SOFT SPECIALISTS: α-RÉNYI ENSEMBLES FOR UNCERTAINTY-AWARE LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing training approaches for large language models typically learn a single set of parameters, based on large volumes of data, which is typically heterogeneous, conflicting and often outright contradictory. As a result, the model is forced to compress conflicting goals, and inherent uncertainties into a single, averaged pattern of behaviour. We propose an α-Rényi variational framework for learning distributions over post-training parameters, offering an uncertainty-aware alternative to deep ensemble approaches. The studied variational objective interpolates between classical variational Bayes and predictively oriented posterior learning, balancing between globally plausible individual models against systems of complementary specialists. We identify local stability criteria, demonstrating how model misspecification can make non-degenerate posterior spread locally favourable. We apply our framework to LLM post-training, learning an ensemble of LoRA adapters attached to a shared, frozen base model, providing a scalable training procedure for both supervised fine-tuning and preference optimisation. Our approach enables training examples to be softly routed across ensemble members, promoting model specialisation and providing actionable uncertainty estimates across different tasks.

1 INTRODUCTION

Large language models are adapted to downstream tasks and user preferences through post-training techniques such as Supervised Fine-Tuning (SFT) (Qi et al., 2024), Direct Preference Optimisation (DPO), (Rafailov et al., 2023) or Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022; Dai et al., 2024; Ouyang et al., 2022). These approaches return a single adapted parameter vector. When post-training data are heterogeneous, ambiguous or partially conflicting, this single vector must compress multiple plausible behaviours into one compromise. This is especially problematic for alignment and high-stakes deployment. A post-trained model is expected to preserve useful base capabilities, remain helpful under ambiguous requests but still avoid harmful behaviour (Anwar et al., 2024; Ganguli et al., 2022). However, these goals need not be simultaneously well accommodated by a single set of weights: a model pushed toward safer operation may over-refuse benign requests or degrade capability, while a model tuned to remain helpful may become brittle under malicious prompts (Röttger et al., 2023; Cui et al., 2025).

The classical way to move beyond a single parameter vector is to learn a distribution over parameters. Bayesian methods such as Bayesian neural networks, variational inference, Laplace approximations, dropout-based approximations, and sampling-based methods have all been used to this end, seeking to characterise uncertainty in deep learning (Arbel et al., 2026; Kingma & Welling, 2013; Blundell et al., 2015; Kingma et al., 2015; MacKay, 1992; Yang et al., 2024; Neal, 2012; Chen et al., 2014; Gal & Ghahramani, 2016). However, the Bayesian posterior distribution is designed to address a different problem from the one identified above: it weights each parameter according to how well it individually explains the entire observed dataset under the assumed model and prior. When data is heterogeneous or conflicting, no single parameter can explain it all, which is essentially a form of model misspecification, under which the Bayesian posterior predictive distributions is known to perform arbitrarily poorly (Grünwald & van Ommen, 2017).

Deep ensembles (Fort et al., 2019; Lakshminarayanan et al., 2017) offer a practical second approach. By combining independently trained models at inference time, they often yield useful heuristic uncertainty, particularly when combined with post-hoc calibration strategies (Angelopoulos et al.,

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.

2025; Rivera et al., 2024). As their diversity is usually induced through random initialisation, data order and optimisation noise, ensemble members are not coordinated as samples of a posterior distribution. However, introducing appropriate coordination during training allows these ensembles to effectively approximate a Bayesian posterior (D'Angelo & Fortuin, 2021; Wild et al., 2023).

A third approach arises from generalised Bayesian methods (Grünwald & Langford, 2007; Zellner, 1988; Aitchison, 1975; Bissiri et al., 2016; Knoblauch et al., 2022), designed to better handle misspecification. Methods such as Gibbs posteriors (Jiang & Tanner, 2008; Martin & Syring, 2022), tempered or fractional posteriors (Bhattacharya et al., 2019; Wenzel et al., 2020), and safe-Bayesian approaches (Grünwald, 2012) adapt posterior concentration to mitigate brittleness arising from the mismatch between the (unseen) data-generating distribution and the parametrised model family. These methods are highly relevant to training large-scale deep learning models, where the distribution is rarely homogeneous. Nevertheless, they retain the same basic structure where posterior mass is assigned according to the averaged loss of individual parameter choices over the full dataset.

Recent work on predictively oriented posteriors (PrO) (McLatchie et al., 2025) builds upon the generalised Bayesian approach, by evaluating a distribution Q through the predictive quality of the mixture it induces. This is effective under heterogeneity or misspecification: no single parameter may adequately explain the data, but a mixture of complementary predictors might. In practice, one may want both: individually meaningful and regularised parameter settings, but also enough predictive cooperation to represent conflicting or heterogeneous data. This motivates the need for a principled approach to interpolating between classical and prediction oriented posteriors.

We study a one-parameter family of variational objectives that interpolates between these endpoints. Each observation is scored through the α-power mean of its likelihood under Q. As $\alpha \to 0$, this reduces to generalised variational Bayes, and when $\alpha = 1$ it recovers the predictively oriented posterior. The parameter α directly influences the geometry of the variational objective: small α favours globally plausible individual models. Larger α increasingly rewards complementary specialists whose utility emerges through aggregation. Intermediate values of α trade-off between the two extremes. Similar objectives appear in Futami et al. (2022) as an entropic-risk variational objective for parametric mean-field families, and in Li & Gal (2017) and Li & Turner (2016) as an approximation to black-box α-divergence minimisation (Hernandez-Lobato et al., 2016). Here we introduce this loss over general distributions Q, and ask when and in which directions it favours spread over concentration.

Our main application is distributional LLM post-training: instead of compressing heterogeneous data into a single adapter, we learn a structured ensemble of LoRA adapters (Hu et al., 2022). The α-Renyi objective induces soft routing: for each training example, the objective assigns responsibilities to particles according to their current likelihood on that example. Thus examples that are well explained by a subset of adapters send larger gradients to those adapters, allowing specialization to emerge from the variational objective itself rather than from a separate gating network.

Main contributions. First, we study an α-indexed variational objective over distributions of post-training parameters, interpolating between classical variational Bayes and predictively oriented posterior learning. Second, we show that under misspecification, posterior spread can become locally favourable. Third, we derive finite-particle objectives for supervised fine-tuning and direct preference optimization, yielding a scalable soft-routing procedure for LoRA ensembles, and demonstrate its empirical effectiveness for SFT and DPO.

2 Background

We recall the variational formulations underlying classical Bayesian inference and predictively oriented posteriors, and introduce the α-indexed family of objectives studied in this paper.

2.1 Interpolating between classical and predictively oriented posteriors

Let $\Theta \subseteq \mathbb{R}^d$ denote a parameter space, and let $\{p_\theta(y \mid x) : \theta \in \Theta\}$ be a parametric family of conditional predictive models. Consider a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, of independent samples from an unknown data-generating distribution $P^\star$. For a distribution $Q \in \mathcal{P}(\Theta)$, define the mixture

predictor

$$p_Q(y \mid x) = \int_\Theta p_\theta(y \mid x)\, Q(d\theta). \tag{1}$$

When Q is the Bayesian posterior $\pi(\cdot \mid \mathcal{D})$, the distribution p_Q becomes the usual Bayesian posterior predictive, thus recovering Bayesian model averaging (Masegosa, 2020).

The two families of posteriors we wish to connect differ only in how they score Q on a single observation. Classical, and generalised, variational Bayes scores each parameter draw separately and averages the resulting log-losses. Predictively oriented objectives (Masegosa, 2020; Morningstar et al., 2022; McLatchie et al., 2025) instead score the log-loss of the mixture predictor:

$$\ell_0(Q; x, y) = -\mathbb{E}_{\theta\sim Q}\big[\log p_\theta(y \mid x)\big], \qquad \ell_1(Q; x, y) = -\log p_Q(y \mid x).$$

Both endpoints are the negative logarithm of an average over member likelihoods $p_\theta(y \mid x)$ under Q: ℓ_1 takes the arithmetic mean, and ℓ_0 the geometric mean $\exp \int \log p_\theta(y \mid x)\, Q(d\theta)$, which allows for the natural interpolation for $\alpha \in (0, 1]$

$$\ell_\alpha(Q; x, y) = -\log\Big(\int_\Theta p_\theta(y \mid x)^\alpha\, Q(d\theta)\Big)^{1/\alpha} = -\tfrac{1}{\alpha}\log\int_\Theta p_\theta(y \mid x)^\alpha\, Q(d\theta). \tag{2}$$

Since power means are non-decreasing in their order, ℓ_α decreases monotonically from ℓ_0 to ℓ_1, so

$$-\log p_Q(y \mid x) \;\le\; \ell_\alpha(Q; x, y) \;\le\; -\mathbb{E}_{\theta\sim Q}\log p_\theta(y \mid x),$$

with equality only when $p_\theta(y \mid x)$ is Q-almost surely constant (Lemma 4, App. A).

Because the geometric mean scores Q as a product of experts, each member must assign reasonable likelihood to every observation. The arithmetic mean instead scores Q as a mixture, so it is sufficient for at least one member to explain each observation well. The parameter α thus controls how far a member's contribution may depend on the rest of the ensemble. The empirical variational objective is

$$\mathcal{L}_\alpha(Q) = \textstyle\sum_{n=1}^N \ell_\alpha(Q; x_n, y_n) + \lambda\,\mathrm{KL}(Q\|\pi_0), \tag{3}$$

where π_0 is a reference prior and $\lambda > 0$ controls regularisation. As $\alpha \to 0$, (3) is the generalised variational Bayes objective with learning rate $1/\lambda$, minimised by the Gibbs posterior.

Writing $L_\theta = -\log p_\theta(y \mid x)$, the loss (2) equals $-\frac{1}{\alpha}\log \mathbb{E}_Q[e^{-\alpha L_\theta}]$ for $\alpha > 0$. This functional is familiar from risk-sensitive optimisation and tilted empirical risk minimisation (Li et al., 2021). In this form, (3) was studied by Futami et al. (2022) as a risk-seeking objective over parametric variational families, with emphasis on consistency and an implicit bias towards flat minima. Related approaches based on second-order refinements of Jensen's inequality were developed by Masegosa (2020) and Futami et al. (2021). We take the complementary view of ℓ_α as a rule for aggregating likelihoods across the members of Q, and study the objective over general distributions $Q \in \mathcal{P}(\Theta)$. This perspective places the coupling between members at the centre of the analysis. In Section 3, it yields soft-routing responsibilities, a self-consistent characterisation of the minimiser and a precise local criterion for when posterior spread is favoured.

The loss (2) remains well defined for $\alpha > 1$. As $\alpha \to \infty$, a finite ensemble credits each observation only to its best-fitting member, recovering multiple choice learning Guzman-Rivera et al. (2012); Lee et al. (2016). Implicit MLE Li & Malik (2018) and explorative modelling Gladstone et al. (2026) correspond to $\alpha = 1$ and $\alpha = \infty$, but take ensembles over generations of a single parameter rather than over parameters. Related best-of-N and pass@k appear in LLM post-training. We restrict attention to $\alpha \in (0, 1]$, since for $\alpha > 1$ a concentrated posterior is unstable even when the model is well specified (Rem. 1).

3 The α-Rényi variational framework

3.1 Properties of the α-Rényi loss

We now study the geometry induced by the α-Rényi loss. As discussed in Section 2, α interpolates between the expected log-loss of individual parameter values and the log-loss of the predictive mixture induced by Q. Understanding this interpolation is key to identifying when the loss favours individual good predictors and when it rewards diversity across the support of Q.

For a fixed observation (x, y), ℓ_α aggregates the likelihoods $p_\theta(y \mid x)$ nonlinearly under Q. Increasing α gives more weight to parameters that explain the observation well. Thus, the gain over ℓ_0 is tied to predictive diversity: if all members assign similar likelihoods, nonlinear aggregation has little effect, if different members explain the observation to different degrees, it can reduce the loss. The following lemma formalizes this. See App. B.1 for the proof.

Lemma 1 (Variance lower bound)**.** *Let $(x, y) \in \mathcal{D}$, $Q \in \mathcal{P}(\Theta)$, and $\alpha \in (0, 1]$. Assume that*

$$0 < p_\theta(y \mid x) \le M_{x,y} < \infty \qquad \text{for } Q\text{-a.e. } \theta,$$

where $M_{x,y} := \operatorname{ess\,sup}_{\theta \sim Q} p_\theta(y \mid x)$. Then

$$\ell_0(Q; x, y) - \ell_\alpha(Q; x, y) \ge \frac{1}{2\alpha M_{x,y}^{2\alpha}} \operatorname{Var}_{\theta \sim Q}(p_\theta(y \mid x)^\alpha)\,.$$

Lemma 1 extends the result in Masegosa (2020) and shows that α-Rényi loss yields a guaranteed improvement over the expected log-loss. In particular, the gain vanishes when all members assign the same likelihood to the observation.

Soft-routing responsibilities To better understand how ℓ_α induces different behaviour when $\alpha > 0$, it is convenient to rephrase the loss ℓ_α in terms of observation-specific responsibilities. Intuitively, these can be interpreted as governing how learning signals are distributed among different parameter settings. For a given data point (x, y) and distribution Q, define the α-responsibility of a parameter θ

$$w^{(\alpha)}(\theta; x, y, Q) := \frac{p_\theta(y|x)^\alpha}{\int_\Theta p_{\theta'}(y|x)^\alpha \, Q(d\theta')}. \tag{4}$$

For a fixed (x, y), the responsibility in (4) induces a probability density with respect to $Q(d\theta)$ for every α. This reveals how $\alpha < 1$ differs from $\alpha = 1$: the objective gains an explicit entropy regularization term that favours diffuse responsibility distributions and hence softer specialization, whereas $\alpha = 1$ leads to more concentrated assignments.

