Bayes-Routed Sparse Interpolated Experts: PAC-Bayes and Tail-Risk Guarantees for Uncertainty-Aware Meta-Tuning

Liz Lemma Future Detective

January 20, 2026

Abstract

Sparse MetA-Tuning (SMAT) improves few-shot OOD generalization by routing tasks to sparse interpolated experts, but its routing is typically a point estimate that can catastrophically misroute when the support set is ambiguous, corrupted, or adversarial—precisely the regime faced by 2026 deployments. We formalize routing as approximate Bayesian model selection over a finite (or low-dimensional) family of expert mixtures and propose Bayes-Routed SMAT: an amortized posterior $q(\alpha \mid S)$ over expert weights conditioned on a task's support set. We train this posterior under a distributionally robust meta-objective that targets tail risk (CVaR) and regularizes routing complexity via KL to a cost-aware prior. On the theory side, we provide PAC-Bayes generalization bounds for the resulting Gibbs routed predictor, extend them to CVaR risk, and prove posterior concentration and a matching indistinguishability lower bound for misrouting probability in a mixture-of-routings task model. On the systems side (experimental component), we outline how posterior sampling/ensembling and entropy-based abstention yield safer OOD behavior and improved worst-quantile accuracy on heavy shift and corrupted-support benchmarks, building directly on the SMAT sparse interpolated expert formulation.

Table of Contents

3. 3. Problem Formulation: Bayes-routed meta-tuning with sparsity constraints and distributionally robust tail objective; define Gibbs predictor, priors, and abstention/fallback option.

4. 4. Method: Posterior Families and Training Objective: discrete routings vs continuous simplex; Concrete/logistic-normal parameterization; KL-to-prior regularization; empirical CVaR estimator; optional teacher distillation integration.

5. 5. Algorithms: Meta-training with stochastic CVaR + reparameterization; meta-testing via posterior mean, sampling/ensembling, and abstention thresholds; implementation notes (required experiments flagged).

6. 6. Theory I — Generalization (PAC-Bayes): bounds for expected risk of Gibbs routed predictor; extension to CVaR tail risk; discussion of dependence on number of tasks and routing KL.

7. 7. Theory II — Routing Identifiability and Concentration: misrouting probability upper bound under TV separation and finite routing set; calibration and abstention guarantees; matching lower bound via Le Cam/Fano indistinguishability.

8. 8. Computational Complexity and Hardness: routing MAP/subset selection hardness; when greedy/approximation applies; implications for practical inference choices.

9. 9. Experimental Plan (to strengthen): heavy OOD suites + corrupted supports; tail-quantile metrics; calibration plots; abstention risk-coverage curves; comparisons to point routing and heuristic gradient-free search.

10. 10. Discussion: limitations (shift model, discretization), extensions (structured sparsity, multi-source priors, certified abstention), and deployment guidance.

1 Introduction

Few-shot meta-tuning with routed experts offers an appealing compromise between full task-specific adaptation and fully frozen inference: we preserve the representational breadth of a pretrained backbone while enabling task-dependent specialization through a small number of trainable degrees of freedom. In this regime, however, performance is often limited not by a lack of capacity in the expert family, but by the brittleness of the routing decision made from the support set. Concretely, when a learned router maps a small labeled support set S to a point estimate $\alpha = h(S)$, the induced predictor can change abruptly under small perturbations of S (label noise, class imbalance, missing classes, atypical examples, or distribution shift), even if the underlying expert deltas are well-behaved. This phenomenon is not captured by sparsity constraints alone: sparsity controls *which parameters* can move, while routing controls *which move* actually occurs. Consequently, the principal failure mode in deployed few-shot adaptation is frequently a *misrouting event* rather than an overfitting event, and such events manifest most clearly in the tail of the task-loss distribution.

Our starting point is the observation that the support set S is itself a random object and, in the few-shot limit, is intrinsically noisy information about the task. Therefore, even an optimal router should represent uncertainty about which expert combination is appropriate. The deterministic choice $\alpha = h(S)$ discards this uncertainty and turns a moderate ambiguity into a hard commitment, which in turn can induce large losses on the query set. If we are interested in robustness—particularly robustness under unseen or corrupted episodes—we should not optimize only the average meta-risk but also the upper tail of the task-loss distribution. In the language of risk measures, we should control a tail functional such as CVaR_ρ, which directly penalizes the worst ρ fraction of episodes. This shifts the meta-training objective from "doing well on most tasks" to "preventing catastrophic errors on the hard tasks", a distinction that becomes operationally important once routing errors can be severe.

We propose to treat routing as approximate Bayesian inference conditioned on the support set, replacing point routing by an amortized posterior distribution $q_\psi(\alpha \mid S)$. This choice has two immediate consequences. First, it yields a *Gibbs routed predictor* whose loss on a task is the posterior expectation $\mathbb{E}_{\alpha\sim q_\psi(\cdot|S)}[\ell(T;\theta(S,\alpha,z))]$, which can be approximated by a small Monte Carlo sample at meta-test time. Second, it provides a quantitative uncertainty signal (e.g., entropy or posterior mass concentration) that can be used to abstain and fall back to a conservative predictor when the support set is uninformative. The resulting mechanism is not merely an implementation detail: it creates a tractable interface between a modern sparse expert parameterization and classical generalization tools (PAC-Bayes) that naturally reason about randomized predictors and KL-regularized posteriors.

In this work we focus on the SMAT-style construction in which a frozen pretrained parameter vector is augmented by a shared dense delta, with per-expert sparsity enforced by masks. This family is attractive precisely because it decouples two issues: (i) *how much* adaptation capacity we allocate (via sparsity budgets and shared deltas), and (ii) *how we choose* among adaptation modes (via routing). Our thesis is that existing work addresses (i) thoroughly but leaves (ii) under-regularized: when S is ambiguous, the router can select a mode that is locally plausible yet globally harmful. By introducing a posterior $q_\psi(\alpha \mid S)$, a cost-aware prior $p(\alpha)$, and a tail-risk objective, we impose a principled regularization on routing decisions and obtain both theoretical control and practical mechanisms (ensembling, abstention) that target the failure modes observed in practice.

Our contributions are as follows.

- **Problem formulation (uncertainty-aware routed sparse meta-tuning).** We formalize meta-tuning with sparse masked experts and uncertainty-aware routing as the optimization of a distributionally robust tail objective. The meta-learner jointly trains (i) the shared delta parameters and mask distributions (subject to per-expert capacity constraints) and (ii) an amortized posterior router $q_\psi(\alpha \mid S)$, with a KL penalty to a prior $p(\alpha)$ that can encode deployment constraints such as expected compute or latency. The objective is expressed in terms of CVaR_ρ of the posterior-routed per-episode loss, thereby explicitly targeting catastrophic task outcomes.

- **PAC-Bayes generalization bounds for routed meta-tuning (mean and tail risk).** We provide generalization guarantees for the Gibbs routed predictor that depend on the expected KL divergence $\mathbb{E}_S[\mathrm{KL}(q_\psi(\cdot \mid S)\|p)]$ and the number of observed episodes. For mean risk we obtain a standard conditional PAC-Bayes inequality for bounded episodic losses. For tail risk we extend the analysis to CVaR_ρ, yielding bounds with explicit ρ dependence and showing the unavoidable difficulty of estimating extreme tail behavior from finite meta-training data. These results justify the KL-regularized posterior training objective and clarify the sample complexity trade-offs in robust meta-learning.

- **Routing identifiability and posterior concentration.** Under a realizability model with a finite routing family and separated support distributions, we show that routing error decays exponentially in the number of support points n_s, with rate governed by a separation parameter Δ. This establishes that posterior routing can, in principle, become reliable in the few-shot regime when tasks are identifiable, and it isolates a concrete statistical bottleneck: if supports induced by different routings are not distinguishable, routing cannot be made uniformly accurate.

- **Matching lower bounds under indistinguishability.** We complement the upper bound with a lower bound based on two-point testing, showing that no router (deterministic or randomized) can achieve substantially faster decay of misrouting probability when the support distributions are close. This clarifies that catastrophic routing errors are information-theoretically unavoidable in the small-n_s or small-Δ regime unless additional information is supplied (more shots, richer task descriptors, or conservative abstention).

- **Empirical protocol targeting tail failures (required to validate the thesis).** We specify an evaluation protocol designed to measure precisely the robustness claims suggested by the theory: (i) report not only mean accuracy but also quantile/CVaR-style episode metrics, (ii) evaluate under support corruptions (label noise, outliers, missing classes) that primarily stress the router rather than the backbone, (iii) compare point routing, posterior-mean routing, and small-sample ensembling over α, and (iv) assess abstention via risk–coverage curves using the router's uncertainty statistic. The expected outcome is not merely improved average performance but a reduction in tail losses and fewer catastrophic misrouting episodes, especially under OOD task distributions.

The organizing principle of the paper is therefore deductive: we begin by isolating routing as the dominant brittleness mechanism in sparse meta-tuning, we introduce uncertainty-aware posterior routing as the minimal modification that exposes both a tail-risk objective and a calibrated abstention signal, and we justify this modification by generalization and identifiability results that make explicit what can and cannot be guaranteed from few-shot support information. The subsequent sections develop the setup, the learning objective, the theoretical bounds, and an empirical methodology aligned with the failure modes implied by the analysis.

