sampling under the envelope proposal q_{w^*}.

F Comprehensive Limitations and Societal Impact

In this section, we provide a detailed analysis of the technical limitations, potential failure modes, and broader societal implications of our robust modular learning framework.

F.1 Technical Limitations and Failure Modes

While the $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ families offer significant computational advantages, they are not without trade-offs.

- **Mass Explosion in $\mathcal{G}_0$:** A primary failure mode for the locally-normalized family is *mass explosion*. Unlike $\mathcal{G}_1$ or $\mathcal{G}_{\infty,1}$, the global mass Z_g in $\mathcal{G}_0$ is not naturally bounded by unity. Although $Z_g \le p$ (since each expert sums to 1), if the quadratic regularization parameter β (Section C.6) is tuned too loosely, the partition function can reach $\mathcal{O}(p)$. This results in a Volume Penalty in the robustness bound and renders native Importance Sampling (SIR) practically impossible due to extreme variance.
- **The Distillation Bottleneck:** Our *autoregressive* inference pathway relies on Structural Distillation. While we provide inheritance theorems (Theorems 10 and 29), the student router's performance is ultimately limited by its own capacity. If the causal router γ_ϕ is significantly under-parameterized, it may fail to capture the teacher's optimal sequence-level geometry, leading to a *robustness leak* where the student performs worse than a monolithic baseline.
- **Rejection Sampling Efficiency:** While Proposition 9 guarantees that the acceptance rate for $\mathcal{G}_{\infty,1}$ is bounded below by $\exp(-(C_T + 1))$, this bound can degrade in regimes where domain overlap is high, causing the envelope mass M_T to significantly exceed Z_{g_T}. In practice, the sparsity of specialized experts in our experiments yields acceptance rates near unity, but deployments with highly redundant experts should monitor this metric.
- **Parametric Non-Convexity:** Our convergence guarantees (Theorems 6 and 27) are established for function-space optimization. In practical deep learning settings, the gating functions are parameterized by non-convex neural networks. While we observe stable convergence in our experiments, standard adversarial training challenges, such as vanishing gradients or mode collapse, could still occur in high-dimensional or poorly initialized regimes.
- **Dependence on Frozen Expert Quality:** The modular framework's performance is fundamentally bounded by the quality of the base experts $\widehat{\pi}_k$. If the pre-trained experts are weak or poorly calibrated on their respective domains, the ensemble inherits these deficiencies regardless of gate optimization. The minimax guarantees (Theorems 5 and 12) bound the excess risk relative to the expert errors ϵ_k, which cannot be improved by gating alone.
- **Inference Cost of the Causal Student:** The distilled student π_γ mixes the next-token distributions of all p experts at every position, so autoregressive generation requires one forward pass per expert per token (or a shared-trunk implementation with p output heads). This is the same per-token cost as dense mixture-of-experts inference and is p times that of a single expert; sparsifying the student router (e.g., top-k routing) would reduce this cost but is not covered by our inheritance guarantees.
- **Training Budget of the Largest Models:** The 1B+ experts and baselines are trained on a fixed budget of 120,000 sequences per domain (Appendix G.1) and are therefore far from convergence by the standards of production-scale pretraining. Our conclusions at that scale concern the *relative* behavior of the gating families under a matched compute budget, not the absolute quality of the resulting language models.
- **Assumption of Known Domain Boundaries:** The current framework assumes that the training data can be partitioned into known domain-specific datasets $D_1, \ldots, D_p$, each associated with a pre-trained expert. In practice, domain boundaries may be ambiguous (e.g., a medical document

containing legal terminology), and automatically discovering the optimal domain decomposition remains an open problem.

F.2 Environmental Impact: Modularity for Greener AI

The current paradigm of training monolithic LLMs is characterized by a task mixture that is energetically unsustainable (Strubell et al., 2019; Schwartz et al., 2020). Our framework offers a principled path toward Sustainable AI:

- **Resource Efficiency:** By freezing high-capacity experts (e.g., 7B parameter modules) and training only a lightweight gating network, we reduce the number of active gradients by orders of magnitude.
- **Model Reuse:** Our framework transforms pre-trained models into *reusable resources*. Instead of discarding a generalist model when new domains emerge, users can simply "plug-and-play" new experts into an existing robust ensemble, significantly extending the lifecycle of specialized modules.

F.3 Privacy and Data Sovereignty

The synchronization bottleneck of $\mathcal{G}_1$ is not just a computational hurdle; it is a privacy barrier. The need for global coordination often requires centralizing datasets, which violates many modern privacy regulations.

- **Decentralized Data Access:** Since $\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$ allow for decoupled updates, our framework is uniquely compatible with Federated Learning. Domain-specific owners (e.g., medical or financial institutions) can train their own specialized experts $\widehat{\pi}_k$ locally. They only need to provide the robust gate with likelihood access (inference-only) to update the routing policy, without ever sharing the raw, sensitive training data. (This assumes that frozen expert weights are distributed to data owners, who compute likelihoods locally. If instead data must be sent to remote expert APIs, standard privacy-preserving mechanisms such as differential privacy or secure multi-party computation are required.)
- **Modular Marketplaces:** This work provides the statistical foundation for modular AI marketplaces, where specialized expertise can be traded and combined securely, ensuring that the aggregated output remains robust regardless of the adversarial shifts in the data mixture.

