Estimating Rare Events in Language Models with Proper Evaluation

Anonymous authors
Paper under double-blind review

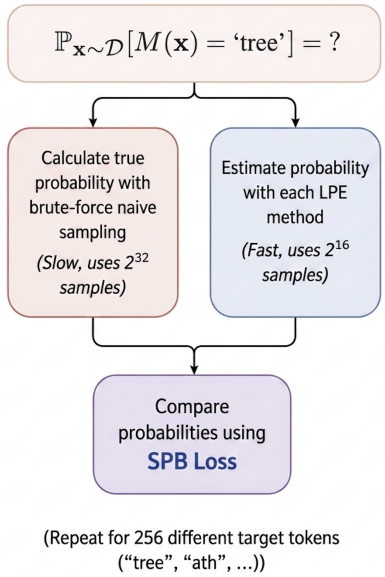
Abstract

Quantifying the risk of rare failures in language models, such as those triggered by adversarial distribution shifts or very large-scale deployments, requires estimating probabilities far too small for random sampling. While recent work has formalized Low Probability Estimation, existing pipelines remain fragile in the rarest regimes: estimators can suffer zero-estimate collapse or systematic bias, and standard evaluation losses can become unstable or poorly matched to asymmetric safety costs. In this work, we introduce Gradient Activation Adaptive Multi-Level Splitting (GA-AMLS), which adapts rare-event Monte Carlo methods to the continuous activation space of language models. Specifically, GA-AMLS uses a gradient-based MCMC kernel to navigate activation space, eliminating the zero-estimate collapse of input-space search and replacing the independence assumptions of prior activation-space estimators with conditional sampling under an explicit, heavier-tailed activation prior. We also propose the Shifted-Power Bregman (SPB) Loss, a proper scoring rule that remains finite for zero-estimates and offers tunable asymmetry between underestimation and overestimation penalties. Experiments on small transformer models reveal a bias-variance tradeoff: GA-AMLS achieves the lowest loss under symmetric evaluation, reducing average log-space squared error relative to the strongest baseline across model sizes, while methods with overestimation bias prevail under asymmetric penalties. Our findings highlight that estimator choice should be matched to deployment context. More broadly, our work establishes activation space as a tractable domain for rare-event estimation in language models, circumventing the brittleness of discrete input-space search.

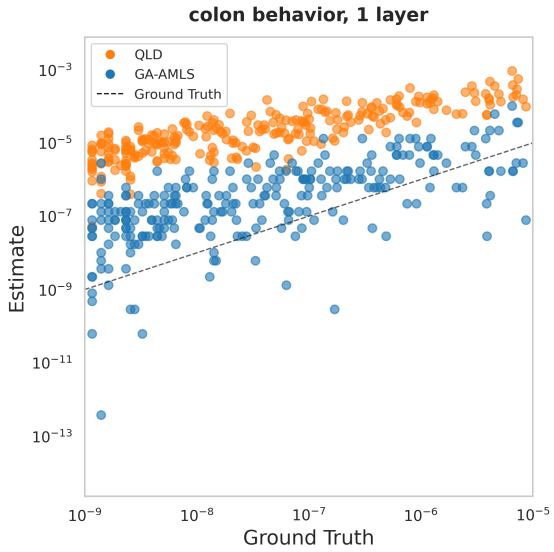
1 Introduction

Modern large language models (LLMs) are trained to generalize to new inputs by minimizing expected loss during training. However, this averaging objective leaves them susceptible to producing highly undesirable outputs on a small subset of rare inputs. While these events may appear statistically negligible during training, the rarity of a failure is often relative. In deployment, distribution shifts, whether manufactured through adversarial jailbreaks or incidental due to goal misgeneralization (Shah et al., 2022), can make these catastrophic failures significantly more common. Furthermore, in safety-critical applications such as autonomous vehicles or medical infrastructure, even a single failure can be unacceptable. Additionally, the large-scale use of language models can make low-probability events eventual certainties. Therefore, there is a growing need for accurate low probability estimation (LPE) methods: the ability to quantify the likelihood of a model producing a specific harmful output, even when that probability is extremely small, such as between 10^{-5} and 10^{-9} , ranges where naive sampling is computationally infeasible. Following the problem setting introduced by Wu & Hilton (2025), we study LPE in the context of argmax sampling, where our goal is to estimate the probability that a specific target token will have the largest output logit within a computationally constrained budget as shown in Figure 1a.

The status quo established by Wu & Hilton (2025) comprises two classes of methods. Input-space Importance Sampling searches the discrete token space; for the rarest targets it often fails to find any triggering input, collapsing estimates to zero, and it additionally requires expensive full backward passes and knowledge of exact input-sequence probabilities. The activation-space alternative, Quadratic Logit Decomposition (QLD),



(a) Low-probability estimation setup.



(b) GA-AMLS and QLD estimates on the colon distribution and 1-layer model. Each point represents a different target token.

Figure 1: Overview of the low-probability estimation setup and GA-AMLS estimates.

avoids zero estimates by recombining whitened activation components as if they were independent, producing n^2 synthetic candidates from n empirical activations. But whitening guarantees only uncorrelatedness, not independence, and language-model activations carry heavy-tailed, outlier-dominated structure (Sun et al., 2024; Dettmers et al., 2022) that can survive whitening; empirically, the resulting distortion surfaces mainly as overestimation on rare events. Finally, the metrics used to compare these methods, Itakura-Saito (IS) loss and squared error loss in log space, are undefined at zero estimates, which Wu & Hilton (2025) patch with an affine calibration that itself requires known ground-truth probabilities.

In this paper, we address both the algorithmic and evaluative limitations of current LPE approaches through two coupled contributions: an activation-space estimation algorithm, and a numerically stable evaluation metric that makes comparing all estimators, including those whose estimates collapse to zero, well-posed.

First, we propose *Gradient Activation Adaptive Multi-Level Splitting* (GA-AMLS), which adapts the Adaptive Multi-Level Splitting (AMLS) algorithm (Guyader et al., 2011) to operate entirely within the continuous activation space of language models. While Webb et al. (2019) previously applied AMLS to computer vision models in the input space, ours is, to our knowledge, the first method to apply it in the activation space of language models. Figure 2 gives a schematic overview of GA-AMLS, while Figure 1b previews its comparison with the strongest activation-space baseline, QLD, on a representative distribution. By working in activation space, we can estimate tail probabilities without the expensive full backward passes required by Importance Sampling. We also avoid needing to compute the input sequence probability under the original distribution, which Importance Sampling methods require for importance weights. Instead, we decompose the search for activations that trigger the target token, into a sequence of intermediate levels, using a gradient-based Metropolis-adjusted Langevin algorithm (MALA) kernel (Roberts & Tweedie, 1996; Srinivasan et al., 2025) to guide the sampler within each level. The MALA kernel's drift term allows the evolved activations to stay "close" to the typical activations of the input distribution, while the adaptive levels progressively steer towards the failure region. This replaces QLD's independence-based recombination with conditional sampling under an explicit prior. Idealized AMLS is an unbiased estimator of the tail probability of its sampling distribution (Cérou et al., 2019); because our MCMC kernel targets a fitted activation prior rather than the true activation distribution, GA-AMLS inherits this guarantee only up to prior mismatch (Section 7).

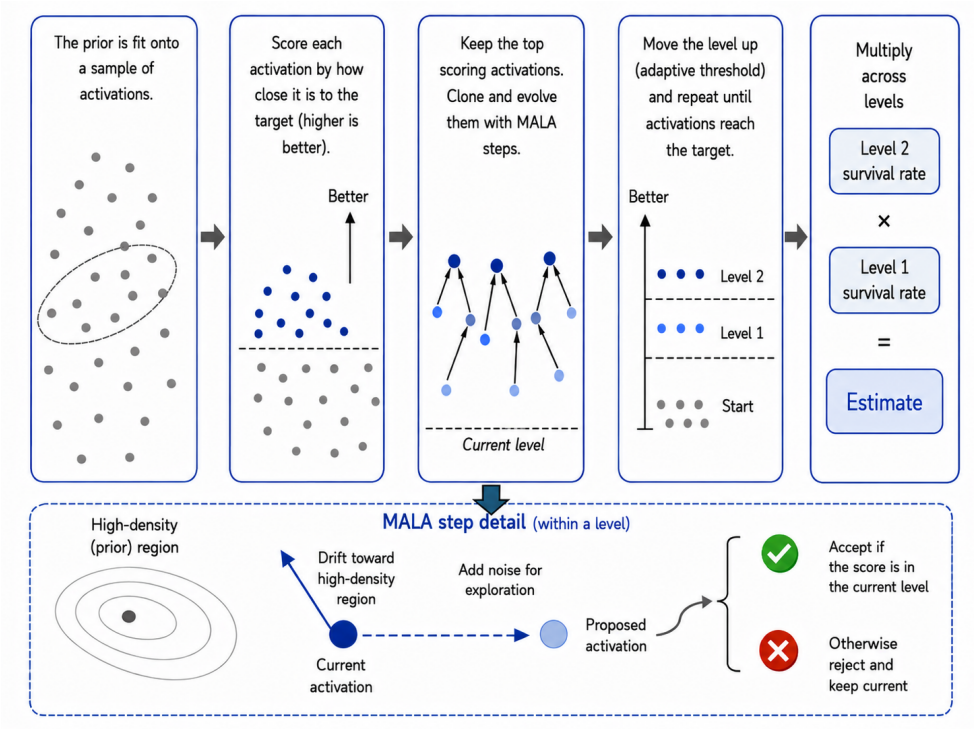


Figure 2: **GA-AMLS in activation space.** The rare-event probability is decomposed into a product of conditional probabilities over adaptive score levels. At each level, activations below the threshold are discarded and survivors are resampled. MALA rejuvenation uses the fitted activation-prior gradient to improve mixing within the constrained level.

Second, we introduce *Shifted Power Bregman* (SPB) Loss, a family of proper scoring-rule divergences designed for evaluating rare-probability estimates. SPB remains finite when an estimator returns exactly zero, supports tunable asymmetry between underestimation and overestimation through an interpretable asymmetry parameter, and approximately preserves scale invariance across several orders of magnitude of the ground-truth probability. The unshifted power family also recovers standard losses as special cases, including squared error and the Itakura–Saito loss used by Wu & Hilton (2025) (Appendix B.8). Thus SPB Loss is a strict generalization of existing evaluation losses, with an explicit finiteness correction and a tunable asymmetry knob. A comparison with standard losses is presented in Table 1.

Empirically, the estimator ranking depends on the evaluation cost model. Under approximately symmetric penalties, including log-space squared error and corresponding SPB settings, GA-AMLS has lower bias and outperforms the baselines. Under strongly asymmetric penalties that heavily punish underestimation, including Itakura–Saito loss and corresponding SPB settings, QLD can outperform GA-AMLS because its positive bias acts as a conservative hedge against false negatives. This rank inversion is visible under pre-existing metrics as well as under SPB; SPB makes the tradeoff explicit while avoiding the numerical instability caused by zero estimates. Consequently, the appropriate estimator depends on the deployment context: conservative catastrophic-risk audits may prefer overestimation, whereas applications sensitive to excessive false alarms may prefer the lower-bias GA-AMLS diagnostic.

2 Background and Problem Setup

2.1 Problem Setup

Low-probability estimation. We adopt the low-probability estimation (LPE) setting of Wu & Hilton (2025). Let M be a language model with vocabulary $\mathcal{V}$, and let $x \sim \mathcal{D}$ be an input sequence drawn from a

Property	Squared Error	IS Loss	Log Squared Error	SPB Loss
Scale invariance	✗	✓	✓	✓ ≈ exact
Defined when estimate = 0	✓	✗	✗	✓
Asymmetric: FN > FP penalty	✗	✓	✗	✓ tunable via α
Proper scoring rule	✓	✓	✗	✓

Table 1: SPB Loss is the only loss satisfying all four desiderata for LPE evaluation: scale invariance, finiteness at zero estimates, tunable asymmetry between underestimation and overestimation penalties via α, and strict propriety.

specified input distribution. We consider deterministic argmax decoding: $M(x)$ is the token with maximal output logit. For a target token $t \in \mathcal{V}$, the rare-event probability of interest is $q_t = \Pr_{x\sim\mathcal{D}}[M(x) = t]$. In our experiments, q_t lies in the extreme-tail regime $10^{-9} \leq q_t \leq 10^{-5}$, where naive Monte Carlo sampling is infeasible.

Activation-space formulation. GA-AMLS operates on internal activations rather than discrete text. Let $f : \mathcal{V}^* \to \mathbb{R}^d$ map an input sequence to its final-layer pre-LayerNorm, pre-unembedding activation $a = f(x)$. Given unembedding matrix W_U, the logits are $z(a) = \text{LayerNorm}(a)W_U$. The event $M(x) = t$ can therefore be written as an activation-space constraint: $q_t = \Pr_{x\sim\mathcal{D}}\left[z_t(f(x)) > \max_{i\neq t} z_i(f(x))\right]$.

Existing estimators. We compare against the methods introduced by Wu & Hilton (2025). Input-space importance sampling methods such as ITGIS and MHIS search over discrete token sequences and reweight samples to estimate q_t. These estimators can be unbiased in principle, but empirically in the extreme-tail regime they often fail to find triggering inputs, producing zero estimates. QLD avoids this failure mode by operating in activation space. It whitens pre-unembedding activations, decomposes each activation as $u = a + b$ into a target-direction component and an orthogonal residual, and estimates q_t by recombining components $a^{(i)} + b^{(j)}$. This converts n activations into n^2 candidate activations, but relies on treating the whitened components as independent. Since whitening guarantees uncorrelatedness rather than independence except under stronger assumptions such as joint Gaussianity, non-Gaussian activation structure can introduce bias.

2.2 Sampling Methods

The Adaptive Multilevel Splitting (AMLS) Algorithm. Adaptive Multi-Level Splitting (AMLS) is a rare-event Monte Carlo method that estimates a small probability by decomposing it into a product of larger conditional probabilities (Guyader et al., 2011; Cérou et al., 2019; Webb et al., 2019). Let $V \sim G$ be a random variable, let $s : \mathbb{R}^d \to \mathbb{R}$ be a score function, and let the rare event be $\{s(V) \geq \tau\}$. For levels $-\infty = L_0 < L_1 < \cdots < L_K = \tau$, $\Pr(s(V) \geq \tau) = \prod_{k=1}^{K} \Pr(s(V) \geq L_k \mid s(V) \geq L_{k-1})$.

AMLS estimates these factors using a population of N particles. At each level, the threshold L_k is chosen adaptively from the empirical score distribution; particles below L_k are discarded, survivors are resampled, and the resulting duplicates are rejuvenated with an MCMC kernel targeting $G(\cdot \mid s(V) \geq L_k)$. The final estimate is $\hat{q}_{\text{AMLS}} = \prod_{k=1}^{K} \hat{p}_k, \qquad \hat{p}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{s(V_i) \geq L_k\}$. Thus, AMLS replaces one extremely rare event with a sequence of more frequent conditional events.

MALA Proposal Kernel. The efficiency of AMLS depends on the MCMC kernel used to rejuvenate particles. We use the Metropolis-adjusted Langevin algorithm (MALA), which proposes moves using local gradient information. For a target density π, MALA proposes (Roberts & Tweedie, 1996; Srinivasan et al., 2025) $x' = x + h\nabla_x \log \pi(x) + \sqrt{2h}\,\xi, \qquad \xi \sim \mathcal{N}(0, I)$, and accepts or rejects the proposal using the standard Metropolis-Hastings correction.

3 Gradient Activation AMLS (GA-AMLS) for Estimating Rare Events

We adapt the general AMLS algorithm to work in the activation space of language models, and use a MALA kernel to guide the search (Section 2.2). We call this method Gradient Activation AMLS (GA-AMLS), presented in Appendix C as Algorithm 1 and outlined below.