2 Background and Setup

We recall the sparse meta-tuning (SMAT) parameterization in which a frozen pretrained backbone is augmented by a small trainable perturbation that is *task-modulated* but *parameter-efficient*. Let f_θ denote the backbone and let $\theta_{\text{pre}} \in \mathbb{R}^d$ be its pretrained parameters. SMAT introduces a shared dense vector $\theta_\delta \in \mathbb{R}^d$ and a collection of expert masks $\{z_m\}_{m=1}^M$, with $z_m \in \{0,1\}^d$ (or a structured analogue, e.g. per-layer or per-block gates). A task-specific parameter vector is then obtained by mixing masked versions of the same delta:

$$\theta(S, \alpha, z) \;=\; \theta_{\text{pre}} + \sum_{m=1}^{M} \alpha_m \, (z_m \odot \theta_\delta), \tag{1}$$

where $\alpha \in \Delta_M$ are routing weights (or, in a discrete specialization, $\alpha \in \mathcal{A} \subseteq \{0,1\}^M$ encodes a budgeted subset of experts). The capacity control is enforced through per-expert constraints such as $\|z_m\|_0 \leq (1-\tau)d$, with τ specifying the target sparsity level. The salient point for our purposes is that the *same* trainable delta vector is reused across experts, while the masks determine *which coordinates* of θ_δ are exposed by each expert; hence expert diversity is achieved without training M independent dense deltas.

We work in the episodic meta-learning protocol. A task episode $T = (S, Q)$ consists of a support set S (few-shot labeled examples used to select adaptation) and a query set Q (labeled during meta-training, unlabeled at test time) used to compute the episode loss. Meta-training samples i.i.d. episodes $T \sim \mathcal{P}_{\mathrm{ID}}$ from an in-distribution task environment. Given episode-specific parameters $\theta(S, \alpha, z)$, we write the task loss as $\ell(T; \theta(S, \alpha, z))$, typically the query cross-entropy of a fixed or lightly parameterized head on top of the backbone features. For theoretical control we assume $\ell \in [0,1]$, which can be enforced by bounded losses or by scaling and clipping; this assumption is standard in concentration-based analyses and does not materially change the optimization procedure.

The remaining degree of freedom is the *routing* map that selects α from S. In conventional routed meta-tuning one posits a deterministic router h_ζ (often a small hypernetwork or set encoder) and sets $\alpha = h_\zeta(S)$. Architecturally, h_ζ must be permutation-invariant in the elements of S and capable of digesting labeled examples; common designs embed each support pair (x, y) via a shared feature extractor, aggregate across shots by averaging or attention, and map the aggregated statistic to α via an MLP and softmax. When α lies on the simplex, (1) yields a convex combination of masked deltas; when α is discrete, it produces a hard selection among a finite set of candidate mixtures. In either case, the router is trained jointly with θ_δ and the masks (or their parameterization), using episodic gradients through the backbone.

This point-routing paradigm is effective when the support set is sufficiently informative to identify the appropriate adaptation mode. The few-shot regime, however, is precisely the regime in which S is a noisy statistic of the task. Even if we condition on an underlying task identity, S is random due to sampling variability, and it is routinely corrupted by practical nuisances: mislabeled shots, class imbalance, missing or spurious classes, outliers, near-duplicates, or distribution shift between the support and query distributions. These perturbations may be small in the sense of affecting only one or two examples, yet they may alter the router output substantially because h_ζ is typically trained to be *decisive* (to commit to an expert or mixture) rather than *calibrated* (to express uncertainty). Consequently, the induced predictor can exhibit discontinuous behavior as a function of S, even when the family (1) is stable for a fixed α.

To isolate the phenomenon, it is useful to distinguish two sources of error.

The first is *within-routing* generalization: for a fixed routing α, the adapted predictor may overfit to the support and perform poorly on the query. The second is *routing error*: the mechanism selecting α from S may choose an adaptation mode that is inappropriate for the task, so that the query loss is high regardless of how well the expert family is trained. The SMAT sparsity constraints primarily address the first issue by limiting the effective degrees of freedom exposed at test time; they do not directly regularize the second issue, because a sparse but wrong adaptation can still be catastrophically wrong. In practice, these two errors can interact: a highly specialized expert can be benign when correctly routed and highly damaging when misrouted.

This motivates evaluating meta-tuning not only by the *mean* episode loss but also by *tail* metrics that emphasize rare failures. Concretely, if we define the per-episode loss random variable

$$X(T) \;=\; \ell(T; \theta(S, \alpha, z)),$$

then average risk $\mathbb{E}[X]$ treats a small probability of large losses as acceptable when compensated by many easy episodes. In deployment, however, one often cares about bounding the frequency and severity of high-loss episodes (e.g. tasks with unusual supports, corrupted labels, or OOD classes), because these correspond to user-visible failures. A standard tail functional is conditional value at risk (CVaR). For a level $\rho \in (0, 1)$, $\mathrm{CVaR}_\rho(X)$ is the expected loss in the worst ρ fraction of episodes. Equivalently, using the variational representation,

$$\mathrm{CVaR}_\rho(X) \;=\; \min_{t \in \mathbb{R}} \left\{ t + \frac{1}{\rho}\, \mathbb{E}\big[(X - t)_+\big] \right\}, \tag{2}$$

where $(u)_+ = \max\{u, 0\}$. This form makes clear that CVaR interpolates between quantile control (via t) and tail expectation (via the hinge term), and it is amenable to stochastic optimization because it expresses a single expectation of a Lipschitz function of X for any fixed t.

In the episodic setting, CVaR can be estimated empirically by sorting the per-episode losses in a meta-batch and averaging the top ρ fraction (or, more smoothly, by optimizing (2) over t jointly with model parameters). The operational interpretation is straightforward: minimizing CVaR discourages configurations that perform well on typical episodes but fail badly on a minority. For routed sparse meta-tuning, this is particularly aligned with the routing-error failure mode, since misrouted episodes tend to dominate the upper tail. Thus, tail-risk analysis provides a lens through which routing brittleness becomes visible and measurable, even when mean accuracy changes little.

Finally, we emphasize that the above setup does not yet specify *how* uncertainty in S should be represented; it only clarifies where the uncertainty enters the pipeline and why mean-risk training can be misaligned with robustness goals. The key structural feature we will exploit in the subsequent

formulation is that routing is a decision made from a small random sample S, and therefore admits an inference perspective: rather than committing to a single α, one may maintain a distribution over plausible routings conditioned on S and couple this with a tail-risk objective. Section 3 makes this precise by replacing point routing with a posterior over α, introducing a prior to encode constraints, and integrating these choices into a distributionally robust episodic objective.

3 Problem Formulation

We formalize uncertainty-aware routing as a conditional inference problem in which the support set S induces not a single routing decision but a *posterior distribution* over routings. Concretely, rather than setting $\alpha = h_\zeta(S)$, we posit a family of conditional distributions

$$q_\psi(\alpha \mid S),$$

parameterized by ψ, from which the routing used in (1) is drawn. This viewpoint separates (i) the representation class of candidate routings (e.g. simplex-valued mixtures or a finite discrete family) from (ii) the epistemic uncertainty arising from having only n_s labeled examples in S. The resulting predictor is a Gibbs (randomized) routed model: conditioned on S, we sample $\alpha \sim q_\psi(\cdot \mid S)$, instantiate $\theta(S, \alpha, z)$, and then evaluate on Q. At the level of the episode loss, this induces the posterior-routed (Gibbs) loss random variable

$$X_q(T) \;:=\; \mathbb{E}_{\alpha \sim q_\psi(\cdot|S)}\big[\ell(T; \theta(S, \alpha, z))\big], \tag{3}$$

where the expectation is with respect to routing randomness only (and $T = (S, Q)$ is itself random under the task environment).

A central modeling choice is the reference distribution $p(\alpha)$, which plays two roles. First, it encodes *deployment preferences* such as favoring sparse or cheap routings, or restricting attention to a feasible candidate set. Second, it provides the reference measure required for a PAC-Bayes analysis of conditional posteriors. We therefore treat $p(\alpha)$ as a *routing prior* and regularize the learned posterior $q_\psi(\alpha \mid S)$ towards $p(\alpha)$ via a conditional KL penalty. Formally, we measure routing complexity by

$$\mathcal{K}(\psi) \;:=\; \mathbb{E}_S\big[\mathrm{KL}(q_\psi(\cdot \mid S) \,\|\, p)\big], \tag{4}$$

where the expectation is taken over S induced by the episode sampler $T \sim \mathcal{P}_{\mathrm{ID}}$ (equivalently over the marginal of supports under $\mathcal{P}_{\mathrm{ID}}$). This term is the natural conditional analogue of the standard PAC-Bayes complexity term, and it will appear explicitly in our generalization bounds.

We couple posterior routing with a tail-risk objective that penalizes rare high-loss episodes. Let $\rho \in (0,1)$ be a tail fraction. Our meta-training criterion minimizes the conditional value at risk of the posterior-routed loss (3):

$$\min_{\theta_\delta, \Phi, \psi} \quad \mathrm{CVaR}_\rho(X_q(T)) \; + \; \lambda \mathcal{K}(\psi), \tag{5}$$

subject to the per-expert mask constraints encoded by Φ (e.g. $\|z_m\|_0 \leq (1-\tau)d$ for each m under hard masks, or the corresponding expected-budget constraints under a stochastic mask parameterization). Here $\lambda > 0$ controls the strength of the KL-to-prior regularization. The use of CVaR is intended to align optimization with the routing-error failure mode described in Section 2: when a small fraction of tasks are misrouted, their losses concentrate in the upper tail, and a mean-risk objective may underweight them. The variational representation (2) yields an equivalent constrained-free form of (5),

$$\min_{\theta_\delta, \Phi, \psi} \; \min_{t \in \mathbb{R}} \left\{ t + \frac{1}{\rho} \, \mathbb{E}_{T \sim \mathcal{P}_{\mathrm{ID}}} \left[(X_q(T) - t)_+ \right] \right\} \; + \; \lambda \mathcal{K}(\psi), \tag{6}$$

which isolates the dependence on $X_q(T)$ through a single expectation of a Lipschitz function and is therefore amenable to stochastic approximation.