F.4 Potential Negative Societal Impacts

While the efficiency gains from unnormalized gating are broadly beneficial, they also lower the computational barrier to training large-scale generative models. This democratization carries dual-use risks: malicious actors could more easily train models for generating disinformation, deepfakes, or other harmful content. Additionally, modular architectures that combine domain-specific experts may amplify biases present in individual expert training corpora, particularly if the domain-specific datasets are imbalanced or reflect systemic biases. We encourage practitioners deploying modular LLMs to adopt standard safety measures, including output filtering, bias auditing, and responsible disclosure practices.

G Experimental Results

In this section, we provide comprehensive experimental results alongside a detailed setup, expanding upon the findings presented in Section 5.

To empirically validate the theoretical phase transitions and generative capabilities of the unnormalized gating families, we evaluate our proposed architectures ($\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$) against the globally normalized baseline ($\mathcal{G}_1$) and monolithic models on real-world text data. To ensure a fair comparison, we adopt the multi-domain experimental setup established in prior work (Cortes et al., 2026).

G.1 Experimental Setup

Model Architecture and Scaling. We conduct our experiments across distinct capacity regimes to evaluate both training dynamics and large-scale efficacy:

- **Lightweight Regime (~6.5M Parameters):** To allow for rapid prototyping and to directly observe the optimization trajectories, we train specialized expert models, one for each domain. Each expert is a causal Transformer decoder (Vaswani et al., 2017) containing 2 layers, 2 attention heads, and a hidden embedding dimension of 64 (~6.5M parameters). The *Monolithic* and standard MoE baselines are capacity-matched to the combined modular system: their parameter count matches the sum of the three experts plus the gating network.
- **20-Task Scaling Regime (~130M Parameters total):** To empirically isolate the $\mathcal{O}(\sqrt{\log p})$ capacity cost and observe the theoretical adversarial limit-cycle dynamics without catastrophic overfitting, we constructed a 20-expert ensemble. Each of the 20 experts matches the ~6.5M parameter architecture, yielding an ensemble of ~130M parameters. The Monolithic and standard MoE baselines were correspondingly scaled to ~132M parameters to ensure strict capacity matching, accounting for both the experts and the additional parameters of the gating networks.
- **1B+ Parameter Regime (~1.6B Parameters):** To validate our findings at a scale reflective of modern large language models, we evaluate massive specialized experts. Each uses a GPT-2 XL-style configuration (Radford et al., 2019) with 48 layers and an embedding dimension of 1600 (~1.6B parameters). As in the other regimes, the Monolithic and standard MoE baselines are scaled to match the combined capacity of the three experts plus the gating network.
- **Pre-Trained Frozen Experts Regime (~124M Parameters):** To evaluate the framework on established foundation models (Bommasani et al., 2021), we instantiate the base GPT-2 model twice to serve as both the Wikipedia and FineWeb experts, alongside CodeParrot-small (von Werra, 2021) for the Code domain. This deliberately creates a stress test: with parametrically identical experts, the gating network cannot exploit differential expert competence and must rely solely on the structural properties of the framework. The fact that the modular ensemble still matches or exceeds the monolithic baseline in this adversarial setup validates the robustness of the approach. We used exactly 15,000 training and 4,000 test sequences per domain to maximize the dataset capacity and evaluate large-scale performance within the realistic constraints of contemporary LLM components.

The gating networks (g_θ) for all three families ($\mathcal{G}_1$, $\mathcal{G}_{\infty,1}$, $\mathcal{G}_0$) are parameterized as bidirectional Transformer encoders. For the causal student router, we employ a unidirectional Transformer decoder with the same hidden dimension. We evaluate two configurations: a lightweight scale consisting of 2 layers, 2 heads, and an internal dimension of 128 (~290K parameters), and a larger scale configured with 4 layers, 2 heads, and an internal dimension of 256 (~2.5M parameters). These encoders apply an initial projection from the expert embedding space, followed by global mean pooling and a final projection layer corresponding to the specific routing constraint. The gating networks are randomly initialized and trained directly by the robust minimax loop, with no classifier warm start.

Datasets and Tokenization. We draw our data from distinct, large-scale open-source repositories. For the natural language domains, we use Wikipedia (`wikimedia/wikipedia`) (Wikimedia Foundation, 2024) and FineWeb-Edu (`HuggingFaceFW/fineweb-edu`) (Penedo et al., 2024). For the code domain, the datasets vary by setting: the pre-trained frozen expert setting exclusively uses CodeAlpaca-20k (`sahil2801/CodeAlpaca-20k`) (Chaudhary, 2023), whereas all other settings (Lightweight, 1B+, and the 20-task code generation) use The Stack Smol (`bigcode/the-stack-smol`) (Kocetkov et al., 2022). For the Lightweight and 1B+ regimes, we use 120,000 training and 30,000 test sequences per domain. Note that these sizes are larger than the 80,000 training and 20,000 test sequences per domain used in the otherwise identical multi-domain setup of Cortes et al. (2026): together with varying only the training splits across runs while keeping the test sets fixed, this substantially reduces run-to-run variance (e.g., the Code expert's standard deviation drops from ± 0.788 to ± 0.025). Absolute NLL values are therefore not directly comparable to those reported in Table 1 of Cortes et al. (2026) (whose *Retrained* model corresponds to our Monolithic model). For the 20-task setting specifically, we used 20 distinct programming languages from The Stack Smol, extracting exactly 8,000 training and 2,000 testing sequences per language. Crucially, all schema metadata (e.g., `lang`, `repository_name`) was explicitly stripped to force the gating network to gate purely based on code syntax. To ensure zero data leakage and reduce