Notation. Let $a \in \mathbb{R}^d$ denote the raw pre-LayerNorm pre-unembedding activation, and let u denote its whitened coordinate. Given mean μ and covariance Σ, choose a de-whitening matrix A such that $AA^\top = \Sigma$, and define $u = (a - \mu)A^{-\top}, a(u) = uA^\top + \mu$. GA-AMLS and the MALA kernel operate over u, while model logits are evaluated by inserting the corresponding unwhitened activation $a(u)$ back into the model.

Prepare the Activation Space: We use activations from the pre-LayerNorm pre-unembedding layer, and whiten them. Raw transformer activations a are characteristically anisotropic, with a few outlier dimensions whose scale, in magnitude and in across-distribution variance, is orders of magnitude larger than the rest. This outlier structure is documented in large language models (Sun et al., 2024; Dettmers et al., 2022); we verify empirically that it also holds at the smaller models we study in the Appendix in Figure 12. To improve mixing of MCMC kernels in this high-dimensional space, we operate in whitened coordinates $u = (a-\mu)A^{-\top}$. This global preconditioning makes isotropic Euclidean steps in u-space correspond to Mahalanobis-scaled steps in the original activation space. We sample pre-LayerNorm activations so that the evolved activations do not need to satisfy the LayerNorm constraint directly.

Define the Score function: We define the score function s that directly describes how "close" an activation is to an activation producing the target token. Given whitened pre-LayerNorm pre-unembedding activation $\mathbf{u}$ as input, A as the de-whitening matrix, μ as the mean, t as the target token index, the score function computes:

$$s(\mathbf{u}) = \Big(\mathrm{LayerNorm}(\mathbf{u}A^\top + \mu)\, W_U\Big)_t - \max_{j \neq t}\Big(\mathrm{LayerNorm}(\mathbf{u}A^\top + \mu)\, W_U\Big)_j .$$

That is, $s(\mathbf{u})$ is the margin by which the target token's logit exceeds the maximum logit among all other tokens. Using the general AMLS algorithm from Section 2.2 , we define the rare event as $P\big(s(\mathbf{u}) \geq \tau\big)$. In our setting $\tau = 0$, since this is when the logit for the target token t is the largest, and thus t will be the next token predicted under our assumption of argmax sampling.

Determine the prior distribution: The prior π is defined over whitened activations u and is used to guide the MCMC kernel. In whitened space, we use a diagonal product Student-t prior. This captures heavier-than-Gaussian marginal tails. This choice is supported by the tuning ablations in Appendix D.1, where the Student-t prior outperforms a Gaussian prior.

Choose the MCMC kernel K: To efficiently evolve the clones' activations within the high-dimensional activation space, we use the Metropolis-adjusted Langevin algorithm (MALA) as outlined in Section 2.2, with the below correction to take into account the constraint induced by the score mechanism of AMLS. We must strictly enforce the level constraints $s(\mathbf{u}) \geq L_i$. This is addressed through a Metropolis-Hastings filter.

The proposal density $q(\mathbf{u}' \mid \mathbf{u})$ is Gaussian, while the prior distribution is Student's t distribution.

The acceptance probability r is computed in two stages to handle the hard constraint:

1. Hard score check: If $s(\mathbf{u}') < L_i$, the proposal is invalid and immediately rejected ($r = 0$);
2. Ratio Computation: If valid, we compute the standard Metropolis Hastings (MH) acceptance ratio r:

$$r = \min\left(1, \frac{\pi(\mathbf{u}')\, q(\mathbf{u} \mid \mathbf{u}')}{\pi(\mathbf{u})\, q(\mathbf{u}' \mid \mathbf{u})}\right).$$

We accept the proposal $\mathbf{u}'$ with probability r; otherwise, the activation remains at $\mathbf{u}$.

Standard Random Walk Metropolis-Hastings (RWMH) proposes isotropic steps blind to the geometry of the distribution, leading to slow mixing in this high dimensional activation space. MALA improves upon

RWMH by using gradient information from the underlying prior distribution to guide the proposal towards the activation space's typical set for the specific input domain $\mathcal{D}$. In GA-AMLS this MALA kernel is applied to the restricted target $\pi_k(u) \propto \pi(u)\,\mathbf{1}\{s(u) \geq L_k\}$, so proposals below the current level L_k are rejected by the MH filter. Thus the Langevin drift improves exploration within the current constrained level set, while the adaptive thresholds L_k move the population toward the rare target-token regions. This choice is also supported by the tuning ablations in Appendix D.1, where MALA outperforms RWMH despite using half as many MH proposal steps.

To ensure optimal mixing, we adapt the step size h (from MALA under 2.2) during a burn-in period of T_{burn} steps. We employ a multiplicative update rule to tune h toward a target acceptance rate of $r^* = 0.57$, which is theoretically optimal for MALA in high dimensions (Roberts & Rosenthal, 1998): $h_{t+1} = h_t \cdot \exp\left(\omega(\bar{r}_t - r^*)\right)$, where $\omega = 0.1$ is the adaptation rate and $\bar{r}_t$ is the empirical acceptance rate at step t. After T_{burn} steps, h is frozen to ensure the Markov chain satisfies detailed balance condition during the sampling.

4 Shifted Power Bregman (SPB) Loss for Evaluation

We introduce Shifted-Power Bregman (SPB) Loss, a proper scoring rule for evaluating rare-probability estimates. SPB remains finite when an estimator returns zero and provides tunable asymmetry between underestimation and overestimation. We define both the pointwise loss and a dataset-level aggregation that controls scale dependence across rare-event probabilities.

Desiderata for an ideal loss for LPE. To evaluate an estimator for LPE, an ideal loss would satisfy the following requirements:

Scale Invariance: As mentioned in Wu & Hilton (2025), it should penalize relative errors similarly across the scale of ground truth probabilities, as we care about performance across a wide range of rarities (e.g., 10^{-5} to 10^{-9}).

Asymmetric Sensitivity: It should support tunable asymmetry between underestimation and overestimation penalties, as different deployment contexts may prioritize avoiding underestimation over overestimation, or vice versa.

Numerical Stability: It must be defined when the estimate is exactly zero, avoiding the need for calibration steps that require ground truth knowledge.

Statistical Rigor: It should be the divergence of a proper scoring rule, so that an estimator minimizing expected loss is incentivized to report its unbiased best estimate rather than a systematically distorted one (Gneiting & Raftery, 2007).

The Itakura–Saito (IS) loss used in Wu & Hilton (2025) satisfies most criteria but fails on numerical stability: it is undefined (infinite) when the estimate is 0. Wu & Hilton (2025) overcome this by fitting an affine transformation $x \mapsto ax^c + b$ to the outputs, where parameters are chosen to minimize IS loss via leave-one-out-cross-validation (LOOCV) for each method, input, distribution pair. A significant drawback is that this transformation relies on knowing the ground truth probabilities to calculate the IS loss which is to be minimized during this calibration. In a realistic deployment setting, establishing this ground truth for rare tokens is computationally infeasible using naive sampling, as estimating these probabilities is the very problem we are trying to solve. SPB uses ground truth only as the reference value in the final score, as any evaluation metric does. By contrast, the affine calibration used in prior evaluation procedures assumes that ground-truth probabilities are available to *transform each estimator's outputs* before scoring. This is a stronger requirement than ordinary evaluation against ground truth, since such calibration information would generally not be available in deployment.

Pointwise SPB Loss. We propose pointwise Shifted-Power Bregman (SPB) Loss, derived from the framework of proper scoring rules. It overcomes this limitation by accepting a trade-off: if the estimate is 0, the loss will not be infinite, but rather capped by a large finite constant. We also introduce parameters α: to control the degree of asymmetric sensitivity, and γ: a dataset level rarity premium.

Following Buja et al. (2005) , every smooth proper scoring rule for binary probability estimation induces a Bregman divergence determined by a nonnegative weight function $\omega(t) \geq 0$ on $(0, 1)$. If q denotes the true probability and p the forecast, the associated divergence can be written as $B(q \mid p) = \int_q^p (t - q)\omega(t)\, dt$.

We define the shifted-power weight function as: $\omega_\varepsilon(t) = (t + \varepsilon)^{-\alpha}$, where $\alpha > 0, \varepsilon > 0$. The shift parameter ε ensures the penalty remains finite at $p = 0$. Geometrically, this integral accumulates cost along the one-dimensional probability axis between the truth q and the prediction p. At an intermediate probability level t, the local contribution is the distance from the truth, scaled by the weight $\omega_\varepsilon(t) = (t+\varepsilon)^{-\alpha}$. Since this weight is largest near zero, errors that push the prediction toward smaller probabilities are penalized more strongly. For an overestimate $p = mq$, the loss integrates over $t \in [q, mq]$, where the weight decreases as t increases. For a reciprocal underestimate $p = q/m$, the loss integrates over $t \in [q/m, q]$, which includes smaller probabilities where $\omega_\varepsilon(t)$ is larger. Increasing α steepens this growth near zero and therefore increases the relative penalty for underestimation compared with overestimation. Thus, α provides a tunable asymmetry parameter, as formalized in Theorem 2.

Dataset-Level SPB Loss and the Rarity Premium γ. To evaluate an estimator across a diverse set of rare events, we aggregate the pointwise divergences into a single dataset-level loss. As derived in Appendix A, the *unshifted* pointwise divergence ($\varepsilon = 0$) scales exactly proportionally to $q^{2-\alpha}$ for a fixed relative error $m = p/q$. While the shift parameter $\varepsilon > 0$ slightly perturbs this geometry, the shifted divergence $B_\varepsilon(q \mid p)$ remains dominated by this leading-order $q^{2-\alpha}$ scaling. Left uncorrected, an aggregate loss would be dominated by the absolute scale of the ground-truth probabilities rather than the relative accuracy of the estimators.

To neutralize this and establish controlled scale invariance, we define the dataset-level SPB Loss as the arithmetic mean[1] of the *weighted* pointwise divergences across n samples: $\mathcal{L}_{\text{SPB}} = \frac{1}{n}\sum_{i=1}^{n} w(q_i)\, B_\varepsilon(q_i \mid p_i)$. We define the weight function as: $w(q) = q^{\alpha-2-\gamma}$, where $\gamma \geq 0$. The parameter γ allows us to toggle between two distinct evaluation modes:

Scale Neutrality ($\gamma = 0$): By setting $\gamma = 0$, the weight becomes $q^{\alpha-2}$. This term cancels the leading $q^{2-\alpha}$ scaling of the divergence. Consequently, the final metric approximates scale invariance, depending almost entirely on the relative multiplicative error ($m = p/q$) rather than the absolute rarity of the event. The residual distortion due to ε is formalized and bounded in the Appendix in Theorem 1.

The Rarity Premium ($\gamma > 0$): In specific safety-critical deployments, evaluating performance symmetrically across scales is less important than severely punishing errors on the deepest tail events. Setting $\gamma > 0$ intentionally breaks scale neutrality by introducing an extra $q^{-\gamma}$ factor. This acts as a tunable *rarity premium*, disproportionately upweighting the smallest-q events in the aggregate score.

Relation to other Losses The unshifted power family ($\varepsilon = 0$) unifies standard losses: $\alpha = 0$ recovers squared error and $\alpha = 2$ is *exactly* the Itakura–Saito loss (Appendix B.8). Therefore SPB comes from a strict generalization of the metric used by Wu & Hilton (2025), adding finiteness at zero estimates (with $\varepsilon > 0$) and a tunable asymmetry knob.

4.1 Practical parameter selection (α, γ, ε)

The metric has three user-facing parameters: 1) α sets the FN-vs-FP asymmetry: $\alpha > 1.5$ penalizes underestimation more than overestimation and $\alpha < 1.5$ the reverse (Theorem 2, Eq. 20 in Appendix; fine-grained control of α is described in Appendix B.6.1). 2) $\gamma \geq 0$ is an optional rarity premium: $\gamma = 0$ gives scale invariance up to a tolerance η (Theorem 1 in Appendix), while $\gamma > 0$ multiplies each example's contribution to the aggregate loss by $q^{-\gamma}$, so at a fixed relative error a $10\times$ rarer event carries $10^\gamma\times$ the weight. 3) ε is the floor that makes $p = 0$ finite at the cost of controlled distortion; a heuristic default is $\varepsilon = 0.01\, q_{\min}/m_{\max}$, and a guaranteed choice is any fraction of the bound given in Appendix by Corollary 1 (interpretation in Appendix B.5). Table 2 (Appendix B.6) suggests settings by evaluation use case.

[1]Crucially, we divide by the sample size n rather than the sum of the weights $\sum_i w(q_i)$. Since the weight $w(q)$ is specifically designed to perform a pointwise cancellation of the $q^{2-\alpha}$ scale factor, applying a normalized weighted average would reintroduce a dependence on the empirical distribution of the ground truth probabilities q, undoing the cancellation.

5 Experimental Results

This section empirically evaluates both components of our proposal. Section 5.1 tests GA-AMLS on the real-data LPE benchmark of Wu & Hilton (2025), comparing it against input-space Importance Sampling methods and QLD using both point-estimate plots and aggregate SPB Loss. Section 5.2 uses synthetic experiments to validate the behavior predicted by the SPB Loss theory, including approximate scale invariance, bounded distortion from the ε-shift, and tunable asymmetry between underestimation and overestimation.

5.1 GA-AMLS: Real Data Experiments

Dataset: We use 8 distributions and the ground-truth probabilities as used in Wu & Hilton (2025). They generated the ground-truth probabilities for each of the 8 distributions by running forward passes on 2^{32} random samples. We select a random set of 256 tokens among those with ground-truth probabilities between 10^{-9} and 10^{-5}, and test the methods on these tokens.

Methodology: We follow the experimental setup of Wu & Hilton (2025). To prevent overfitting, the method was only run on the first four distributions during development, and finalized before testing on the last four distributions. We test the methods on 3 models: a 1-layer, a 2-layer, and a 4-layer transformer from Nanda & Bloom (2022). All models have a hidden dimension of $d = 512$, a vocabulary size of $|V| = 48262$, GELU non-linearities Hendrycks & Gimpel (2023), and were trained on the C4 dataset Raffel et al. (2023) and CodeParrot Tunstall et al. (2022). We use the Shifted-Power Bregman (SPB) Loss as our evaluation loss. By varying the asymmetry parameter α and the dataset-level rarity premium γ , we demonstrate how different penalization profiles (detailed in Appendix in Table 2) alter the apparent ranking of the methods. For hyperparameter tuning details please refer to Appendix D.1.

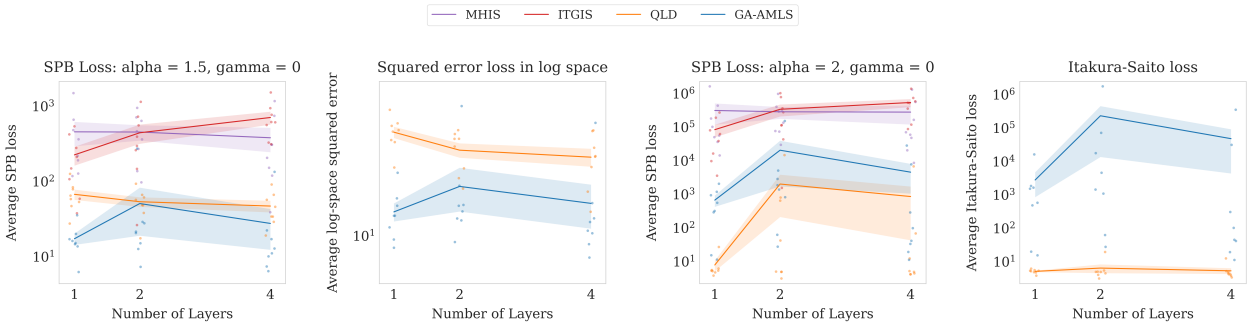


Figure 3: **Multi-metric evaluation:** Plotted side-by-side, we observe a rank inversion depending on the metric's sensitivity to underestimation. Under symmetric penalties (SPB $\alpha = 1.5$ and Squared Error in log space), GA-AMLS's lower bias leads to outperformance. Under severe asymmetric penalties (SPB $\alpha = 2$ and IS Loss), QLD's strategy of systematic overestimation guards it against false-negative penalties, resulting in lower loss. SPB Loss safely evaluates input-space methods that collapse to zero, overcoming the numerical instability of uncalibrated IS loss and squared error loss in log space. (All plotted with log transformed y axis.)