The parameterization of the task-conditioned weights follows (1), but we emphasize that the routing posterior is the *only* adaptation mechanism that depends on S at test time: θ_δ is shared across tasks, and $z = \{z_m\}_{m=1}^M$ is shared (up to mask sampling if one uses a stochastic mask estimator). Thus, for any fixed (θ_δ, z), posterior routing defines a randomized family of predictors indexed by α. In the discrete specialization, we restrict to a finite routing family $\mathcal{A}$ (e.g. a catalog of k-expert subsets or a set of hand-designed mixtures), and $q_\psi(\cdot \mid S)$ becomes a categorical distribution over $\mathcal{A}$. In the continuous specialization, $\alpha \in \Delta_M$ and $q_\psi(\cdot \mid S)$ is a distribution on the simplex. We postpone the choice of these posterior families, and their differentiable relaxations, to Section 4; for the present section it suffices that $q_\psi(\cdot \mid S)$ is a valid conditional distribution and that the induced loss $X_q(T)$ is in $[0,1]$ whenever $\ell \in [0,1]$.

The sparsity constraints are enforced at the level of the masks z_m. Since (5) is stated abstractly over Φ, it covers both hard-constraint and soft-constraint implementations. In a hard-constraint view, one requires that each realized mask obey $\|z_m\|_0 \leq (1-\tau)d$, and meta-training optimizes over a constrained set of masks (or over parameters that generate masks in that set). In a soft-constraint view, one introduces Lagrange multipliers to penalize violations of an expected budget (e.g. under hard-concrete gates), thereby yielding an unconstrained surrogate objective whose minimizers satisfy the target sparsity in the controlled-sparsity regime. Importantly, the routing posterior and the sparsity mechanism interact: a prior $p(\alpha)$ can be used to bias the router towards routings that activate fewer experts or otherwise

reduce expected compute, while the masks control per-expert capacity by limiting the number of exposed coordinates of θ_δ.

Posterior routing also suggests a principled abstention/fallback mechanism. We define a scalar uncertainty score $U(S)$ computed from $q_\psi(\cdot \mid S)$, for instance the Shannon entropy $U(S) = H(q_\psi(\cdot \mid S))$, the posterior gap $1 - \max_\alpha q_\psi(\alpha \mid S)$ in the discrete case, or a divergence-to-prior measure such as $\mathrm{KL}(q_\psi(\cdot \mid S)\|p)$. Given a threshold $\eta \geq 0$, we consider the decision rule

$$\text{if } U(S) > \eta \text{ then abstain and use a fallback predictor; otherwise use posterior routing.} \tag{7}$$

The fallback predictor may be the unadapted model θ_{pre} (equivalently $\alpha = 0$), or a conservative routing such as the prior mean routing under p. This design ensures that when S is uninformative or corrupted, the system can revert to a baseline known to be stable. While abstention is not required to define (5), it will be convenient both for practice (to prevent catastrophic failures) and for analysis (to decompose risk into covered and abstained episodes, as in Theorem 5). At test time one may also replace the Gibbs predictor by a deterministic decision rule derived from $q_\psi(\cdot \mid S)$, such as MAP routing, posterior mean routing (when $\alpha \in \Delta_M$), or an ensemble that averages predictions over a small number of sampled routings; these are standard conversions from a Gibbs predictor to implementable predictors, and our bounds will be stated at the Gibbs level to preserve convexity and enable PAC-Bayes control.

In summary, the problem is to learn three coupled components: (i) the shared delta θ_δ, (ii) the mask mechanism Φ satisfying the sparsity/capacity constraints, and (iii) the amortized routing posterior $q_\psi(\alpha \mid S)$ that trades off empirical tail performance against adherence to the prior $p(\alpha)$. The objective (5) makes this trade-off explicit and yields a formulation in which both robustness (via CVaR_ρ) and generalization (via $\mathcal{K}(\psi)$) can be analyzed. Section 4 instantiates q_ψ (discrete versus continuous), describes differentiable parameterizations and estimators, and gives the empirical CVaR estimator used in meta-training.

4 Method: Posterior Families and Training Objective

We now instantiate the conditional routing posterior $q_\psi(\alpha \mid S)$ and the resulting meta-training objective in forms that admit (i) principled regularization to a cost-aware prior $p(\alpha)$ and (ii) low-variance gradient estimation under stochastic routing. Throughout, we treat the mask mechanism $z = \{z_m\}_{m=1}^M$ as implemented by a SMAT-style sparsifier (e.g. hard masks or hard-concrete gates), and focus on the remaining design degrees of freedom: the routing

family (discrete versus continuous), its differentiable parameterization, and the empirical estimator of the CVaR_ρ objective.

Discrete routing families. In the discrete specialization we posit a finite catalog of feasible routings $\mathcal{A}$. Each element $A \in \mathcal{A}$ may represent, for example, a k-hot subset of experts (binary selection), a small mixture over experts (sparse simplex vectors), or a structured routing template consistent with a deployment budget. We identify α with A and take

$$q_\psi(\alpha \mid S) \equiv q_\psi(A \mid S) \qquad \text{for } A \in \mathcal{A},$$

where $q_\psi(\cdot \mid S)$ is a categorical distribution produced by a routing network that embeds the support set. The KL regularizer assumes a particularly simple form:

$$\mathrm{KL}(q_\psi(\cdot \mid S) \,\|\, p) = \sum_{A \in \mathcal{A}} q_\psi(A \mid S) \, \log \frac{q_\psi(A \mid S)}{p(A)}.$$

The prior p can be uniform when we wish to avoid structural bias, or can encode compute preferences by setting $p(A) \propto \exp(-\beta \, \mathrm{cost}(A))$ for a user-defined routing cost. When $\mathcal{A}$ is not too large, discrete routing has two advantages: (i) posterior uncertainty is directly interpretable (e.g. via the maximum posterior mass), and (ii) the KL term is tractable and stable. The main disadvantage is that sampling from a categorical distribution is not naively reparameterizable; we address this below via relaxations that preserve the intended discrete semantics at test time.

Continuous simplex families. In the continuous specialization we allow $\alpha \in \Delta_M$ to be a convex mixture of experts. This removes the need to pre-enumerate a catalog and permits interpolation between experts when beneficial. We require $q_\psi(\cdot \mid S)$ to be a distribution on the simplex, and we consider two standard choices.

First, a *logistic-normal* posterior: let $u \in \mathbb{R}^M$ follow a Gaussian distribution whose parameters are predicted from S,

$$u \sim \mathcal{N}(\mu_\psi(S), \Sigma_\psi(S)), \qquad \alpha = \mathrm{softmax}(u) \in \Delta_M.$$

This yields a flexible posterior with a fully reparameterizable sampling path via $u = \mu_\psi(S) + L_\psi(S)\,\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$. The induced density on Δ_M has no closed-form KL to most priors; we therefore either (a) choose a logistic-normal prior and estimate $\mathrm{KL}(q\|p)$ by Monte Carlo using the corresponding log-densities, or (b) adopt a simpler proxy regularizer such as $\mathrm{KL}(\mathcal{N}(\mu_\psi, \Sigma_\psi) \,\|\, \mathcal{N}(\mu_0, \Sigma_0))$ on the pre-softmax logits, which empirically controls posterior spread while remaining analytically tractable.

Second, a *Concrete / Gumbel–Softmax* posterior (primarily useful when we want near-discrete routings with straight-through gradients). For logits $\pi_\psi(S) \in \mathbb{R}^M$, temperature $\gamma > 0$, and i.i.d. Gumbel noise g_m, we set

$$\alpha_m = \frac{\exp\big((\pi_{\psi,m}(S) + g_m)/\gamma\big)}{\sum_{j=1}^{M} \exp\big((\pi_{\psi,j}(S) + g_j)/\gamma\big)}. \tag{8}$$

As $\gamma \to 0$ this concentrates near vertices of the simplex and approximates categorical routing; for larger γ it behaves like a smooth mixture. This parameterization is compatible with both discrete and continuous viewpoints: during training we propagate gradients through (8), while at test time we may either sample (Gibbs), use the posterior mean, or take a discrete proxy (e.g. $\arg\max_m \alpha_m$) when deployment requires a single expert.