evaluation variance across multiple runs, we keep the test sets strictly fixed at a specific offset across the 5 parallel runs. For the Lightweight and 1B+ regimes, we report two protocols: varying the initialization seeds (Diff Seed) or the training data splits (Diff Data). For the 20-task and pre-trained frozen experts regimes, the 5 runs vary only the initialization seeds. All data is tokenized using the standard GPT-2 byte-pair encoding (BPE) tokenizer (Sennrich et al., 2016), padded and truncated to a maximum sequence length of 128 tokens. Unless otherwise specified, figures illustrate representative training trajectories or averages across 5 independent runs. All datasets and models are used under their respective open-source licenses: Wikipedia (CC-BY-SA 3.0), FineWeb-Edu (ODC-BY 1.0), The Stack Smol (BigCode OpenRAIL-M), CodeAlpaca-20k (Apache 2.0), and GPT-2/CodeParrot (MIT license).

Training Details. The base experts are pre-trained for 4 epochs using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 10^{-3} and a batch size of 32. The Monolithic and standard MoE models are trained for 4 epochs with a learning rate of 5×10^{-4}, where the MoE uses standard load-balancing heuristics with a coefficient of 0.01 (Shazeer et al., 2017). This design reflects the practical deployment scenario: domain experts are pre-existing, frozen models with their own training history, while the baselines in the Lightweight, 20-task, and 1B+ regimes train from random initialization (capacity-matched architecture, no pre-trained weights) to simulate the scenario where a pre-trained monolithic model covering the full domain union is unavailable. In the pre-trained frozen experts regime, the Monolithic and MoE baselines are instead initialized from pretrained weights and fine-tuned for one epoch on the union of the three domains: the Monolithic baseline is a larger pretrained GPT-2 variant whose parameter count matches the combined capacity of the three experts and the gating network, which is the natural monolithic counterpart to reusing frozen pretrained experts (a from-scratch baseline would not be meaningful at the ~5.8M-token budget of that regime).

The four-epoch budget gives the from-scratch baselines the same number of passes over every training sequence as the experts receive on their own domains, so the monolithic comparison is not confounded by training compute. The gate models are optimized using the no-regret and Primal-Dual algorithms detailed in Sections B.9 (Algorithm 1) and C.5 (Algorithm 4). We use a learning rate of 10^{-4} for the gate parameters, $\eta_\lambda = 0.05$ for the adversarial mixture player, and $\eta_\mu = 0.02$ for the dual constraint player (in $\mathcal{G}_1$). For the scaling law experiments, we strictly held the training sample size to $m_k = 500$ with a batch size of 8. During the minimax optimization, we apply an entropy regularization coefficient of 0.1. The global mass Z_{g_θ} that enters the penalized $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective is estimated at every step from sequences sampled from the frozen experts, using the identity $Z_{g_\theta} = \sum_{k=1}^{p} \mathbb{E}_{x\sim\widehat{\pi}_k}[g_\theta(x,k)]$, which requires no partition-function computation over $\mathcal{X}_0$; the gradient $\nabla_\theta Z_{g_\theta}$ flows through this Monte Carlo estimate. Since the mass term enters the $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective linearly, its mini-batch gradient is unbiased. The same estimator and gradient path are used for the $\mathcal{G}_0$ mass regularizer $\beta(Z_g - 1)^2$ and for the squared capacity penalty of Algorithm 1; for these squared terms the mini-batch gradient is consistent but not unbiased, since $\mathbb{E}[\nabla_\theta(\widehat{Z} - 1)^2] = \nabla_\theta(Z - 1)^2 + \nabla_\theta \mathrm{Var}(\widehat{Z})$, with a bias of order $1/n$ in the number n of sampled sequences per step. An exponential moving average (EMA) of the estimate with decay 0.9 is maintained only for logging; it does not enter the loss. The evaluation-time normalizer $\log \widehat{Z}_g$ used for all reported normalized NLL values is computed separately for each trained gate with the rejection-sampling estimator $\widehat{Z}_g^{\mathrm{RS}}$ of Appendix E (Method 2). The capacity penalty enforces the structural ceiling $\sum_k \max_x g(x,k) \le 1$ (which implies $Z_g \le 1$), while the mass term supplies the downward pressure on Z_g that distinguishes the $\widetilde{\mathsf{D}}_{\mathrm{KL}}$ objective from a plain penalized NLL. For the $\mathcal{G}_{\infty,1}$ formulation, we apply a penalty coefficient of $\alpha = 10.0$, and for the $\mathcal{G}_0$ model, we use a global mass regularization factor of $\beta = 2.0$. Although the capacity penalty is enforced over mini-batch maxima during training, the trained $\mathcal{G}_{\infty,1}$ gates satisfy $\sum_k \max_x g_\theta(x,k) < 1$ when the maxima are evaluated on held-out data. The architecturally bounded gate of deployment mode (b) (Section 3.4) enforces the ceiling, and hence $Z_g \le 1$, for all inputs by construction.