Baselines: We compare against QLD, ITGIS and MHIS as described in Section 2.1.

Performance under Symmetric Penalties: Figure 3 compares the overall evaluation losses of all methods across different model sizes grouped by symmetric and asymmetric metrics, while Figure 16 (Appendix F) details the SPB loss for each of the 8 individual data distributions. Under evaluation metrics that penalize overestimation and underestimation symmetrically, such as log-space Squared Error and SPB Loss with $\alpha = 1.5$, we observe that GA-AMLS significantly outperforms all baseline methods (Figure 3, first two panels). This outperformance is consistent across 7 of the 8 individual distributions (Figure 16).

Performance under Asymmetric Penalties: Focusing on the latter half of Figure 3, the figure illustrates the methods' performance under asymmetric metrics that severely penalize underestimation (SPB Loss with

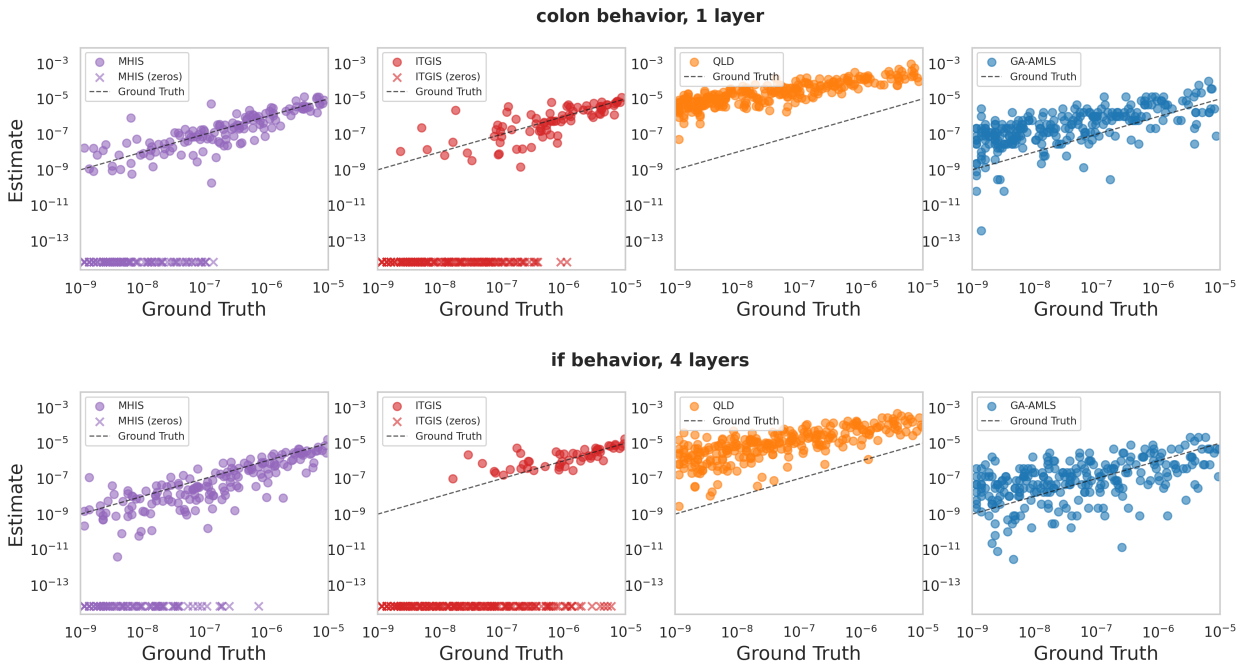


Figure 4: **Examples of probability estimate vs ground truth on two input distributions and models.** Input Space Methods (ITGIS/MHIS) frequently fail to find valid triggers, causing their estimates to collapse to zero (placed at the bottom of each graph for visibility). While QLD shows a systematic overestimation bias, GA-AMLS centers closest to the ground truth, at the cost of higher variance.

$\alpha = 2$ and IS Loss). When evaluating with these metrics, we observe a clear inversion in method rankings, where QLD outperforms GA-AMLS under these conditions. This occurs because QLD is biasing its estimates toward overestimation. While less accurate overall, this systematic positive bias acts as a natural safeguard against the severe false-negative penalties imposed by asymmetric losses.

Deep-Tail Overestimation and Rarity Premium γ: Figure 4 plots the methods' individual point estimates against the true probabilities to visualize estimator bias, and Figure 17 (Appendix F) shows how the aggregate SPB loss shifts across methods as the rarity premium γ increases. The scatter plots (Figure 4 and Appendix E) reveal that QLD's overestimation bias is not uniform; it becomes significantly more pronounced at the deepest tail (the rarest events). Consequently, as seen in Figure 17, as we increase the dataset-level rarity premium ($\gamma > 0$) to disproportionately upweight the rarest events, the performance gap between QLD and GA-AMLS narrows, demonstrating the cost of QLD's deep-tail overbias.

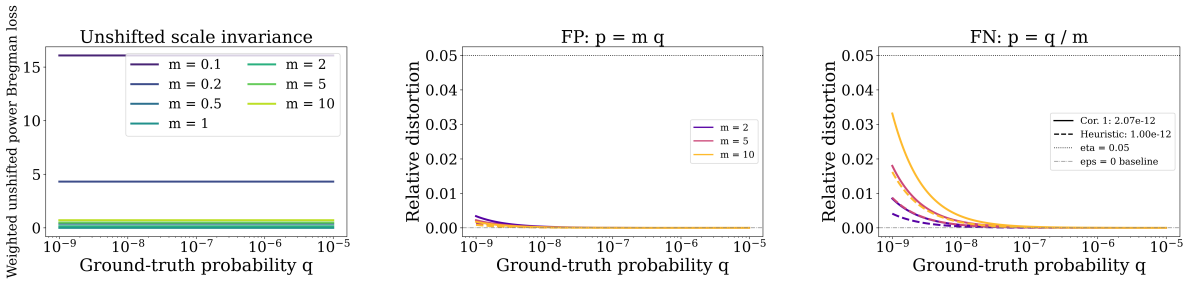
The Zero-Estimate Collapse and Bias-Variance Tradeoff: The scatter plots of method estimates (Figure 4 and Appendix E) also directly capture the distribution and structural failure modes of each estimator's predictions. The distinct performance profiles of these methods stem from a fundamental bias-variance tradeoff, visually observed in these plots. The input-space Importance Sampling methods (MHIS/ITGIS) suffer from frequently collapsing to zero for the rarest events. QLD avoids zero-estimates but shifts the entire distribution of predictions upward. GA-AMLS shows lower bias and is more centred around the ground truth line, but exhibits higher variance across tokens than QLD (as seen visually in scatterplots in Appendix E).

5.2 SPB Loss: Synthetic Data Experiments

We now isolate the behavior of SPB Loss itself, independently of any estimator. The goal of these synthetic experiments is to illustrate the loss geometry discussed in Section 4.1 and formalized in Appendix B: approximate scale invariance, controlled distortion from the finite shift ε, and tunable asymmetry between underestimation and overestimation.

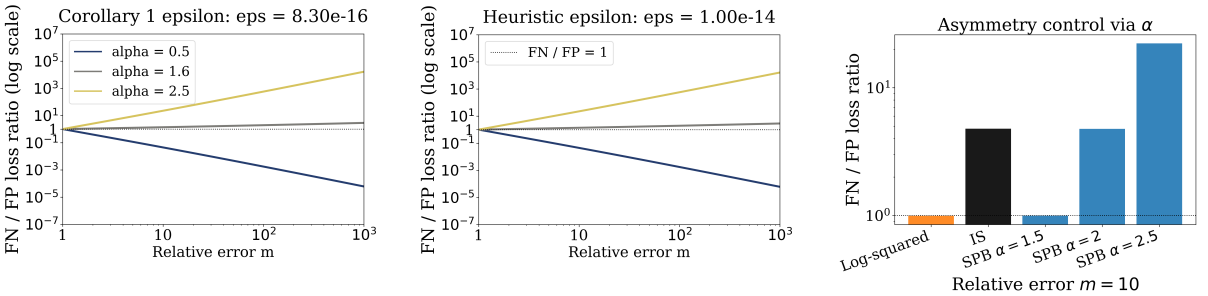
Setup. We evaluate rare probabilities $q \in [10^{-9}, 10^{-5}]$ on a log-spaced grid and we perturb each probability by fixed multiplicative factors. For scale-invariance experiments, predictions range from $0.1q$ to $10q$. For asymmetry experiments, we compare paired reciprocal errors: an overestimate/False-Positive (FP) $p = mq$ and an underestimate/False-Negative (FN) $p = q/m$, with m ranging up to 1000. We use the scale-neutral setting $\gamma = 0$.

For finite-shift $\varepsilon > 0$ plots, we compare two choices of ε: one chosen according to the sufficient bound in Corollary 1, and a simple heuristic choice $\varepsilon = 0.01(q_{\min}/m_{\max})$. The tolerance for the scale-distortion plots is $\eta = 0.05$.



(a) Unshifted scale invariance. (b) FP distortion. (c) FN distortion.

Figure 5: **Scale invariance and finite-shift distortion of SPB Loss. (a)** With $\varepsilon = 0$, the weighted loss is flat across q for each fixed multiplicative error, showing exact scale neutrality. **(b)–(c)** Introducing $\varepsilon > 0$ makes the loss finite at zero estimates but slightly perturbs the unshifted geometry. The observed distortion remains below the target tolerance $\eta = 0.05$ for both overestimation (FP) and underestimation (FN).



(a) Corollary ε. (b) Heuristic ε. (c) Comparison with standard losses.

Figure 6: **Asymmetry control via α. (a)–(b)** FN/FP loss ratios for reciprocal multiplicative errors under the Corollary-based and heuristic choices of ε. Ratios above one indicate that underestimation is penalized more strongly than overestimation. **(c)** At $m = 10$, SPB Loss provides a tunable continuum between approximately symmetric behavior and strongly FN-averse behavior, compared with log-space squared error and IS Loss.

Scale invariance and shift distortion. Figure 5 examines whether SPB Loss treats the same relative error similarly across different rarity levels. In Panel 5a, the unshifted weighted loss is flat as a function of q: for a fixed multiplicative error, the loss is essentially independent of whether the true probability is 10^{-5} or 10^{-9}. This is the desired scale-neutral behavior.

Panels 5b and 5c then show what changes when the finite shift $\varepsilon > 0$ is introduced. The shift is needed to keep the loss finite when an estimator outputs zero, but it slightly perturbs the ideal unshifted geometry. The plots report this perturbation separately for overestimation and underestimation. For both the Corollary-based shift and the heuristic shift, the observed distortion remains below the target tolerance $\eta = 0.05$ over

the displayed range. Thus, in this regime, the finite correction preserves the practical scale-neutral behavior of the loss.

Asymmetry control via α. Figure 6 studies the relative penalty for FN versus FP. Panels 6a and 6b plot the FN/FP loss ratio for reciprocal errors: a value above one means that underestimating q by a factor m is penalized more than overestimating q by the same factor. The curves show that α controls this asymmetry. Values above the threshold $\alpha = 1.5$, such as $\alpha = 1.6$ and $\alpha = 2.5$, produce FN/FP ratios above one across the plotted range, consistent with Theorem 2. The $\alpha = 0.5$ curve is included as a contrasting setting outside this FN-averse regime.

Panel 6c compares this behavior with standard losses at a fixed multiplicative error $m = 10$. Log-space squared error is approximately symmetric between underestimation and overestimation, while IS Loss is FN-averse. SPB Loss interpolates between these behaviors through α: $\alpha \approx 1.5$ is close to symmetric, while larger α increasingly penalizes underestimation. This gives SPB Loss an explicit asymmetry knob while retaining finiteness at zero estimates.

6 Related Work

Low Probability Estimation in LLMs. Low probability estimation for language models was formalized by Wu & Hilton (2025), who introduced input-space Importance Sampling methods and the activation-space QLD baseline used in our experiments. We adopt their benchmark setting, but replace discrete input-space search with gradient-guided sampling in activation space and introduce SPB Loss to handle zero estimates and asymmetric safety costs. In a complementary line of work, Jones et al. (2025) address rare behavior estimation from a forecasting perspective rather than direct probability estimation. Their approach leverages elicitation probabilities defined as the probability that a given query produces a target behavior under repeated sampling, and uses extreme value theory to forecast how deployment-scale risks grow from smaller evaluation sets. Their method is particularly suited for anticipating when rare failures might emerge at scale, whereas our approach provides precise probability estimates for individual rare events. The two approaches are complementary: GA-AMLS can provide accurate point estimates of elicitation probabilities that could serve as inputs to the forecasting framework of Jones et al. (2025), potentially improving the precision of deployment-scale risk predictions. Concurrent work by Dorman et al. (2026) develops an end-to-end framework for rare event analysis in LLMs, extending importance sampling to full autoregressive completions in the discrete token space. Because their method relies on generating modified text sequences, it must calculate the exact probability of those generated sequences under the original model to compute importance weights. GA-AMLS avoids scoring full text sequences by operating in the continuous activation space.

Activation-Space Methods. A growing interpretability literature supports treating language-model activations as a structured domain for analysis and intervention. The linear representation hypothesis (Park et al., 2024), representation engineering (Zou et al., 2023a), and activation steering methods (Turner et al., 2024; Panickssery et al., 2024) show that model behaviors can often be manipulated through directions in residual-stream space. Recent work further suggests that such behavioral structure need not be captured by a single global direction: heterogeneity-aware steering methods model activations as clustered distributions and use optimal transport to construct input-dependent interventions (Abdullaev et al., 2026). Related work finds low-dimensional structure for refusal behavior (Arditi et al., 2024), while sparse autoencoders reveal richer monosemantic feature structure in activations (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024). The Tuned Lens (Belrose et al., 2025) further motivates operating near the pre-unembedding layer, where activations are predictive of output tokens. GA-AMLS builds on this view of activation space, but treats it as a probability space: instead of identifying causal directions, interventions, or features, we estimate the probability mass of rare-event regions.

Rare-Event Sampling. Our method adapts Adaptive Multi-Level Splitting (AMLS), a rare-event Monte Carlo technique from statistical physics and particle simulation (Cérou et al., 2019). Webb et al. (2019) applied AMLS to neural-network robustness verification in computer vision using a random-walk Metropolis-Hastings kernel in input space. Directly applying this approach to language models is difficult because text

is discrete. GA-AMLS instead moves the estimator to continuous activation space, where gradient-based MCMC kernels such as MALA can be used efficiently.