Stochastic objective and Monte Carlo estimation. Given an episode $T = (S, Q)$, we recall the posterior-routed loss

$$X_q(T) = \mathbb{E}_{\alpha \sim q_\psi(\cdot|S)}\big[\ell(T; \theta(S, \alpha, z))\big] \in [0, 1].$$

In practice we estimate this expectation with K_α routing samples,

$$\widehat{X}_q(T) = \frac{1}{K_\alpha} \sum_{k=1}^{K_\alpha} \ell\Big(T; \theta(S, \alpha^{(k)}, z)\Big), \qquad \alpha^{(k)} \sim q_\psi(\cdot \mid S), \tag{9}$$

optionally sharing the same sampled mask z across the K_α routings to reduce variance. The KL term is evaluated per support set, either exactly (categorical) or by a small Monte Carlo estimate in continuous families. The meta-training objective is the regularized tail risk

$$\mathrm{CVaR}_\rho(X_q(T)) + \lambda\, \mathbb{E}_S\big[\mathrm{KL}(q_\psi(\cdot \mid S) \| p)\big],$$

and we implement CVaR_ρ using its variational form. Concretely, for a scalar threshold $t \in \mathbb{R}$,

$$\mathrm{CVaR}_\rho(X_q(T)) = \min_{t \in \mathbb{R}} \left\{ t + \frac{1}{\rho}\, \mathbb{E}\big[(X_q(T) - t)_+\big] \right\}. \tag{10}$$

Given a meta-batch of B i.i.d. episodes $\{T_i\}_{i=1}^B$, we form per-episode estimates $\widehat{X}_q(T_i)$ as in (9) and compute an empirical CVaR estimator by selecting t as the empirical $(1-\rho)$-quantile threshold (equivalently, the smallest value such that approximately a ρ fraction of episodes lie above it), yielding

$$\widehat{\mathrm{CVaR}}_\rho = t + \frac{1}{\rho B} \sum_{i=1}^{B} \big(\widehat{X}_q(T_i) - t\big)_+. \tag{11}$$

This estimator is consistent and matches the form used in our generalization analysis; operationally it amounts to sorting the B losses (or using a selection algorithm) and averaging those in the upper tail. The meta-training loss for one batch is then

$$\mathcal{L}_{\text{meta}} = \widehat{\text{CVaR}}_\rho + \lambda \frac{1}{B} \sum_{i=1}^{B} \text{KL}(q_\psi(\cdot \mid S_i) \,\|\, p) + \mathcal{L}_{\text{mask}}(\Phi), \tag{12}$$

where $\mathcal{L}_{\text{mask}}(\Phi)$ denotes whichever controlled-sparsity penalty or projection is used to enforce the per-expert capacity constraints.

Gradients via reparameterization and relaxations. When $q_\psi(\alpha \mid S)$ is reparameterizable (logistic-normal or Concrete), we backpropagate through the sampling path used to form $\alpha^{(k)}$ in (9). In the purely categorical discrete case, we either apply a Gumbel–Softmax relaxation during training and anneal γ to approach discreteness, or use a score-function estimator with baselines if exact discreteness is essential. Since K_α is small in typical deployments, we emphasize low-variance estimators; in our experiments we therefore default to reparameterizable families.

Optional teacher distillation. A practical enhancement is to incorporate a teacher signal that stabilizes routing early in training or transfers a compute-heavy oracle to a light amortized router. We consider two compatible variants. (i) *Routing distillation:* given a teacher posterior $q_{\text{teach}}(\alpha \mid S)$ (e.g. obtained by evaluating candidate routings on S or by a larger router), we add

$$\lambda_{\text{dist}} \, \mathbb{E}_S \big[\text{KL}(q_\psi(\cdot \mid S) \,\|\, q_{\text{teach}}(\cdot \mid S)) \big]$$

to (12), optionally combined with the prior KL. (ii) *Prediction distillation:* we match the predictive distribution of a teacher ensemble over routings by minimizing a cross-entropy (or KL) between student predictions under $\alpha \sim q_\psi(\cdot \mid S)$ and teacher predictions, which indirectly shapes q_ψ toward routings that reproduce the teacher's behavior. Both forms preserve our primary objective and can be viewed as additional regularizers; we treat them as optional, since our theoretical analysis is stated for the prior-regularized objective.

These design choices yield a concrete family of uncertainty-aware routers and a tractable empirical approximation to the regularized CVaR_ρ objective. Section 5 specifies the resulting meta-training and meta-testing procedures, including the deterministic decision rules derived from $q_\psi(\cdot \mid S)$ and the abstention mechanism based on posterior uncertainty.

5 Algorithms: Meta-training and Meta-testing

We now make the training and inference procedures explicit. Our goal is to optimize the regularized tail-risk objective introduced in Section 4 using episodic meta-training, while ensuring that the learned router admits practical meta-test decision rules (deterministic, stochastic, and abstaining) with predictable compute.

Meta-training loop (stochastic CVaR with posterior routing). We assume access to i.i.d. episodes $T_i = (S_i, Q_i) \sim \mathcal{P}_{\mathrm{ID}}$ in meta-batches of size B. For each episode T_i, we form a Monte Carlo estimate $\widehat{X}_q(T_i)$ of the posterior-routed loss $X_q(T_i)$ as in (9), using K_α samples $\alpha_i^{(k)} \sim q_\psi(\cdot \mid S_i)$. In parallel, we apply the SMAT sparsifier to obtain masks $z_{i,m}$ (either sampled from a gate distribution or taken as an expected/thresholded mask), subject to the per-expert capacity constraints. Concretely, the per-episode loss estimate takes the form

$$\widehat{X}_q(T_i) = \frac{1}{K_\alpha} \sum_{k=1}^{K_\alpha} \ell\left(T_i; \theta_{\mathrm{pre}} + \sum_{m=1}^{M} \alpha_{i,m}^{(k)} \left(z_{i,m} \odot \theta_\delta \right) \right).$$

We then compute the empirical CVaR_ρ across the batch by choosing a threshold t corresponding to the empirical $(1-\rho)$-quantile of $\{\widehat{X}_q(T_i)\}_{i=1}^B$ and forming (11). The resulting meta-objective for one batch is the differentiable surrogate

$$\mathcal{L}_{\mathrm{meta}} = t + \frac{1}{\rho B} \sum_{i=1}^{B} (\widehat{X}_q(T_i) - t)_+ + \lambda \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL}(q_\psi(\cdot \mid S_i) \,\|\, p) + \mathcal{L}_{\mathrm{mask}}(\Phi).$$

The only subtlety is that the quantile selection defining t is not differentiable. In implementation, we treat t as a stop-gradient statistic of the current batch (equivalently, we optimize a piecewise-smooth objective in which gradients flow through the hinge $(\cdot - t)_+$ but not through the sorting operation). This choice preserves the intended semantics of empirical tail averaging and is consistent with common practice in CVaR optimization; if desired, one may replace the hard quantile by a smooth approximation (e.g. soft-sorting) at additional computational cost.

Gradient estimation and reparameterization. When $q_\psi(\alpha \mid S)$ is reparameterizable, we write $\alpha = g_\psi(S, \varepsilon)$ for noise ε from a fixed distribution and backpropagate through g_ψ when differentiating $\widehat{X}_q(T_i)$ w.r.t. ψ. This covers the logistic-normal and Concrete families described in Section 4. In the Concrete case (8), we typically anneal the temperature γ from a moderate value to a smaller value, trading bias for variance while gradually approaching near-discrete routings. When a strictly discrete catalog $\mathcal{A}$ is

required at training time (rather than only at test time), we may instead use a score-function estimator; however, in few-shot regimes and for small K_α, we have found the reparameterized relaxations to be materially more stable.

The KL regularizer is computed per support set. In the categorical case it is exact; in continuous families it is either a Monte Carlo estimate of log-density ratios (when a matching prior is available) or an analytic proxy in the pre-softmax space. We emphasize that the KL term is not merely a complexity penalty: in deployment it can encode explicit cost preferences through $p(\alpha) \propto \exp(-\beta \operatorname{cost}(\alpha))$, thereby shaping $q_\psi(\cdot \mid S)$ toward routings that are likely to be both accurate and feasible.

Mask enforcement and controlled sparsity. We treat the sparsity mechanism as orthogonal to routing, but we must ensure that the per-expert capacity constraints are satisfied throughout training. Operationally, $\mathcal{L}_{\text{mask}}(\Phi)$ is implemented either (i) by a primal–dual penalty that drives $\mathbb{E}[\|z_m\|_0]$ below $(1-\tau)d$ for each expert m, or (ii) by projection/thresholding into the feasible set after gradient steps when hard masks are used. To reduce variance in the nested stochasticity (mask sampling and routing sampling), a practical default is to share the same mask sample z_i across all K_α routing samples for episode i.

Meta-testing decision rules (mean, sampling/ensembling, MAP). At meta-test time, we are given a new support set S and must predict labels on queries Q without updating θ_δ. We first compute the routing posterior $q_\psi(\alpha \mid S)$. We then instantiate one of the following decision rules.

1. *Posterior mean routing:* use $\bar{\alpha}(S) := \mathbb{E}_{q_\psi(\cdot|S)}[\alpha]$ (available in closed form for categorical distributions and estimated by a small number of samples otherwise), and predict with $\theta(S, \bar{\alpha}, z)$. This yields a single-forward inference path and typically the best compute/accuracy trade-off.

2. *Gibbs sampling / ensembling:* draw $\alpha^{(1)}, \ldots, \alpha^{(K_\alpha)} \sim q_\psi(\cdot \mid S)$ and average predictions (or logits) across routings. This is the direct operational analog of the Gibbs predictor analyzed in Section 6 and often improves tail performance by reducing misrouting sensitivity.

3. *MAP / top-k routing:* in discrete families, take $\hat{\alpha}(S) = \arg\max_\alpha q_\psi(\alpha \mid S)$, or restrict to the top-k routings by posterior mass and ensemble only within that subset. This interpolates between single-route inference and full posterior averaging.

