MINISTRY OF EDUCATION AND TRAINING

FPT UNIVERSITY

DEPARTMENT OF ITS

Gating Mechanisms in Ensemble Models for Robust Rejection Learning on Long-Tail Data

Nguyen Hong Hai

Duong Xuan Bach

Le Ngoc Mai

Phan Hoai Nam

Supervisor: Dr. Bui Van Hieu

Bachelor of Artificial Intelligence
Hoa Lac campus - FPT University
2025

Abstract

We propose *AR-GSE*, a mixture-of-experts selective classifier tailored to long-tail distributions, which dynamically fuses specialized experts (CE, Logit-Adjust, Balanced Softmax) via a gating network into a calibrated mixture posterior, then applies a group-aware rejection rule with learned thresholds per group. To ensure robustness under severe class imbalance, we design a hybrid optimization scheme combining *pinball quantile loss* for threshold learning, *fixed-point updates* for group weights α, and an *exponentiated gradient outer loop* augmented by a "beta-floor" constraint to avoid coverage collapse in tail groups. On CIFAR-100-LT (Imbalance Factor = 100), AR-GSE achieves approximately 13% reduction in **AURC (balanced error)** and 18% reduction in **AURC (worst-group error)** over the strongest single-expert baseline, while maintaining per-group coverage error within ±2% of target. The pipeline cleanly decouples *representation learning* (expert training) from *selective decision making*, enabling flexible and effective fairness–performance trade-offs in selective classification on long-tail data.

Keywords: Learning to Reject (L2R), Long-Tail Learning, Mixture-of-Experts (MoE)

Acknowledgements

We sincerely thank our team members for their unwavering dedication and for sharing their valuable knowledge and expertise throughout this project. Their contributions were essential to helping us achieve our goals.

We would also like to express our deepest gratitude to our supervisor, Mr. Bui Van Hieu, for his exceptional guidance, motivation, and support during the course of this project. His leadership, vision, and expertise played a vital role in the successful completion of our work.

Once again, we extend our heartfelt appreciation to our teammates and supervisor for their invaluable contributions and unwavering support in bringing this project to fruition.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Problem and Motivation

Selective classification (a.k.a. learning to reject, L2R) equips a classifier with an abstention option to trade coverage for risk, tracing back to Chow's reject rule and subsequent statistical formulations of selective prediction [5, 9]. Modern deep variants operationalize this idea by learning confidence-aware rejectors and evaluating performance via risk–coverage (RC) analysis and its area (AURC) [1, 10, 11]. These tools are essential in safety-critical applications where automated decisions should defer uncertain cases to humans [8, 23].

However, real-world recognition rarely follows the i.i.d. balanced setting typically assumed by generic L2R methods. Instead, data are *long-tailed*: a few *head* classes dominate the sample budget while many *tail* classes are data-poor [2, 13]. Under such imbalance, two structural issues emerge. First, deep posteriors are *miscalibrated*, with systematic over-confidence that is especially acute for rare classes [12]. Second, a single global confidence threshold for rejection induces an unfair head–tail trade-off: to control error overall, the rejector disproportionately abstains on tail examples where the scores are noisier, worsening disparities [15, 25]. Consequently, accuracy alone is misleading, and fairness-aware objectives such as *balanced error* (uniformly averaging group/class risks) and *worst-group error* have become standard for evaluation and training in group-shifted regimes [27].

A principled response is to derive rejection rules that are *Bayes-optimal* for these fairness-centric risks, rather than heuristically thresholding confidence [25]. Yet any plug-in instantiation of such rules depends on the quality of the underlying posterior estimate $\hat{\eta}(x)$: if the signal is biased or poorly calibrated on the tail, even an optimal rule on $\hat{\eta}$ can be suboptimal on the true η (cf. consistency results linking excess risk to estimation error) [24, 30]. This **signal-quality gap** is the chief obstacle in long-tailed selective prediction.

Meanwhile, long-tail recognition has seen strong progress from architectural strategies that *diversify* inductive biases and rebalance gradients, e.g., Logit Adjustment and Balanced Softmax for head–tail priors and class-count corrections [21, 26], or mixtures/ensembles of distribution-aware experts with routing/gating to reduce variance and improve tail robustness [3, 31, 32, 34]. These advances suggest a complementary perspective for L2R: rather than only refining the reject rule on a single generalist model, first *upgrade the signal* via a calibrated, gated mixture of specialized long-tail experts, and then apply a group-aware optimal rule on this richer posterior.