Adversarial Search and Defenses. Red-teaming and jailbreaking methods search for individual inputs that trigger undesired behavior, using methods ranging from discrete gradient-based attacks such as GCG (Zou et al., 2023b) to activation-space attacks such as Subspace Rerouting (Winninger et al., 2025). These methods are useful for finding failures and generating adversarial training data (Goodfellow et al., 2015; Madry et al., 2019), including in latent adversarial training settings (Casper et al., 2024; Sheshadri et al., 2025). However, finding one failure is not the same as estimating its probability, and transfer across search methods can be unreliable (Kang et al., 2019; Wei et al., 2023). GA-AMLS is therefore diagnostic rather than defensive: it estimates how much probability mass lies in the rare-event region.

7 Discussion and Limitations

While GA-AMLS provides a principled activation-space estimator for rare model behaviours, several limitations remain.

Residual bias and activation-space approximation. GA-AMLS is designed as an unbiased rare-event estimator under the assumed activation prior and exact MCMC sampling. In practice, residual bias can arise from three sources. First, we approximate the true activation distribution with a fitted Student's t distribution, which introduces distributional mismatch. Second, finite MCMC budgets may prevent full mixing between AMLS levels, leaving residual correlations among particles. Third, because GA-AMLS operates in continuous activation space, extreme activations found by the sampler may not correspond to any valid discrete token sequence. We partially mitigate this activation pre-image issue by fitting the prior to activations induced by samples from the input distribution $\mathcal{D}$, encouraging samples to remain near the empirical activation manifold. However, the gap between continuous activations and discrete inputs remains a fundamental limitation of activation-space methods, including QLD.

Computational cost. GA-AMLS has a different computational profile from input-space importance sampling. It avoids full-model backward passes and operates from cached activations, but it requires many sequential activation-space scorer evaluations across AMLS levels. GA-AMLS is computationally heavier than QLD: QLD reduces estimation to fast closed-form matrix operations, whereas GA-AMLS requires sequential Markov-chain evolution across adaptive levels. This additional cost is the price of replacing QLD's independence-based recombination approximation with conditional sampling under a fitted activation prior. It may therefore limit very large sweeps over many target tokens, layers, or models.

Evaluation scope. For comparability with prior low-probability estimation work, our experiments follow the benchmark setting of Wu & Hilton (2025): the input distribution $\mathcal{D}$ consists of independent tokens, and rare events are defined by a single target token under deterministic argmax conditions. This controlled setting enables direct comparison with ITGIS, MHIS, and QLD, but it does not capture correlated natural prompts, instruction-following contexts, or multi-step autoregressive failures. Extending activation-space rare-event estimation to realistic prompt distributions and trajectory-level events is an important direction for future work. Finally, all of our evidence comes from the 1–4 layer, $d = 512$ transformers of this benchmark; whether the bias–variance profile of GA-AMLS, and the adequacy of a diagonal Student-t prior in whitened coordinates, persist at modern scale is untested.

8 Conclusion

We introduced GA-AMLS, an activation-space adaptive multilevel splitting method for estimating rare LLM behaviors that are infeasible to measure accurately with naive sampling. Empirically, GA-AMLS reduces zero-estimate failures and under symmetric penalties improves rare-probability estimation relative to baselines. We also proposed SPB Loss to evaluate rare-event estimates under asymmetric costs for underestimation and overestimation. While the method depends on the fidelity of the activation-space model, the results support activation-space rare-event simulation as a useful tool for measuring low-probability LLM behavior.

References

Laziz U. Abdullaev, Noelle Y. L. Wong, Ryan T. Z. Lee, Shiqi Jiang, Khoi N. M. Nguyen, and Tan M. Nguyen. Concept heterogeneity-aware representation steering, 2026. URL https://arxiv.org/abs/2603.02237.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, University of Pennsylvania, The Wharton School, 2005.

Stephen Casper, Lennart Schulze, Oam Patel, and Dylan Hadfield-Menell. Defending Against Unforeseen Failure Modes with Latent Adversarial Training, August 2024. URL http://arxiv.org/abs/2403.05030. arXiv:2403.05030 [cs].

Frédéric Cérou, Arnaud Guyader, and Mathias Rousset. Adaptive multilevel splitting: Historical perspective and recent results. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(4):043108, April 2019. ISSN 1054-1500, 1089-7682. doi: 10.1063/1.5082247. URL https://pubs.aip.org/cha/article/29/4/043108/1027309/Adaptive-multilevel-splitting-Historical.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale, 2022. URL https://arxiv.org/abs/2208.07339.

Jake McAllister Dorman, Edward Gillman, Dominic C. Rose, Jamie F. Mair, and Juan P. Garrahan. Rare event analysis of large language models, 2026. URL https://arxiv.org/abs/2602.06791.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Arnaud Guyader, Nicolas Hengartner, and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. *Applied Mathematics Optimization*, 64:171–196, 10 2011. doi: 10.1007/s00245-011-9135-z.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Erik Jones, Meg Tong, Jesse Mu, Mohammed Mahfoud, Jan Leike, Roger Grosse, Jared Kaplan, William Fithian, Ethan Perez, and Mrinank Sharma. Forecasting rare language model behaviors, 2025. URL https://arxiv.org/abs/2502.16797.

Daniel Kang, Yi Sun, Tom Brown, Dan Hendrycks, and Jacob Steinhardt. Transfer of adversarial robustness between perturbation types, 2019. URL https://arxiv.org/abs/1905.01034.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Neel Nanda and Joseph Bloom. TransformerLens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):255–268, January 1998. ISSN 1369-7412, 1467-9868. doi: 10.1111/1467-9868.00123. URL https://academic.oup.com/jrsssb/article/60/1/255/7083121.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 13507265. URL http://www.jstor.org/stable/3318418.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022. URL https://arxiv.org/abs/2210.01790.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs, July 2025. URL http://arxiv.org/abs/2407.15549. arXiv:2407.15549 [cs].

Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. High-accuracy sampling from constrained spaces with the Metropolis-adjusted Preconditioned Langevin Algorithm, February 2025. URL http://arxiv.org/abs/2412.18701. arXiv:2412.18701 [stat].

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models, 2024. URL https://arxiv.org/abs/2402.17762.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

L. Tunstall, L. von Werra, and T. Wolf. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, 2022. ISBN 9781098103248. URL https://books.google.ch/books?id=7hhyzgEACAAJ.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. A Statistical Approach to Assessing Neural Network Robustness, February 2019. URL http://arxiv.org/abs/1811.07209. arXiv:1811.07209 [stat].

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483.

Thomas Winninger, Boussad Addad, and Katarzyna Kapusta. Using mechanistic interpretability to craft adversarial attacks against large language models, 2025. URL https://arxiv.org/abs/2503.06269.

Gabriel Wu and Jacob Hilton. Estimating the probabilities of rare outputs in language models, 2025. URL https://arxiv.org/abs/2410.13211.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a. URL https://arxiv.org/abs/2310.01405.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. URL https://arxiv.org/abs/2307.15043.

Code Availability. An anonymized code repository is available at: `https://anonymous.4open.science/r/rare-event-estimation-EDE0/` The repository contains the implementation and example notebook.

Appendix outline. Appendix A develops the SPB loss theory from its Bregman-divergence representation, derives the unshifted power family, and analyzes the resulting false-negative/false-positive asymmetry. Appendix B gives the shifted-power parameterization used in the paper, proves bounded scale distortion, explains how to choose ε, α, and γ, provides a closed-form computation, connects SPB to standard losses, and empirically verifies the intended scaling behavior. Appendix C gives the full GA-AMLS procedure in Algorithm 1. Appendix D reports experimental and implementation details, including hyperparameters, baseline implementation notes, hyperparameter sweeps and ablations, and the activation-whitening diagnostic. Appendix E provides the full set of estimate versus ground-truth scatterplots across model sizes and input distributions. Appendix F breaks down SPB losses by distribution, model size, method, and evaluation profile.

A SPB Loss Theory

A.1 Bregman integral representation

Let $q \in (0,1)$ denote the true probability and $p \in [0,1]$ the predicted probability. A one-dimensional Bregman divergence generated by a twice differentiable convex potential with second derivative $\omega(t) > 0$ can be written as

$$B_\omega(q \mid p) = \int_q^p (t-q)\omega(t)\,dt. \tag{1}$$

This is an oriented integral. Equivalently,

$$B_\omega(q \mid p) = \begin{cases} \displaystyle\int_q^p (t-q)\omega(t)\,dt, & p \geq q, \\ \displaystyle\int_p^q (q-t)\omega(t)\,dt, & p < q. \end{cases}$$

Thus the divergence is nonnegative, and if $\omega(t) > 0$ almost everywhere, then $B_\omega(q \mid p) = 0$ if and only if $p = q$, whenever the integral is finite.

A.2 The unshifted power family

Consider the unshifted power weight

$$\omega_0(t) = t^{-\alpha}, \qquad \alpha > 0. \tag{2}$$

For a multiplicative overestimate $p = mq$, with $m > 1$, the change of variables $t = qu$ gives

$$\begin{aligned} B_0(q \mid mq) &= \int_q^{mq} (t-q)t^{-\alpha}\,dt \\ &= q^{2-\alpha}\int_1^m (u-1)u^{-\alpha}\,du \\ &= q^{2-\alpha} I_{FP}(\alpha, m), \end{aligned} \tag{3}$$

where

$$I_{FP}(\alpha, m) := \int_1^m (u-1)u^{-\alpha}\,du. \tag{4}$$

Similarly, for a multiplicative underestimate $p = q/m$,

$$\begin{aligned} B_0(q \mid q/m) &= \int_{q/m}^q (q-t)t^{-\alpha}\,dt \\ &= q^{2-\alpha}\int_{1/m}^1 (1-u)u^{-\alpha}\,du \\ &= q^{2-\alpha} I_{FN}(\alpha, m), \end{aligned} \tag{5}$$

where

$$I_{FN}(\alpha, m) := \int_{1/m}^{1} (1-u)u^{-\alpha}\, du. \tag{6}$$

Therefore both false-positive and false-negative multiplicative errors factor into a ground-truth scale term $q^{2-\alpha}$ and a relative-error term. The dataset-level weight

$$w(q) = q^{\alpha-2} \tag{7}$$

cancels the $q^{2-\alpha}$ prefactor exactly:

$$w(q)B_0(q \mid mq) = I_{FP}(\alpha, m), \qquad w(q)B_0(q \mid q/m) = I_{FN}(\alpha, m).$$

Thus the weighted unshifted power family is exactly multiplicatively scale-invariant: for fixed relative error m, the weighted loss is independent of the absolute scale q.

More generally, if one wants to add an explicit rarity premium $\gamma \geq 0$, one can use

$$w_\gamma(q) = q^{\alpha-2-\gamma},$$

in which case the weighted unshifted loss scales as $q^{-\gamma}$ times the relative-error term. The formal scale-invariance results below correspond to the scale-neutral case $\gamma = 0$.

A.3 Unshifted FN/FP asymmetry

Define the unshifted FN-to-FP ratio

$$R_\alpha(m) := \frac{I_{FN}(\alpha, m)}{I_{FP}(\alpha, m)}. \tag{8}$$

The following lemma shows that for $\alpha > 3/2$, underestimation is penalized more than overestimation, and the asymmetry increases with the multiplicative error size.

Lemma 1. *For $\alpha > 3/2$, $R_\alpha(m)$ is strictly increasing on $(1, \infty)$. Moreover, $R_\alpha(m) > 1$ for all $m > 1$.*

Proof. Starting from the definition of I_{FN}, apply the substitution $u = 1/v$. Then $du = -v^{-2}\, dv$, and as u goes from $1/m$ to 1, v goes from m to 1. Hence

$$\begin{aligned} I_{FN}(\alpha, m) &= \int_{1/m}^{1} (1-u)u^{-\alpha}\, du \\ &= \int_{m}^{1} \left(1 - \frac{1}{v}\right) v^{\alpha}(-v^{-2})\, dv \\ &= \int_{1}^{m} \left(1 - \frac{1}{v}\right) v^{\alpha-2}\, dv \\ &= \int_{1}^{m} (v-1)v^{\alpha-3}\, dv. \end{aligned}$$

Renaming v as u,

$$I_{FN}(\alpha, m) = \int_{1}^{m} (u-1)u^{\alpha-3}\, du = \int_{1}^{m} (u-1)u^{-\alpha}u^{2\alpha-3}\, du.$$

Let

$$f(u) := (u-1)u^{-\alpha}, \qquad g(u) := u^{2\alpha-3}.$$

Then $f(u) > 0$ for $u > 1$, and because $\alpha > 3/2$, $g(u)$ is strictly increasing on $(1, \infty)$. Moreover,

$$I_{FP}(\alpha, m) = \int_{1}^{m} f(u)\, du, \qquad I_{FN}(\alpha, m) = \int_{1}^{m} f(u)g(u)\, du.$$

Thus

$$R_\alpha(m) = \frac{\int_1^m f(u)g(u)\,du}{\int_1^m f(u)\,du},$$

so $R_\alpha(m)$ is the f-weighted average of the increasing function g over $[1, m]$.

To prove strict monotonicity, define

$$A(m) := \int_1^m f(u)g(u)\,du, \qquad C(m) := \int_1^m f(u)\,du.$$

Then $R_\alpha(m) = A(m)/C(m)$, and

$$A'(m) = f(m)g(m), \qquad C'(m) = f(m).$$

Therefore

$$\begin{aligned} R'_\alpha(m) &= \frac{A'(m)C(m) - A(m)C'(m)}{C(m)^2} \\ &= \frac{f(m)}{C(m)}\left[g(m) - \frac{A(m)}{C(m)}\right] \\ &= \frac{f(m)}{C(m)}\left[g(m) - R_\alpha(m)\right]. \end{aligned}$$

Since g is strictly increasing, its weighted average over $[1, m]$ is strictly smaller than its endpoint value $g(m)$. Hence

$$g(m) - R_\alpha(m) > 0.$$

Also $f(m) > 0$ and $C(m) > 0$ for $m > 1$. Therefore

$$R'_\alpha(m) > 0.$$

So R_α is strictly increasing on $(1, \infty)$.

Finally, because $g(u) = u^{2\alpha-3} > 1$ for all $u > 1$, its weighted average is also greater than 1. Hence $R_\alpha(m) > 1$ for all $m > 1$. □

At the boundary $\alpha = 3/2$, we have $g(u) \equiv 1$, so $R_{3/2}(m) = 1$ for all $m > 1$. Thus $\alpha = 3/2$ is the symmetric point, while $\alpha > 3/2$ makes false negatives more costly than false positives; the mirrored argument, with g strictly decreasing, shows that $\alpha < 3/2$ makes false positives more costly.

B SPB Loss Guarantees and Parameterization

B.1 The shifted power family

The unshifted power weight $t^{-\alpha}$ gives exact scale invariance after dataset weighting, but it is singular at $t = 0$. In particular, for $\alpha \geq 1$, predicting $p = 0$ can produce an infinite penalty. To make the loss finite at zero, define the shifted power weight

$$\omega_\varepsilon(t) := (t + \varepsilon)^{-\alpha}, \qquad \alpha > 0, \quad \varepsilon > 0. \tag{9}$$

The corresponding shifted divergence is

$$B_\varepsilon(q \mid p) = \int_q^p (t - q)(t + \varepsilon)^{-\alpha}\,dt. \tag{10}$$

Because $(t + \varepsilon)^{-\alpha} \leq \varepsilon^{-\alpha} < \infty$ on $[0, 1]$, the shifted divergence is finite even when $p = 0$.

The price of this shift is that exact multiplicative scale invariance is lost: after the change of variables $t = qu$, the ratio

$$x := \frac{\varepsilon}{q}$$

remains inside the integral. The following results show that the desired properties are preserved on a finite evaluation box

$$q \in [q_{\min}, q_{\max}], \qquad m \in (1, m_{\max}],$$

provided ε is sufficiently small relative to $q_{\min}$.