Lemma 2 (Soft specialisation)**.** *Let $(x, y) \in \mathcal{D}$, $Q \in \mathcal{P}(\Theta)$, and $\alpha \in (0, 1]$. The following holds for ℓ_α in (2)*

$$\ell_\alpha(x, y) = \ell_1(x, y) - \tfrac{1-\alpha}{\alpha} H_\alpha(w^{(1)}; Q), \quad \textit{where} \quad H_\alpha(w; Q) = \tfrac{1}{1-\alpha} \log \textstyle\int_\Theta w(\theta)^\alpha Q(\theta)$$

denotes the Rényi entropy of order α of the responsibility distribution with respect to $Q(d\theta)$ instead of the Lebesgue measure.

See App. B.2 for the proof. Since $H_\alpha(w^{(1)}; Q)$ is negative, it follows that $\ell_\alpha \ge \ell_1$. The entropy term vanishes only when the responsibilities $w^{(1)}$ are uniform, and is largest when the responsibilities collapse to a point mass, thereby explicitly penalising responsibility collapse. This introduces a competing pressure against the predictive mixture negative log-likelihood term ℓ_1 which is seeking to make hard assignments of data points to parameters selected by Q. Consequently, intermediate values $0 < \alpha < 1$ add an entropic regularisation effect on the responsibilities, promoting soft specialisation.

3.2 Properties of the α-Rényi posterior

To understand how the α-Rényi loss influences posterior selection, we can examine the optimal finite-sample posterior. The Bayesian variational objective $\mathcal{L}_{\alpha \to 0}$ in (3) admits a unique minimiser which has the usual product form

$$\tfrac{dQ_0^*}{d\pi_0}(\theta) = \tfrac{1}{Z_0} \textstyle\prod_{i=1}^N p_\theta(y_i \mid x_i)^{\frac{1}{\lambda}},$$

so that each parameter θ is weighted depending on its individual performance across all the data. Increasing $\alpha > 0$ introduces a non-linear coupling over the parameter space. We first establish convexity of (3), see App. B.3 for the proof.

Lemma 3 (Convexity of the α-Rényi variational objective)**.** *Fix $\alpha > 0$. Define for each data point $z = (x, y)$*

$$A_z(Q) := \textstyle\int_\Theta p_\theta(y \mid x)^\alpha \, Q(d\theta), \qquad \ell_\alpha(Q; z) := -\tfrac{1}{\alpha} \log A_z(Q),$$

on the domain where $0 < A_z(Q) < \infty$. Then $Q \mapsto \ell_\alpha(Q; z)$ is convex on this domain. Consequently, for a fixed dataset $\mathcal{D} = \{z_i\}_{i=1}^N$, a prior $\pi_0 \in \mathcal{P}(\Theta)$ and $\lambda \ge 0$, the functional $\mathcal{L}_\alpha$ in (3) is convex on the convex domain $\mathcal{Q}_\alpha := \big\{Q \in \mathcal{P}(\Theta) : \mathrm{KL}(Q\|\pi_0) < \infty,\ 0 < A_{z_i}(Q) < \infty \text{ for } i = 1, \dots, N\big\}$. In particular, if $\lambda > 0$, then $\mathcal{L}_\alpha$ is strongly convex with respect to total variation on $\mathcal{Q}_\alpha$. Hence $\mathcal{L}_\alpha$ has at most one minimiser on $\mathcal{Q}_\alpha$.

The following proposition characterises a minimiser with strictly positive density through a self-consistency equation. The proof is provided in App. B.4.

Proposition 1 (Self-consistent α-posterior)**.** *Let $\mathcal{D} = \{z_i\}_{i=1}^N$, with $z_i = (x_i, y_i)$, be a fixed dataset, and fix $\alpha > 0$ and $\lambda > 0$. Consider $\mathcal{L}_\alpha$ defined by (3) on $\mathcal{Q}_\alpha$. Suppose that it admits a minimiser $Q_\alpha^\star \in \mathcal{Q}_\alpha$ whose density*

$$q_\alpha^\star(\theta) := \tfrac{dQ_\alpha^\star}{d\pi_0}(\theta)$$

is strictly positive π_0-almost everywhere. Then $Q_\alpha^\star$ is the unique minimiser of $\mathcal{L}_\alpha$ on $\mathcal{Q}_\alpha$, and its density satisfies, π_0-almost everywhere,

$$q_\alpha^\star(\theta) = \tfrac{1}{Z_\alpha} \exp\Big(\tfrac{1}{\lambda\alpha} \textstyle\sum_{i=1}^N w_i^{(\alpha)}(\theta; Q_\alpha^\star) \Big), \tag{5}$$

where $Z_\alpha \in (0, \infty)$ is the normalising constant, and

$$w_i^{(\alpha)}(\theta; Q_\alpha^\star) := \tfrac{p_\theta(y_i|x_i)^\alpha}{\int_\Theta p_\vartheta(y_i|x_i)^\alpha Q_\alpha^\star(d\vartheta)}.$$

Conversely, any $Q \in \mathcal{Q}_\alpha$ with strictly positive density $q = dQ/d\pi_0$ satisfying (5), with Q in place of $Q_\alpha^\star$, is the unique minimiser.

From (5), we can see that each observation induces a local reweighting of the current posterior toward those parameter settings that explain it well. The final posterior reconciles all such local predictive tilts simultaneously. Thus $Q_\alpha^\star$ is best understood as a *self-consistent predictive posterior*, in contrast to the generalised Bayes posterior, which concentrates on parameter values that are individually plausible explanations of the observed data. An example demonstrating the effect of the non-linear dependence on Q is provided in App B.5.

While the α-Rényi posterior differs in structure compared to generalised Bayesian posteriors over finitely many observations, we can show that they are consistent with respect to each other at the population level. In the idealised setting where the true data-generating distribution lies strictly within our model class (i.e. well-specified case), the α-Rényi population loss is still minimised by the true parameter, so that the ensemble is not forced to diversify even when a single, perfect predictor exists. This is formalised in the following proposition, which builds on Masegosa (2020), see App. B.6.

Proposition 2 (Well-specified population minimisers)**.** *Assume the model is well-specified, so that there exists $\theta^\star \in \Theta$ with $p_{\theta^\star}(\cdot \mid x) = p^\star(\cdot \mid x)$ for $P_X^\star$-a.e. x. Then $\delta_{\theta^\star}$ is a minimiser of the population risk $\mathcal{R}_\alpha(Q) := \mathbb{E}_{P^\star}[\ell_\alpha(Q; X, Y)]$, over $Q \in \mathcal{P}(\Theta)$, for every $\alpha \in (0, 1]$.*

For $\alpha = 1$, every minimiser satisfies $p_Q = p^\star$, whereas for $0 < \alpha < 1$, every minimiser is supported on the exact-fit set $\Theta^\star := \{\theta \in \Theta : p_\theta(\cdot \mid x) = p^\star(\cdot \mid x)$ for $P_X^\star$-a.e. $x\}$.

3.3 Stability under misspecification

To better understand how the α-objective departs from the classical variational Bayes regime under misspecification, we first study its local behaviour near $\alpha = 0$. Let $L_\theta(x, y) := -\log p_\theta(y \mid x)$ denote the negative log-likelihood. Assume $L_\theta(z)$ is bounded below and $\mathbb{E}_Q|L_\theta(z)|^3 < \infty$. For sufficiently small $\alpha > 0$, we can show that the lower bound in Lemma 1 is tight, by taking a Taylor expansion of the cumulant generating function. This yields

$$\ell_\alpha(Q; x, y) = \ell_0(Q; x, y) - \tfrac{\alpha}{2} \operatorname{Var}_Q\big(L_\theta(x, y)\big) + \mathcal{O}(\alpha^2). \tag{6}$$

Eq. (6) makes clear that positive α introduces a first-order correction to the classical variational objective, which permits dispersion in the per-example loss under Q. Similar bounds have been derived in previous works, seeking to introduce second order corrections to Jensen's inequality, Masegosa (2020); Liao & Berg (2019); Becker (2012). Intuitively, this expansion shows that amongst distributions with comparable mean loss, positive α will favour ones whose ensemble members make different contributions to prediction.

The following result formalises this, showing that if Q is close to a Dirac measure δ_θ, then centred local spread can increase the population α-risk, governed by a local stability operator.

Proposition 3 (Local expansion under centred perturbations)**.** *Let $\Theta \subseteq \mathbb{R}^d$ be open, fix $\theta \in \Theta$ and $\alpha \in [0, 1]$ and write $L_\vartheta(z) = -\log p_\vartheta(y \mid x)$ for $z = (x, y)$. Let ξ be a random vector, independent*

of $Z \sim P^\star$, such that $\mathbb{E}[\xi] = 0$, $\mathrm{Cov}(\xi) = \Sigma$ and $\|\xi\| \le b$ almost surely for some $b > 0$. Assume that there exists $r > 0$ with $\overline{B}(\theta, r) \subset \Theta$ such that, for $P^\star$-almost every z, the map $\vartheta \mapsto L_\vartheta(z)$ is finite and twice continuously differentiable on a neighbourhood of $\overline{B}(\theta, r)$. For $0 \le t < r/b$, define $Q_t = \mathrm{Law}(\theta + t\xi)$ and $\Sigma_t = \mathrm{Cov}(Q_t)$. Then,

$$\ell_\alpha(Q_t; z) = L_\theta(z) + \tfrac{1}{2}\,\mathrm{Tr}\big[\big(\nabla^2_\theta L_\theta(z) - \alpha \nabla_\theta L_\theta(z) \nabla_\theta L_\theta(z)^\top\big)\Sigma_t\big] + o(t^2), \qquad t \downarrow 0, \quad (7)$$

where $\ell_0(Q; z) = \mathbb{E}_Q[L_\vartheta(z)]$. Suppose additionally that there are measurable nonnegative functions G and K such that, for $P^\star$-almost every z,

$$\sup_{\vartheta \in \overline{B}(\theta,r)} \|\nabla_\vartheta L_\vartheta(z)\| \le G(z), \quad \sup_{\vartheta \in \overline{B}(\theta,r)} \|\nabla^2_\vartheta L_\vartheta(z)\|_{\mathrm{op}} \le K(z),$$
$$\mathbb{E}_{P^\star}\big[|L_\theta(Z)| + G(Z)^2 + K(Z)\big] < \infty.$$

Then the population risk $\mathcal{R}_\alpha(Q) = \mathbb{E}_{P^\star}[\ell_\alpha(Q; Z)]$ satisfies

$$\mathcal{R}_\alpha(Q_t) = \mathcal{R}_\alpha(\delta_\theta) + \tfrac{1}{2}\,\mathrm{Tr}[(V(\theta) - \alpha J(\theta))\Sigma_t] + o(\|\Sigma_t\|_{\mathrm{op}}), \qquad t \downarrow 0, \quad (8)$$

where

$$V(\theta) = \mathbb{E}_{P^\star}[\nabla^2_\theta L_\theta(Z)], \qquad J(\theta) = \mathbb{E}_{P^\star}[\nabla_\theta L_\theta(Z) \nabla_\theta L_\theta(Z)^\top].$$

Remark 1 (The $V - \alpha J$ stability matrix and α_{critical})**.** *Eq. (8) shows that $V(\theta) - \alpha J(\theta)$ governs the local stability of a concentrated population minimiser against infinitesimal spread in Q-space. In the well specified case, Bartlett's identity implies $V(\theta^\star) = J(\theta^\star) = I(\theta^\star)$, so $V(\theta^\star) - \alpha J(\theta^\star) = (1 - \alpha) I(\theta^\star)$, and hence Dirac posteriors are stable to first order iff $\alpha < 1$. Under misspecification, stability is lost once $V(\theta^\star) - \alpha J(\theta^\star) \preceq 0$, hence small non-degenerate perturbations of the Dirac posterior may reduce the population risk.*

We can explicitly characterise the critical threshold α_{critical} at which posterior spread becomes locally favourable under general misspecification. A concentrated posterior at $\theta^\star$ becomes unstable to spread in a direction $u \in \mathbb{R}^d$ if $u^\top (V(\theta^\star) - \alpha J(\theta^\star)) u < 0$. Assuming the Fisher information matrix $J(\theta^\star) \succ 0$, the threshold is determined by the minimum of the generalised Rayleigh quotient

$$\alpha_{\text{critical}} = \min_{u \ne 0} \tfrac{u^\top V(\theta^\star) u}{u^\top J(\theta^\star) u}. \quad (9)$$

Remark 2 (Invariance under reparametrisation)**.** *Let $\theta = h(\eta)$ be a C^2 local reparametrisation with $h(\eta^\star) = \theta^\star$ and nonsingular Jacobian $A = Dh(\eta^\star)$. At a stationary point $\theta^\star$ satisfying $\mathbb{E}_{P^\star}[\nabla_\theta L_{\theta^\star}(Z)] = 0$, the corresponding matrices in the η-coordinates satisfy*

$$\widetilde{V}(\eta^\star) = A^\top V(\theta^\star) A, \qquad \widetilde{J}(\eta^\star) = A^\top J(\theta^\star) A.$$

Thus the existence of a strictly negative direction of $V(\theta^\star) - \alpha J(\theta^\star)$ is invariant under reparametrisation. If $J(\theta^\star) \succ 0$, the generalised Rayleigh quotient also gives the same threshold $\alpha_{\mathrm{critical}}$.