When masks are stochastic, we analogously either fix z to its expected/thresholded value for deterministic inference or sample and ensemble if uncertainty in sparsity is part of the intended model.

Abstention and fallback via routing uncertainty. We additionally expose a meta-test abstention mechanism to guard against catastrophic errors under ambiguous or corrupted support sets. We define a scalar uncertainty score $U(S)$ derived from $q_\psi(\cdot \mid S)$, and abstain (fall back) whenever $U(S) > \eta$ for a chosen threshold η. In the categorical case, simple choices include

$$U_{\max}(S) = 1 - \max_{\alpha} q_\psi(\alpha \mid S), \qquad U_{\text{ent}}(S) = -\sum_{\alpha} q_\psi(\alpha \mid S) \log q_\psi(\alpha \mid S),$$

and in continuous families we may use the entropy of the induced categorical proxy (e.g. after discretization), or the sample variance of α under $q_\psi(\cdot \mid S)$. Upon abstention we use θ_{pre} (or a conservative fixed routing chosen to be cheap and robust), and we report both the prediction and the abstention indicator. The threshold η is selected on a validation distribution to realize a target coverage (fraction of non-abstained episodes) or to minimize a user-specified risk–coverage objective.

Implementation notes and required empirical checks. We use a permutation-invariant set encoder for S (e.g. DeepSets-style pooling over embedded support points) to parameterize $q_\psi(\alpha \mid S)$, since the ordering of shots is immaterial. The CVaR_ρ estimator in (11) requires sorting B per-episode losses; thus B should be large enough that ρB is not too small (otherwise the tail average is dominated by a few episodes and gradients are noisy). Typical stable settings are $\rho \in [0.05, 0.2]$ and $K_\alpha \in \{1, 2, 4\}$.

Since our contribution is primarily about tail robustness and uncertainty-aware routing, we mark as *required* the following experiments: (a) tail accuracy curves as a function of ρ (not only mean accuracy), including OOD task shifts; (b) a comparison of meta-test decision rules (posterior mean vs. Gibbs ensembling vs. MAP) under a fixed compute budget; (c) an ablation demonstrating that abstention improves worst-case performance relative to always routing, with risk–coverage plots and calibration diagnostics for $U(S)$; and (d) sensitivity to shot count n_s, verifying the predicted monotone decrease in misrouting and the empirical benefit of increased support. These checks align the implementation with the guarantees and failure modes formalized in our theory.

6 Theory I: Generalization via PAC-Bayes for Posterior Routing

We analyze the generalization behavior of the *Gibbs routed predictor* induced by the learned conditional posterior $q_\psi(\alpha \mid S)$. Throughout, we treat the pretrained parameters θ_{pre} as fixed and view $(\theta_\delta, \Phi, \psi)$ as the learned components; the stochasticity of routing is explicit, while the sparsity masks

z may be random or deterministic (for the bounds below, it suffices that the induced episode loss is bounded).

Episode-wise routed loss and risks. Given an episode $T = (S, Q) \sim \mathcal{P}_{\mathrm{ID}}$, define the posterior-routed loss random variable

$$X_q(T) \;:=\; \mathbb{E}_{\alpha\sim q(\cdot|S)}\big[\ell(T;\theta(S,\alpha,z))\big] \in [0,1].$$

We denote by

$$R(q) \;:=\; \mathbb{E}_{T\sim\mathcal{P}_{\mathrm{ID}}}\big[X_q(T)\big] \qquad \text{and} \qquad \widehat{R}(q) \;:=\; \frac{1}{N}\sum_{i=1}^{N} X_q(T_i)$$

the population and empirical *mean* risks over N i.i.d. training episodes $T_1, \ldots, T_N \sim \mathcal{P}_{\mathrm{ID}}$. The essential point is that $q(\cdot \mid S)$ is *data-dependent* through S, and therefore complexity control must be expressed in terms of an expected conditional divergence $\mathbb{E}_S[\mathrm{KL}(q(\cdot \mid S)\|p)]$ to a prior $p(\alpha)$.

A conditional PAC-Bayes bound for mean risk. When the routing family is finite (or discretized) to $\mathcal{A}$, we may invoke a conditional variant of the PAC-Bayes theorem for Gibbs predictors. The next statement makes explicit the role of the conditional posterior and the KL term used in our meta-objective.

Theorem 6.1 (PAC-Bayes bound for Gibbs-routed meta-tuning)**.** *Assume $\ell(T;\theta) \in [0,1]$ and a finite routing set $\mathcal{A}$. Fix any prior $p(\alpha)$ on $\mathcal{A}$. Let $q(\alpha \mid S)$ be any learned conditional posterior. With probability at least $1-\delta$ over N i.i.d. episodes, we have*

$$R(q) \;\le\; \widehat{R}(q) + \sqrt{\frac{\mathbb{E}_S[\mathrm{KL}(q(\cdot \mid S)\|p)] + \log\frac{2\sqrt{N}}{\delta}}{2(N-1)}}.$$

Interpretation. Theorem 6.1 yields a direct reading of our regularization choice: for fixed empirical risk $\widehat{R}(q)$, any posterior with smaller expected divergence to the prior admits a tighter bound. In particular, if p is chosen to prefer cheap routings (e.g. $p(\alpha) \propto \exp(-\beta\,\mathrm{cost}(\alpha))$), then the bound formalizes a trade-off between predictive performance and deployment cost. We also emphasize that the bound controls the *Gibbs* predictor (randomly sample $\alpha \sim q(\cdot \mid S)$); deterministic rules such as MAP routing may be analyzed via standard Gibbs-to-MAP conversions (e.g. bounding MAP risk by Gibbs risk plus a term depending on the posterior sharpness), but the Gibbs form is the natural object for PAC-Bayes analysis and aligns with our ensembling decision rule.

Extension to tail risk via CVaR_ρ. Mean risk bounds do not directly address catastrophic episodes. To capture worst-quantile behavior, we consider the population tail objective

$$\mathrm{CVaR}_\rho(X_q) \;:=\; \inf_{t\in\mathbb{R}} \left\{ t + \frac{1}{\rho}\,\mathbb{E}\big[(X_q(T)-t)_+\big] \right\}, \qquad \rho \in (0,1),$$

and its empirical analog (computed from $X_q(T_1),\dots,X_q(T_N)$ or Monte Carlo estimates thereof). Using the variational representation above, we reduce tail control to uniform control of the class of bounded hinge losses $(X-t)_+$ indexed by t.

Theorem 6.2 (Generalization for CVaR_ρ tail risk of Gibbs routing)**.** *Let $X_q(T) \in [0,1]$ be defined as above and let $\widehat{\mathrm{CVaR}}_\rho(X_q)$ denote the empirical CVaR_ρ computed from N i.i.d. episodes. Then, with probability at least $1-\delta$,*

$$\mathrm{CVaR}_\rho(X_q) \;\le\; \widehat{\mathrm{CVaR}}_\rho(X_q) + O\left(\sqrt{\frac{\mathbb{E}_S[\mathrm{KL}(q(\cdot \mid S)\|p)] + \log(1/\delta)}{N\,\rho^2}} \right).$$

Moreover, there exist bounded task families for which the ρ^{-2} dependence is unavoidable up to constants.

Consequences for sample complexity. Theorem 6.2 makes precise an expected phenomenon in tail-risk learning: as ρ decreases, the effective number of informative tail samples behaves like $N\rho$, and concentration degrades accordingly. In particular, to keep the tail generalization gap below ε one needs, in the worst case,

$$N \;\gtrsim\; \frac{\mathbb{E}_S[\mathrm{KL}(q(\cdot \mid S)\|p)] + \log(1/\delta)}{\rho^2\,\varepsilon^2}.$$

This scaling justifies two practical design choices already reflected in our algorithm: (i) one should not take ρ extremely small unless the number of meta-training episodes is correspondingly large, and (ii) one should actively control the routing posterior complexity via the KL term, since any uncontrolled growth in $\mathbb{E}_S[\mathrm{KL}(q\|p)]$ directly enlarges the tail generalization gap.

Relation to the training objective. Our meta-training objective may be viewed as minimizing a proxy for the right-hand side of Theorem 6.2: the empirical tail term $\widehat{\mathrm{CVaR}}_\rho$ is optimized directly, while $\lambda\,\mathbb{E}_S[\mathrm{KL}(q\|p)]$ plays the role of a tunable surrogate for the complexity term in the bound. Although the theorem itself does not prescribe the optimal λ, it provides a principled interpretation: increasing λ biases training toward posteriors that generalize better (especially in the tail) at the cost of potentially higher empirical loss. In addition, because $q(\cdot \mid S)$ is amortized, the KL penalty is applied per support set, which encourages the router to be *simple on average* while still allowing confident deviations from the prior when the support evidence is decisive.

Discussion and limitations. The bounds above are stated for a finite routing family $\mathcal{A}$ to obtain sharp complexity measures. In implementation, we may train with continuous relaxations (logistic-normal or Concrete) and interpret the theory as applying to a discretized approximation (e.g. by quantizing to a finite catalog at test time, or by restricting attention to a top-k set of routings). Finally, these generalization results do not by themselves guarantee that the router selects the *correct* routing for a given task; they control only the routed predictor's population risk relative to its empirical risk and posterior complexity. To connect support size to routing correctness, we next study identifiability and posterior concentration under separation conditions on the support distributions.