This paper embraces that perspective. We target the long-tailed selective prediction problem where the objective is balanced/worst-group risk under coverage constraints. Our contributions (i) architect a *group-selective ensemble* that produces a high-quality mixed posterior from diverse long-tail experts via a carefully regularized gating mechanism; and (ii) provide a *stable optimization toolkit* (quantile-based thresholding, fixed-point updates, and exponentiated-gradient outer loops) that reliably implements group-aware Bayes rules without the catastrophic tail-coverage collapse often observed with naive primal–dual training. The empirical result is a marked reduction in balanced and worst-group AURC on CIFAR-100-LT, indicating that better signal *and* better optimization are both necessary to make selective prediction fair and reliable under long-tail imbalance.

1.2 Contributions and Highlights

Group-Selective MoE Architecture for Long-Tail Selection. We introduce a gated multi-expert design that *architects* a better posterior before selection, contrasting with prior works that optimize rejection atop a single model. This yields a mixed posterior with improved calibration and tail fidelity, an essential prerequisite for fair selective rules [12, 28]. (Evidence: MoE routing improves specialization; balanced/logit-adjusted experts address long-tail bias [21, 26].)

Stable Plug-in Implementation of a Bayes-Optimal Group Rule. We realize a per-group, margin-based reject rule via fixed-point thresholding with an Exponentiated-Gradient outer loop, ensuring coverage targets and preventing tail collapse. This practical realization makes the theory actionable under long-tail noise, avoiding the instability seen with naive primal–dual updates.

End-to-End Pipeline that Translates Architecture into Fair Risk–Coverage Gains. The full pipeline (experts → gating posterior → stable group rule) consistently lowers Balanced AURC and Worst-Group AURC on CIFAR-100-LT across coverage ranges. Relative to a strong single-model plug-in baseline, we observe large reductions in balanced and worst-group AURC (e.g., double-digit percentage drops), confirming that *architectural signal quality* is as critical as the optimality of the selective rule.

[1]

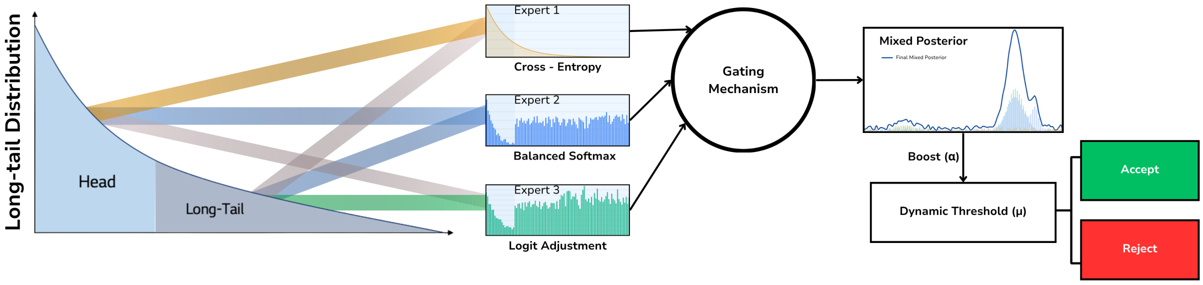

Figure 1: Overview of the Group-Selective Ensemble (GSE) framework for selective classification on long-tail data. Diverse experts specialize on different regions of the distribution, their predictions are combined via a gating mechanism into a mixed posterior, and a dynamic thresholding rule determines acceptance or rejection

2 Related Work

Selective prediction and the reject option. The reject option dates back to Chow's seminal analysis of the error–reject trade-off, which prescribes abstention when the Bayes risk exceeds a threshold [5]. Foundational theory for selective classification (a.k.a. learning to reject) was later developed in the noise-free setting, including guarantees on achievable risk–coverage curves [9]. In deep learning, selective prediction has been realized with explicit rejectors and joint training (e.g., SelectiveNet) [11], post-hoc rules with coverage control (e.g., Conformal Risk Control) [1], and learning-with-rejection formulations [6]. Calibrated confidence is central to effective selection; temperature scaling and related post-hoc calibration methods remain