Corollary 1 (Local expansion with prior-potential regularisation)**.** *Under the assumptions of Proposition 3, assume $\Sigma \ne 0$ and let π_0 be a prior density that is strictly positive and has twice continuously differentiable log-density in a neighbourhood of θ. For $\lambda \ge 0$, define*

$$U(\vartheta) = -\log \pi_0(\vartheta), \qquad \mathcal{F}_\alpha(Q) = \mathcal{R}_\alpha(Q) + \lambda\, \mathbb{E}_{\vartheta \sim Q}[U(\vartheta)].$$

Then, for $Q_t = \mathrm{Law}(\theta + t\xi)$ and $\Sigma_t = t^2 \Sigma$,

$$\mathcal{F}_\alpha(Q_t) = \mathcal{F}_\alpha(\delta_\theta) + \tfrac{1}{2}\,\mathrm{Tr}\big[\big(V(\theta) - \alpha J(\theta) + \lambda \nabla^2_\theta U(\theta)\big)\Sigma_t\big] + o(\|\Sigma_t\|_{\mathrm{op}}), \qquad t \downarrow 0, \quad (10)$$

where $\mathcal{F}_\alpha(\delta_\theta) = \mathbb{E}_{P^\star}[L_\theta(Z)] + \lambda U(\theta)$.

The matrix $V(\theta) - \alpha J(\theta) + \lambda \nabla^2_\theta U(\theta)$ governs the quadratic effect of centred local spread for this prior-potential surrogate. If $\nabla^2_\theta U(\theta) \succeq 0$, the prior curvature acts as an additional force favouring concentration.

To illustrate the local stability mechanism in Proposition 3, we consider a linear-Gaussian model for which the critical value of α and the resulting predictive uncertainty can be computed explicitly.

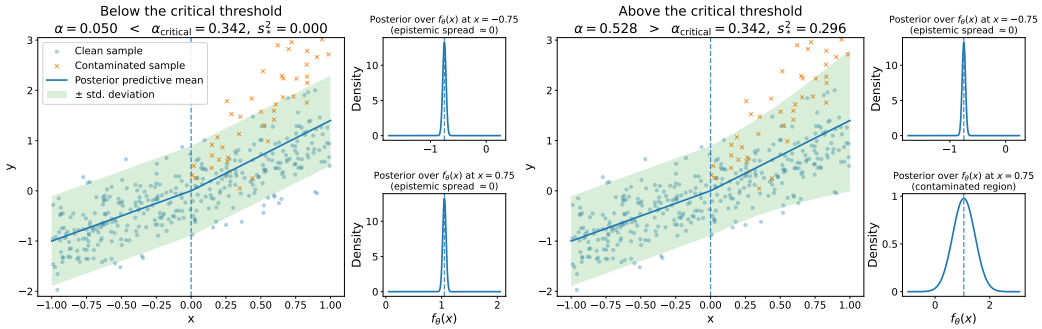

Figure 1: Behaviour of the model in Example 1 for α above and below the critical threshold.

Example 1 (Localised uncertainty under contamination)**.** *Let* $X \sim \mathrm{Unif}[-1,1]$ *and consider*

$$Y \mid X = x, \theta \sim \mathcal{N}\big(\phi(x)^\top \theta, \sigma^2\big), \qquad \phi(x) := (x, x_+)^\top, \qquad x_+ := \max(x, 0). \tag{11}$$

The first feature controls the global slope, whereas x_+ *modifies the prediction only on the positive half-line. Suppose that the true data-generating distribution is*

$$Y \mid X = x \sim \begin{cases} \mathcal{N}(\beta x, \sigma^2), & x < 0, \\ (1-\varepsilon)\mathcal{N}(\beta x, \sigma^2) + \varepsilon \mathcal{N}((\beta + a)x, \sigma^2), & x \geq 0, \end{cases}$$

where $\varepsilon \in (0,1)$ *and* $a \neq 0$*. Its conditional mean* $\mathbb{E}[Y \mid X = x]$ *is represented by* $m^\star = (\beta, \varepsilon a)^\top$*, but its conditional variance*

$$\mathrm{Var}(Y \mid X = x) = \sigma^2 + \varepsilon(1-\varepsilon)a^2 x_+^2$$

is heteroscedastic and therefore misspecified by the model in (11). For a Gaussian posterior ansatz $Q = \mathcal{N}(m^\star, \Sigma)$*, the local stability matrix in Proposition 3 is*

$$M_\alpha := V(m^\star) - \alpha J(m^\star) = \tfrac{1-\alpha}{\sigma^2}\mathbb{E}\big[\phi(X)\phi(X)^\top\big] - \tfrac{\alpha\varepsilon(1-\varepsilon)a^2}{\sigma^4}\mathbb{E}\big[X_+^2 \phi(X)\phi(X)^\top\big].$$

The second term acts only on $x \geq 0$*, selectively destabilising the parameter direction associated with* x_+*. The first direction to become unstable is* $u = e_2$*, and the corresponding critical value is*

$$\alpha_{\mathrm{critical}} = \tfrac{\sigma^2 \mathbb{E}[X_+^2]}{\sigma^2 \mathbb{E}[X_+^2] + \varepsilon(1-\varepsilon)a^2 \mathbb{E}[X_+^4]}.$$

Thus, for $\alpha > \alpha_{\mathrm{critical}}$*, posterior spread along the contamination direction locally decreases the population* α*-risk: for* $\Sigma = \mathrm{diag}(0, s^2)$*, the predictive variance becomes*

$$\mathrm{Var}_Q(Y \mid X = x) = \sigma^2 + s^2 x_+^2.$$

Hence the $s^2 x_+^2$ *component vanishes on the clean region* $x < 0$ *and grows only on the contaminated region* $x \geq 0$*. This example shows that the eigenstructure of* $V - \alpha J$ *determines not only when posterior spread becomes favourable, but also where the resulting uncertainty is expressed in the input space. See Figure 1 for an illustration of posterior uncertainties in each regime.*

4 Finite-particle LoRA ensembles

The preceding sections define the α-Rényi objective as a variational problem over probability measures $Q \in \mathcal{P}(\Theta)$. In practice, we do not optimise over arbitrary measures directly. Instead, we approximate Q using a finite ensemble of M LoRA adapters, $Q_M = \frac{1}{M}\sum_{i=1}^M \delta_{\theta_i}$, where each θ_i parametrises a model state. Particle i induces an effective model $W(\theta_i) = W_0 + \Delta W(\theta_i)$ built on the frozen pretrained model W_0. This reduces the variational problem to a finite-dimensional optimisation over $\theta_1, \ldots, \theta_M$, which we solve using mini-batch AdamW. Direct substitution of Q_M into $\mathcal{L}_\alpha$, however, is ill-posed, since the divergence term $\mathrm{KL}(Q_M \| \pi_0)$ is singular whenever the prior π_0 admits a continuous density. We therefore consider two practical approximations: a smoothed empirical-density estimate of the KL term, and a simpler prior-potential surrogate which drops the entropy term. The full algorithm is provided in Alg. 1.

Supervised fine-tuning. For a minibatch $\mathcal{B} = \{(x_b, y_b)\}_{b=1}^{B}$, the supervised objective is

$$\widehat{\mathcal{L}}_{\alpha}^{\mathrm{SFT}}(\theta_{1:M}) = \tfrac{N}{|\mathcal{B}|} \textstyle\sum_{(x_b,y_b)\in\mathcal{B}} \ell_{\alpha}^{(M)}(\theta_{1:M}; x_b, y_b) + \mathcal{R}(\theta_{1:M}), \tag{12}$$

where $\ell_{\alpha}^{(M)}(\theta_{1:M}; x, y) = -\frac{1}{\alpha}\log(\frac{1}{M}\sum_i p_{\theta_i}(y|x)^{\alpha})$ and $\mathcal{R}$ is a regulariser. Differentiating gives

$$\nabla_{\theta_i}\widehat{\mathcal{L}}_{\alpha}^{\mathrm{SFT}} = -\tfrac{N}{|\mathcal{B}|}\textstyle\sum_{b\in\mathcal{B}} w_{i,b}^{(\alpha)} \nabla_{\theta_i}\log p_{\theta_i}(y_b \mid x_b) + \nabla_{\theta_i}\mathcal{R}, \qquad w_{i,b}^{(\alpha)} = \frac{p_{\theta_i}(y_b|x_b)^{\alpha}}{\sum_{j=1}^{M} p_{\theta_j}(y_b|x_b)^{\alpha}}.$$

Each example contributes to particle i in proportion to its responsibility $w_{i,b}^{(\alpha)}$. Note that, the responsibilities can be computed from the current particle likelihoods and treated as fixed weights during backpropagation. Thus, each optimisation step reduces to a weighted SFT update per particle: one forward pass to compute the responsibilities, followed by standard gradient computations.

Preference optimisation. For pairwise preference data (x, y^+, y^-), each particle defines a DPO-style (Bradley-Terry) preference likelihood $r_i(x, y^+, y^-)$ relative to a frozen reference model (see App. C.2 for details). The α-aggregated preference loss is

$$\ell_{\alpha}^{(M)}(\theta_{1:M}; x_b, y_b^+, y_b^-) = -\tfrac{1}{\alpha}\log\left(\tfrac{1}{M}\textstyle\sum_{i=1}^{M} r_i(x_b, y_b^+, y_b^-)^{\alpha}\right).$$

This induces the same responsibility mechanism, now routing preference pairs toward particles whose current policy-reference margin explains the preference well.

For the regularisation term, recall that if Q admits a density $q = dQ/d\theta$ and π_0 denotes the prior density, then $\mathrm{KL}(Q\|\pi_0) = \int q(\theta)\log q(\theta)\,d\theta - \int q(\theta)\log\pi_0(\theta)\,d\theta$, i.e. the negative differential entropy of Q plus the expected prior potential. For the empirical measure Q_M, the entropy term is not well defined, we therefore consider the prior potential surrogate $\mathcal{R}_{\mathrm{prior}}(\theta_{1:M}) = -\frac{\lambda}{M}\sum_{i=1}^{M}\log\pi_0(\theta_i)$.

Appendix C.3 connects the responsibility-weighted updates to measure-valued gradient flows and distinguishes the full KL objective from the finite-particle surrogates used in practice.

5 Experiments

We investigate how α-Rényi LoRA ensembles affect specialisation and predictive disagreement in LLM post-training. The experiments test whether increasing α (i) promotes specialisation across particles, and (ii) increases refusal disagreement more strongly on difficult benign requests than on a harmful comparison set. See App. D for experimental details.

Specialisation on MMLU. We first test whether increasing the Rényi parameter induces specialisation across particles while preserving ensemble accuracy. We evaluate Nemotron-3-8B, Mistral-7B-Instruct-v0.3, Granite-3.1-8B-Instruct, and OLMo-2-1124-7B-Instruct using ensembles of eight rank-4 LoRA adapters. Each ensemble first undergoes independent pre-adaptation, followed by subject-balanced fine-tuning on MMLU. For each model and seed, all $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ start from the same pre-adapted checkpoint. We evaluate five training seeds on 14042 test questions.

Fig. 2 compares ensemble accuracy and conditional particle-correctness mutual information (MI) with independently trained deep ensembles and a post-hoc Laplace-LoRA baseline. MI measures disagreement about the probability of the correct option through the difference between ensemble and average particle binary entropy, as defined in (17), providing a proxy for specialisation. Mean conditional MI increases monotonically with α in every model family and exceeds the deep-ensemble baseline at $\alpha = 1$, while ensemble accuracy remains stable. Laplace-LoRA produces greater disagreement (Fig. 4), but its ensemble accuracy is below the base-model. These results support a favourable trade-off between specialisation and collective accuracy for the Rényi sweep.

Refusal disagreement after preference fine-tuning. We study how the Rényi parameter affects disagreement in refusal behaviour after preference fine-tuning. We use Qwen2-1.5B-Instruct, Phi-3-mini-4k-instruct and Nemotron-3-8B-base-4k, each with an ensemble of $M = 10$ rank-4 LoRA adapters and frozen pretrained weights. We train on `trl-lib/ultrafeedback_binarized` using the Rényi DPO objective in App. C.2 with $\alpha \in \{0, 0.5, 1\}$. We use five training seeds, matching adapter initialisations, minibatch order, and optimisation settings across α.