G.2 Real-World Robustness Results (Lightweight Models)

We evaluate the models based on their Negative Log-Likelihood (NLL) per token across two distinct experimental settings: one testing stability across different initialization seeds, and another testing stability across distinct training data splits while keeping the test set strictly fixed. We report the average test NLL and standard deviation over 5 runs.

Table 5: Performance comparison on real-world datasets across different seeds and training data splits for the lightweight regime (~6.5M parameters). Individual experts are evaluated on their respective in-domain test sets; all other models (Monolithic, MoE, Gate) are evaluated on the union test mixture $\widehat{\mathsf{p}}_{\lambda^*}$. The gate model NLL values correspond to the non-causal teacher (see Section 5); causal student results are in Figure 9.

Model	NLL per token (Diff Seed)	NLL per token (Diff Data)
Wiki Expert	4.938 ± 0.008	4.934 ± 0.002
Code Expert	5.638 ± 0.022	5.109 ± 0.025
FineWeb Expert	5.471 ± 0.002	5.467 ± 0.003
Monolithic Model	5.311 ± 0.017	5.155 ± 0.017
Standard MoE Model	5.271 ± 0.014	5.117 ± 0.014
Gate Model $\mathcal{G}_1$ (Baseline)	5.143 ± 0.005	4.983 ± 0.009
Gate Model $\mathcal{G}_{\infty,1}$ (Ours)	$\mathbf{5.123 \pm 0.004}$	$\mathbf{4.962 \pm 0.009}$
Gate Model $\mathcal{G}_0$ (Ours)	5.129 ± 0.008	4.967 ± 0.008

As shown in Table 5, combining highly disjoint domains such as natural language and source code introduces severe interference for the Monolithic model, forcing it to compromise performance across domains (the per-task gap discussed in Appendix B.7). While the standard MoE baseline partially mitigates this by routing at the token level, it still falls short of the sequence-level modular gates.

Most importantly, both unnormalized gating families ($\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$) successfully outperform the strictly normalized $\mathcal{G}_1$ gate in both experimental setups. While individual performance margins between gating families are tight, the consistent ordering across diverse seeds and data splits aligns with our theoretical hierarchy. The bounded family $\mathcal{G}_{\infty,1}$ consistently achieves the lowest overall NLL, corroborating our theoretical finding that it represents a highly effective architectural choice. By removing the strict unity constraint, $\mathcal{G}_{\infty,1}$ effectively sidesteps the capacity term in the $\mathcal{G}_1$ analysis while retaining the Separability Bonus necessary to optimally gate diverse data. Post-training mass estimation empirically validates our geometric bounds: the $\mathcal{G}_{\infty,1}$ model maintains a sub-normalized mass penalty ($\log Z_{g^*} \approx -0.197$), actively providing a Normalization Bonus, whereas the $\mathcal{G}_0$ family incurs a slight Volume Penalty ($\log Z_{g^*} \approx 0.049$). The fully decoupled $\mathcal{G}_0$ model achieves highly competitive results, demonstrating that its local ensembling mechanisms efficiently mitigate mixture ambiguity while offering the maximum degree of distributed training scalability.

G.3 Scaling to the 1B+ Parameter Regime

To confirm that the theoretical capacity bounds established in Theorem 19 translate to massive, state-of-the-art architectures, we replicated the evaluation using 1B+ parameter models. Table 6 reports the test NLL across varied seeds and data splits.

Table 6: Performance comparison scaled to the 1B+ parameter regime. Individual experts are evaluated on their in-domain test sets; all other models are evaluated on the union test mixture. The proposed unnormalized gate models ($\mathcal{G}_{\infty,1}$ and $\mathcal{G}_0$) maintain their theoretical advantages at scale, consistently outperforming the 1B+ Monolithic model, the standard MoE model, and the strictly normalized $\mathcal{G}_1$ baseline.

Model (1B+ Parameter Regime)	NLL per token (Diff Seed)	NLL per token (Diff Data)
Wiki Expert	4.756 ± 0.005	4.743 ± 0.023
Code Expert	5.215 ± 0.007	4.938 ± 0.036
FineWeb Expert	5.390 ± 0.005	5.387 ± 0.002
Monolithic Model	4.987 ± 0.006	4.886 ± 0.021
Standard MoE Model	4.982 ± 0.005	4.879 ± 0.012
Gate Model $\mathcal{G}_1$ (Baseline)	4.895 ± 0.010	4.769 ± 0.017
Gate Model $\mathcal{G}_{\infty,1}$ (Ours)	$\mathbf{4.835 \pm 0.005}$	$\mathbf{4.749 \pm 0.015}$
Gate Model $\mathcal{G}_0$ (Ours)	4.859 ± 0.002	4.758 ± 0.015

As predicted by our theory, the robust gating models successfully scale to the 1B+ regime and consistently outperform both the Monolithic model and standard MoE baselines. The scale of the base models does not diminish the geometric benefits of the unnormalized relaxations; in fact, the $\mathcal{G}_{\infty,1}$ model provides an even starker improvement over the Monolithic and standard MoE models by efficiently leveraging the distinct specializations of the massive pre-trained experts without facing global synchronization bottlenecks.