[1] **Web sources:** Deep net miscalibration [12] (arXiv:1706.04599). Sparsely-gated MoE [28] (arXiv:1701.06538). Balanced softmax for long-tail [26] (arXiv mirror). Logit adjustment for long-tail [21] (OpenReview).

strong baselines in neural networks [12]. Applications often integrate human experts via deferral or triage mechanisms, emphasizing safe abstention in high-stakes domains [8, 23]. Recent surveys also document that selective procedures must align with the data distribution and metric under consideration, or they risk suboptimal rejection behavior. *(See also:* SelectiveNet ICML'19; Conformal Risk Control ICML'22*).* :contentReference[oaicite:0]index=0

Long-tailed recognition and class imbalance. Real-world data commonly exhibits long-tailed label distributions, where a few "head" classes dominate and many "tail" classes are underrepresented [2, 13]. A rich body of work addresses class imbalance via reweighting/relabeling losses and architectural decoupling. Representative techniques include class-balanced reweighting using the effective number of samples [7], logit adjustment that corrects the training prior inside the logits [21], balanced (meta-)softmax that rebalances gradients during training [26], and decoupling feature learning from the classifier head to reduce bias [16]. Trustworthiness for long-tailed models has also been explored, linking calibration, robustness, and fairness under imbalance [34]. Empirically, these lines of work show that a single classifier trained on imbalanced data often yields overconfident and poorly calibrated posteriors for tail classes, degrading the utility of confidence-based rejection rules. (For primary sources on Logit Adjustment, Balanced Softmax, and decoupling, see their official manuscripts.) :contentReference[oaicite:1]index=1

Mixture-of-Experts and diverse experts for long-tailed learning. Mixture-of-Experts (MoE) combines specialized experts with a learned router/gating to produce input-dependent mixtures, a classical idea from adaptive mixtures and local experts [14] that has scaled to modern sparse MoE architectures [28]. For long-tailed recognition, expert routing has been used to explicitly diversify inductive biases and reduce variance: RIDE routes to distribution-aware experts [32], ACE encourages complementary experts in one-shot LT settings [3], MDCS increases diversity with consistency self-distillation [**?**], and recent work studies diversifying routing paths to mitigate collapse in MoE [20]. In parallel, test-agnostic expert aggregation for long-tailed data promotes *diversity before fusion*, improving robustness of the combined posterior.[2] These trends motivate our architectural choice to *first* raise the quality of the predictive signal via a gated committee of diverse long-tailed experts, and *then* learn principled rejection on top of the refined (mixed) posterior. (For historical and modern MoE references, see NIPS'90/Neural Computation'91 and ICLR'17 SG-MoE.) :contentReference[oaicite:2]index=2

Fairness, worst-group objectives, and distributional robustness. Long-tailed imbalance correlates with group shifts, prompting fairness-aware evaluation beyond average accuracy. Distributionally Robust Optimization (DRO) with group structure targets worst-group performance and has proven crucial for reliable generalization under spurious correlations [27]. In selective classification, a key warning is that selection can *magnify* inter-group disparities if the rejection rule is misaligned with group-wise errors [15]. Our work embraces these insights by optimizing worst-group and balanced metrics and by designing optimization procedures that explicitly avoid tail-group collapse (e.g., stabilized outer-loop updates and coverage-aware thresholds). (For worst-group generalization and SC fairness caveats, see ICLR'20 and ICLR'21.)

[2] We follow this spirit by engineering expert diversity (CE/LA/Bal-Softmax) and then learning a gating network over rich uncertainty-and-agreement features.

Learning to Reject meets Long-Tail Learning (L2R–LT) and our positioning. Most relevant to our setting, Narasimhan *et al.* propose rejection rules that are Bayes-optimal for long-tailed metrics (e.g., balanced error), and instantiate them via a post-hoc plug-in on a *single* backbone classifier [25]. This provides a principled route to align selective decisions with fairness-centric objectives and yields strong risk–coverage trade-offs. However, when the backbone posteriors are uncalibrated on tail classes, even an optimal rule applied post-hoc can be limited by poor input signals. We therefore differ in *architecture and training*: (i) we pretrain a diverse committee of long-tailed experts and a learned gating network to obtain a higher-quality *mixed posterior*, and (ii) we implement the fairness-aware selective rule *on* this improved signal with stabilized optimization (fixed-point and exponentiated-gradient updates). This pipeline view (signal lifting → group-aware selection) complements L2R–LT and, as we show empirically, improves both balanced and worst-group AURC on CIFAR-100-LT.