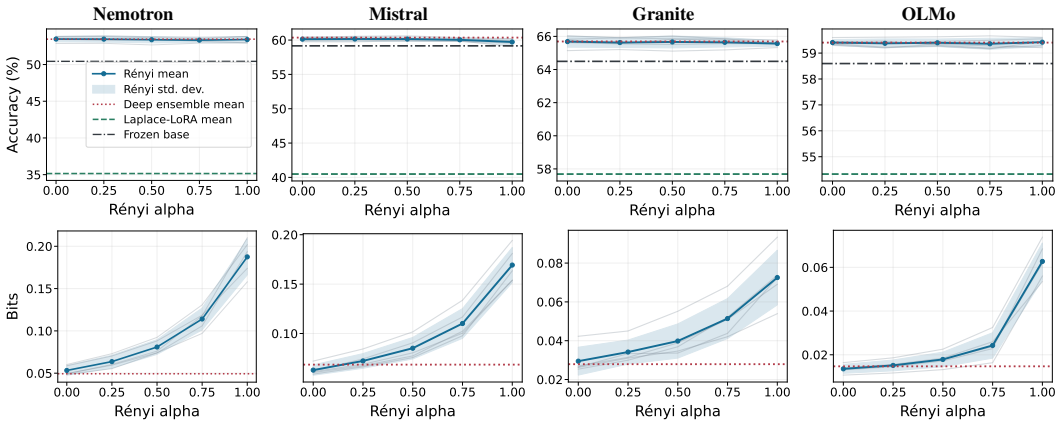

Figure 2: **MMLU specialisation.** Columns: Nemotron, Mistral, Granite, and OLMo. Top: ensemble accuracy, bottom: conditional particle-correctness MI. Given its substantially lower ensemble accuracy, the Laplace-LoRA reference is omitted from the MI panels to improve readability.

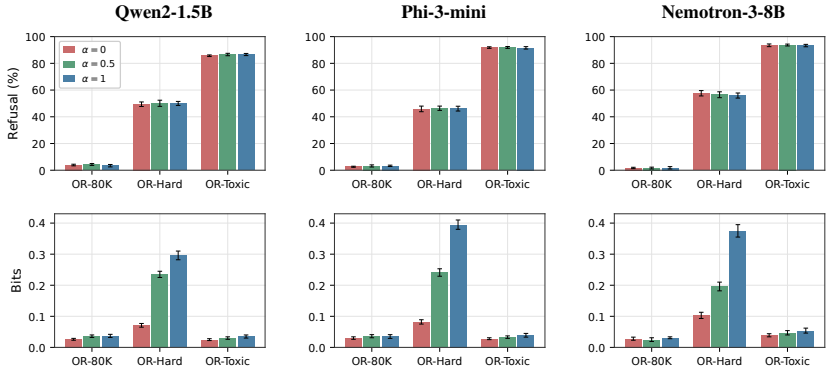

Figure 3: **Refusal rate (top) and conditional particle-refusal MI (bottom) across α.** Bars show the mean $\pm$ standard deviation across training seeds. Full results shown in Table 1 in App. D.3

We evaluate on three OR-Bench sets: easy and difficult benign requests from OR-80k and OR-Hard, and harmful requests from OR-Toxic. For each prompt and adapter, we generate $K = 60$ responses at temperature 0.7, and apply a fixed refusal judge. From the resulting refusal frequencies, we compute a plug-in estimate of conditional particle-refusal mutual information (Eq. 18), which measures the dependence between particle identity and refusal outcome given the prompt, quantifying between-adapter disagreement in refusal behaviour. Fig. 3 shows that increasing α produces substantially larger increases in conditional particle–refusal MI on OR-Hard than on OR-80K or OR-Toxic, while mean ensemble refusal probabilities change comparatively little. This suggests that ensemble disagreement could serve as a signal for identifying difficult benign requests that warrant further review.

6 Discussion and Limitations

The α-Rényi objective provides a principled way to train a distribution over post-training adapters rather than a single adapter. Its soft-routing responsibilities connect the statistical objective to an implementable LoRA ensemble training procedure. The local stability analysis explains why misspecification can make posterior spread useful, and the experiments show that this spread corresponds to specialization and alignment-relevant disagreement.

The approach also has limitations. The uncertainty signal depends on maintaining sufficient particle diversity; if all particles assign uniformly poor likelihood to an out-of-support example, responsibilities alone need not identify a safe response. Training multiple adapters increases compute relative to a single adapter, although the frozen backbone is shared. Finally, α is treated as a hyperparameter here; adaptive or data-dependent selection of α is a promising direction for future work.

AI USE STATEMENT

We acknowledge the use of Large Language Models (LLMs) for preparing this manuscript. Specifically, LLMs were used to improve grammar and clarity and generate boilerplate code snippets for our experiments and testing scripts. The authors have carefully reviewed and edited all LLM-generated outputs and take full responsibility for the final content and scientific integrity of this work.

REPRODUCIBILITY STATEMENT

Details of our experimental setup and implementation are provided in Section 5 and Appendix D. We will release the code to facilitate reproduction of our results.

REFERENCES

James Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on Bayesian neural networks: review and debates. *Statistical Science*, 41(2):316–353, 2026.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Robert A Becker. The variance drain and Jensen's inequality. *CAEPR Working Paper, No. 2012-004*, 2012.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *The Annals of Statistics*, 47(1):39–66, 2019.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78 (5):1103–1130, 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, pp. 1613–1622. PMLR, 2015.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian Monte carlo. In *International Conference on Machine Learning*, pp. 1683–1691. PMLR, 2014.

Sinho Chewi, Atsushi Nitanda, and Matthew S Zhang. Uniform-in-log-sobolev inequality for the mean-field langevin dynamics with convex energy. *SIAM Journal on Mathematical Analysis*, 58 (4):3652–3668, 2026.

Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. In *International Conference on Machine Learning*, pp. 11515–11542. PMLR, 2025.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Futoshi Futami, Tomoharu Iwata, Issei Sato, Masashi Sugiyama, et al. Loss function based second-order jensen inequality and its application to particle variational inference. *Advances in Neural Information Processing Systems*, 34:6803–6815, 2021.

Futoshi Futami, Tomoharu Iwata, Naonori Ueda, Issei Sato, and Masashi Sugiyama. Predictive variational Bayesian inference as risk-seeking optimization . In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5051–5083. PMLR, 28–30 Mar 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Alexi Gladstone, Heng Ji, and Yilun Du. Explorative modeling: Unlocking a third pretraining axis and end-to-end generation. *arXiv preprint arXiv:2607.27372*, 2026.

Peter Grünwald. The safe Bayesian: learning the learning rate via the mixability gap. In *International Conference on Algorithmic Learning Theory*, pp. 169–183. Springer, 2012.

Peter Grünwald and John Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, 66(2):119–149, 2007.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4):1069 – 1103, 2017.

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pp. 1511–1520. PMLR, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wenxin Jiang and Martin A. Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207 – 2231, 2008.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research*, 23:1–109, 2022.

Yunbum Kook, Matthew S Zhang, Sinho Chewi, Murat A Erdogdu, and Mufan Bill Li. Sampling from the mean-field stationary distribution. In *The thirty seventh annual conference on learning theory*, pp. 3099–3136. PMLR, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. *Advances in Neural Information Processing Systems*, 29, 2016.

Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted Empirical Risk Minimization. In *International Conference on Learning Representations*, 2021.

Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In *International Conference on Machine Learning*, pp. 2052–2061. PMLR, 2017.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.

Jason G Liao and Arthur Berg. Sharpening Jensen's inequality. *The American Statistician*, 2019.

David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Ryan Martin and Nicholas Syring. Direct Gibbs posterior inference on risk minimizers: Construction, concentration, and calibration. In *Handbook of Statistics*, volume 47, pp. 1–41. Elsevier, 2022.

Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.

Yann McLatchie, Badr-Eddine Cherief-Abdellatif, David T Frazier, and Jeremias Knoblauch. Predictively oriented posteriors. *arXiv preprint arXiv:2510.01915*, 2025.

Warren R Morningstar, Alex Alemi, and Joshua V Dillon. Pacm-bayes: Narrowing the empirical risk gap in the misspecified bayesian regime. In *International Conference on Artificial Intelligence and Statistics*, pp. 8270–8298. PMLR, 2022.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex analysis of the mean field langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pp. 9741–9757. PMLR, 2022.

Yann Ollivier, Hervé Pajot, and Cédric Villani. *Optimal transport: Theory and applications*, volume 413. Cambridge University Press, 2014.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, volume 2024, pp. 30988–31043, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Eduardo Ochoa Rivera, Yash Patel, and Ambuj Tewari. Conformal prediction for ensembles: Improving efficiency via score-based aggregation. *arXiv preprint arXiv:2405.16246*, 2024.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. *arXiv preprint arXiv:2308.01263*, 2023.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Convergence of mean-field langevin dynamics: time-space discretization, stochastic gradient, and variance reduction. *Advances in Neural Information Processing Systems*, 36:15545–15577, 2023.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In *International Conference on Machine Learning*, pp. 10248–10259. PMLR, 2020.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) Bayesian methods. *Advances in Neural Information Processing Systems*, 36:39782–39811, 2023.

Adam Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 1812–1842, 2024.

Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42 (4):278–280, 1988.

A ADDITIONAL DETAILS FOR SECTION 2

We formalise how the parameter α induces a smooth interpolation between classical parameter-based variational inference and predictive mixture training.

Lemma 4 (Limits, interpolation, and monotonicity of the α-loss)**.** *Fix a data point (x, y) and a distribution $Q \in \mathcal{P}(\Theta)$. Assume that*

$$p_\theta(y \mid x) > 0 \qquad \text{for } Q\text{-a.e. } \theta,$$

and that $\log p_\theta(y \mid x)$ and $p_\theta(y \mid x)^\alpha$ are Q-integrable for the values of α under consideration. Then the following hold.

1. *(*Variational objective limit as $\alpha \downarrow 0$*) The one-sided limit $\alpha \downarrow 0$,*
$$\ell_0(Q; x, y) := \lim_{\alpha \downarrow 0} \ell_\alpha(Q; x, y) = -\mathbb{E}_{\theta \sim Q}\big[\log p_\theta(y \mid x)\big].$$

2. *(*Predictive mixture at $\alpha = 1$*) At $\alpha = 1$,*
$$\ell_1(Q; x, y) = -\log \mathbb{E}_{\theta \sim Q}\big[p_\theta(y \mid x)\big] = -\log p_Q(y \mid x).$$

3. *(*Interpolation inequality*) For every $\alpha \in (0, 1]$,*
$$-\log \mathbb{E}_{\theta \sim Q}\big[p_\theta(y \mid x)\big] \;\le\; \ell_\alpha(Q; x, y) \;\le\; -\mathbb{E}_{\theta \sim Q}\big[\log p_\theta(y \mid x)\big].$$

4. *(*Monotonicity in α*) For fixed $Q \in \mathcal{P}(\Theta)$, $(x, y) \in \mathcal{D}$, the map*
$$\alpha \mapsto \ell_\alpha(Q; x, y)$$
is non-increasing on $(0, \infty)$. In particular, for $0 < \alpha_1 \le \alpha_2 \le 1$,
$$\ell_{\alpha_1}(Q; x, y) \ge \ell_{\alpha_2}(Q; x, y).$$
The inequality is strict when $\alpha_1 < \alpha_2$ unless $p_\theta(y \mid x)$ is Q-a.s. constant.

Proof. Let
$$Z := p_\theta(y \mid x), \qquad \theta \sim Q.$$
By assumption, $Z > 0$ Q-a.s. The α-loss may be written as
$$\ell_\alpha(Q; x, y) = -\tfrac{1}{\alpha} \log \mathbb{E}_Q[Z^\alpha] = -\tfrac{1}{\alpha} \log \mathbb{E}_Q[\exp(\alpha \log Z)].$$

1. *Limit as $\alpha \to 0$.* Define $\varphi(\alpha) := \log \mathbb{E}_Q[\exp(\alpha \log Z)]$. Since $\varphi(0) = 0$ and, under the stated integrability assumptions, $\varphi'(0) = \mathbb{E}_Q[\log Z]$, we obtain
$$\ell_\alpha(Q; x, y) = -\tfrac{\varphi(\alpha)}{\alpha} \longrightarrow -\varphi'(0) = -\mathbb{E}_Q[\log Z].$$
This proves the first claim.

2. *Predictive mixture at $\alpha = 1$.* At $\alpha = 1$,
$$\ell_1(Q; x, y) = -\log \mathbb{E}_Q[Z] = -\log \textstyle\int_\Theta p_\theta(y \mid x)\, Q(d\theta) = -\log p_Q(y \mid x).$$

3. *Interpolation inequality.* For $0 < \alpha \le 1$, the standard power-mean inequality for the positive random variable Z gives
$$\exp\{\mathbb{E}_Q[\log Z]\} \le \big(\mathbb{E}_Q[Z^\alpha]\big)^{1/\alpha} \le \mathbb{E}_Q[Z].$$
Taking $-\log$ throughout reverses the inequalities and yields
$$-\log \mathbb{E}_Q[Z] \le -\tfrac{1}{\alpha} \log \mathbb{E}_Q[Z^\alpha] \le -\mathbb{E}_Q[\log Z],$$
which is the desired interpolation inequality.