G.4 Scaling Modularity: 20-Task Code Generation

To thoroughly investigate the performance of unnormalized gating across highly diverse and specialized distributions, we scaled our evaluation to a 20-task ensemble. As the number of experts p scales, the global capacity cost of $\mathcal{O}(\log p)$ becomes a pronounced penalty, rendering this an ideal testbed for evaluating the theoretical advantages of the unconstrained families. As reported in Table 7, despite the tight performance margins, the unnormalized gating models effectively circumvent this capacity cost, successfully scaling robustness to massive mixtures and maintaining their dominance over both the capacity-matched Monolithic model and the strictly normalized $\mathcal{G}_1$ baseline. The aggregate NLL is computed per token over the pooled 20-language test set (token-weighted); it therefore differs slightly from the unweighted mean of the per-language values in Table 8.

Table 7: Aggregate Test NLL across the 20-language code mixture. The $\mathcal{G}_{\infty,1}$ gate achieves the lowest NLL and tightest variance, confirming its ability to offset the $\mathcal{O}(\log p)$ capacity cost.

Model (20-Task Mixture)	**Average Test NLL**
Monolithic Model	2.467 ± 0.006
Standard MoE Model	2.400 ± 0.006
Gate Model $\mathcal{G}_1$ (Baseline)	2.365 ± 0.004
Gate Model $\mathcal{G}_0$ (Ours)	2.356 ± 0.003
Gate Model $\mathcal{G}_{\infty,1}$ (Ours)	$\mathbf{2.353 \pm 0.002}$

Distributional Robustness. To demonstrate true distributional robustness, we evaluated the models on each of the 20 individual language test sets. A monolithic model typically minimizes the average empirical risk, inevitably sacrificing performance on inherently harder distributions (such as Julia) to overfit the easier ones. In contrast, as shown in Table 8, our robust modular framework actively suppresses outlier failures by propagating dynamically across specialized experts, preserving uniform performance stability across arbitrary task variations. We note that the gain over the $\mathcal{G}_1$ gate is realized in the aggregate and on 19 of the 20 languages, whereas on the single hardest language (Julia) the worst-case NLL differs by only 0.001 between $\mathcal{G}_1$ and $\mathcal{G}_{\infty,1}$, within the seed standard deviation (≈ 0.01); the substantial worst-case improvement in this regime is over the Monolithic model ($3.593 \to 3.485$).

Verification of Theoretical Scaling Laws. To strictly isolate the $\mathcal{O}(\sqrt{\log p})$ scaling dependence inherent to the gating class Rademacher complexity (Theorems 19 and 8), we measured the *Excess Gating Risk* (Gate NLL minus the Oracle Expert NLL). By perfectly holding the total training subset size constant at $m_k = 500$ per domain and evaluating across $p \in \{2, 5, 10, 15, 20\}$, we removed data scale confounds to purely observe the theoretical structural capacity scaling. As shown in Figure 6, the excess gating risk scales with the $\mathcal{O}(\sqrt{\log p})$ dependence on p of the generalization bound (this excess risk is the quantity we can measure with a held-out oracle, not the train–test gap itself). Moreover, the minimax regret trajectories reach plateaus whose levels separate according to the $\log p$ dependence of the rate in Theorem 6; the residual oscillations are the limit cycles of the adversarial dynamics, so the plot supports the p-dependence of the bound rather than the $T^{-1/2}$ decay itself.

G.5 Scaling with Pre-Trained Frozen Experts

To evaluate the framework on established foundation models, we used the base GPT-2 model (~124M parameters) for both the Wiki and FineWeb domains, alongside CodeParrot-small for the Code domain. We maximized dataset capacity by using 15,000 training and 4,000 testing sequences per domain. As illustrated in Figure 7 and detailed in the main text, the test NLL across the union mixture validates our theory. Over five independent runs (varying the initialization seeds, with the test set fixed), the unnormalized $\mathcal{G}_{\infty,1}$ gate achieves a mean NLL of 2.396 ± 0.003 and the $\mathcal{G}_0$ gate

Table 8: Distributional Robustness across 20 individual language test sets. By performing a dynamic soft mixture over the 20-expert ensemble, the gating mechanisms successfully mitigate the massive geometric interference suffered by the Monolithic model on intrinsically difficult tasks.