7 Theory II: Routing Identifiability and Posterior Concentration

The PAC-Bayes bounds in Section 6 control the *risk* of the routed predictor as a randomized (Gibbs) procedure, but they do not by themselves imply that the router selects an appropriate expert configuration for a given episode. We now isolate a complementary question: when does the support set S contain enough information to *identify* a good routing, and how fast does the induced posterior $q(\cdot \mid S)$ concentrate?

A realizable finite-routing model. We assume that there is a finite set $\mathcal{A}$ of candidate routings (e.g. a catalog of sparse expert mixtures), and that each episode admits a latent routing variable $A \in \mathcal{A}$ which governs the distribution of the support set. Formally, for each $A \in \mathcal{A}$ there is a distribution P_A over labeled supports S of size n_s (often $S = \{(x_j, y_j)\}_{j=1}^{n_s}$ i.i.d.), and the meta-training (or meta-testing) episode first samples A and then samples $S \sim P_A$. We emphasize that this assumption concerns identifiability of *routing* from the support, not correctness of the backbone; it is compatible with arbitrary query distributions and arbitrary downstream heads, provided the routing is the principal latent choice.

A necessary condition for identification is that different routings induce distinguishable support distributions. We capture this by a separation condition: there exists $\Delta > 0$ such that

$$\mathrm{TV}(P_A, P_{A'}) \;\geq\; \Delta \qquad \text{for all } A \neq A',$$

where TV denotes total variation. This is a strong but transparent assumption: if Δ is small, then no support-based router can reliably discriminate routings without many shots.

MAP routing as multi-hypothesis testing. Given a learned posterior $q(\alpha \mid S)$, a natural deterministic decision rule is MAP routing

$$\hat{A}(S) \;\in\; \arg\max_{A \in \mathcal{A}} q(A \mid S),$$

while the Gibbs decision samples $\tilde{A} \sim q(\cdot \mid S)$. In the realizable setting where $q(\cdot \mid S)$ approximates the Bayes posterior for the generative family $\{P_A\}_{A \in \mathcal{A}}$, the MAP rule is an optimal test for the latent routing A (up to approximation error), and its error probability decays exponentially in n_s under separation.

Theorem 7.1 (Misrouting probability under TV separation)**.** *Assume $|\mathcal{A}| < \infty$ and pairwise separation $\mathrm{TV}(P_A, P_{A'}) \geq \Delta$ for all $A \neq A'$. Assume further that S consists of n_s i.i.d. draws from P_A conditional on the latent routing A. Then there exists a universal constant $c > 0$ such that the MAP misrouting probability satisfies*

$$\Pr\big[\hat{A}(S) \neq A\big] \;\leq\; (|\mathcal{A}| - 1)\exp(-c\, n_s \Delta^2).$$

Proof sketch. We reduce routing to multi-hypothesis testing. For any fixed A, by a union bound it suffices to control $\Pr[\text{choose } A' \neq A]$. Under i.i.d. sampling, the likelihood ratio between $P_A^{\otimes n_s}$ and $P_{A'}^{\otimes n_s}$ concentrates, and standard inequalities (e.g. Bretagnolle–Huber or a Chernoff bound) relate the optimal testing error to a divergence between P_A and $P_{A'}$. Total variation separation implies a nontrivial gap in these divergences, yielding an exponential decay rate in n_s of the form $\exp(-\Omega(n_s\Delta^2))$. The factor $|\mathcal{A}| - 1$ accounts for competing alternatives.

Posterior concentration and a usable uncertainty score. Theorem 7.1 is most useful when translated into a statement about the *posterior itself*. In the well-specified case, the Bayes posterior satisfies a concentration phenomenon: for typical supports sampled under P_A, the posterior mass assigned to A approaches 1 at an exponential rate. Operationally, this suggests that simple uncertainty measures derived from $q(\cdot \mid S)$ can serve as proxies for misrouting risk. A canonical choice is

$$U(S) \;:=\; 1 - \max_{A \in \mathcal{A}} q(A \mid S),$$

which is the Gibbs probability of *not* sampling the MAP routing, conditional on S. Indeed, if $\tilde{A} \sim q(\cdot \mid S)$, then $\Pr[\tilde{A} \neq \hat{A}(S) \mid S] = U(S)$ holds identically. Moreover, under posterior concentration, $U(S)$ is small precisely when the support provides decisive evidence.

While $U(S)$ does not automatically upper-bound $\Pr[\hat{A}(S) \neq A \mid S]$ for an arbitrary amortized router, it is the correct quantity under the Bayesian model and remains a meaningful calibration target in practice: one may post-train $q(\cdot \mid S)$ (e.g. via temperature scaling) so that supports with large $U(S)$ empirically correlate with misrouting events.

Abstention and fallback guarantees. Given an uncertainty score $U(S)$, we may introduce an abstention rule: if $U(S) > \eta$ we decline to route (or choose a conservative default), and instead evaluate using θ_{pre}. Let $L_{\text{post}}(S)$ denote the posterior-routed loss on the episode and L_{pre} the loss under θ_{pre}. By decomposing on the abstention event, we obtain the deterministic risk–coverage trade-off

$$\mathbb{E}[L] \;\le\; \Pr[U(S) \le \eta] \cdot \mathbb{E}\big[L_{\text{post}}(S) \mid U(S) \le \eta\big] \;+\; \Pr[U(S) > \eta] \cdot L_{\text{pre}}.$$

Consequently, if $U(S)$ is calibrated so that large values indeed correspond to likely misrouting (as predicted by Theorem 7.1 in the separated regime), then abstention prevents catastrophic losses from exceeding the baseline performance of θ_{pre} on the abstained episodes. This mechanism is particularly relevant when the support set is corrupted, ambiguous, or drawn from $\mathcal{P}_{\text{OOD}}$ where separation may fail.

A matching lower bound: indistinguishability is unavoidable. The previous results rely on separation. Without it, routing is information-theoretically hard: if two routings induce nearly indistinguishable support distributions, no algorithm can reliably infer which routing generated the support, regardless of computation. The next theorem formalizes this by a standard two-point testing argument.

Theorem 7.2 (Lower bound via Le Cam for binary routing)**.** *Let $\mathcal{A} = \{0, 1\}$ and suppose S consists of n_s i.i.d. samples from P_A given A. For any (possibly randomized) router $\pi(\cdot \mid S)$,*

$$\sup_{A \in \{0,1\}} \Pr_{S \sim P_A}\big[\pi(S) \neq A\big] \;\ge\; \frac{1}{2}\Big(1 - \mathrm{TV}\big(P_0^{\otimes n_s}, P_1^{\otimes n_s}\big)\Big).$$

In particular, for families with $\mathrm{TV}(P_0, P_1) = \Delta$ in the small-Δ regime, there exist constants $C, c' > 0$ such that

$$\inf_{\pi} \sup_{A \in \{0,1\}} \Pr[\pi(S) \neq A] \;\ge\; c' \exp(-C\, n_s \Delta^2),$$

showing that exponential-in-$n_s\Delta^2$ rates are tight up to constants.

Implications. Taken together, Theorems 7.1 and 7.2 provide a sharp picture: support-based routing is feasible when routings are separated at the level of support distributions, and impossible to make uniformly reliable when they are not. Thus, increasing n_s improves routing only to the extent that the induced Δ is non-negligible; conversely, when Δ is small (or when $\mathcal{P}_{\text{OOD}}$ breaks the generative assumptions), calibrated uncertainty and abstention are not merely heuristics but necessary safeguards. In Section 8 we turn to computational considerations, where even in identifiable regimes, exact deterministic routing can be intractable under realistic budget constraints.

8 Computational Complexity and Hardness

The identifiability results of Section 7 address when the support set *contains information* sufficient to discriminate routings. They do not, however, resolve the separate question of whether selecting an optimal routing is computationally tractable under realistic routing constraints. We therefore isolate the computational bottleneck at meta-test time: given a support set S, and a (learned) scoring rule or posterior $q_\psi(\alpha \mid S)$, how hard is it to compute a *deterministic* routing decision such as MAP, or more generally to optimize a routing objective under budgets?

MAP routing as discrete optimization. If α is restricted to a finite catalog $\mathcal{A}$ (as in the theory), MAP routing is trivial once we can evaluate $q_\psi(A \mid S)$ for each $A \in \mathcal{A}$. The computational issue is then moved to the construction of $\mathcal{A}$: if $\mathcal{A}$ is large (e.g. all k-sparse subsets of M experts), brute-force enumeration is infeasible. In deployed systems, the natural constraint is not finiteness per se, but a *budget* such as "activate at most k experts" or "expected cost no more than c," which yields a combinatorial feasible set.

To make this explicit, consider the common binary-selection restriction $\alpha \in \{0,1\}^M$ with $\|\alpha\|_0 \le k$ (or the simplex with α supported on at most k indices). Even if we fix masks $\{z_m\}$ and the shared delta θ_δ, the mapping

$$\alpha \;\mapsto\; \ell\left(T; \theta_{\text{pre}} + \sum_{m=1}^{M} \alpha_m (z_m \odot \theta_\delta)\right)$$

is an arbitrary nonconvex set function in general, since it is induced by a deep network. Thus, selecting an optimal routing by minimizing loss over feasible α is a priori a hard combinatorial problem.