4. *Monotonicity in* α. Writing $\varphi(\alpha) = \log \mathbb{E}_Q[Z^\alpha]$, then $\ell_\alpha(Q; x, y) = -\frac{\varphi(\alpha)}{\alpha}$. For $\alpha > 0$, define the tilted distribution Q_α by

$$\frac{dQ_\alpha}{dQ}(\theta) = \frac{Z(\theta)^\alpha}{\mathbb{E}_Q[Z^\alpha]}.$$

Differentiating φ gives

$$\varphi'(\alpha) = \mathbb{E}_{Q_\alpha}[\log Z].$$

Therefore

$$\frac{d}{d\alpha}\ell_\alpha(Q; x, y) = \frac{\varphi(\alpha) - \alpha\varphi'(\alpha)}{\alpha^2}.$$

On the other hand,

$$\mathrm{KL}(Q_\alpha \| Q) = \mathbb{E}_{Q_\alpha}\left[\log \frac{dQ_\alpha}{dQ}\right] = \mathbb{E}_{Q_\alpha}\left[\alpha \log Z - \varphi(\alpha)\right] = \alpha\varphi'(\alpha) - \varphi(\alpha).$$

Thus

$$\frac{d}{d\alpha}\ell_\alpha(Q; x, y) = -\frac{1}{\alpha^2}\,\mathrm{KL}(Q_\alpha \| Q) \le 0.$$

Hence $\alpha \mapsto \ell_\alpha(Q; x, y)$ is non-increasing.

Finally, equality in the monotonicity derivative occurs if and only if $\mathrm{KL}(Q_\alpha \| Q) = 0$, i.e. $Q_\alpha = Q$. This holds if and only if Z^α is Q-a.s. constant, equivalently if $p_\theta(y \mid x)$ is Q-a.s. constant. This proves the strictness claim. □

B Additional details for section 3

B.1 Proof of Lemma 1

Proof. Let

$$Z := p_\theta(y \mid x), \qquad \theta \sim Q,$$

and define $U := Z^\alpha$. By assumption,

$$0 < U \le M_{x,y}^\alpha \qquad Q\text{-a.s.}$$

Moreover,

$$\ell_0(Q; x, y) = -\mathbb{E}_Q[\log Z] = -\frac{1}{\alpha}\mathbb{E}_Q[\log U],$$

and

$$\ell_\alpha(Q; x, y) = -\frac{1}{\alpha}\log \mathbb{E}_Q[U].$$

Therefore

$$\ell_\alpha(Q; x, y) - \ell_0(Q; x, y) = \frac{1}{\alpha}\left(\mathbb{E}_Q[\log U] - \log \mathbb{E}_Q[U]\right).$$

The function $f(u) = \log u$ is $1/M_{x,y}^{2\alpha}$-strongly concave on $(0, M_{x,y}^\alpha]$, since

$$f''(u) = -\frac{1}{u^2} \le -\frac{1}{M_{x,y}^{2\alpha}}.$$

Thus, applying the strong-concavity Jensen gap bound to U,

$$\mathbb{E}_Q[\log U] - \log \mathbb{E}_Q[U] \le -\frac{1}{2M_{x,y}^{2\alpha}}\,\mathrm{Var}_Q(U).$$

Substituting $U = Z^\alpha = p_\theta(y \mid x)^\alpha$ gives

$$\ell_\alpha(Q; x, y) - \ell_0(Q; x, y) \le -\frac{1}{2\alpha M_{x,y}^{2\alpha}}\,\mathrm{Var}_{\theta\sim Q}\left(p_\theta(y \mid x)^\alpha\right),$$

thus proving the result. □

B.2 Proof of Lemma 2

Let $Z = \int_\Theta p_{\theta'}(y \mid x)\, Q(d\theta')$ and $w(\theta) = w^{(1)}(\theta; x, y, Q)$, we have the following auxiliary expression

$$p_\theta(y \mid x)^\alpha = (w^{(1)}(\theta; x, y, Q)Z)^\alpha = (w(\theta)Z)^\alpha.$$

Using this, we can write

$$\begin{aligned}\ell_\alpha(x, y) &= -\tfrac{1}{\alpha} \log \int_\Theta Z^\alpha w(\theta)^\alpha Q(d\theta) = -\log Z - \tfrac{1}{\alpha} \log \int_\Theta w(\theta)^\alpha Q(d\theta) \\ &= \ell_1(x, y) - \tfrac{1-\alpha}{\alpha} H_\alpha(w; Q),\end{aligned}$$

where

$$H_\alpha(w; Q) = \tfrac{1}{1-\alpha} \log \int_\Theta w(\theta)^\alpha Q(\theta).$$

B.3 Proof of Lemma 3

Proof. For a fixed data point $z = (x, y)$, let $Q_0, Q_1 \in \mathcal{P}(\Theta)$ satisfy $0 < A_z(Q_0), A_z(Q_1) < \infty$. Let $t \in [0, 1]$ and define

$$Q_t := (1 - t)Q_0 + tQ_1.$$

Linearity of integration gives

$$A_z(Q_t) = (1 - t)A_z(Q_0) + tA_z(Q_1),$$

so $0 < A_z(Q_t) < \infty$. Since $u \mapsto -\alpha^{-1} \log u$ is convex on $(0, \infty)$, it follows that

$$\ell_\alpha(Q_t; z) \le (1 - t)\ell_\alpha(Q_0; z) + t\ell_\alpha(Q_1; z).$$

Thus $Q \mapsto \ell_\alpha(Q; z)$ is convex on its domain, and the finite-sample data term, being a sum of convex functionals, is convex.

Now let $Q_0, Q_1 \in \mathcal{Q}_\alpha$. The preceding argument and convexity of KL in its first argument imply that $Q_t \in \mathcal{Q}_\alpha$. Since $\lambda \ge 0$, they also imply convexity of $\mathcal{L}_\alpha$ on this domain.

To obtain strong convexity, it suffices to consider $t \in (0, 1)$, since the claimed inequality is an equality at $t = 0, 1$. Because Q_0 and Q_1 have finite KL relative to π_0,

$$(1 - t)\,\mathrm{KL}(Q_0 \| \pi_0) + t\,\mathrm{KL}(Q_1 \| \pi_0) - \mathrm{KL}(Q_t \| \pi_0) = (1 - t)\,\mathrm{KL}(Q_0 \| Q_t) + t\,\mathrm{KL}(Q_1 \| Q_t).$$

By Pinsker's inequality, using

$$\begin{aligned}\mathrm{TV}(Q_0, Q_t) &= t\ \mathrm{TV}(Q_0, Q_1), \\ \mathrm{TV}(Q_1, Q_t) &= (1 - t)\,\mathrm{TV}(Q_0, Q_1),\end{aligned}$$

we obtain

$$(1 - t)\,\mathrm{KL}(Q_0 \| Q_t) + t\,\mathrm{KL}(Q_1 \| Q_t) \ge 2t(1 - t)\,\mathrm{TV}(Q_0, Q_1)^2.$$

Therefore,

$$\mathrm{KL}(Q_t \| \pi_0) \le (1 - t)\,\mathrm{KL}(Q_0 \| \pi_0) + t\,\mathrm{KL}(Q_1 \| \pi_0) - 2t(1 - t)\,\mathrm{TV}(Q_0, Q_1)^2.$$

Combining this with convexity of the data term yields

$$\mathcal{L}_\alpha(Q_t) \le (1 - t)\mathcal{L}_\alpha(Q_0) + t\mathcal{L}_\alpha(Q_1) - 2\lambda t(1 - t)\,\mathrm{TV}(Q_0, Q_1)^2.$$

Thus $\mathcal{L}_\alpha$ is strongly convex in total variation when $\lambda > 0$. In particular, it can have at most one minimiser on $\mathcal{Q}_\alpha$. □

B.4 Proof of Proposition 1

Proof. By Lemma 3, the functional $\mathcal{L}_\alpha$ is strongly convex on $\mathcal{Q}_\alpha$ when $\lambda > 0$. Hence it has at most one minimiser. It remains to derive the first-order condition.

Let

$$A_i(q) := \int_\Theta p_\theta(y_i \mid x_i)^\alpha q(\theta)\, \pi_0(d\theta).$$

For $Q(d\theta) = q(\theta)\pi_0(d\theta)$, write $\mathcal{L}_\alpha(q)$ for $\mathcal{L}_\alpha(Q)$. The objective is

$$\mathcal{L}_\alpha(q) = -\tfrac{1}{\alpha}\sum_{i=1}^N \log A_i(q) + \lambda \int_\Theta q(\theta)\log q(\theta)\, \pi_0(d\theta),$$

subject to

$$\int_\Theta q(\theta)\, \pi_0(d\theta) = 1.$$

Let q be a strictly positive density corresponding to $Q \in \mathcal{Q}_\alpha$, and let v be any bounded measurable function satisfying

$$\int_\Theta v(\theta) q(\theta)\, \pi_0(d\theta) = 0.$$

For $|\varepsilon|\|v\|_\infty < 1$, set

$$q_\varepsilon(\theta) = q(\theta)(1 + \varepsilon v(\theta)).$$

These perturbations preserve positivity and unit mass, and the corresponding measures remain in $\mathcal{Q}_\alpha$. Finite KL implies $\int_\Theta q|\log q|\, d\pi_0 < \infty$. Together with $A_i(q) < \infty$ and boundedness of v, this justifies differentiating at $\varepsilon = 0$, giving

$$\tfrac{d}{d\varepsilon}\mathcal{L}_\alpha(q_\varepsilon)\big|_{\varepsilon=0} = \int_\Theta \left[\lambda(\log q(\theta) + 1) - \tfrac{1}{\alpha}\sum_{i=1}^N \tfrac{p_\theta(y_i|x_i)^\alpha}{A_i(q)}\right] v(\theta) q(\theta)\, \pi_0(d\theta).$$

At the minimiser $q^\star_\alpha$, this derivative vanishes for every such v. The bracketed expression is $Q^\star_\alpha$-integrable. Taking $v = \mathbf{1}_B - Q^\star_\alpha(B)$ for measurable B shows that it is constant $Q^\star_\alpha$-almost everywhere. Since $q^\star_\alpha > 0$ π_0-almost everywhere, it is also constant π_0-almost everywhere. Thus there is a finite constant C such that

$$\lambda(\log q^\star_\alpha(\theta) + 1) - \tfrac{1}{\alpha}\sum_{i=1}^N \tfrac{p_\theta(y_i|x_i)^\alpha}{A_i(q^\star_\alpha)} = C.$$

Rearranging,

$$\log q^\star_\alpha(\theta) = C' + \tfrac{1}{\lambda\alpha}\sum_{i=1}^N \tfrac{p_\theta(y_i|x_i)^\alpha}{A_i(q^\star_\alpha)}.$$

Exponentiating and using $\int_\Theta q^\star_\alpha\, d\pi_0 = 1$ gives

$$q^\star_\alpha(\theta) = \tfrac{1}{Z_\alpha}\exp\!\left(\tfrac{1}{\lambda\alpha}\sum_{i=1}^N \tfrac{p_\theta(y_i|x_i)^\alpha}{A_i(q^\star_\alpha)}\right),$$

where $Z_\alpha = e^{-C'} \in (0, \infty)$. Since

$$A_i(q^\star_\alpha) = \int_\Theta p_\vartheta(y_i \mid x_i)^\alpha Q^\star_\alpha(d\vartheta),$$

this is the stated self-consistency equation.

Conversely, let $Q \in \mathcal{Q}_\alpha$ have a strictly positive density q satisfying the fixed-point equation. For any $P \in \mathcal{Q}_\alpha$, define

$$r_i := \tfrac{A_{z_i}(P)}{A_{z_i}(Q)} > 0.$$

The fixed-point equation and finiteness of $A_{z_i}(P)$ imply $\log q \in L^1(P)$. Since $P \ll Q$, we may therefore write

$$\mathrm{KL}(P\|\pi_0) = \mathrm{KL}(P\|Q) + \int_\Theta \log q\, dP.$$

Substituting the fixed-point expression for $\log q$ yields

$$\mathcal{L}_\alpha(P) - \mathcal{L}_\alpha(Q) = \lambda\,\mathrm{KL}(P\|Q) + \tfrac{1}{\alpha}\sum_{i=1}^N (r_i - 1 - \log r_i) \geq 0.$$

The inequality follows from nonnegativity of KL and $r - 1 - \log r \geq 0$ for $r > 0$. Thus Q is a global minimiser on $\mathcal{Q}_\alpha$, and uniqueness follows from Lemma 3. □

B.5 A GENERALIST TO SPECIALIST TRANSITION IN THE α-RÉNYI POSTERIOR

Consider a finite parameter space $\Theta = \{g, a, b\}$, with a uniform prior π_0. Suppose we have a dataset comprising two observations z_1, z_2, with associated likelihoods

	z_1	z_2
g	m	m
a	h	ε
b	ε	h

$$0 < \varepsilon < m < h < 1.$$

Thus we can consider the model associated with g is a generalist, while a and b are specialists for z_1 and z_2, respectively. Let Q assign probabilities

$$Q(g) = q_g, \qquad Q(a) = q_a, \qquad Q(b) = q_b.$$

Because the likelihood table is invariant under simultaneously swapping $a \leftrightarrow b$ and $z_1 \leftrightarrow z_2$, any fixed point of (5) would be symmetric, so we can just assume that

$$Q(a) = Q(b) = s, \qquad Q(g) = 1 - 2s, \qquad 0 < s < \frac{1}{2}.$$

For this symmetric posterior candidate, the two normalising denominators in the responsibility weights are equal

$$A_\alpha(Q) := \textstyle\sum_{\theta\in\Theta} Q(\theta) p_\theta(z_1)^\alpha = \sum_{\theta\in\Theta} Q(\theta) p_\theta(z_2)^\alpha = q_g m^\alpha + s(h^\alpha + \varepsilon^\alpha).$$