Dataset (Task)	**Monolithic**	**Gate $\mathcal{G}_1$**	**Gate $\mathcal{G}_0$**	**Gate $\mathcal{G}_{\infty,1}$**
html	1.5411 ± 0.0084	1.3144 ± 0.0053	1.3088 ± 0.0055	**1.3069 ± 0.0060**
java	1.8437 ± 0.0062	1.7189 ± 0.0209	1.6996 ± 0.0197	**1.6981 ± 0.0202**
c-sharp	1.9195 ± 0.0092	1.7501 ± 0.0067	1.7216 ± 0.0050	**1.7190 ± 0.0042**
go	1.9338 ± 0.0054	1.8168 ± 0.0114	1.7940 ± 0.0053	**1.7936 ± 0.0056**
css	2.0649 ± 0.0074	1.9379 ± 0.0139	1.9163 ± 0.0103	**1.9102 ± 0.0097**
scala	2.0646 ± 0.0058	1.9522 ± 0.0063	1.9273 ± 0.0080	**1.9234 ± 0.0092**
php	2.0772 ± 0.0098	2.0358 ± 0.0125	2.0210 ± 0.0076	**2.0201 ± 0.0070**
c++	2.1703 ± 0.0033	2.1885 ± 0.0100	**2.1689 ± 0.0103**	2.1711 ± 0.0101
haskell	2.2582 ± 0.0021	1.9792 ± 0.0034	1.9563 ± 0.0058	**1.9531 ± 0.0026**
c	2.2670 ± 0.0033	2.2171 ± 0.0141	**2.2005 ± 0.0134**	2.2033 ± 0.0130
perl	2.3713 ± 0.0105	2.2984 ± 0.0062	**2.2892 ± 0.0038**	2.2901 ± 0.0035
sql	2.8004 ± 0.0046	2.4078 ± 0.0061	2.4024 ± 0.0013	**2.4009 ± 0.0022**
rust	2.5741 ± 0.0077	2.4794 ± 0.0043	2.4621 ± 0.0062	**2.4590 ± 0.0059**
typescript	2.5853 ± 0.0092	2.5213 ± 0.0086	2.5031 ± 0.0055	**2.5000 ± 0.0059**
ruby	2.7741 ± 0.0071	2.5212 ± 0.0058	2.5151 ± 0.0035	**2.5134 ± 0.0036**
javascript	2.9397 ± 0.0084	2.9272 ± 0.0098	2.9174 ± 0.0055	**2.9155 ± 0.0061**
python	2.9470 ± 0.0085	2.9422 ± 0.0108	2.9396 ± 0.0050	**2.9375 ± 0.0044**
shell	3.3939 ± 0.0087	2.9637 ± 0.0035	**2.9609 ± 0.0020**	2.9611 ± 0.0019
lua	3.4590 ± 0.0061	**3.3417 ± 0.0102**	3.3438 ± 0.0089	3.3418 ± 0.0093
julia	3.5928 ± 0.0053	3.4855 ± 0.0105	3.4860 ± 0.0102	**3.4845 ± 0.0106**
Worst-Case Risk	**3.5928**	**3.4855**	**3.4860**	**3.4845**

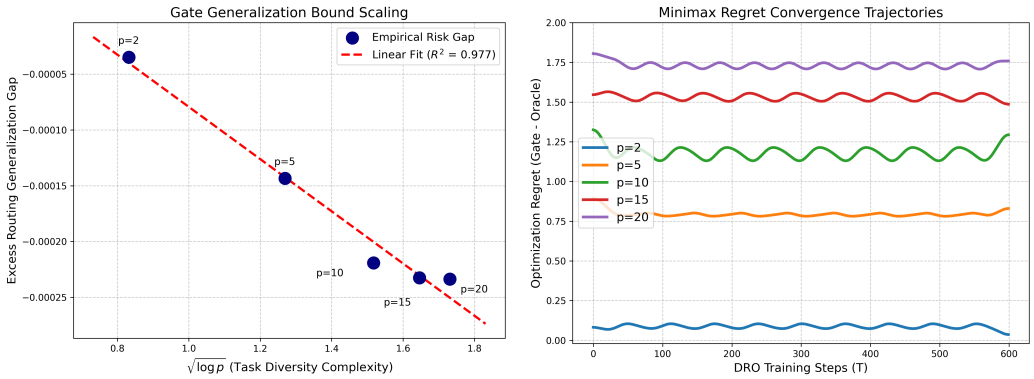


Figure 6: Empirical Verification of Scaling Laws and Convergence. (Left) The excess gating risk scales with the theoretical $\mathcal{O}(\sqrt{\log p})$ dependence on the number of experts. (Right) Minimax regret trajectories plateau at levels that separate according to the $\log p$ dependence of the rate in Theorem 6. Limit-cycle oscillations inherent to adversarial minimax training have been smoothed using a zero-phase low-pass filter over deterministic epoch boundaries.

achieves 2.446 ± 0.004, both comfortably outperforming the $\mathcal{G}_1$ baseline (2.466 ± 0.005), the Standard MoE (2.586 ± 0.007), and the Monolithic model (2.736 ± 0.008). Here the Monolithic baseline is a capacity-matched pretrained GPT-2 variant fine-tuned for one epoch on the union of the three domains (Appendix G.1), so the gates outperform a monolithic model of equal capacity that has itself been adapted to the full domain mixture.