B.2 Bounded scale distortion

Theorem 1 (Bounded Scale Distortion)**.** *Fix $\alpha > 0$, $\eta \in (0,1)$, and $m_{\max} > 1$. Let $w(q) = q^{\alpha-2}$. If*

$$\varepsilon \le \frac{q_{\min}}{m_{\max}} \left((1-\eta)^{-1/\alpha} - 1 \right), \tag{11}$$

then, uniformly over $q \in [q_{\min}, q_{\max}]$ and $m \in (1, m_{\max}]$,

$$0 \le \frac{I_{FP}(\alpha, m) - w(q) B_\varepsilon(q \mid mq)}{I_{FP}(\alpha, m)} \le \eta, \tag{12}$$

$$0 \le \frac{I_{FN}(\alpha, m) - w(q) B_\varepsilon(q \mid q/m)}{I_{FN}(\alpha, m)} \le \eta. \tag{13}$$

Proof. Fix $q \in [q_{\min}, q_{\max}]$ and $m \in (1, m_{\max}]$, and define

$$x := \frac{\varepsilon}{q}.$$

First consider the overestimate $p = mq$. Using $t = qu$,

$$\begin{aligned} B_\varepsilon(q \mid mq) &= \int_q^{mq} (t-q)(t+\varepsilon)^{-\alpha}\, dt \\ &= q^{2-\alpha} \int_1^m (u-1)(u+x)^{-\alpha}\, du \\ &= q^{2-\alpha} I_{FP,\varepsilon}(\alpha, m; x), \end{aligned} \tag{14}$$

where

$$I_{FP,\varepsilon}(\alpha, m; x) := \int_1^m (u-1)(u+x)^{-\alpha}\, du.$$

For $u \in [1, m]$,

$$u \le u + x \le u(1+x).$$

Since $\alpha > 0$, raising to the power $-\alpha$ reverses the inequalities:

$$u^{-\alpha} \ge (u+x)^{-\alpha} \ge u^{-\alpha}(1+x)^{-\alpha}.$$

Multiplying by $(u-1) \ge 0$ and integrating gives

$$(1+x)^{-\alpha} I_{FP}(\alpha, m) \le I_{FP,\varepsilon}(\alpha, m; x) \le I_{FP}(\alpha, m). \tag{15}$$

Therefore

$$0 \le \frac{I_{FP}(\alpha, m) - I_{FP,\varepsilon}(\alpha, m; x)}{I_{FP}(\alpha, m)} \le 1 - (1+x)^{-\alpha}. \tag{16}$$

Now consider the underestimate $p = q/m$. Again using $t = qu$,

$$\begin{aligned} B_\varepsilon(q \mid q/m) &= \int_{q/m}^{q} (q-t)(t+\varepsilon)^{-\alpha}\, dt \\ &= q^{2-\alpha} \int_{1/m}^{1} (1-u)(u+x)^{-\alpha}\, du \\ &= q^{2-\alpha} I_{FN,\varepsilon}(\alpha, m; x), \end{aligned} \tag{17}$$

where

$$I_{FN,\varepsilon}(\alpha, m; x) := \int_{1/m}^{1} (1-u)(u+x)^{-\alpha}\, du.$$

For $u \in [1/m, 1]$, we have $1/u \le m$. Hence

$$u + x = u\left(1 + \frac{x}{u}\right) \le u(1 + mx),$$

and also $u + x \ge u$. Therefore

$$u \le u + x \le u(1 + mx).$$

Raising to the power $-\alpha$ gives

$$u^{-\alpha} \ge (u+x)^{-\alpha} \ge u^{-\alpha}(1+mx)^{-\alpha}.$$

Multiplying by $(1-u) \ge 0$ and integrating gives

$$(1+mx)^{-\alpha} I_{FN}(\alpha, m) \le I_{FN,\varepsilon}(\alpha, m; x) \le I_{FN}(\alpha, m). \tag{18}$$

Therefore

$$0 \le \frac{I_{FN}(\alpha, m) - I_{FN,\varepsilon}(\alpha, m; x)}{I_{FN}(\alpha, m)} \le 1 - (1+mx)^{-\alpha}. \tag{19}$$

Now assume (11). Since $q \ge q_{\min}$ and $m \le m_{\max}$,

$$mx = \frac{m\varepsilon}{q} \le \frac{m_{\max}\varepsilon}{q_{\min}} \le (1-\eta)^{-1/\alpha} - 1.$$

Thus

$$1 + mx \le (1-\eta)^{-1/\alpha}.$$

Equivalently,

$$(1+mx)^{-\alpha} \ge 1 - \eta,$$

so

$$1 - (1+mx)^{-\alpha} \le \eta.$$

This proves the FN distortion bound. Since $x \le mx$, the same condition also implies

$$1 - (1+x)^{-\alpha} \le \eta,$$

and therefore proves the FP distortion bound.

Finally, because $w(q) = q^{\alpha-2}$,

$$w(q) B_\varepsilon(q \mid mq) = I_{FP,\varepsilon}(\alpha, m; x),$$

and

$$w(q) B_\varepsilon(q \mid q/m) = I_{FN,\varepsilon}(\alpha, m; x).$$

Substituting these identities into the relative distortion bounds proves the claim. □

B.3 Asymmetric sensitivity

Theorem 2 (Asymmetric Sensitivity)**.** *Fix $\alpha > 3/2$ and $1 < m_0 \leq m_{\max}$. Let*

$$R_\alpha(m) := \frac{I_{FN}(\alpha, m)}{I_{FP}(\alpha, m)}.$$

If

$$\varepsilon < \frac{q_{\min}}{m_{\max}} \left[R_\alpha(m_0)^{1/\alpha} - 1 \right], \tag{20}$$

then

$$B_\varepsilon(q \mid q/m) > B_\varepsilon(q \mid mq) \tag{21}$$

for all $q \in [q_{\min}, q_{\max}]$ and all $m \in [m_0, m_{\max}]$.

Proof. Fix $q \in [q_{\min}, q_{\max}]$ and $m \in [m_0, m_{\max}]$, and define

$$x := \frac{\varepsilon}{q}.$$

For the overestimate $p = mq$,

$$B_\varepsilon(q \mid mq) = q^{2-\alpha} I_{FP,\varepsilon}(\alpha, m; x),$$

where

$$I_{FP,\varepsilon}(\alpha, m; x) = \int_1^m (u-1)(u+x)^{-\alpha}\, du.$$

Since $x \geq 0$ and $\alpha > 0$,

$$(u+x)^{-\alpha} \leq u^{-\alpha}.$$

Therefore

$$I_{FP,\varepsilon}(\alpha, m; x) \leq I_{FP}(\alpha, m). \tag{22}$$

For the underestimate $p = q/m$,

$$B_\varepsilon(q \mid q/m) = q^{2-\alpha} I_{FN,\varepsilon}(\alpha, m; x),$$

where

$$I_{FN,\varepsilon}(\alpha, m; x) = \int_{1/m}^1 (1-u)(u+x)^{-\alpha}\, du.$$

For $u \in [1/m, 1]$, we have

$$u + x = u\left(1 + \frac{x}{u}\right) \leq u(1 + mx).$$

Hence

$$(u+x)^{-\alpha} \geq u^{-\alpha}(1+mx)^{-\alpha}.$$

Multiplying by $(1-u) \geq 0$ and integrating gives

$$I_{FN,\varepsilon}(\alpha, m; x) \geq (1+mx)^{-\alpha} I_{FN}(\alpha, m). \tag{23}$$

It is therefore enough to ensure

$$(1+mx)^{-\alpha} I_{FN}(\alpha, m) > I_{FP}(\alpha, m).$$

Equivalently, using

$$R_\alpha(m) = \frac{I_{FN}(\alpha, m)}{I_{FP}(\alpha, m)},$$

it is enough to have

$$(1+mx)^\alpha < R_\alpha(m),$$

or

$$1 + mx < R_\alpha(m)^{1/\alpha}.$$

By Lemma 1, $R_\alpha(m)$ is strictly increasing in m for $\alpha > 3/2$. Since $m \geq m_0$,

$$R_\alpha(m) \geq R_\alpha(m_0).$$

Also, since $m \leq m_{\max}$ and $q \geq q_{\min}$,

$$mx = \frac{m\varepsilon}{q} \leq \frac{m_{\max}\varepsilon}{q_{\min}}.$$

The assumed condition (20) implies

$$1 + \frac{m_{\max}\varepsilon}{q_{\min}} < R_\alpha(m_0)^{1/\alpha}.$$

Therefore

$$1 + mx < R_\alpha(m_0)^{1/\alpha} \leq R_\alpha(m)^{1/\alpha}.$$

Hence

$$(1 + mx)^{-\alpha} I_{FN}(\alpha, m) > I_{FP}(\alpha, m).$$

Combining this strict inequality with (22) and (23), we obtain

$$I_{FN,\varepsilon}(\alpha, m; x) > I_{FP,\varepsilon}(\alpha, m; x).$$

Multiplying by the positive factor $q^{2-\alpha}$ gives

$$B_\varepsilon(q \mid q/m) > B_\varepsilon(q \mid mq).$$

This proves the theorem. □

B.4 Simultaneous finite-error guarantee

Corollary 1 (Simultaneous finite-error guarantee)**.** *Fix $\alpha > 3/2$, $\eta \in (0, 1)$, and*

$$1 < m_0 \leq m \leq m_{\max}.$$

Let

$$R_\alpha(m) := \frac{I_{FN}(\alpha, m)}{I_{FP}(\alpha, m)}.$$

If

$$0 \leq \varepsilon < \min\left\{ \frac{q_{\min}}{m_{\max}} \left((1-\eta)^{-1/\alpha} - 1 \right),\ \frac{q_{\min}}{m_{\max}} \left[R_\alpha(m_0)^{1/\alpha} - 1 \right] \right\}, \tag{24}$$

then the shifted weighted loss simultaneously satisfies:

1. *approximate scale invariance within relative tolerance η:*

$$0 \leq \frac{I_{FP}(\alpha, m) - w(q) B_\varepsilon(q \mid mq)}{I_{FP}(\alpha, m)} \leq \eta,$$

   and

$$0 \leq \frac{I_{FN}(\alpha, m) - w(q) B_\varepsilon(q \mid q/m)}{I_{FN}(\alpha, m)} \leq \eta;$$

2. *finite-error asymmetric sensitivity:*

$$B_\varepsilon(q \mid q/m) > B_\varepsilon(q \mid mq).$$

Both guarantees hold uniformly for $q \in [q_{\min}, q_{\max}]$ and $m \in [m_0, m_{\max}]$.

Proof. The first term in the minimum in (24) is exactly the sufficient condition of Theorem 1. The second term is exactly the sufficient condition of Theorem 2. Therefore both conclusions hold simultaneously. □

B.5 Interpretation of the shift bound

Setting $\varepsilon = 0$ gives exact multiplicative scale invariance and, for $\alpha > 3/2$, strict FN-over-FP asymmetry. However, the unshifted loss can be infinite when a model predicts $p = 0$. Introducing $\varepsilon > 0$ removes this singularity, but it also introduces an additive floor into the local weight $(t + \varepsilon)^{-\alpha}$. If ε is too large relative to $q_{\min}$, this floor overwhelms the relevant probability scale and destroys both approximate scale invariance and asymmetric sensitivity.

The bounds above quantify the permissible size of ε. They are sufficient rather than necessary, and are intentionally conservative. A simple practical choice is to set

$$\varepsilon = c\frac{q_{\min}}{m_{\max}}, \qquad c \ll 1,$$

and then verify that this value lies below the theorem-specific upper bounds.

B.6 Parameter Guidance

Table 2 summarizes suggested (α, γ) settings for the evaluation use cases referenced in Section 4.1. The parameter γ controls any additional dataset-level rarity premium, while α controls the asymmetry between reciprocal overestimation and underestimation. We now describe a more fine-grained way to choose α by specifying a desired worst-case FN/FP ratio.

B.6.1 Choosing α via a worst-case FN/FP ratio

A natural way to calibrate the asymmetry parameter α is to specify a desired lower bound $r_\star > 1$ on the cost of reciprocal underestimation relative to reciprocal overestimation. In the unshifted geometry,

$$\frac{B_0(q \mid q/m)}{B_0(q \mid mq)} = \frac{I_{FN}(\alpha, m)}{I_{FP}(\alpha, m)} = R_\alpha(m),$$

so this ratio is independent of the base probability q. The same is true after multiplying both losses by any dataset-level weight depending only on q, since the weight cancels in the ratio.

Such a guarantee cannot hold uniformly as $m \downarrow 1$. Indeed, for every fixed α,

$$\lim_{m \downarrow 1} R_\alpha(m) = 1.$$

Thus no finite α can guarantee $R_\alpha(m) \geq r_\star > 1$ for arbitrarily small multiplicative errors. A meaningful worst-case calibration must therefore specify a minimum multiplicative error $m_0 > 1$.

Fix $m_0 > 1$ and $m_{\max} \geq m_0$. We require

$$R_\alpha(m) \geq r_\star \qquad \forall m \in [m_0, m_{\max}].$$

For $r_\star > 1$, any feasible choice must have $\alpha > 3/2$. By Lemma 1, $R_\alpha(m)$ is then increasing in m, so the worst case occurs at $m = m_0$. Therefore, in the unshifted geometry, the uniform calibration condition over $m \in [m_0, m_{\max}]$ is equivalent to

$$R_\alpha(m_0) \geq r_\star.$$

Using the identity $I_{FN}(\alpha, m) = I_{FP}(3 - \alpha, m)$, we can write

$$R_\alpha(m) = \frac{I_{FP}(3 - \alpha, m)}{I_{FP}(\alpha, m)}.$$

Here

$$I_{FP}(\beta, m) = \int_1^m (u - 1)u^{-\beta}\, du = \begin{cases} \dfrac{m^{2-\beta} - 1}{2 - \beta} - \dfrac{m^{1-\beta} - 1}{1 - \beta}, & \beta \notin \{1, 2\}, \\ m - 1 - \log m, & \beta = 1, \\ \log m + \frac{1}{m} - 1, & \beta = 2. \end{cases}$$

For fixed $m_0 > 1$, the map $\alpha \mapsto R_\alpha(m_0)$ is strictly increasing. Indeed, defining

$$F(\beta, m) = I_{FP}(\beta, m),$$

we have

$$\partial_\beta F(\beta, m) = -\int_1^m (u-1)u^{-\beta} \log u \, du < 0.$$

Thus $F(\beta, m)$ is strictly decreasing in β, and

$$R_\alpha(m_0) = \frac{F(3-\alpha, m_0)}{F(\alpha, m_0)}$$

is strictly increasing in α. Moreover,

$$R_{3/2}(m_0) = 1, \qquad \lim_{\alpha\to\infty} R_\alpha(m_0) = \infty.$$

Hence for every $r_\star > 1$ there is a unique threshold

$$\alpha_\star = \alpha_\star(m_0, r_\star) > 3/2$$

satisfying

$$R_{\alpha_\star}(m_0) = r_\star.$$

Choosing $\alpha \geq \alpha_\star$ guarantees

$$\min_{m\in[m_0, m_{\max}]} R_\alpha(m) \geq r_\star$$

for the unshifted loss.