The self-consistency equation gives

$$Q^\star_\alpha(\theta) \propto \exp\left\{ \tfrac{1}{\lambda\alpha} \textstyle\sum_{i=1}^2 \frac{p_\theta(z_i)^\alpha}{A_\alpha(Q^\star_\alpha)} \right\}.$$

Therefore the ratio of the posterior mass assigned to a specialist and to the generalist is

$$\tfrac{Q^\star_\alpha(a)}{Q^\star_\alpha(g)} = \exp\left\{ \tfrac{1}{\lambda\alpha A_\alpha(Q^\star_\alpha)} \left[h^\alpha + \varepsilon^\alpha - 2m^\alpha\right] \right\}.$$

The same expression holds for b. Hence

$$Q^\star_\alpha(a) > Q^\star_\alpha(g) \qquad \Longleftrightarrow \qquad h^\alpha + \varepsilon^\alpha > 2m^\alpha.$$

Thus, at the posterior level, the $\alpha > 0$ fixed point assigns more mass to each specialist than to the generalist exactly when the α-power mean of the specialist likelihoods exceeds the generalist likelihood. In the $\alpha \to 0$ limit, the objective reduces to the usual Gibbs posterior,

$$Q^\star_0(\theta) \propto \pi_0(\theta) \exp\left\{ \tfrac{1}{\lambda} \textstyle\sum_{i=1}^2 \log p_\theta(z_i) \right\}.$$

With a uniform prior, the specialist-to-generalist posterior/likelihood ratio is therefore

$$\tfrac{Q^\star_0(a)}{Q^\star_0(g)} = \left(\tfrac{h\varepsilon}{m^2}\right)^{1/\lambda}.$$

Consequently, if $m^2 > h\varepsilon$, then the classical Gibbs posterior favours the generalist. On the other hand, for $\alpha > 0$, the self-consistent posterior favours the specialists when

$$m^2 > h\varepsilon \qquad \text{and} \qquad h^\alpha + \varepsilon^\alpha > 2m^\alpha.$$

For example, with

$$h = 0.9, \qquad \varepsilon = 0.01, \qquad m = 0.3,$$

we have $m^2 = 0.09 > 0.009 = h\varepsilon$, so the $\alpha = 0$ Gibbs posterior favours the generalist. At the other extreme when $\alpha = 1$,

$$h + \varepsilon = 0.91 > 0.6 = 2m,$$

so the self-consistent $\alpha = 1$ posterior favours the specialists. The transition occurs when

$$h^\alpha + \varepsilon^\alpha = 2m^\alpha,$$

which for these values gives $\alpha_{\text{critical}} \approx 0.56$.

B.6 Proof of Proposition 2

Proof. First, we show that if the model is well-specified, then $\delta_{\theta^\star}$ is a minimiser of the population risk $\mathcal{R}_\alpha(Q)$. Using Lemma 4, we have the following bound

$$\mathcal{R}_\alpha(Q) = \mathbb{E}_{P^\star}[\ell_\alpha(Q;X,Y)] \geq \mathbb{E}_{P^\star}[\ell_1(Q;X,Y)] = \mathbb{E}_{P^\star}[-\log p_Q(Y|X)] = \mathcal{R}_1(Q).$$

Using the non-negativity of the KL divergence, the cross-entropy between the true conditional distribution $p^*(y|x)$ and any other distribution $p_Q(y|x)$ is strictly minimised when the two distributions are identical. Therefore,

$$\mathbb{E}_{Y\sim p^*(\cdot|X)}[-\log p_Q(Y|X)] \geq \mathbb{E}_{Y\sim p^*(\cdot|X)}[-\log p^*(Y|X)].$$

Taking the expectation over X, this gives us a lower bound of the population risk $\mathcal{R}_\alpha(Q)$ for any Q

$$\mathcal{R}_\alpha(Q) \geq \mathcal{R}_1(Q) \geq \mathbb{E}_{P^*}[-\log p^*(Y|X)].$$

The lower bound $\mathbb{E}_{P^*}[-\log p^*(Y|X)]$ is attained when $Q = \delta_{\theta^*}$ centred at the true parameter. Therefore, δ_{θ^*} is a global minimiser of the population risk for every $\alpha \in (0,1]$.

Now we characterise the minimisers. Any minimiser Q must achieve this lower bound, requiring both inequalities above to hold with equality. The second inequality is an equality if and only if $\mathcal{R}_1(Q) = \mathbb{E}_{P^\star}[-\log p^\star(Y|X)]$, which implies the KL divergence between $p^\star(\cdot|x)$ and $p_Q(\cdot|x)$ is zero for $P^\star$-a.e. x. Thus, for every $\alpha \in (0,1]$, any minimiser must satisfy

$$p_Q(\cdot|x) = p^\star(\cdot|x) \quad \text{for } P^\star\text{-a.e. } x.$$

This establishes the claim for $\alpha = 1$.

For $0 < \alpha < 1$, achieving the lower bound additionally requires the first inequality to hold with equality: $\mathcal{R}_\alpha(Q) = \mathcal{R}_1(Q)$. Because $\ell_\alpha(Q;x,y) \geq \ell_1(Q;x,y)$ pointwise, equality in expectation requires $\ell_\alpha(Q;X,Y) = \ell_1(Q;X,Y)$ almost surely. The bound in Lemma 4 relies on Jensen's inequality applied to the strictly concave function $t \mapsto t^\alpha$. Therefore, equality holds if and only if the random variable is constant almost surely. Thus, $p_\theta(\cdot|x)$ must be constant for Q-a.e. θ. In particular, this constant must be exactly $p^\star(\cdot|x)$. Therefore, for Q-a.e. θ:

$$p_\theta(\cdot|x) = p^\star(\cdot|x) \quad \text{for } P^\star\text{-a.e. } x.$$

This means that Q must be supported entirely on the exact-fit set $\Theta^\star$.

□

B.7 Proof of Proposition 3

Proof. Define

$$g_\theta(z) = \nabla_\theta L_\theta(z), \qquad H_\theta(z) = \nabla^2_\theta L_\theta(z).$$

Fix z satisfying the smoothness assumptions and first let $\alpha > 0$. Unless indicated otherwise, expectations below are taken with respect to the law of ξ. The α-loss can be written as

$$\ell_\alpha(Q_t;z) = -\tfrac{1}{\alpha}\log \mathbb{E}_{\vartheta\sim Q_t}[\exp(-\alpha L_\vartheta(z))].$$

Let $f(\vartheta) = \exp(-\alpha L_\vartheta(z))$. Since f is C^2 near θ, a second-order Taylor expansion gives

$$f(\theta + t\xi) = f(\theta) + t\nabla f(\theta)^\top \xi + \tfrac{t^2}{2}\xi^\top \nabla^2 f(\theta)\xi + r_t(z,\xi),$$

where

$$\sup_{\|u\|\leq b} |r_t(z,u)| = o(t^2).$$

Indeed, continuity of $\nabla^2 f$ implies that

$$|r_t(z,u)| \leq \tfrac{t^2 b^2}{2} \sup_{\|\vartheta-\theta\|\leq tb} \|\nabla^2 f(\vartheta) - \nabla^2 f(\theta)\|_{\mathrm{op}} = o(t^2)$$

uniformly over $\|u\| \leq b$. Taking expectations, the linear term vanishes because $\mathbb{E}[\xi] = 0$, and

$$\mathbb{E}_{\vartheta\sim Q_t}[f(\vartheta)] = f(\theta) + \tfrac{1}{2}\operatorname{Tr}(\nabla^2 f(\theta)\Sigma_t) + o(t^2),$$
$$\nabla^2 f(\theta) = f(\theta)\big[\alpha^2 g_\theta(z) g_\theta(z)^\top - \alpha H_\theta(z)\big].$$

Since $f(\theta) > 0$, substitution yields

$$\mathbb{E}_{\vartheta\sim Q_t}[f(\vartheta)] = f(\theta)\left(1 + \tfrac{1}{2}\operatorname{Tr}\big[\big(\alpha^2 g_\theta(z)g_\theta(z)^\top - \alpha H_\theta(z)\big)\Sigma_t\big] + o(t^2)\right).$$

Using $\log(1+u) = u + O(u^2)$ as $u \to 0$, and $\Sigma_t = t^2\Sigma$, we obtain

$$\ell_\alpha(Q_t; z) = L_\theta(z) + \tfrac{1}{2}\operatorname{Tr}(H_\theta(z)\Sigma_t) - \tfrac{\alpha}{2} g_\theta(z)^\top \Sigma_t g_\theta(z) + o(t^2).$$

This proves the pointwise expansion. For $\alpha = 0$, the same result follows directly by expanding $L_{\theta+t\xi}(z)$ and taking expectations.

It remains to justify averaging the expansion over $P^\star$. For z also satisfying the derivative-envelope bounds, Taylor's theorem gives

$$L_{\theta+t\xi}(z) - L_\theta(z) = tW + \eta_t, \qquad W = g_\theta(z)^\top \xi, \qquad |\eta_t| \le \tfrac{t^2 b^2}{2} K(z),$$

where $\mathbb{E}[W] = 0$ and $|W| \le bG(z)$. For any centred random variable W with $|W| \le c$, Jensen's inequality and convexity of the exponential, together with $\cosh(u) \le e^{u^2/2}$, give

$$1 \le \mathbb{E}[e^{sW}] \le \cosh(sc) \le e^{s^2c^2/2}, \qquad s \in \mathbb{R}.$$

Applying this with $s = -\alpha t$ and $c = bG(z)$, and using the bound on η_t, yields, for $\alpha > 0$,

$$-\tfrac{t^2 b^2}{2}\big(K(z) + \alpha G(z)^2\big) \le \ell_\alpha(Q_t; z) - L_\theta(z) \le \tfrac{t^2 b^2}{2} K(z).$$

Hence, since $\alpha \in [0, 1]$,

$$\left|\tfrac{\ell_\alpha(Q_t;z) - L_\theta(z)}{t^2}\right| \le \tfrac{b^2}{2}\big(K(z) + G(z)^2\big), \qquad 0 < t < r/b.$$

For $\alpha = 0$, this bound follows directly from $\mathbb{E}[W] = 0$ and the bound on η_t. The right-hand side is $P^\star$-integrable by assumption. Together with the integrability of $L_\theta(Z)$, this also ensures that $\mathcal{R}_\alpha(Q_t)$ is finite. Dominated convergence applied to the normalised pointwise expansion results in

$$\lim_{t\downarrow 0} \mathbb{E}_{P^\star}\left[\tfrac{\ell_\alpha(Q_t;Z) - L_\theta(Z)}{t^2}\right] = \tfrac{1}{2}\operatorname{Tr}[(V(\theta) - \alpha J(\theta))\Sigma].$$

Hence,

$$\mathcal{R}_\alpha(Q_t) = \mathbb{E}_{P^\star}[L_\theta(Z)] + \tfrac{1}{2}\operatorname{Tr}[(V(\theta) - \alpha J(\theta))\Sigma_t] + o(t^2).$$

Finally, $\Sigma \neq 0$ and $\|\Sigma_t\|_{\mathrm{op}} = t^2\|\Sigma\|_{\mathrm{op}}$ show that $o(t^2) = o(\|\Sigma_t\|_{\mathrm{op}})$, concluding the proof. □

C Additional details for section 4

C.1 Algorithm

The full algorithm is provided in Algorithm 1.

C.2 α-Rényi ensemble training for direct preference optimisation

This appendix provides the full derivation of the preference optimisation variant of the α-Rényi ensemble objective introduced in Section 4.

A standard probabilistic model for pairwise comparisons is the Bradley-Terry model. Given a latent utility function $r(x, y)$, it assumes that the probability that response y^+ is preferred to response y^- satisfies

$$\mathbb{P}(y^+ \succ y^- \mid x) = \tfrac{\exp(r(x,y^+))}{\exp(r(x,y^+)) + \exp(r(x,y^-))} = \sigma\big(r(x, y^+) - r(x, y^-)\big), \tag{13}$$

where $\sigma(u) = 1/(1 + e^{-u})$ is the logistic sigmoid. The corresponding negative log-likelihood for one preference pair is

$$-\log \sigma\big(r(x, y^+) - r(x, y^-)\big).$$

In the DPO framework, the latent utility is identified, up to an additive constant, with a scaled log-density ratio between the policy and a fixed reference model

$$r_\theta(x, y) = \beta\Big(\log p_\theta(y \mid x) - \log p_{\mathrm{ref}}(y \mid x)\Big), \tag{14}$$

Algorithm 1 Direct AdamW training of an α-Rényi LoRA ensemble

Require: supervised dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, frozen base model W_0, number of LoRA particles M, parameter $\alpha \in [0, 1]$, prior regularisation weight $\lambda \geq 0$, number of training steps T, AdamW optimiser.