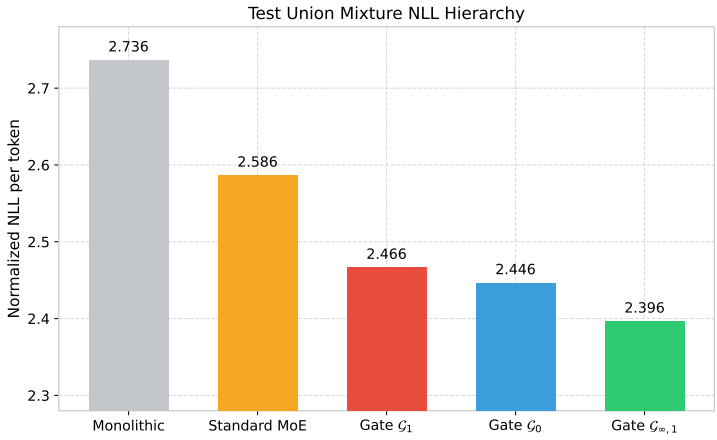


Figure 7: Test Union Mixture NLL Hierarchy for the Pre-Trained Frozen Experts Regime. The unnormalized gating relaxations successfully outperform both the monolithic and MoE baselines, as well as the strictly normalized $\mathcal{G}_1$ model.

G.5.1 Computational Efficiency: Wall-Clock Time

The central practical motivation of unnormalized gating is the resolution of synchronization bottlenecks inherent to massive-scale distributed training. Figure 2 in Section 5 provides an empirical wall-clock comparison of training time per iteration. The globally synchronized $\mathcal{G}_1$ gate is severely bottlenecked, requiring approximately 0.24 seconds per iteration due to sequential expert evaluation and global partition function estimation. In stark contrast, the asynchronous $\mathcal{G}_{\infty,1}$ gate reduces this to ~ 0.10 seconds, and the fully decoupled $\mathcal{G}_0$ gate operates at maximum throughput (~ 0.08 seconds).

G.5.2 Generative Exact Inference and Acceptance Rates

To empirically verify our claims of exact and efficient generative sampling (Section B.12), we tracked the acceptance rate during exact rejection sampling. As shown in Figure 8, the empirical acceptance rate ($Z_{\overline{g}_T}/M_T$) stably converges to a highly efficient regime, validating our structural bounds from Proposition 9 and proving near-perfect self-normalization for the $\mathcal{G}_{\infty,1}$ objective. Since the pre-trained experts are highly specialized across distinct domains, the unnormalized mixture mass $Z_{\overline{g}_T}$ naturally aligns with its capacity envelope M_T. This tight convergence yields highly efficient sampling with no approximation bias: these runs use the architecturally bounded gate of deployment mode (b) (Section 3.4), whose envelope $\sum_k W_k \widehat{\pi}_k$ holds for all inputs by construction, so the sampler is exact. This shows that the $\mathcal{G}_{\infty,1}$ gate organically uses sub-normalization to yield exact generative inference.

G.5.3 Causal Structural Distillation Dynamics

We implemented the structural distillation protocol (Algorithm 3) to map the non-causal $\mathcal{G}_{\infty,1}$ teacher into a causal student. Figure 9 (a) shows the token-level distillation KL $\mathrm{D}_{\mathrm{KL}}(p_{g^\star} \| \pi_\gamma)$ smoothly converging to a small positive residual, consistent with the structural mismatch bound of Proposition 21. More importantly, Figure 9 (b) evaluates the Teacher vs. Causal Student NLL across all constituent domains. The realized performance gap quantitatively validates our Distillation Performance Inheritance guarantee (Theorem 10), confirming that the student inherits the teacher's routing geometry without exceeding the $\delta + C\sqrt{2\delta}$ theoretical divergence bound. On the union test mixture of the frozen-experts regime, the causal $\mathcal{G}_{\infty,1}$ student attains an NLL of 2.401 ± 0.002 over five seeds (token-weighted over the pooled test set), versus 2.396 ± 0.003 for the non-causal teacher, so the ordering of Table 3 is unchanged under a strict autoregressive-vs-autoregressive comparison: the

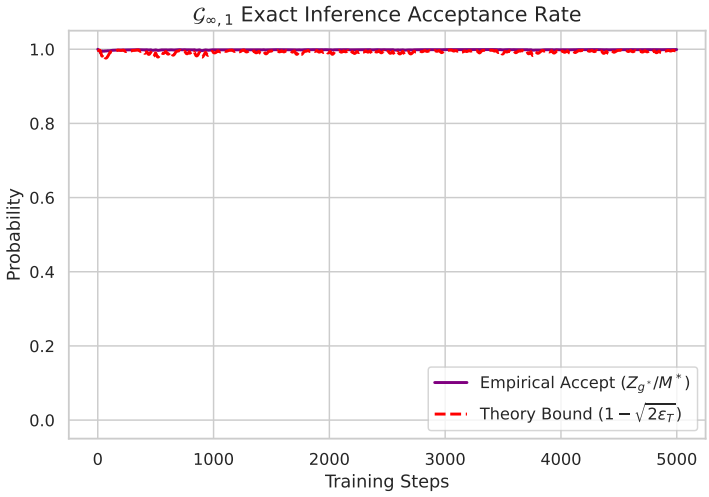
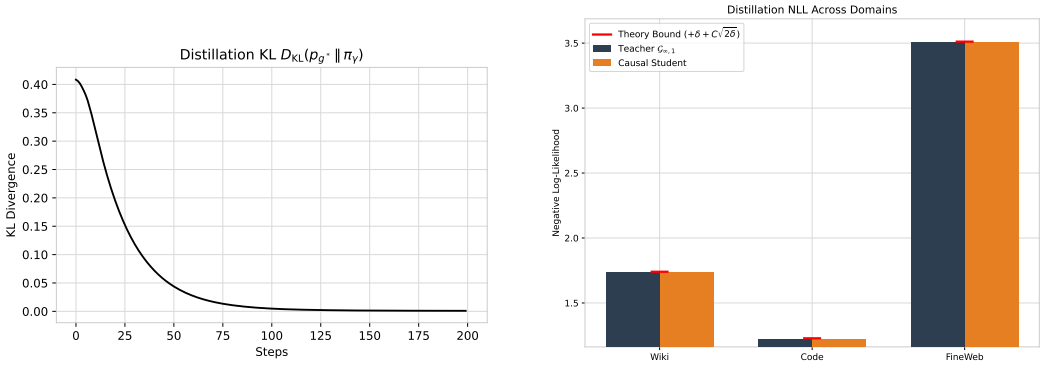