Computing $\alpha_\star$. The threshold $\alpha_\star$ generally has no closed-form inverse, but it can be computed by bisection because $\alpha \mapsto R_\alpha(m_0)$ is continuous and strictly increasing. A practical procedure is:

1. Choose the minimum relevant multiplicative error $m_0 > 1$, the desired worst-case ratio $r_\star > 1$, and a numerical tolerance $\tau > 0$.
2. Set $\alpha_{\text{lo}} = \frac{3}{2}$. Then $R_{\alpha_{\text{lo}}}(m_0) = 1 < r_\star$.
3. Choose an initial upper bracket $\alpha_{\text{hi}} > 3/2$. Increase it, for example by repeatedly doubling,$\alpha_{\text{hi}} \leftarrow 2\alpha_{\text{hi}}$, until $R_{\alpha_{\text{hi}}}(m_0) \geq r_\star$.
4. While $\alpha_{\text{hi}} - \alpha_{\text{lo}} > \tau$, set $\alpha_{\text{mid}} = \frac{\alpha_{\text{lo}}+\alpha_{\text{hi}}}{2}$. If $R_{\alpha_{\text{mid}}}(m_0) < r_\star$, set $\alpha_{\text{lo}} \leftarrow \alpha_{\text{mid}}$. Otherwise set $\alpha_{\text{hi}} \leftarrow \alpha_{\text{mid}}$.
5. Return α_{hi}. This gives a conservative numerical choice satisfying $R_{\alpha_{\text{hi}}}(m_0) \geq r_\star$.

Effect of the shift. For the shifted loss, the exact unshifted ratio no longer holds. However, the proof of Theorem 2 gives the conservative lower bound

$$\frac{B_\varepsilon(q \mid q/m)}{B_\varepsilon(q \mid mq)} \geq (1 + m\varepsilon/q)^{-\alpha} R_\alpha(m).$$

Therefore, uniformly over $q \in [q_{\min}, q_{\max}]$ and $m \in [m_0, m_{\max}]$,

$$\frac{B_\varepsilon(q \mid q/m)}{B_\varepsilon(q \mid mq)} \geq \left(1 + \frac{m_{\max}\varepsilon}{q_{\min}}\right)^{-\alpha} R_\alpha(m_0).$$

Thus a sufficient condition for the shifted loss to achieve worst-case ratio at least $r_\star$ is

$$\left(1 + \frac{m_{\max}\varepsilon}{q_{\min}}\right)^{-\alpha} R_\alpha(m_0) \geq r_\star.$$

Equivalently, after choosing α, it is sufficient to choose

$$\varepsilon \le \frac{q_{\min}}{m_{\max}} \left[\left(\frac{R_\alpha(m_0)}{r_\star} \right)^{1/\alpha} - 1 \right].$$

If a strict ratio $> r_\star$ is desired, use a strict inequality. This bound is meaningful for positive ε only when

$$R_\alpha(m_0) > r_\star.$$

Consequently, if one intends to use a positive shift $\varepsilon > 0$, one should choose α with some margin above the unshifted threshold $\alpha_\star$, rather than choosing exactly $\alpha = \alpha_\star$. Theorem 2 corresponds to the special case $r_\star = 1$.

Use case	**What you want the metric to emphasize**	$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$
Unbiased method benchmarking	Scale-neutral comparison across a wide ground truth range where FN and FP treated approximately symmetrically.	1.5	0
Product safety oversight	Mild FN aversion (similar to Itakura-Saito Loss) but avoid excessive false alarms; performance should still matter across the full ground truth range.	1.7–2.2	0
Deep-tail stress testing (tail-focused but not fully worst-case)	Put substantially more weight on the smallest q examples (rarest events), while still comparing methods on estimation quality (not just conservatism).	1.5–2.5	0.5–1
Catastrophic-risk audit	Extreme-tail priority and willing to tolerate many false alarms if it reduces underestimation in the deepest tail.	3–6	1–2

Table 2: Suggested shifted-power metric parameters by evaluation use case. Here α controls FN-vs-FP asymmetry (higher α penalizes underestimation more), and γ controls dataset-level rarity premium (higher γ upweights smaller ground-truth probabilities q).

B.7 Computation: closed form

Closed-form $B_\varepsilon(q \mid p)$. For $\alpha \neq 1, 2$, the shifted power family admits the closed form

$$B_\varepsilon(q \mid p) = \frac{(p+\varepsilon)^{2-\alpha} - (q+\varepsilon)^{2-\alpha}}{2-\alpha} - \frac{(\varepsilon+q)\left[(p+\varepsilon)^{1-\alpha} - (q+\varepsilon)^{1-\alpha}\right]}{1-\alpha}. \tag{25}$$

For $\alpha = 1$:

$$B_\varepsilon(q \mid p) = (p-q) - (\varepsilon+q)\log\left(\frac{p+\varepsilon}{q+\varepsilon}\right). \tag{26}$$

For $\alpha = 2$:

$$B_\varepsilon(q \mid p) = \log\left(\frac{p+\varepsilon}{q+\varepsilon}\right) + (\varepsilon+q)\left(\frac{1}{p+\varepsilon} - \frac{1}{q+\varepsilon}\right). \tag{27}$$

B.8 Connections to standard losses

For the unshifted family $\omega_0(t) = t^{-\alpha}$, the divergence is

$$B_0(q \mid p) = \int_q^p (t-q)t^{-\alpha}\, dt.$$

For $\alpha = 0$,

$$B_0(q \mid p) = \int_q^p (t-q)\, dt = \frac{(p-q)^2}{2},$$

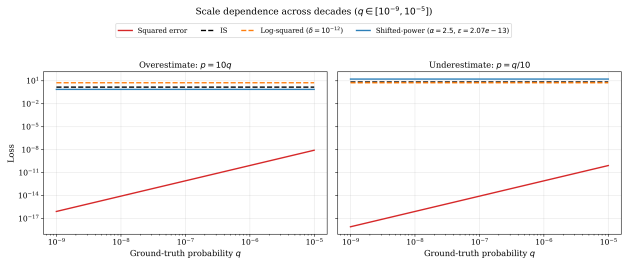


Figure 7: Scale dependence of each metric for a fixed relative error (overestimate $p = 10q$, left; underestimate $p = q/10$, right) across ground-truth probabilities $q \in [10^{-9}, 10^{-5}]$. Squared error varies by orders of magnitude with q, while IS loss, log-space squared error, and the weighted shifted-power loss are flat (scale invariant).

so the family recovers squared error up to a constant factor 1/2.

For $\alpha = 2$,

$$B_0(q \mid p) = \int_q^p \left(\frac{1}{t} - \frac{q}{t^2}\right) dt = \log\frac{p}{q} + \frac{q}{p} - 1 = \frac{q}{p} - \log\frac{q}{p} - 1,$$

which is exactly the Itakura–Saito divergence $D_{\mathrm{IS}}(q, p)$.

B.9 Empirical verification of SPB scaling behavior

Figure 7 contrasts the scale behavior of the metrics for a fixed relative error across the ground-truth range.

C Pseudocode

Algorithm 1 gives the full GA-AMLS procedure described in Section 3.

D Experimental and Implementation Details

D.1 Hyperparameters

MAPLA tuning candidate. In the tuning sweep we also considered MAPLA, the Metropolis-adjusted Preconditioned Langevin Algorithm (Srinivasan et al., 2025). MAPLA generalizes MALA by replacing the isotropic proposal covariance with a position-dependent metric. For a target density π, its proposal has the form

$$u' \sim \mathcal{N}\big(u + hG(u)^{-1}\nabla_u \log \pi(u), 2hG(u)^{-1}\big)\,,$$

followed by the usual Metropolis–Hastings correction. In our AMLS setting this kernel is applied to the constrained target

$$\pi_k(u) \propto \pi(u)\mathbf{1}\{s(u) \geq L_k\}.$$

We used MAPLA only as a hyperparameter-tuning candidate: although it can reduce rejections near level-set boundaries by adapting proposal geometry, its extra metric computations did not yield a sufficient improvement over MALA under a matched compute budget. Therefore all main GA-AMLS results use MALA.

Tuning protocol. Like Wu & Hilton (2025), to prevent overfitting our method was only run on the first four distributions during development, and architectural choices were finalized before testing on the last four distributions. We chose the MH kernel and the prior as those which minimized squared error loss and IS loss on 100 randomly chosen tokens with ground-truth probabilities in the range $[10^{-5}, 10^{-3}]$; probabilities in this range are directly estimable by naive Monte Carlo with a modest budget, so tuning does not require privileged knowledge of the extreme tail. The selected prior and kernel were consistent across metrics, distributions

Algorithm 1 GA-AMLS with MALA moves and a Student-t prior

Require: Input distribution $\mathcal{D}$, model M, target token t, population size N, quantile level ρ, target score τ, MALA step size h, MCMC steps T

Ensure: Probability estimate $\widehat{P}$

Draw calibration inputs from D and compute raw activations $\{a_j\}_{j=1}^{n_{\text{cal}}}$ at the chosen site.

Estimate μ, Σ, choose A such that $AA^\top = \Sigma$, and transform activations by

$$u_j \leftarrow (a_j - \mu)A^{-\top}.$$

Fit the prior in whitened coordinates as an independent Student-t distribution

$$\pi(u) = \prod_{d=1}^{D} t_{\nu=5}(u_d; 0, 1).$$

Define the score $s(u)$ by inserting the unwhitened activation $a(u) = uA^\top + \mu$ into the model.

Initialize particles $\{u_i\}_{i=1}^{N}$ by sampling from the whitened calibration activations.

$k \leftarrow 0$, $L_0 \leftarrow -\infty$, $\widehat{P} \leftarrow 1$

while $L_k < \tau$ **do**

  Compute scores $\{s(u_i)\}_{i=1}^{N}$.

  Set the next level

$$L_{k+1} \leftarrow \min\left\{\tau, \text{Quantile}_\rho\left(\{s(u_i)\}_{i=1}^{N}\right)\right\}.$$

  Estimate the conditional survival probability

$$\widehat{p}_k \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{s(u_i) \geq L_{k+1}\}, \qquad \widehat{P} \leftarrow \widehat{P}\widehat{p}_k.$$

  Keep the survivors $\mathcal{S} \leftarrow \{u_i : s(u_i) \geq L_{k+1}\}$ and resample with replacement from $\mathcal{S}$ until the population again has size N.

  for $r = 1, \ldots, T$ **do**

    for each particle u_i **do**

      Compute the prior score gradient

$$g_i \leftarrow \nabla_u \log \pi(u_i).$$

      Propose a MALA move

$$u_i' \leftarrow u_i + hg_i + \sqrt{2h}\,\xi_i, \qquad \xi_i \sim \mathcal{N}(0, I).$$

      if $s(u_i') < L_{k+1}$ **then**

        Reject the proposal.

      else

        Compute $g_i' \leftarrow \nabla_u \log \pi(u_i')$ and proposal densities

$$q(u_i' \mid u_i) = \mathcal{N}(u_i';\, u_i + hg_i,\, 2hI), \qquad q(u_i \mid u_i') = \mathcal{N}(u_i;\, u_i' + hg_i',\, 2hI).$$

        Accept u_i' with probability

$$a_i = \min\left\{1,\, \frac{\pi(u_i')q(u_i \mid u_i')}{\pi(u_i)q(u_i' \mid u_i)}\right\}.$$

      end if

    end for

    Adapt h during burn-in toward the MALA target acceptance rate.

  end for

  $k \leftarrow k + 1$

end while

return $\widehat{P}$

$\mathcal{D}$, and model sizes. Both MALA and MAPLA outperformed the random-walk kernel (RWMH), and the Student-t prior outperformed a Gaussian prior; since MAPLA's higher computational cost did not yield a commensurate performance gain over MALA, we chose the MALA kernel. To ensure a fair comparison across MH kernels, we equalized the total computational budget per rejuvenation step: since MALA and MAPLA require gradient computations, we allocate 1500 MH steps to RWMH and 750 steps to the gradient-based kernels.

GA-AMLS configuration All GA-AMLS estimates are from a single run per (token, distribution, model). We use a population of $N = 2000$ particles, threshold quantile $\rho = 0.6$, $T = 900$ MALA steps per level with a burn-in of $T_{\text{burn}} = 126$ steps and initial step size $h_0 = 10^{-3}$, and a diagonal Student-t prior with $\nu = 5$ degrees of freedom fit on $n_{\text{cal}} = 2^{16} = 65{,}536$ calibration activations drawn from $\mathcal{D}$. The calibration activations are generated once per (model,D) pair and reused across target tokens.

Baselines. For the importance-sampling baselines ITGIS and MHIS we used the temperatures tuned by Wu & Hilton (2025) .We identified and fixed a numerical-tolerance bug in the public QLD implementation: in the randomized projection routine for finding the shortest accepting vector, the tolerance was meant to ignore only slightly negative constraint violations $Px + b \geq -\texttt{tol}$, but the original code used $Px + b < \texttt{tol}$, incorrectly marking feasible near-boundary constraints as violated.; all QLD results in this paper use the corrected implementation, which improves QLD's performance relative to the numbers reported by Wu & Hilton (2025).

Ablation studies. Figures 8–11 report the per-distribution SPB loss for every kernel–prior–hyperparameter configuration explored during tuning, under each of the four evaluation profiles of Table 2. Two patterns hold consistently across distributions and model sizes: (i) the MALA kernel outperforms RWMH despite using half as many MH proposal steps, and (ii) the Student-t prior outperforms a Gaussian prior.

D.2 Activation Whitening Diagnostic

Figure 12 plots the held-out marginal variance of each pre-LayerNorm, pre-unembedding activation dimension, sorted by variance, before and after whitening; the whitening transform used by GA-AMLS substantially flattens this marginal variance profile.

E Scatterplots

Figures 13–15 show estimates against ground truth for all four methods on all 8 input distributions, for each model size, complementing the two examples in Figure 4.

F All Methods Loss by Distributions

Figure 16 breaks the aggregate SPB results down by input distribution, Figure 17 reports the aggregate SPB loss across model sizes for all four (α, γ) evaluation profiles, and Tables 3 and 4 report per-distribution SPB loss.

Table 3: SPB loss comparison across model sizes ($\alpha = 1.5$, $\gamma = 0$, $\varepsilon = \varepsilon_{\text{corr}}$). Lower is better.