Initialise LoRA particles $\theta_1, \ldots, \theta_M$.
Freeze base model parameters W_0.
for $t = 1, \ldots, T$ **do**
  Sample minibatch $\mathcal{B} = \{(x_b, y_b)\}_{b=1}^{B} \subset \mathcal{D}$.
  for $i = 1, \ldots, M$ **do**
    Compute teacher-forced sequence log-likelihoods

$$s_{i,b} := \log p_{\theta_i}(y_b \mid x_b) = \sum_{\tau=1}^{T_b} \log p_{\theta_i}(y_{b,\tau} \mid x_b, y_{b,<\tau}) \qquad b = 1, \ldots, B.$$

  if $\alpha = 0$ **then**
    Compute the average particle loss

$$\mathcal{L}_{\text{data}} = -\frac{1}{BM} \sum_{b=1}^{B} \sum_{i=1}^{M} s_{i,b}.$$

  else
    Compute the α-Rényi data loss using log-sum-exp:

$$\mathcal{L}_{\text{data}} = -\frac{1}{B\alpha} \sum_{b=1}^{B} \left[\text{logsumexp}_{i=1}^{M}(\alpha s_{i,b}) - \log M \right].$$

  Compute a finite-particle approximation to the prior/KL regulariser: $\mathcal{R} = \mathcal{R}_{\text{prior}}$.

$$\mathcal{R}_{\text{prior}} = -\frac{\lambda}{M} \sum_{i=1}^{M} \log q_0(\theta_i).$$

  Form the finite-particle objective

$$\mathcal{J}_\alpha = \mathcal{L}_{\text{data}} + \frac{1}{N} \mathcal{R}.$$

  Update $\theta_1, \ldots, \theta_M$ by performing one AdamW step on $\nabla_{\theta_{1:M}} \mathcal{J}_\alpha$.
Output: LoRA ensemble $\{\theta_i\}_{i=1}^{M}$, posterior approximation $Q_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{\theta_i}$.

where $\beta > 0$ is an inverse-temperature parameter and p_{ref} is a frozen reference model. Substituting (14) into (13) yields the standard DPO likelihood

$$\mathbb{P}_\theta(y^+ \succ y^- \mid x) = \sigma\Big(\beta \left[\log \tfrac{p_\theta(y^+|x)}{p_\theta(y^-|x)} - \log \tfrac{p_{\text{ref}}(y^+|x)}{p_{\text{ref}}(y^-|x)} \right] \Big).$$

In the LoRA ensemble setting, each particle θ_i induces its own preference score. For a preference triple (x, y^+, y^-), define the particle-wise preference margin

$$\Delta_i(x, y^+, y^-) := \log p_{\theta_i}(y^+ \mid x) - \log p_{\theta_i}(y^- \mid x) - \log p_{\text{ref}}(y^+ \mid x) + \log p_{\text{ref}}(y^- \mid x).$$

Each sequence log-likelihood is the usual autoregressive teacher-forced sum over target tokens. The corresponding Bradley-Terry preference likelihood for particle i is

$$r_i(x, y^+, y^-) := \sigma\big(\beta\, \Delta_i(x, y^+, y^-)\big). \tag{15}$$

Substituting this into the per-example α-Rényi preference loss and expanding over a minibatch $\mathcal{B} = \{(x_b, y_b^+, y_b^-)\}_{b=1}^{B}$, the finite-particle objective becomes

$$\widehat{\mathcal{L}}_\alpha^{\text{DPO}}(\theta_{1:M}) = -\tfrac{N}{|\mathcal{B}|} \textstyle\sum_{b=1}^{B} \frac{1}{\alpha} \log \Big(\frac{1}{M} \sum_{i=1}^{M} r_i(x_b, y_b^+, y_b^-)^\alpha \Big) + \mathcal{R}(\theta_{1:M}).$$

The α-objective induces responsibilities across particles. For each minibatch example b, define

$$w_{i,b}^{(\alpha)} = \tfrac{r_i(x_b, y_b^+, y_b^-)^\alpha}{\sum_{j=1}^{M} r_j(x_b, y_b^+, y_b^-)^\alpha}.$$

The gradient of $\widehat{\mathcal{L}}_\alpha^{\mathrm{DPO}}$ with respect to particle θ_i takes the form

$$\nabla_{\theta_i}\widehat{\mathcal{L}}_\alpha^{\mathrm{DPO}} = -\tfrac{N}{B}\sum_{b=1}^{B} w_{i,b}^{(\alpha)}\nabla_{\theta_i}\log r_i(x_b, y_b^+, y_b^-) + \nabla_{\theta_i}\mathcal{R}.$$

Using (15), we have

$$\nabla_{\theta_i}\log r_i = \beta\big(1 - r_i\big)\nabla_{\theta_i}\Delta_i,$$

where

$$\nabla_{\theta_i}\Delta_i = \nabla_{\theta_i}\log p_{\theta_i}(y_b^+ \mid x_b) - \nabla_{\theta_i}\log p_{\theta_i}(y_b^- \mid x_b).$$

Therefore, the gradient step remains fully compatible with standard autoregressive LLM training.

C.3 Gradient-flow interpretation

The responsibility-weighted updates induced by the finite-particle objective in (12) admit a formal measure-valued gradient-flow interpretation. This connects the α-Rényi objective to mean-field Langevin dynamics for entropy-regularised optimisation over probability measures, and clarifies the relationship between posterior sampling and the deterministic particle surrogate used in our experiments.

For $0 < \alpha \leq 1$, consider the data-fit term

$$\mathcal{D}_\alpha(Q) := \sum_{n=1}^{N}\ell_\alpha(Q; x_n, y_n) = -\tfrac{1}{\alpha}\sum_{n=1}^{N}\log\textstyle\int_\Theta p_\theta(y_n \mid x_n)^\alpha\, Q(d\theta).$$

Writing $z_n = (x_n, y_n)$ and $p_\theta(z_n) = p_\theta(y_n \mid x_n)$, its first variation, up to an additive constant, is

$$\tfrac{\delta\mathcal{D}_\alpha}{\delta Q}(\theta) = -\tfrac{1}{\alpha}\sum_{n=1}^{N} w^{(\alpha)}(\theta; z_n, Q).$$

Let $U = -\log\pi_0$. The full objective can therefore be written as

$$\mathcal{L}_\alpha(Q) = \mathcal{F}_\alpha(Q) + \lambda\,\mathrm{Ent}(Q), \qquad \mathcal{F}_\alpha(Q) := \mathcal{D}_\alpha(Q) + \lambda\mathbb{E}_Q[U],$$

where $\mathrm{Ent}(Q) = \int q\log q\, d\theta$ for a density q. By Lemma 3, $\mathcal{F}_\alpha$ is convex under linear interpolation of measures. This places the objective within the class of linearly convex, entropy-regularised energies studied in recent analyses of mean-field Langevin dynamics (Nitanda et al., 2022; Chizat, 2022).

Mean-field Langevin dynamics. For the following formal calculations, take $\Theta = \mathbb{R}^d$, assume positive smooth likelihoods and prior density, and let Q_t have finite second moment and a smooth, strictly positive density ρ_t. Assume that the displayed quantities are finite and that there is sufficient decay at infinity to justify integration by parts. For $\lambda > 0$, the Wasserstein gradient flow (Ollivier et al., 2014) of $\mathcal{L}_\alpha$ is

$$\begin{aligned}\partial_t\rho_t &= \nabla_\theta\cdot(\rho_t\nabla_\theta\Psi_t),\\ \Psi_t(\theta) &:= -\tfrac{1}{\alpha}\textstyle\sum_{n=1}^{N} w^{(\alpha)}(\theta; z_n, Q_t) + \lambda\log\frac{\rho_t(\theta)}{\pi_0(\theta)}.\end{aligned}$$

Equivalently, expanding the KL contribution gives the nonlinear Fokker–Planck equation

$$\begin{aligned}\partial_t\rho_t &= -\nabla_\theta\cdot\big(\rho_t b_\alpha(\theta, Q_t)\big) + \lambda\Delta\rho_t,\\ b_\alpha(\theta, Q) &:= \textstyle\sum_{n=1}^{N} w^{(\alpha)}(\theta; z_n, Q)\nabla_\theta\log p_\theta(z_n) + \lambda\nabla_\theta\log\pi_0(\theta).\end{aligned}$$

This is the forward equation of the McKean–Vlasov diffusion

$$d\theta_t = b_\alpha(\theta_t, Q_t)\,dt + \sqrt{2\lambda}\,dB_t, \qquad Q_t = \mathrm{Law}(\theta_t),$$

where B_t is standard Brownian motion. The responsibilities introduce dependence on the current law through the likelihood normalisers. The prior contributes to the drift, while the entropy component of the KL produces diffusion. At $\alpha = 0$, the data term in Ψ_t becomes $-\sum_n\log p_\theta(z_n)$ and $w^{(0)} \equiv 1$, recovering the usual Langevin diffusion targeting the Gibbs posterior. At $\alpha = 1$, the construction specialises to mean-field Langevin computation for the logarithmic-score predictively oriented posterior (McLatchie et al., 2025).

Dissipation and convergence. Along sufficiently regular finite-energy solutions,

$$\frac{d}{dt}\mathcal{L}_\alpha(Q_t) = -\int_\Theta \|\nabla_\theta \Psi_t(\theta)\|^2 Q_t(d\theta) \le 0. \tag{16}$$

A useful connection to the convergence literature is the associated *proximal Gibbs distribution*

$$\nu_Q(d\theta) := \tfrac{1}{Z(Q)} \exp\Big\{ \tfrac{1}{\lambda\alpha} \textstyle\sum_{n=1}^N w^{(\alpha)}(\theta; z_n, Q) \Big\} \pi_0(d\theta),$$

whenever its normalising constant is finite. Up to a spatially constant term, $\Psi_t = \lambda \log(\rho_t/\nu_{Q_t})$, where which must be interpreted in the weak-sense. Thus the self-consistency equation in Proposition 1 is precisely the fixed-point relation $Q^\star_\alpha = \nu_{Q^\star_\alpha}$. Under the regularity and decay assumptions above, a stationary positive density has $\nabla\Psi = 0$ and satisfies this first-order optimality condition, thus if it belongs to $\mathcal{Q}_\alpha$, it becomes the unique minimiser.

Linear convexity alone does not turn (16) into a quantitative convergence guarantee. However, recent works such as Nitanda et al. (2022); Chizat (2022) establish exponential convergence for linearly convex energies under suitable regularity conditions and uniform log-Sobolev inequalities for the associated proximal Gibbs distributions. These results could provide a natural route to analysing the dynamics presented in this work under appropriate assumptions, but we leave this to the scope of future work.

Interacting particles and propagation of chaos. The corresponding interacting Langevin approximation replaces Q_t in the drift by

$$Q^t_M = \tfrac{1}{M} \textstyle\sum_{i=1}^M \delta_{\theta_i(t)}, \qquad d\theta_i(t) = b_\alpha(\theta_i(t), Q^t_M)\, dt + \sqrt{2\lambda}\, dB_i(t),$$

with independent Brownian motions B_i. The continuous responsibility density and discrete weights satisfy

$$w^{(\alpha)}(\theta_i; z_n, Q^t_M) = M\, w^{(\alpha)}_{i,n}.$$

In particular, the drift is the negative gradient of the joint particle potential $M\mathcal{F}_\alpha(Q_M)$. When normalisable and under conditions ensuring invariance, the particle system has stationary law

$$\Pi_{\alpha,M}(d\theta_{1:M}) = \tfrac{1}{Z_{\alpha,M}} \exp\big\{ -\tfrac{M}{\lambda} \mathcal{D}_\alpha(Q_M) \big\} \textstyle\prod_{i=1}^M \pi_0(d\theta_i).$$

This joint law generally couples the particles and is not the product $(Q^\star_\alpha)^{\otimes M}$. Propagation-of-chaos results quantify when fixed-size particle marginals approach independent copies of the mean-field law as M increases, with uniform-in-time or stationary results allowing comparison with the limiting posterior. In particular, Kook et al. (2024) separate approximation of the mean-field stationary distribution from sampling the finite-particle stationary law. Chewi et al. (2026) establish log-Sobolev inequalities for these joint laws with constants independent of the particle count, under linear convexity, confinement and additional derivative bounds. These results explain how interacting particle methods can approximate an implicitly defined posterior such as $Q^\star_\alpha$, subject to their respective assumptions. Analyses incorporating time discretisation and stochastic gradients are provided by Suzuki et al. (2023); establishing such bounds for our likelihood-based interaction is beyond the scope of this work.

Relation to the implemented updates. The interacting diffusion above is distinct from directly substituting an empirical measure into the full variational objective: $\mathrm{KL}(Q_M\|\pi_0) = \infty$ when π_0 has a Lebesgue density. Its finite-particle construction instead uses $\mathcal{F}_\alpha(Q_M)$, with the entropy contribution represented by Brownian noise. In our experiments, we optimise the deterministic prior-potential surrogate

$$\mathcal{R}_{\mathrm{prior}}(\theta_{1:M}) = -\tfrac{\lambda}{M} \textstyle\sum_{i=1}^M \log \pi_0(\theta_i).$$

For a fixed minibatch $\mathcal{B}$ of size B, the corresponding negative Euclidean gradient flow is

$$\dot\theta_i = -\nabla_{\theta_i} \widehat{\mathcal{L}}^{\mathrm{SFT}}_\alpha = \tfrac{N}{B} \textstyle\sum_{b=1}^B w^{(\alpha)}_{i,b} \nabla_{\theta_i} \log p_{\theta_i}(z_b) - \nabla_{\theta_i}\mathcal{R}_{\mathrm{prior}}, \qquad i = 1, \dots, M.$$

For $\alpha = 0$, the discrete weights are $w^{(0)}_{i,b} = 1/M$. With the full dataset, the velocity above equals $b_\alpha(\theta_i, Q_M)/M$; the factor M is a rescaling of time for these deterministic dynamics. Uniform minibatch sampling gives unbiased estimates of the full-data gradients.