NP-hardness via subset selection reductions. We formalize the above intuition by a standard reduction from maximum coverage (equivalently, minimum set cover under a suitable encoding of loss). Fix a universe U of elements, and let each expert m correspond to a subset $E_m \subseteq U$. Construct a synthetic family of episodes in which the query loss decreases as more universe elements are "covered" by the chosen experts, with a budget $\|\alpha\|_0 \le k$. Concretely, define a task-dependent loss of the form

$$L(\alpha) \;=\; 1 - \frac{1}{|U|}\Big|\bigcup_{m:\alpha_m=1} E_m\Big|,$$

which is minimized exactly by choosing the k experts that maximize coverage. We can realize (or approximate) such a set function as a bounded loss $\ell(T; \theta(S, \alpha, z))$ by embedding the coverage signal into a supervised classification task whose error rate matches $L(\alpha)$ up to an arbitrarily small approximation (e.g. by hard-coding features and a shallow head), while treating

the routed model as selecting which feature blocks become active. It follows that, in the worst case, computing

$$\arg \min_{\alpha \in \{0,1\}^M : \|\alpha\|_0 \le k} \ell(T; \theta(S, \alpha, z))$$

is NP-hard. This remains true if we add a cost-aware prior $p(\alpha)$ or an explicit penalty (e.g. $+\gamma \cdot \mathrm{cost}(\alpha)$): the reduction can be arranged so that all feasible selections have the same cost, or so that costs encode the same cardinality constraint.

The implication is not that routing is hopeless, but rather that *exact* deterministic routing under realistic subset constraints admits no general polynomial-time algorithm unless P = NP. Accordingly, a learned amortized router $q_\psi(\alpha \mid S)$ should be viewed as an efficient *inference heuristic* trained end-to-end, not as an algorithm that solves the underlying discrete optimization problem exactly.

When greedy or approximation guarantees apply. The above hardness is worst-case and does not preclude approximation in structured regimes. Indeed, in the coverage construction the objective $-L(\alpha)$ is a monotone submodular function of the selected set, and the classical greedy algorithm achieves a $(1 - 1/e)$ approximation under a cardinality constraint. This observation suggests a conditional message for routing: if, for the episode distribution of interest, the mapping

$$A \subseteq [M] \;\mapsto\; -\mathbb{E}[\ell(T; \theta(S, \mathbf{1}_A, z)) \mid S]$$

behaves *approximately* like a monotone submodular function (diminishing returns of adding experts), then greedy selection of experts based on marginal improvements may be effective. However, verifying submodularity for deep losses is generally infeasible, and interactions between experts (especially under shared deltas θ_δ and overlapping masks) can induce strong non-submodular effects. Thus greedy search should be interpreted as a pragmatic baseline, with guarantees only in special cases (e.g. additive expert contributions, or provably submodular surrogate scores).

A more robust strategy is to restrict the candidate family $\mathcal{A}$ a priori to a tractable size (a "routing catalog"), either by design (e.g. a small set of mixture patterns) or by construction (e.g. generate a pool of candidate routings offline by greedy/beam search on training tasks, then learn q_ψ over this pool). This returns us to the finite $\mathcal{A}$ setting where MAP is computationally easy and where the identifiability theory is directly applicable.

Continuous relaxations and nonconvexity. If we instead allow continuous $\alpha \in \Delta_M$ and optimize a score (e.g. expected loss on S or an ELBO-like objective) by gradient methods, we avoid NP-hardness at the level of

discrete search, but we inherit nonconvex optimization. Concretely, even for fixed S, minimizing $\mathbb{E}_{\alpha \sim q_\psi(\cdot|S)}[\ell(T; \theta(S, \alpha, z))]$ over ψ is a nonconvex problem due to the backbone and the router network. Moreover, enforcing sparsity of α (top-k mixtures) typically reintroduces combinatorial structure; common relaxations (Concrete/Gumbel-softmax, sparsemax/entmax, or top-k straight-through estimators) are computationally efficient but do not provide worst-case optimality guarantees.

Implications for practical inference. The preceding considerations motivate the inference choices emphasized by our algorithmic design.

First, we treat $q_\psi(\alpha \mid S)$ as the primary mechanism for routing, since it yields $O(C_{\text{router}})$ inference without per-task combinatorial search. Second, we distinguish *point* decisions (MAP or posterior mean) from *stochastic* decisions (Gibbs sampling). From a computational perspective, MAP over a large structured space is hard, whereas sampling a small number of routings from an amortized q_ψ is easy. From a statistical perspective, sampling aligns with the PAC-Bayes view (Theorems in Section 6) and provides uncertainty measures such as $U(S) = 1 - \max_A q(A \mid S)$, which we can exploit for abstention.

Third, when deployment requires strict latency/cost constraints, we can encode them either as a prior $p(\alpha)$ (biasing q_ψ toward cheap routings) or by restricting the support of $q_\psi(\cdot \mid S)$ to a small feasible subset (e.g. a handful of routings in $\mathcal{A}$). This is the computational analogue of the identifiability separation condition: even if routings are distinguishable, we must still ensure that the decision space is operationally searchable.

Finally, we note a conceptual complementarity between hardness and abstention. Theorems 7.1–7.2 show that uncertainty is unavoidable when supports are indistinguishable. The hardness discussion shows that, even when supports are informative, exact optimization over routings may be infeasible. In both cases, calibrated posteriors and fallback rules provide a principled way to trade coverage for safety without requiring exact discrete optimization at test time.

9 Experimental Plan

We view the main empirical burden as validating that posterior routing and tail-risk training improve not only *average* few-shot performance but also the *upper tail* of the per-episode loss distribution under realistic distribution shifts, and that the induced uncertainty is sufficiently calibrated to support abstention. Accordingly, our experimental plan is organized around (i) heavy OOD evaluation suites, (ii) controlled corruptions of the support set (the router input), (iii) tail-focused metrics aligned with our objective, and (iv) explicit comparisons to point routing and to search-based heuristics that

attempt to optimize routings at meta-test time.

Benchmarks and OOD suites. We will evaluate on both standard few-shot classification benchmarks and cross-domain/meta-dataset style suites. On the standard side, we will use miniImageNet, tieredImageNet, CIFAR-FS, and FC100 with $n_s \in \{1, 5\}$ shots and N-way episodes (typically 5-way and 10-way). To stress OOD generalization, we will include cross-domain transfer (e.g. train on ImageNet-like sources and test on CUB, Cars, Places, Plantae) and large heterogeneous suites (e.g. Meta-Dataset style evaluation) where test episodes come from a mixture of visually distinct datasets. We will report results separately for ID and each OOD target, but we will treat the primary success criterion as improved tail performance on the OOD mixture distribution.

Corrupted and ambiguous supports (router stress tests). Since routing depends on S, we will create controlled families of support-set corruptions that leave the query distribution unchanged (or corrupted in a known way), thereby isolating *misrouting* and *uncertainty* effects. We will consider at least the following interventions: (i) *label noise in S*: flip each support label independently with probability $\epsilon \in \{0.1, 0.2, 0.4\}$; (ii) *feature corruption in S*: apply common corruptions (Gaussian noise, blur, JPEG compression, occlusion) at multiple severities to the support images only; (iii) *shot imbalance / partial support*: delete a random fraction of support examples from a subset of classes, producing ambiguous S at fixed n_s budget; (iv) *support–query mismatch*: apply a style shift (e.g. color jitter, grayscale, sketch-like filter) to S but not to Q, modeling spurious cues that can mislead the router. In each case we will evaluate (a) the routed predictor without abstention, and (b) the abstaining predictor that can fall back to θ_{pre} (or a conservative fixed routing) when uncertainty is high. These controlled tests complement natural OOD suites by creating regimes where indistinguishability is explicit and where we can measure the degradation curve as a function of corruption strength.

Tail-quantile and robustness metrics. Let L_i denote the per-episode query loss (or $1 -$ accuracy) on episode i. In addition to standard mean accuracy, we will report metrics that directly probe the upper tail:

1. Empirical $\text{CVaR}_\rho(L)$ for $\rho \in \{0.05, 0.1, 0.2\}$, computed over the evaluation set of episodes via the usual "top-ρ fraction" estimator.

2. Empirical quantiles $Q_{0.9}(L)$ and $Q_{0.95}(L)$ (or equivalently the 10th and 5th percentile of accuracy), to make tail behavior interpretable without the auxiliary threshold optimization in CVaR.

3. *Worst-group* performance when the evaluation suite provides natural partitions (dataset identity, corruption type, severity); concretely, $\max_g \mathbb{E}[L \mid g]$ and $\mathrm{CVaR}_\rho(L \mid g)$.

Because posterior routing introduces stochasticity, we will separately report (i) *Gibbs* performance (sample $\alpha \sim q_\psi(\cdot \mid S)$ once), and (ii) *Bayes model averaging* with $K_\alpha \in \{2, 4, 8\}$ routing samples. We expect averaging to improve tail metrics disproportionately, and we will quantify the compute–robustness trade-off by plotting CVaR_ρ versus K_α.

Calibration and misrouting diagnostics. A central claim is that $q_\psi(\alpha \mid S)$ yields usable uncertainty. We will therefore evaluate calibration at two levels. First, *predictive* calibration of the final classifier (ECE, reliability diagrams, Brier score) under both ID and OOD evaluation. Second, *routing* calibration: we will treat $U(S)$ (e.g. $1 - \max_A q(A \mid S)$, posterior entropy, or the prior-posterior KL) as a score for the event "the chosen routing is suboptimal." Operationally, we will approximate "suboptimal" by comparing the achieved query loss under the chosen routing to the best loss among a small candidate set (e.g. top-J routings under q_ψ, or a catalog $\mathcal{A}$), and we will plot calibration curves of $U(S)$ versus observed excess loss. We will also report selective risk curves conditioned on $U(S) \leq \eta$ to verify that uncertainty meaningfully stratifies difficult episodes.