(a) 1-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	117.64	538.82	53.26	**35.44**
camel	660.47	148.67	125.83	**14.62**
colon	127.09	278.54	58.18	**6.24**
if	362.68	211.57	46.96	**15.16**
caps	1489.97	421.47	83.47	**17.05**
english	190.48	58.75	27.49	**16.16**
spanish	480.38	106.99	78.54	**13.67**
icl	218.55	37.67	64.68	**20.05**
Avg.	455.91	225.31	67.30	**17.30**

(b) 2-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	136.82	1141.36	**66.22**	272.59
camel	947.94	725.50	93.04	**27.75**
colon	361.36	555.99	48.02	**12.63**
if	270.16	146.68	48.44	**7.31**
caps	463.91	392.87	56.80	**20.55**
english	708.88	295.81	**17.51**	21.46
spanish	637.53	259.79	61.24	**15.04**
icl	97.04	**26.26**	38.06	28.94
Avg.	452.96	443.03	53.67	**50.78**

(c) 4-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	308.79	1520.17	**45.92**	133.83
camel	1167.25	561.11	90.86	**12.81**
colon	151.61	767.77	34.08	**11.06**
if	119.23	615.02	34.02	**6.41**
caps	205.58	609.63	56.90	**17.07**
english	306.29	327.72	**19.07**	22.30
spanish	743.64	312.04	66.29	**7.39**
icl	38.85	940.25	28.98	**10.07**
Avg.	380.15	706.71	47.01	**27.62**

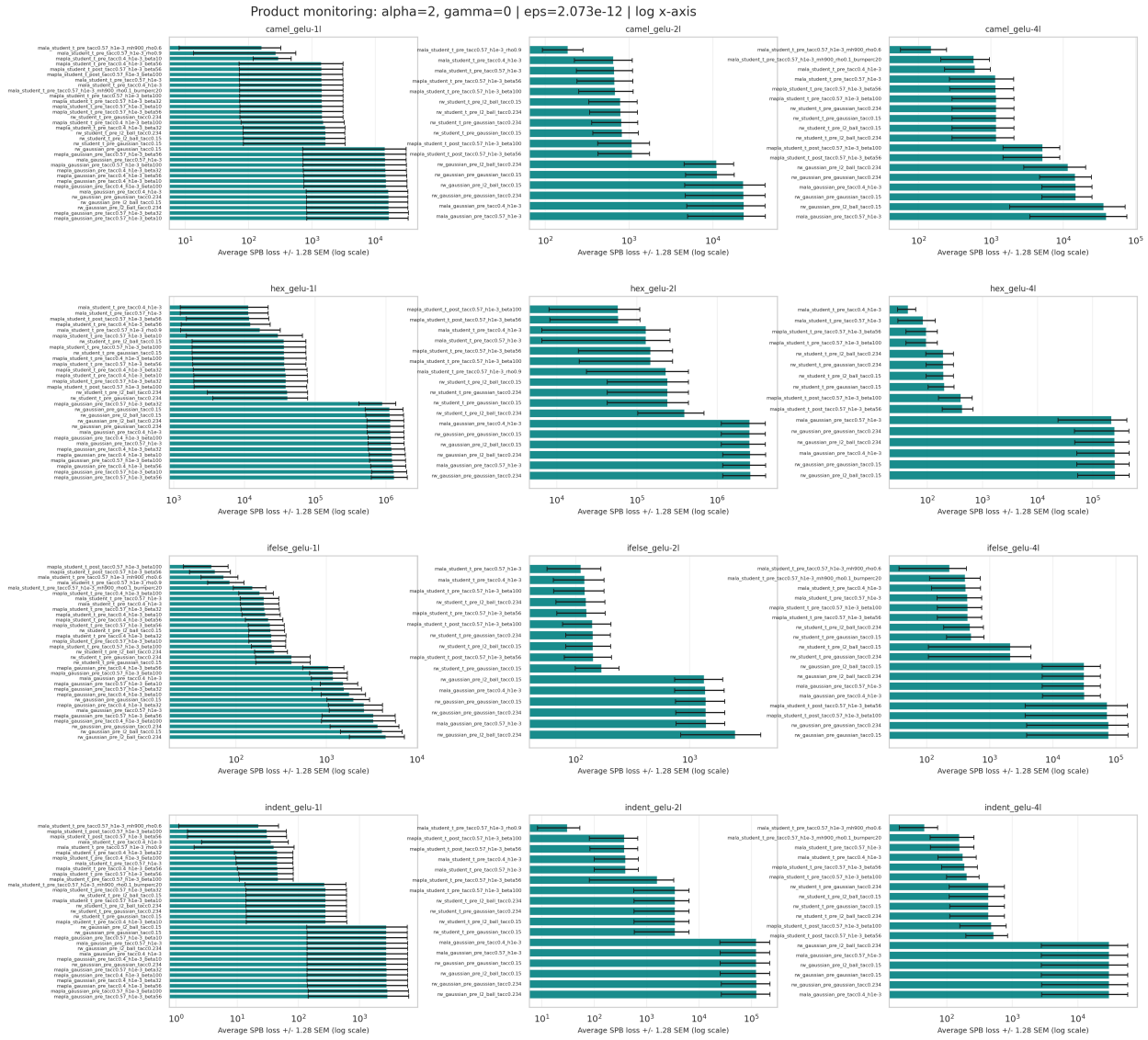


Figure 8: Hyperparameter sweep under the product-monitoring evaluation profile ($\alpha = 2$, $\gamma = 0$, $\varepsilon = 2.073 \times 10^{-12}$). Bars show average SPB loss (± 1.28 SEM, log scale) for each kernel–prior–hyperparameter configuration, per input distribution (rows) and model size (columns). Configuration names encode kernel, prior, activation site, and tuning parameters.

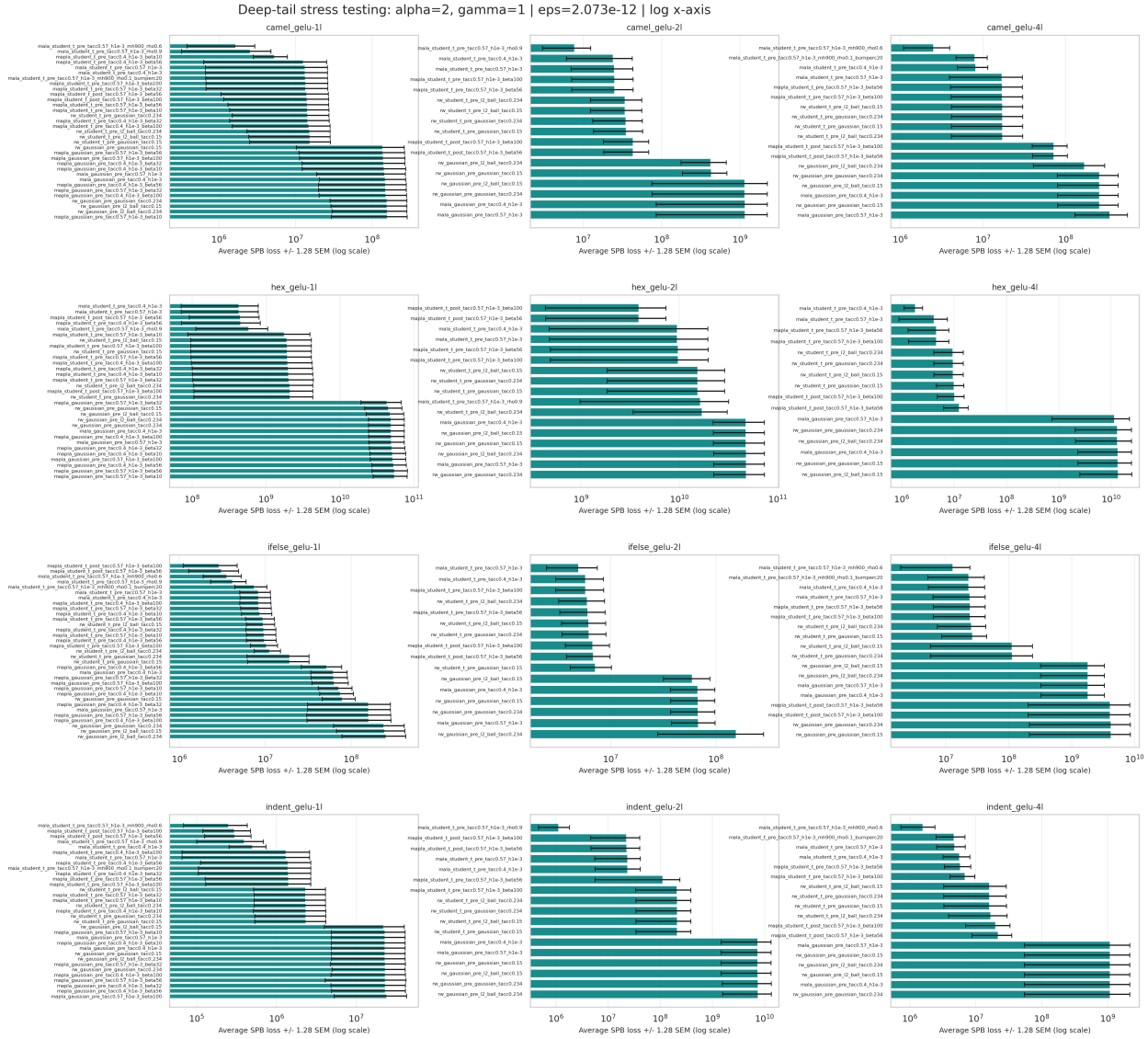


Figure 9: Hyperparameter sweep under the deep-tail stress-testing profile ($\alpha = 2$, $\gamma = 1$); format as in Figure 8.

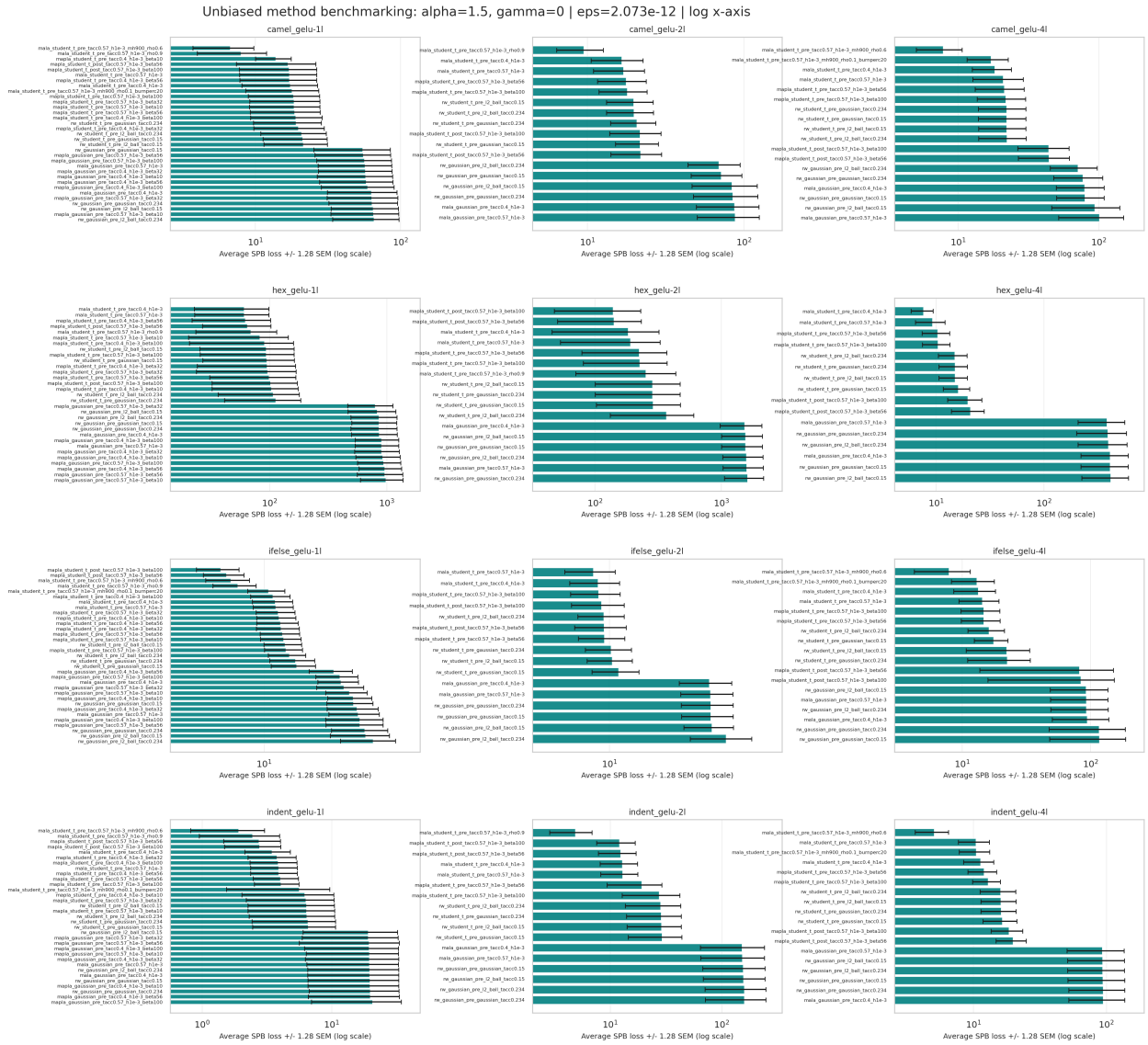


Figure 10: Hyperparameter sweep under the unbiased method benchmarking profile ($\alpha = 1.5$, $\gamma = 0$); format as in Figure 8.

Figure 11: Hyperparameter sweep under the catastrophic-risk audit profile ($\alpha = 4$, $\gamma = 2$); format as in Figure 8.

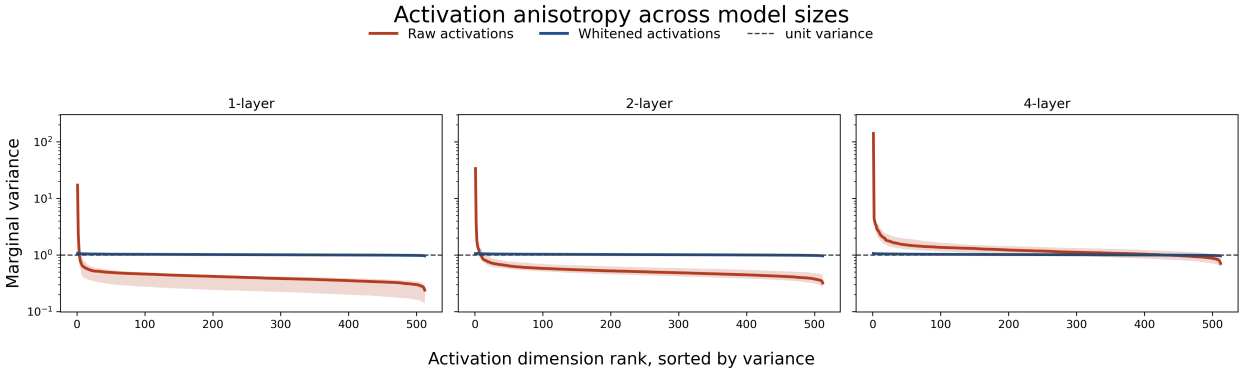


Figure 12: **Marginal activation variances before and after whitening.** Let $a_i \in \mathbb{R}^d$ denote held-out pre-LayerNorm, pre-unembedding activations. For each model size and input distribution, we compute the empirical marginal variance $\mathrm{Var}_i(a_{ij})$ of every activation dimension j, and repeat the calculation after applying the whitening transform used by GA-AMLS. For visualization, dimensions are sorted from largest to smallest marginal variance within each distribution; the x-axis is therefore the sorted dimension rank. Each panel reports the median sorted curve across the eight input distributions, with shaded bands showing the interquartile range. Raw activations exhibit strong marginal anisotropy, while whitening produces a much flatter variance profile, making isotropic MALA proposals in whitened coordinates more reasonable.

Table 4: SPB loss comparison across model sizes ($\alpha = 2$, $\gamma = 0$, $\varepsilon = \varepsilon_{\text{corr}}$). Lower is better.