Figure 8: Exact Inference Acceptance Rate. The empirical sampling efficiency closely aligns with the theoretical structural limits. The correct theoretical lower bound is given by Proposition 9: $\mathbb{P}(\text{accept}) \geq \exp(-(C_T + 1))$.

causal student still outperforms the $\mathcal{G}_1$ teacher (2.466), the Standard MoE (2.586), and the Monolithic model (2.736).



(a) Distillation KL $\mathrm{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$ (b) Distillation NLL Gaps across Domains

Figure 9: Structural Distillation Dynamics. (a) The distillation KL $\mathrm{D}_{\mathrm{KL}}(p_{g^*} \| \pi_\gamma)$ converges to a small residual during training. (b) The empirical sequence-level performance gap between Teacher and Student cleanly respects the $+\delta + C\sqrt{2\delta}$ inheritance bound across all domains.

G.5.4 Ablation Studies: Capacity Ceiling Constraint (α)

We explicitly ablate the capacity constraint parameter α within the $\mathcal{G}_{\infty,1}$ optimization to understand its effect on global mass stability. As demonstrated in Figure 10, applying a weak penalty ($\alpha = 0.1$) results in constraint violations where the ensemble mass floats up to roughly 1.1, artificially lowering the unnormalized training loss (though the normalized evaluation NLL increases by $+\log Z_g$). Conversely, proper parameterization ($\alpha = 10.0$ and $\alpha = 50.0$) strictly enforces the theoretical $\sum_k \max_x g(x,k) \leq 1.0$ boundary, proving the necessity and efficacy of our soft-penalty projection method. This ablation uses the scaling-law setup ($m_k = 500$ per domain, Appendix G.1), so the absolute level of the converged ceiling is specific to that small-sample setting and is not comparable to the post-training mass reported for the main experiments ($Z_{g^*} \approx 0.82$ in the Lightweight regime). The ceiling settling well below unity here ($\sum_k \max_x g(x,k) \approx 0.1$, which upper-bounds Z_g) should be read as a property of the parametric optimization rather than of the robust solution: an exact minimizer of the unrestricted game saturates the capacity, $\sum_k \max_x g(x,k) = 1$ (Remark on

capacity saturation, Appendix B.1), while the batch-maxima penalty is inactive below the ceiling and supplies no upward pressure, so the small-sample gate converges with unallocated capacity. The Normalization Bonus $\Delta(Z_g) = Z_g - 1 - \log Z_g \geq 0$ (Theorem 5) concerns the mass Z_g and is realized in the main experiments ($Z_{g^*} \approx 0.82$), independently of whether the capacity is saturated.

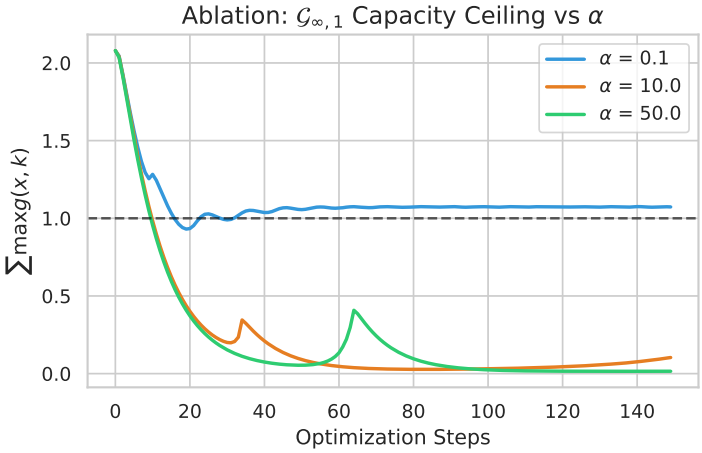


Figure 10: Ablation Study on the Capacity Constraint Coefficient (α), scaling-law setup ($m_k = 500$). Weak penalties allow mass to exceed the theoretical ceiling, whereas appropriate regularization ($\alpha \in \{10.0, 50.0\}$) safely stabilizes the bounding operation.