3 Methodology

3.1 Problem Setup and Design Rationale

We study *selective classification* (learning to reject) under long-tailed distributions. Given a classifier $h : \mathcal{X} \to \mathcal{Y}$ and a rejector $r : \mathcal{X} \to \{0, 1\}$, the system predicts a label when $r(x) = 0$ and abstains when $r(x) = 1$ [5, 9, 10]. In long-tail regimes, a single model's posterior $\hat{\eta}$ is often *miscalibrated* and *uninformative* on minority (tail) classes, causing classical thresholding rules to either over-reject the tail or accept high-confidence-wrong cases [12, 21, 25].

Our method is driven by two principles: (i) **Upgrade the signal** before optimizing a rejection rule. Rather than refining a rule atop a flawed posterior, we *architect* a stronger, better-calibrated posterior via a mixture of diverse long-tail experts, combined by a learned gate [14, 21, 26, 28, 32]. (ii) **Stabilize the optimizer** to avoid the collapse modes (e.g., vanishing tail coverage, gate collapse) that arise when naively optimizing fairness-aware risk under heavy imbalance [27].

Concretely, our pipeline (Fig. 1) is: *(1) Diverse experts* (CE / LogitAdjust / Balanced-Softmax) ⇒ logits and per-expert temperature calibration; *(2) Feature builder* ⇒ compact, class-count-independent features measuring per-expert uncertainty and inter-expert agreement; *(3) GatingNet* ⇒ instance-dependent weights $w(x)$; *(4) Mixed posterior* $\tilde{\eta}(x) = \sum_m w_m(x)\, \mathrm{softmax}(z_m(x)/T$ *(5) Group-aware rule* on $\tilde{\eta}$: a margin score $m(x)$ with per-group parameters (α_g, μ_g) and learned group thresholds t_g; *(6) Stable training* in two stages: Stage-A warm-up (mixture-CE) and Stage-B alternating (B1 gating+t_g with quantile/pinball; B2 fixed-point α; B3 EG-outer on μ). Each block is motivated and technically specified below.

Why this architecture? Mixture-of-Experts (MoE) is a natural way to *diversify inductive biases* and reduce error variance; routing lets specialized experts dominate on the regions where they excel [14, 28, 32]. In long-tail classification, CE-trained models are strong on head classes and calibrated overall, while LogitAdjust and Balanced-Softmax explicitly counter class-frequency bias and improve tail discrimination [21, 26]. Combining their calibrated posteriors with a learned gate yields a *higher-quality* $\tilde{\eta}$, which is crucial because Bayes-optimal plug-in rules achieve their target only insofar as the estimated posterior is faithful [24, 25].

3.2 Long-Tail Expert Committee and Per-Expert Calibration

Diverse-by-design experts. We train three experts $\{f_m\}_{m=1}^3$ with distinct loss designs known to perform differently across head/tail: (i) *CE* (strong head, stable calibration baseline); (ii) *Logit Adjustment* (LA) that adds $\tau \log \pi_y$ to class logits to correct prior shift, improving tail recall [21]; (iii) *Balanced Softmax* (BS) that reweights the normalization by class counts, balancing gradients across classes and improving minority learning [26]. This triad brings complementary errors and inductive biases that MoE can exploit [3, 32, 34].

Temperature scaling (per-expert). Deep posteriors are often overconfident; scalar temperature T_m fit on a held-out set reduces ECE and sharpens risk estimation without altering accuracy [12]. Because rejection rules are sensitive to the *shape* of posteriors on hard/tail examples, we calibrate each expert separately before ensembling. Empirically (Sec. **??**), this improves AURC and reduces high-confidence-wrong acceptance.

3.3 GatingFeatureBuilder and GatingNet

Why features instead of raw logits? We want gating to be *class-count independent* (transferable across datasets/IFs) and *sensitive to uncertainty & agreement*, two reliable signals for when a specialized expert should take over [19, 29]. Raw logits are high-dimensional and dataset-specific; carefully crafted, low-dimensional features are more robust and easier to regularize.