(a) 1-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	3.62e+04	2.70e+05	**26.99**	2053.05
camel	4.34e+05	3.84e+04	**6.15**	318.50
colon	1.43e+04	6.20e+04	**5.32**	11.35
if	1.02e+05	7.99e+04	**4.72**	935.19
caps	1.62e+06	1.90e+05	**5.34**	15.28
english	5.57e+04	1.63e+04	**3.79**	1183.62
spanish	1.79e+05	9693.76	**5.55**	289.98
icl	5.00e+04	3564.92	**5.13**	556.86
Avg.	3.12e+05	8.38e+04	**7.87**	670.48

(b) 2-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	2.96e+04	1.02e+06	**1.46e+04**	1.52e+05
camel	9.71e+05	7.79e+05	**644.61**	5066.22
colon	1.35e+05	2.76e+05	**4.88**	62.13
if	7.62e+04	4.30e+04	**4.92**	27.54
caps	2.89e+05	3.60e+05	**42.00**	1379.45
english	4.02e+05	1.12e+05	**3.12**	2853.15
spanish	3.06e+05	1.14e+05	**4.99**	816.30
icl	7.31e+04	1545.28	**798.60**	1858.51
Avg.	2.85e+05	3.38e+05	**2017.40**	2.05e+04

(c) 4-layer model

Dist.	MHIS	ITGIS	QLD	GA-AMLS
hex	1.64e+05	1.38e+06	**6749.56**	2.81e+04
camel	1.30e+06	5.27e+05	**4.90**	287.16
colon	4.77e+04	5.83e+05	**4.14**	11.17
if	2.03e+04	3.53e+05	**4.16**	97.97
caps	5.92e+04	5.17e+05	**4.42**	41.13
english	7.00e+04	8.72e+04	**68.21**	7759.93
spanish	5.58e+05	1.18e+05	**5.17**	33.23
icl	8595.27	7.34e+05	**12.28**	19.18
Avg.	2.79e+05	5.38e+05	**856.60**	4547.43

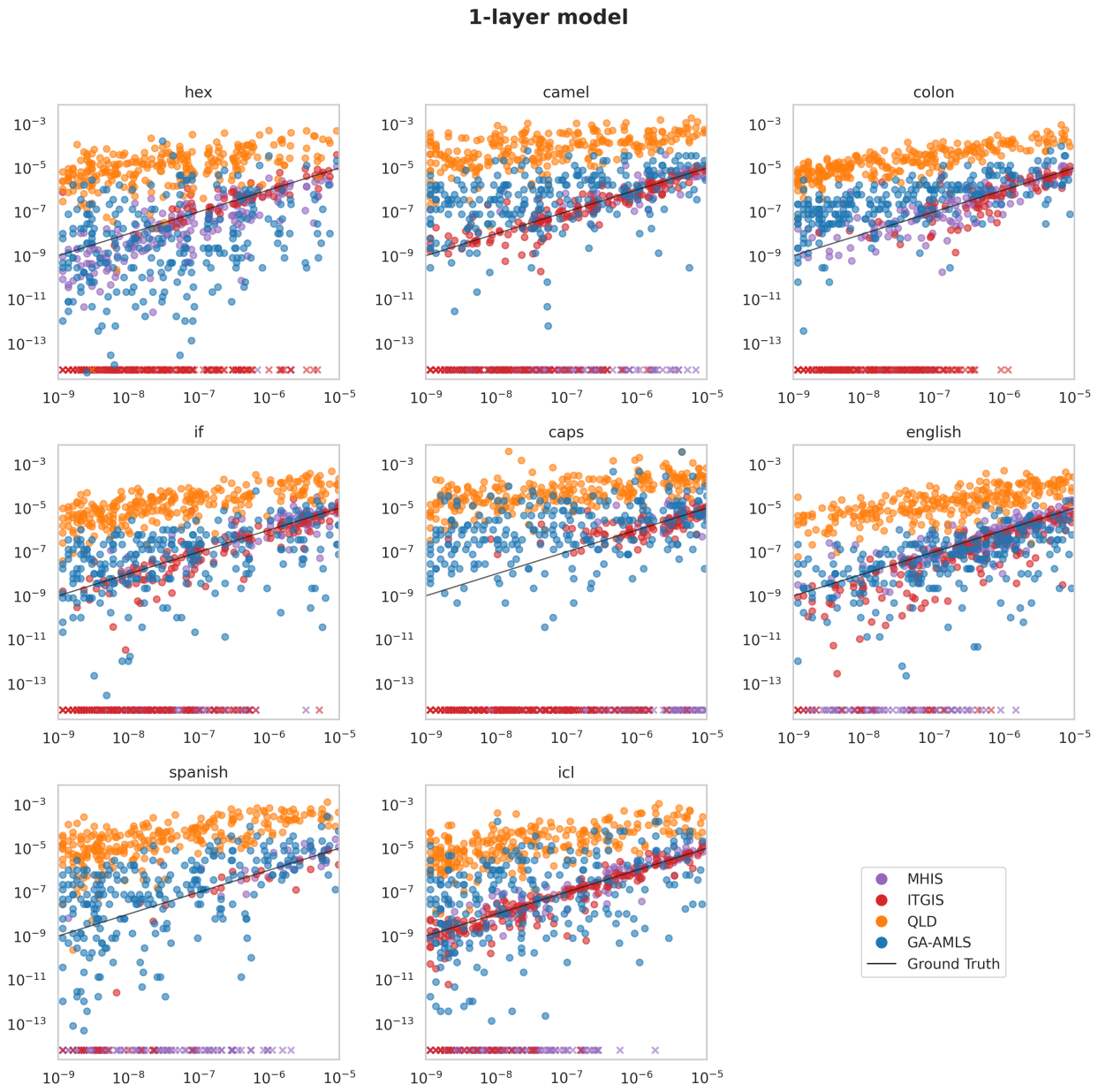


Figure 13: Estimates vs. ground truth for all methods across all 8 input distributions (1-layer model). MHIS/ITGIS zero estimates are plotted as crosses at the bottom of each panel for visibility.

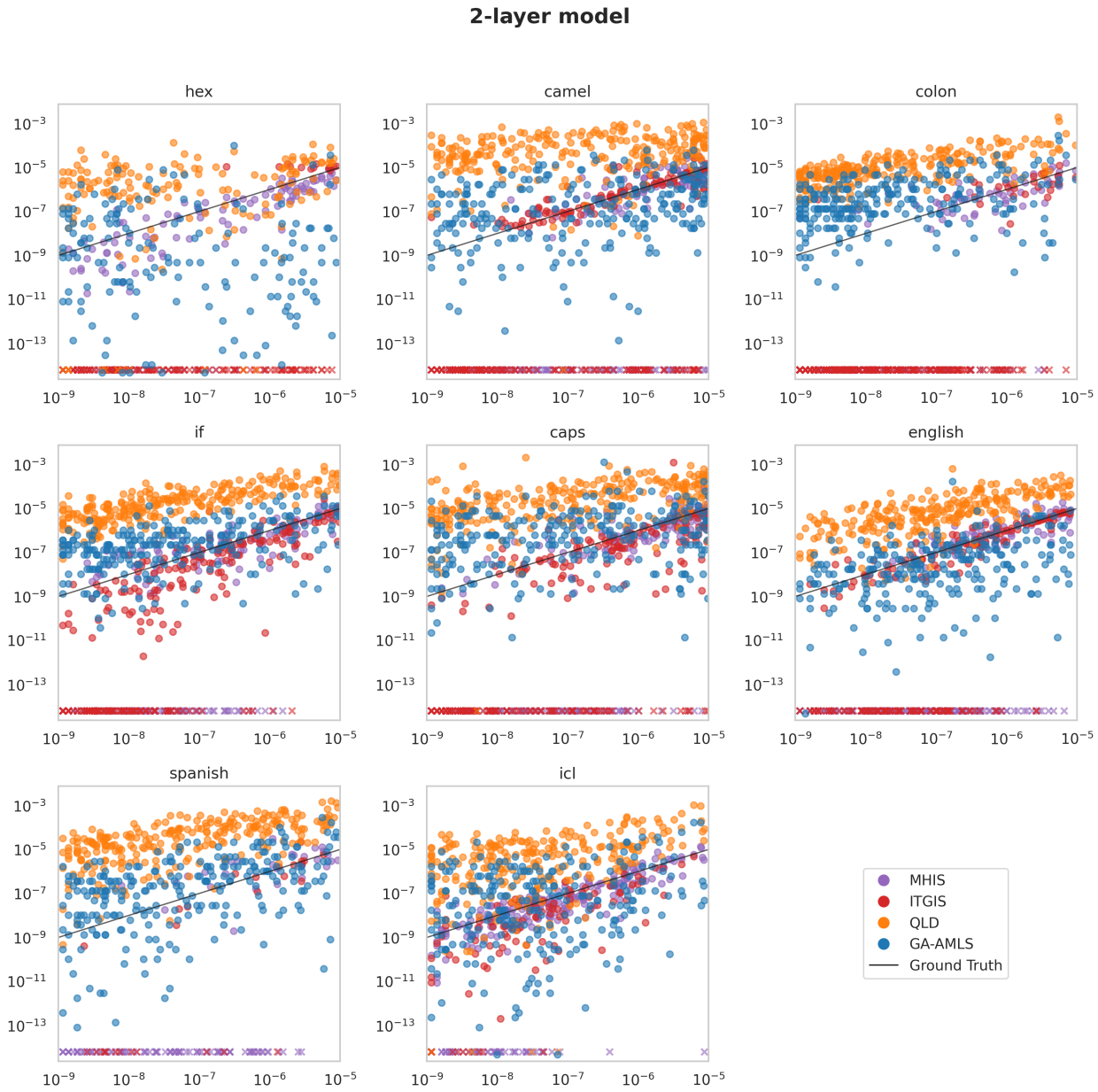


Figure 14: As Figure 13, for the 2-layer model.

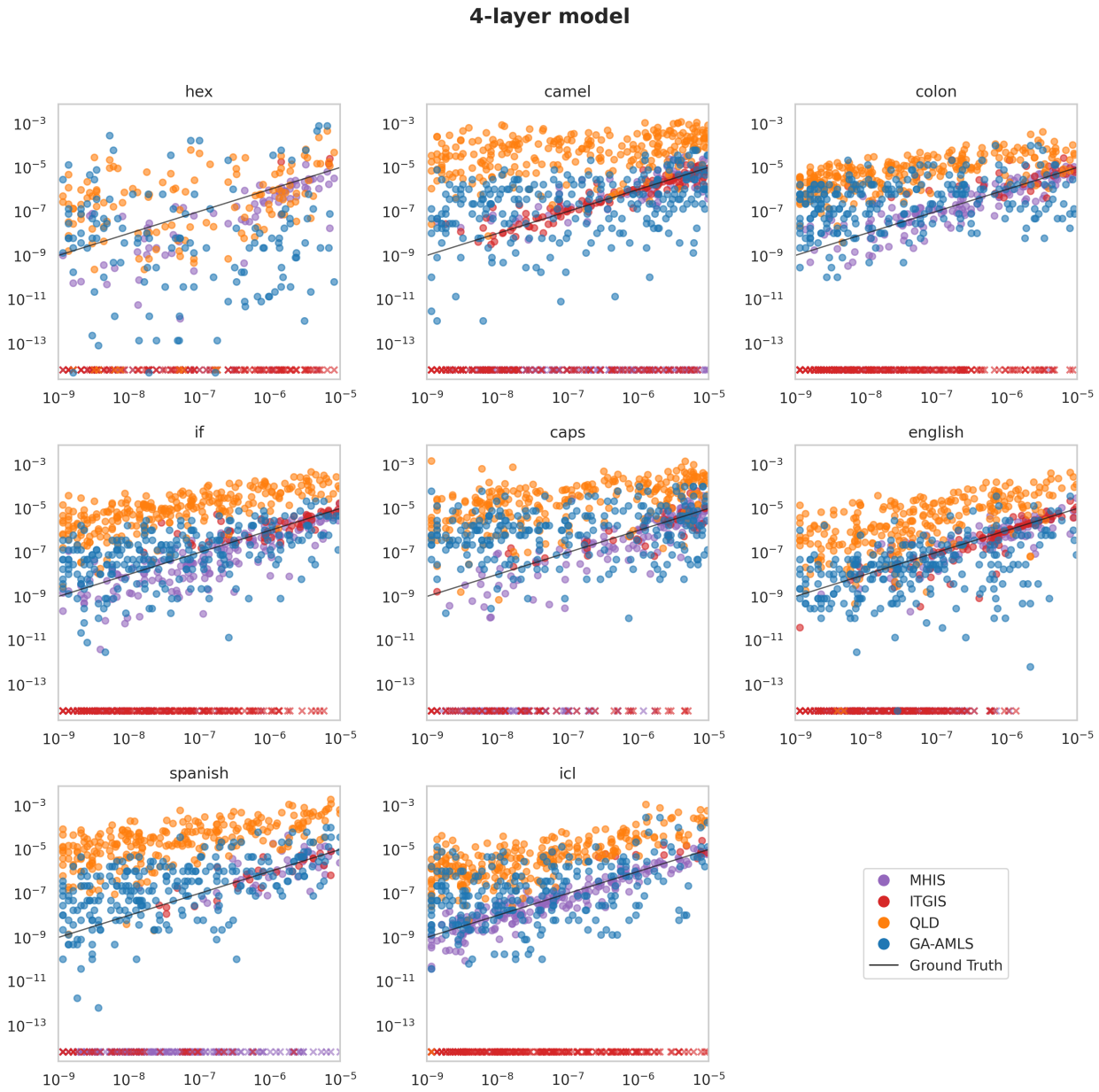


Figure 15: As Figure 13, for the 4-layer model.

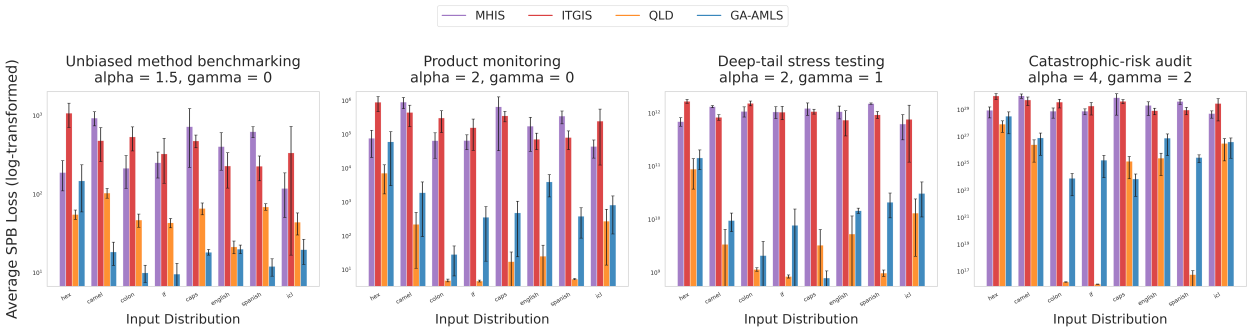


Figure 16: **SPB loss across individual distributions and parameter configurations**. GA-AMLS (blue) consistently outperforms baseline methods across 7 of the 8 text distributions under a symmetric penalty ($\alpha = 1.5$), demonstrating that aggregate improvements are not skewed by outliers. As the penalty for underestimation increases ($\alpha > 1.5$), QLD (orange) gains an advantage due to its systematic overestimation bias. Lower is better; error bars denote standard error across the three model sizes.

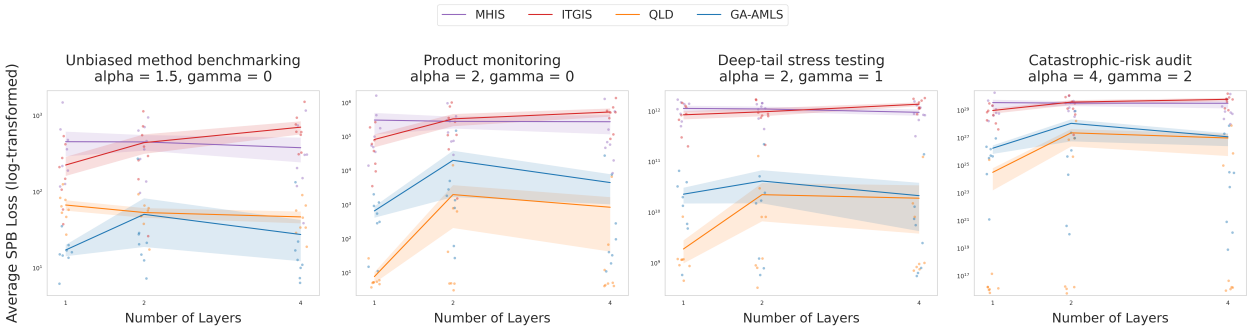


Figure 17: **The SPB loss of all methods across different model sizes and parameter combinations.** The solid lines indicate the loss of each method averaged over all 8 distributions, with bands showing standard error. The colored points indicate the loss on individual distributions, with horizontal jitter added for visibility. Lower is better. We see that GA-AMLS (blue) outperforms when the penalty for FN is approximately the same as FP, and when the weight on the rarest events increases.