We use minibatch AdamW with the normalisation specified in Algorithm 1. These updates retain the responsibility-weighted drift and prior regularisation but omit the explicit diffusion associated with posterior entropy. Minibatch randomness does not in general reproduce this calibrated Brownian noise, so the sampling and propagation-of-chaos guarantees discussed above do not automatically transfer to the implemented optimiser. Moreover, the finite-particle energy is generally nonconvex in $\theta_{1:M}$, and its stationary points do not inherit the global optimality conclusion for regular positive stationary densities of the full KL flow.

D EXPERIMENTAL DETAILS

Code is available in an anonymous repository.

D.1 EFFICIENT IMPLEMENTATION WITH SHARED FROZEN BACKBONES

A central practical advantage of the LoRA ensemble setting is that all particles share the same frozen base model. This induces substantial computational structure.

First, the dominant parameter memory remains that of the single base model W_0, only the low-rank increments $\Delta W(\theta_i)$ are replicated across particles. If the LoRA rank r is small, the memory overhead of the ensemble scales roughly linearly in Mr, rather than in the full model dimension.

Second, the particle dimension can be partially vectorised. Given a minibatch $\{(x_b, y_b)\}_{b=1}^{B}$, one may form an augmented batch indexed by both data point and particle, evaluate all M LoRA variants in parallel, and compute the matrix $\big(\log p_{\theta_i}(y_b \mid x_b)\big)_{1\le i\le M,\, 1\le b\le B}$. From this matrix, the responsibilities are obtained by a softmax across the particle dimension after multiplying by α. In practice, for autoregressive sequence losses, this is done by accumulating token-level log-probabilities over the target continuation.

Third, because all particles are small modifications of a common backbone, one can exploit implementation-level sharing. For example, all non-adapted layers are evaluated identically across particles, and only the adapted projections differ. When memory permits, the LoRA updates can be stacked and applied in parallel using batched low-rank matrix multiplications. When memory is more constrained, particles can be processed in chunks while still sharing the same frozen model weights.

We provide further details regarding our experimental setup. Furthermore, we include a code snippet that demonstrates the lightweight adaptation of our framework into standard LoRA layers.

D.2 MMLU SPECIALISATION EXPERIMENTS

Models and data. We use Nemotron-3-8B-base-4k, Mistral-7B-Instruct-v0.3, Granite-3.1-8B-Instruct, and OLMo-2-1124-7B-Instruct. Each ensemble contains $M = 8$ rank-4 LoRA adapters applied to every query and value projection, with zero dropout and frozen pretrained weights. We use the MMLU benchmark (`cais/mmlu`), comprising multiple-choice questions across 57 subjects. Before training, we remove overlaps between splits. For auxiliary pre-adaptation, we use the `auxiliary_train` split, selecting 2000 unique examples for training and a disjoint set of 1000 examples for checkpoint selection. For the subsequent stages, we use 1502 validation questions for target fine-tuning, 280 development questions for hyperparameter selection, and 14042 test questions for final evaluation.

Auxiliary pre-adaptation. Before target fine-tuning, we train the particles independently for five epochs on the 2000 auxiliary training examples. Each particle optimises its own answer-token likelihood using shared minibatches. We select the checkpoint with the highest ensemble accuracy on the 1000 example auxiliary monitoring set among epochs 2-5. For each model and seed, this checkpoint provides a common initialisation for all values of α and the deep-ensemble baseline.

Optimisation and seeds. For each model and seed, all $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ share the same initial model checkpoint. Fine-tuning uses five epochs and batch size two per particle. We use AdamW with zero weight decay, 5% linear warmup and cosine decay to 10% of the initial rate. The learning rate rate used during the pre-adaptation phase is 2×10^{-4} for all models, while during the

fine-tuning stage the target rates are 10^{-4} for Nemotron/OLMo and 5×10^{-5} for Mistral/Granite. Learning rate were selected based on a validation dataset of 280 questions.

Baselines. Deep ensembles start from the same eight auxiliary adapters as our method. Laplace-LoRA is fitted post hoc around member zero of the final deep ensemble, using its auxiliary adapter as prior mean. Prediction averages eight globally coherent parameter draws (to match the number of ensemble members) after linearising option logits and applying softmax.

Metrics. Let $q_{mj} = p_m(y_j \mid x_j)$ and $\bar{q}_j = M^{-1} \sum_m q_{mj}$. For a uniformly sampled test question X and particle J, correctness C is Bernoulli with parameter q_{mj}, giving

$$\widehat{I}(J; C \mid X) = \frac{1}{N_{\text{test}}} \sum_j \left[h(\bar{q}_j) - \frac{1}{M} \sum_m h(q_{mj}) \right], \tag{17}$$

where $h(q) = -q \log_2 q - (1-q) \log_2 (1-q)$, with arguments clipped to $[10^{-12}, 1 - 10^{-12}]$. This label-dependent binary MI is computed directly from probabilities, without answer sampling. Ensemble accuracy uses $\arg\max_a M^{-1} \sum_m p_m(a \mid x_j)$. Routed accuracy uses

$$m_j^* = \arg\max_m q_{mj}, \qquad \widehat{y}_j^{\text{route}} = \arg\max_a p_{m_j^*}(a \mid x_j).$$

Results. Fig. 4 compares all three diagnostics. From $\alpha = 0$ to 1, routed accuracy rises by 7.62, 4.75, 2.26, and 5.83 percentage points for Nemotron, Mistral, Granite, and OLMo. Seed-averaged MI increases monotonically in all four models. Laplace-LoRA gives higher MI but ensemble accuracy below the pretrained model.

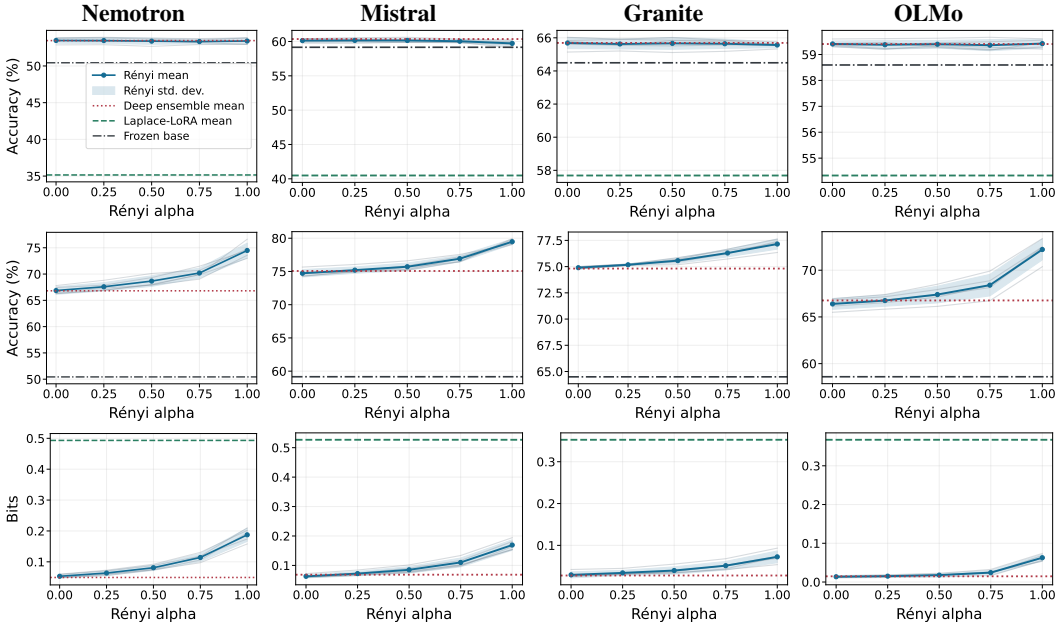

Figure 4: **MMLU diagnostics.** Columns: Nemotron, Mistral, Granite, and OLMo. Top: ensemble accuracy, middle: routed accuracy, bottom: conditional MI, with Laplace-LoRA in the top and bottom rows.

D.3 OR-Bench refusal disagreement experiments

Models and data. We use Qwen2-1.5B-Instruct, Phi-3-mini-4k-instruct, and Nemotron-3-8B-base-4k. Each ensemble contains $M = 10$ rank-4 LoRA adapters with frozen pretrained weights. Preference fine-tuning uses `trl-lib/ultrafeedback_binarized`, retaining pairs for which the chosen response's score exceeds the rejected response's score by more than one, to ensure high-quality preference signals All values of α use the same preference data, and OR-Bench requests are reserved for evaluation.

Preference fine-tuning. We optimise the Rényi DPO objective in App. C.2, with $\beta = 0.1$ and $\alpha \in \{0, 0.5, 1\}$. The starting pretrained checkpoint serves as the frozen reference policy throughout training. For each preference pair, chosen and rejected responses are scored by summing their completion-token log-probabilities, excluding the prompt and padding. At $\alpha = 0$, the data-fit objective reduces to the average of the individual adapters DPO losses, positive α couples their updates through the Rényi aggregation of preference probabilities.

Optimisation and seeds. We use AdamW with a peak learning rate of 10^{-5} and five independent training seeds. For each model and seed, all values of α share the same initial adapter parameters, minibatch order, and optimisation settings. Each seed defines a complete ten-member ensemble. The five ensembles provide the independent training repetitions used to summarise the results. Training uses an effective batch size of 64 preference pairs and five epochs, with weight decay 0.01, 3% linear warmup, and cosine decay to zero. We consider 5% held-out validation set, to select the ensemble checkpoint for evaluation.

Evaluation and response generation. We evaluate on the OR-80K, OR-Hard, and OR-Toxic configurations of `bench-llm/or-bench`. OR-80K provides the broader benign comparison set, OR-Hard contains difficult benign requests, and OR-Toxic contains harmful requests. For the OR-80K comparison, we sample benign requests to match the category proportions of OR-Hard, excluding requests appearing in OR-Hard. We use the same evaluation prompts across models, training seeds, and values of α. For each prompt x and adapter m, we sample $K = 60$ responses at temperature 0.7, keeping the adapter fixed throughout each response. We apply the same refusal judge and judging procedure to all responses. We use `allenai/wildguard` as a fixed judge. Each generated response is evaluated together with its full original prompt using WildGuard's official template and deterministic decoding at temperature 0.

Metrics. For a fixed trained ensemble, let $R_{mk}(x) \in 0, 1$ indicate whether the judge labels the kth response from adapter m to prompt x as a refusal. We estimate the member and ensemble refusal probabilities as

$$\widehat{q}_m(x) = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} R_{mk}(x), \qquad \widehat{\bar{q}}(x) = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \widehat{q}_m(x).$$

For an evaluation set $\mathcal{D}$ containing N prompts, the plug-in estimate of the conditional particle-refusal mutual information is

$$\widehat{I}(J; R \mid X) = \tfrac{1}{N} \textstyle\sum_{x \in \mathcal{D}} \left[h\left(\widehat{\bar{q}}(x)\right) - \tfrac{1}{M} \sum_{m=1}^{M} h\left(\widehat{q}_m(x)\right) \right], \tag{18}$$

where $h(q) = -q \log_2 q - (1-q) \log_2 (1-q)$ is the binary entropy function in bits, with $0 \log_2 0 = 0$. Here X and J are sampled independently and uniformly from $\mathcal{D}$ and the M adapters, respectively, and R denotes the refusal indicator of the resulting response. This quantity measures between-adapter disagreement in refusal behaviour conditional on the prompt.

Results Table 1 reports the numerical values shown in Fig. 3, including conditional particle-refusal MI and mean ensemble refusal probability.

Table 1: **Conditional particle-refusal MI (bits) and mean ensemble refusal probability (%)**, mean ± sample standard deviation across training seeds.

Model	α	OR-80K		OR-Hard		OR-Toxic	
		BALD	Refusal (%)	BALD	Refusal (%)	BALD	Refusal (%)
Qwen2-1.5B	0	0.026 ± 0.003	3.9 ± 0.6	0.071 ± 0.006	49.4 ± 1.7	0.025 ± 0.003	85.7 ± 0.6
	0.5	0.036 ± 0.004	4.3 ± 0.7	0.235 ± 0.010	50.1 ± 2.3	0.030 ± 0.004	86.6 ± 0.9
	1	0.037 ± 0.005	3.4 ± 0.9	0.296 ± 0.014	50.0 ± 1.5	0.035 ± 0.005	86.6 ± 0.8
Phi-3-mini	0	0.030 ± 0.004	2.6 ± 0.4	0.082 ± 0.007	45.9 ± 2.1	0.028 ± 0.003	91.8 ± 0.6
	0.5	0.036 ± 0.005	3.2 ± 0.8	0.241 ± 0.012	46.4 ± 1.6	0.033 ± 0.004	91.9 ± 0.7
	1	0.035 ± 0.006	3.3 ± 0.5	0.395 ± 0.015	46.1 ± 1.8	0.039 ± 0.006	91.6 ± 0.9
Nemotron-3-8B	0	0.028 ± 0.005	1.7 ± 0.5	0.103 ± 0.010	57.6 ± 2.0	0.039 ± 0.005	93.5 ± 1.0
	0.5	0.025 ± 0.006	1.6 ± 0.7	0.196 ± 0.014	56.5 ± 2.2	0.047 ± 0.007	93.7 ± 0.7
	1	0.031 ± 0.003	1.8 ± 0.9	0.375 ± 0.020	55.9 ± 1.9	0.054 ± 0.008	93.3 ± 0.9