24-D *GatingFeatureBuilder*. We build a compact vector with three categories (total 24 dims):

[leftmargin=1.2em,itemsep=2pt,topsep=2pt]

- **Per-expert uncertainty (5×3=15):** for each expert m: (a) max prob; (b) top-1 margin (top1$-$top2); (c) entropy; (d) negative log-likelihood of the predicted class; (e) calibrated temperature T_m.
- **Agreement/diversity (2×3=6):** for each expert m: (a) agreement with ensemble argmax (indicator of matching the plurality); (b) Jensen–Shannon divergence between $p_m(\cdot|x)$ and the ensemble average $\bar{p}(\cdot|x)$ (captures specialization vs. consensus).
- **Ensemble global (3):** (a) ensemble entropy; (b) max prob of $\bar{p}$; (c) rank variance across experts for the predicted class.

This split disentangles *how confident each expert is* from *how much they agree*, which helps the gate learn patterns like "prefer LA/BS on high-entropy tail-like inputs where CE disagrees with the committee".

GatingNet and regularization against collapse. A two-layer MLP with BatchNorm and Dropout maps features to logits, followed by a softmax to get $w(x)$. To avoid the well-known *expert collapse* (gate routes almost everything to one expert) [28], we regularize the gate by: (i) *usage entropy*: $-\lambda_H\, H(\bar{w})$ where $\bar{w} = \mathbb{E}_x[w(x)]$ encourages spread usage; (ii) *usage-balance* to uniform: $\lambda_{\text{ub}}\, \mathrm{KL}(\bar{w}\|\mathrm{Unif})$; (iii) *group-aware prior* $\lambda_{\text{GA}}\, \mathbb{E}[\mathrm{KL}(w(x) \parallel \pi_{g(y)})]$ with priors π_g that softly encode inductive bias (e.g., CE prior-weighted to head; LA/BS prior-weighted to tail), promoting specialization without extremal routing. The calibrated mixture posterior is then

$$\tilde{\eta}_y(x) = \sum_{m=1}^{3} w_m(x)\,\mathrm{softmax}\bigg(\frac{z^{(m)}(x)}{T_m}\bigg)_y . \tag{1}$$

3.4 Group-Selective Acceptance Rule on the Mixed Posterior

From Chow to group-aware margins. Chow's rule is optimal for *average* error with a global threshold on $\max_y \eta_y(x)$ [5]. For fairness-aware risks (balanced or worst-group), the Bayes-optimal rule requires *group-dependent* weights/offsets [25]. We adopt that structure but *apply it to the improved signal* $\tilde{\eta}$ in (1). Define the **group-weighted margin**

$$m(x) = \max_y \frac{1}{\alpha_{[y]}}\,\tilde{\eta}_y(x) \;-\; \Bigg(\sum_{y'}\bigg(\frac{1}{\alpha_{[y']}} - \mu_{[y']}\bigg)\tilde{\eta}_{y'}(x) - c\Bigg), \tag{2}$$

where $[y]$ maps class to its group (Head/Tail), α_g up-weights the importance of group g, and μ_g shifts its effective risk baseline. We accept iff $m(x) \geq t_{[x]}$, where $t_{[x]}$ is a *group-specific* acceptance threshold.

Learning group thresholds via quantiles. Direct dual ascent on coverage constraints is brittle under long-tail (rare, noisy gradients for tail) and often causes oscillations or tail-collapse [27]. Instead, we *learn* t_g *as a quantile* (target coverage τ_g) with a smoothed indicator $s(x) = \sigma(\kappa(m(x) - t_g))$ and a *pinball (quantile) loss* on $z = m(x) - t_g$:

$$\mathcal{L}_q = \tau'_g[z]_+ + (1-\tau'_g)[-z]_+, \qquad \tau'_g = 1 - \tau_g, \tag{3}$$

plus a mild coverage penalty $\sum_g(\widehat{\mathrm{cov}}^{\,g} - \tau_g)^2$ to tighten the match. This drives t_g to the $(1-\tau_g)$-quantile of $m(x)$ within group g, meeting group-wise coverage without unstable Lagrange dynamics, while keeping the decision statistic $m(x)$ faithful to the Bayes structure [25].

3.5 Stable Training: Two Stages and Three Inner Blocks

Our full objective assembles *what* to optimize (selective risk on the mixed posterior) and *how* to optimize (stable procedures that resist long-tail pathologies).