Abstention and risk–coverage curves. We will implement abstention rules of the form "abstain if $U(S) > \eta$" with fallback to θ_{pre} (or to a fixed low-variance routing). For each evaluation distribution we will sweep η and report: (i) coverage $c(\eta) = \Pr[U(S) \leq \eta]$; (ii) selective risk $\mathbb{E}[L \mid U(S) \leq \eta]$; (iii) overall risk with fallback, i.e. $\mathbb{E}[L]$ under abstention. We will also summarize the risk–coverage curve by area-under-curve style aggregates (or by the minimum achievable selective risk at fixed coverage levels, e.g. 80% and 95%). The goal is to empirically instantiate the decomposition in Theorem 5: abstention should cap catastrophic losses on corrupted supports without sacrificing too much coverage on clean episodes.

Baselines and comparisons. We will compare against: (i) *Point routing* baselines: deterministic routers $h_\zeta(S)$ trained with mean risk, and the same architecture trained with a CVaR objective but producing a point estimate; (ii) *Non-routed adaptation*: standard adapters/LoRA or SMAT-style sparse deltas without routing uncertainty; (iii) *All-experts* or uniform mixtures (when feasible) to separate the value of sparsity/routing from raw capacity. To address the question "why amortized posterior routing rather than test-time optimization," we will include heuristic *gradient-free* or search-based routing procedures that attempt to minimize support loss under a budget, such as greedy forward selection of experts (based on marginal improvement

on S), beam search over k-expert subsets from a fixed catalog, and random search with a fixed evaluation budget. We will report not only their mean and tail performance but also their meta-test latency, since these heuristics are plausible competitors precisely in the regime where exact discrete optimization is infeasible.

Ablations aligned with the theory. Finally, we will run ablations targeting the specific mechanisms in the objective: remove the KL regularizer (or vary λ), vary the prior $p(\alpha)$ to encode different cost preferences, vary ρ to test the tail-risk dependence, and vary n_s to empirically probe the concentration phenomenon predicted by Theorem 3. We will also ablate the size/structure of the routing family (catalog size $|\mathcal{A}|$, top-k constraints, structured masks) to determine where tail benefits saturate and where uncertainty becomes essential.

10 Discussion

We conclude by clarifying the scope of our claims, isolating the main limitations of our modeling and analysis choices, and outlining several extensions that appear technically natural within the same formalism.

Limitations of the shift model and the meaning of "robustness." Our objective is distributionally robust only in the specific sense encoded by CVaR_ρ over episodes: we emphasize the upper tail of the loss distribution induced by the meta-training episode sampler (and, at evaluation time, by a chosen test suite). This differs from worst-case robustness over an adversarial uncertainty set, and it does not, by itself, guarantee performance under arbitrary shifts in $\mathcal{P}_{\mathrm{OOD}}$. In particular, if the support sets S under $\mathcal{P}_{\mathrm{OOD}}$ are systematically misleading—e.g. they contain spurious cues that are highly predictive for the router but irrelevant for the query distribution—then even an optimally calibrated $q_\psi(\alpha \mid S)$ relative to $\mathcal{P}_{\mathrm{ID}}$ may route confidently and incorrectly. Our abstention mechanism mitigates this by capping risk at L_{pre} on abstained episodes, but the guarantee is only as meaningful as the deployment-specific choice of fallback and the calibration of the uncertainty score $U(S)$.

Moreover, the analysis assumes episodic i.i.d. sampling and bounded losses $\ell \in [0, 1]$. While boundedness is not restrictive for classification losses after normalization, the i.i.d. assumption can fail in practical evaluation streams (correlated episodes, drifting domains). In such regimes, both our PAC-Bayes-style controls and the empirical CVaR_ρ estimator may become optimistic unless one explicitly accounts for dependence (e.g. via block estimates or martingale variants).

Discretization: finite routing families versus continuous routings. Our sharpest generalization and concentration statements treat the routing family as a finite catalog $\mathcal{A}$, both to obtain explicit dependence on $\mathbb{E}_S[\mathrm{KL}(q(\cdot \mid S)\|p)]$ and to make misrouting well-defined relative to latent routings. In implementation, however, one may use continuous relaxations (simplex-valued α via softmax, Concrete/logistic-normal posteriors, or top-k relaxed gates). This creates an approximation gap in both directions: (i) a continuous router may place mass on routings not representable in a finite catalog, making it unclear how to interpret Δ-separation between "routings"; (ii) conversely, a too-small catalog may artificially discretize a smoothly varying adaptation space, forcing unnecessary routing entropy. Bridging this gap rigorously likely requires complexity measures beyond $|\mathcal{A}|$ (e.g. covering numbers or PAC-Bayes bounds for continuous hypothesis classes), and we do not attempt this here.

The discretization issue is also entangled with computational hardness. When α is constrained to select a subset of experts under a budget, exact MAP routing can encode NP-hard subset selection problems. Our method takes amortization as a design choice: we accept that routing is learned and approximate, and we shift the emphasis from exact optimality of a discrete combinatorial problem to calibrated uncertainty and tail-aware training.

Identifiability and separation assumptions. The posterior concentration claim relies on a separation parameter Δ between support distributions associated with different routings. This is a strong assumption: in realistic few-shot regimes, two routings may induce nearly indistinguishable support statistics, particularly under low n_s, class imbalance, or systematic label noise. Theorem 4 formalizes the consequence: when indistinguishability holds, no router can avoid a nontrivial misrouting probability. Practically, this suggests that improvements should be sought not only in router architecture but also in the *informativeness* of S (more shots, better curation, auxiliary task descriptors), and that abstention is not optional in high-stakes settings.

Tail-risk estimation is statistically expensive. The dependence on ρ in both optimization and generalization control is not merely an artifact: our bound scales as ρ^{-2}, and the lower-bound statement indicates this scaling is unavoidable in worst-case bounded settings. Consequently, very small ρ (e.g. $\rho = 0.01$) can be unstable unless one has a large number N of meta-training episodes and sufficiently large meta-batches B to estimate the empirical tail. This is a deployment-relevant limitation: one can train with a moderate ρ to stabilize learning and still report smaller-ρ metrics at evaluation, but then the training objective is only an approximation to the evaluation desideratum.

Extensions: structured sparsity and hierarchical expertization. Our exposition treats $z_m \in \{0,1\}^d$ abstractly; in practice, structured masks (per-layer blocks, attention heads, channels, or low-rank factors) are often preferable. From an optimization perspective, structured sparsity can reduce gradient variance (fewer independent Bernoulli gates) and can enforce hardware-aligned constraints. Formally, one may replace $\|z_m\|_0 \le (1-\tau)d$ by module-wise budgets, e.g. $\sum_\ell \|z_{m,\ell}\|_0 \le B_m$, and maintain the same meta-objective with a primal–dual enforcement of constraints. A further extension is hierarchical routing: first choose a coarse expert family (domain-level), then a fine expert (task-level), which may improve both identifiability (larger effective Δ at the top level) and interpretability.

Extensions: multi-source and cost-aware priors. The prior $p(\alpha)$ is an explicit lever for deployment constraints (latency, energy, memory bandwidth). In multi-domain settings, it is natural to use a mixture prior $p(\alpha) = \sum_j \pi_j p_j(\alpha)$, where components correspond to operating regimes (e.g. "cheap" versus "accurate") or source domains. More ambitiously, one can place a hyperprior on cost parameters and learn them jointly with the router, effectively performing empirical Bayes while retaining the $\mathrm{KL}(q\|p)$ control. In all cases, we recommend treating p as part of the system specification rather than a purely statistical regularizer.

Extensions: toward certified abstention. Our abstention guarantee (Theorem 5) is a decomposition that becomes meaningful when $U(S)$ is calibrated. A natural next step is to convert the posterior uncertainty into a certificate. One path is PAC-Bayes: for a given deployment distribution and a fixed abstention threshold η, one can attempt to bound the selective risk on the accepted set via a bound on the Gibbs risk plus an estimate of $\Pr[U(S) \le \eta]$. Another path is conformal-style calibration of $U(S)$ against excess loss or misrouting indicators on held-out episodes, yielding finite-sample guarantees on coverage or conditional error under exchangeability assumptions. We regard such certification as feasible but not automatic, and it requires careful dataset construction to avoid leakage between calibration and evaluation episodes.

Deployment guidance. We summarize practical choices that follow logically from our analysis:

1. *Choose ρ to match data and risk appetite.* Smaller ρ targets rarer failures but requires more episodes for stable optimization and evaluation.

2. *Use $K_\alpha > 1$ when tail failures matter.* Bayes model averaging over a small number of routings can reduce variance and disproportionately improve upper-tail metrics, at a linear compute cost in K_α.

3. *Treat abstention as a first-class decision.* Select η by risk–coverage analysis on validation shifts that resemble the anticipated deployment shifts; ensure the fallback θ_{pre} (or conservative routing) is acceptable.

4. *Monitor routing drift.* The quantities $\mathrm{KL}(q_\psi(\cdot \mid S) \| p)$ and the entropy of $q_\psi(\cdot \mid S)$ are inexpensive diagnostics; persistent increases can indicate support corruption or domain shift and can trigger conservative policies.

These recommendations do not remove the fundamental indistinguishability barrier, but they make explicit where the system is expected to fail gracefully rather than catastrophically.