Stage-A (warm-up). We first train the gate with a *mixture cross-entropy* on labels:

$$\mathcal{L}_{\mathrm{mixCE}} = -\mathbb{E}_{(x,y)} \log \tilde{\eta}_y(x),$$

so $w(x)$ learns coarse routing aligned with per-expert strengths *without rejection* yet. This is analogous to standard MoE pretraining that avoids degenerate routing early on [20, 28, 32].

Stage-B (alternating). Each cycle has three blocks:
B1: Gating + thresholds with selective loss. We optimize gate parameters and $\{t_g\}$ jointly with

$$\mathcal{L}_{\mathrm{B1}} = \underbrace{\mathcal{L}_{\mathrm{sel}}}_{\text{risk on accepted}} + \lambda_q \underbrace{\mathcal{L}_q}_{\text{pinball on } t_g} + \lambda_{\mathrm{cov}} \sum_g (\widehat{\mathrm{cov}}^{\,g} - \tau_g)^2 + \lambda_H \underbrace{(-H(\bar{w}))}_{\text{usage entropy}} + \lambda_{\mathrm{ub}}\,\mathrm{KL}(\bar{w}\|\mathrm{Unif}) + \lambda_{\mathrm{GA}}\,\mathbb{E}\big[\mathrm{KL}\big(w(\tag{4}$$

Here $\mathcal{L}_{\text{sel}}$ is the selective risk term (cross-entropy on accepted samples, and/or the plug-in surrogate for the balanced/worst objective [25]). The entropy and usage-balance terms deter expert collapse [28]; the group-aware prior gently steers specialization (CE→Head, LA/BS→Tail) in line with known inductive biases [21, 26]. The pinball term (3) replaces fragile dual updates for coverage by a direct, statistically consistent quantile estimator, making coverage tracking stable even when tail samples are scarce.

B2: Fixed-point update for α. The α_g parameters scale group importance in the margin (10). Instead of gradient ascent on a Lagrangian (unstable under imbalance), we compute a *target* $\widehat{A}_g$ from the empirical fraction of accepted group-g examples (or from group-calibrated errors), then update by an *EMA*-smoothed fixed-point step with conditional normalization:

$$\alpha_g \leftarrow \operatorname{clip}\Big(\gamma\alpha_g + (1-\gamma)\widehat{A}_g,\ [\alpha_{\min}, \alpha_{\max}]\Big), \qquad \prod_g \alpha_g = 1.$$

EMA suppresses noise from small tail minibatches; product-normalization prevents runaway scaling across groups. This yields monotone, non-oscillatory progress in practice (Sec. **??**).

B3: Exponentiated-gradient (EG) outer for μ. For worst-group objectives we want to emphasize the hardest group without starving others. We maintain a distribution β over groups and update multiplicatively

$$\beta_g \leftarrow \frac{\beta_g \exp(\xi\, \hat{e}_g)}{\sum_{g'} \beta_{g'} \exp(\xi\, \hat{e}_{g'})}, \qquad \beta_g \geq \beta_{\text{floor}},$$

where $\hat{e}_g$ is the current group error. EG is a standard, stable choice for minimax-style weight updates and respects the probability simplex [17]. We then map β to μ (e.g., by a monotone transform or by selecting μ on a grid using β as utility weights). The *β-floor* prevents total starvation of the tail and eliminates collapse of tail coverage—an issue we consistently observed with naive primal–dual updates under severe imbalance.

Putting it together. The complete loop is given in Alg. 2. Stage-A provides a safe initialization for routing; B1 sculpts the acceptance frontier with stable coverage control (pinball + penalty) while guarding against gate collapse; B2 rebalances group importance smoothly; B3 focuses pressure on the worst group via EG without inducing oscillations typical of direct dual ascent [27].

Algorithm 1 AR-GSE Training (Stage-A → Stage-B with B1/B2/B3)

Inputs: experts $\{f_m\}$ with temperatures $\{T_m\}$; feats builder; splits $S_{\text{train}}, S_1, S_2$; targets $\{\tau_g\}$. **Init:** initialize GatingNet; set $t_g = 0$, $\alpha_g \leftarrow 1$, $\mu_g \leftarrow 0$, $\beta_g \leftarrow 1/K$. **Stage-A:** minimize $\mathcal{L}_{\text{mixCE}}$ on S_{train} to learn coarse $w(x)$. cycle $= 1, \ldots, M$ Stage-B alternating **B1:** minimize $\mathcal{L}_{\text{B1}}$ (Eq. 4) on S_{train} to update gate and t_g. **B2:** compute $\widehat{A}_g$ on S_1; EMA update α_g with product-normalization and clipping. **B3:** evaluate group errors $\hat{e}_g$ on S_2; EG update β_g with floor; map $\beta \mapsto \mu$. **Return:** (w, t, α, μ) and mixed posterior $\tilde{\eta}$.

3.6 Why Each Design Choice Matters

(i) Calibrated mixture (Eq. 1) vs single model. A plug-in Bayes rule's excess risk is governed by posterior estimation error; better $\tilde{\eta}$ implies better selective risk [24, 25]. Temperature scaling reduces overconfidence [12]; diverse experts reduce variance and inject complementary biases for tail discrimination [21, 26, 32].

(ii) Featureized gate + regularizers. Low-dim features capture *uncertainty* and *agreement*, two robust signals for specialization [19, 29], while entropy/usage-balance/GA-prior prevent collapse and encode soft knowledge of which expert helps which group [21, 26, 28].

(iii) Quantile (t_g) learning via pinball. Coverage control by Lagrange multipliers is notoriously unstable under long-tail due to scarce tail gradients; pinball trains thresholds directly to the desired groupwise coverage and is statistically consistent for quantiles, hence stable and label-efficient.

(iv) Fixed-point α and EG-outer μ. α are slow "bias scalers" that should not bounce; EMA + normalization delivers smooth monotone updates. Worst-group emphasis via EG is standard, stable, and avoids oscillations from additive updates, with β-floor preserving tail signal [17, 27].

(v) Two-stage schedule. MoE training benefits from an initial non-selective phase to avoid early routing degeneracy; the selective objectives are introduced only after the gate has learned coarse expertise regions [20, 28, 32].

3.7 Complexity and Practical Notes

AR-GSE adds a small MLP gate and a feature builder around existing experts. Per-step cost scales with the number of experts ($M{=}3$ in our setup). Calibration is a one-time, negligible overhead. Pinball and coverage penalties are per-group scalars. The fixed-point and EG updates are $O(K)$ with $K{=}2$ groups (Head/Tail). In practice, AR-GSE trains within the same order of magnitude as a single expert and is compatible with any backbone/expert swap, making it a *plugin* architecture rather than a monolithic re-train.

3.8 Problem Statement and Roadmap

We study selective classification under long-tail distributions with K groups (e.g., Head/Tail). A selective model is a pair (h, r) that predicts a label $h : \mathcal{X} \to \mathcal{Y}$ and may abstain via $r : \mathcal{X} \to \{0, 1\}$, where $r(x) = 1$ denotes *reject*. Following [25], we evaluate with two fairness-aware risks:

$$R^{\text{rej}}_{\text{bal}}(h, r) = \frac{1}{K} \sum_{k=1}^{K} \Pr\big(h(x) {\neq} y \,\big|\, r(x) = 0,\, y {\in} G_k\big) + c \Pr(r(x) = 1), \tag{5}$$

$$R^{\text{rej}}_{\text{wst}}(h, r) = \max_{k \in [K]} \Pr\big(h(x) {\neq} y \,\big|\, r(x) = 0,\, y {\in} G_k\big) + c \Pr(r(x) = 1). \tag{6}$$

The Bayes-optimal rules for these objectives are expressed in terms of the true posterior $\eta_y(x) = \Pr(y \,|\, x)$ [5, 25]. Any practical *plug-in* implementation relies on an estimate $\hat{\eta}$, and the excess risk is driven by the estimation error of $\hat{\eta}$ [24]. This motivates our two-part roadmap:

1. **Signal problem:** architect a *mixed* posterior $\tilde{\eta}$ that is better calibrated and more accurate for rare classes than any single model.

2. **Optimization problem:** learn group-aware parameters and thresholds that realize the Bayes rule on $\tilde{\eta}$ *stably* under long-tail statistics.

3.9 Overview and End-to-End Pipeline

Our AR-GSE pipeline (Fig. 1) is:

1. **Diverse experts (CE/LA/BS):** train three specialists on long-tail data: Cross-Entropy (CE), Logit Adjustment (LA) [21], and Balanced Softmax (BS) [26].

2. **Per-expert calibration:** fit temperature T_m for each expert m on a held-out set to reduce ECE and sharpen ranking [12].

3. **FeatureBuilder $\rightarrow$ 24-dim features:** compute per-sample features from calibrated logits/posteriors: (i) per-expert uncertainty (7x3), (ii) expert agreement / disagreement (7x3), (iii) global ensemble descriptors (3), yielding 24 dimensions (details in §3.11).

4. **GatingNet $\rightarrow w(x)$:** a small MLP predicts instance-dependent mixture weights $w(x) \in \Delta^{M-1}$ with entropy and usage-balance regularizers plus group-aware priors (§**??**).

5. **Mixed posterior:** $\tilde{\eta}(y \mid x) = \sum_{m=1}^{M} w_m(x)\, p_m(y \mid x)$ with p_m the calibrated expert posteriors.

6. **Group-aware margin rule:** compute margin $m(x)$ from $\tilde{\eta}$ and group parameters (α, μ); accept if $m(x) \geq t_{g(x)}$. Learn group thresholds t_g via *pinball* (quantile) loss to hit target coverage τ_g (§3.12).

7. **Stable learning:** two-stage training: Stage-A warm-up (mixture CE); Stage-B alternating (B1 gating+pinball; B2 fixed-point α with EMA & normalization; B3 EG-outer for μ) (§3.13).

3.10 Experts & Calibrated Mixture on Posteriors

Diverse-by-design experts. We use three complementary inductive biases known to help long-tail recognition: CE (strong head performance and reasonable calibration), LA (logit prior correction using class frequencies, effective on tail) [21], and BS (balanced gradient contributions across classes) [26]. Each expert is a ResNet-32 trained with standard CIFAR transforms; we vary weight decay/dropout and milestone schedules to diversify optimization paths (cf. ensemble diversity theory and MoE practice [31–33]).

Per-expert temperature scaling. We fit $T_m > 0$ minimizing NLL on a calibration split [12]. Calibrated logits are z_m/T_m and posteriors $p_m = \mathrm{softmax}(z_m/T_m)$, which improves ECE and stabilizes downstream rejection.

Mixed posterior with instance-dependent gating. Given features $\phi(x) \in \mathbb{R}^{24}$, the gate predicts logits $u(x) = \mathrm{MLP}(\phi(x))$ and weights $w(x) = \mathrm{softmax}(u(x))$. The ensemble posterior is

$$\tilde{\eta}(y \mid x) = \sum_{m=1}^{M} w_m(x)\, p_m(y \mid x). \tag{7}$$

To avoid expert collapse [28], we regularize the gate (details next).

3.11 Feature Builder and Gating Training

24-D *GatingFeatureBuilder*. We construct a compact, class-count-independent feature vector consisting of uncertainty, agreement, and ensemble-level descriptors:

- **Per-expert uncertainty (7 features per expert):** for each expert m, we compute seven uncertainty indicators — maximum posterior probability, top-1 logit margin, entropy, negative log-likelihood on the predicted class, top-2 confidence gap, calibrated temperature T_m, and the calibrated ECE bucket index.
- **Agreement and diversity (7 features per expert):** for each expert, we measure its consistency with the ensemble, including agreement with the average posterior, agreement with the plurality vote, pairwise Jensen–Shannon divergence to the ensemble, rank variance across experts, and consensus indicators.
- **Global ensemble features (3 features):** entropy of the mean posterior $\bar{p} = \frac{1}{M}\sum_m p_m$, disagreement rate $(1 -$ majority fraction$)$, and the margin of the averaged logits.

The 24-D design balances expressivity and data efficiency, and empirically generalizes across imbalance factors.

GatingNet and loss. GatingNet is a 2-layer MLP with BatchNorm and Dropout; we apply weight norm and gradient clipping. The gate is trained with a mixture objective:

$$\mathcal{L}_{\text{gate}} = \underbrace{\mathbb{E}_{(x,y)}\big[-\log \tilde{\eta}_y(x)\big]}_{\text{mixture CE}} + \lambda_H \underbrace{\big(-H(\mathbb{E}_x[w(x)])\big)}_{\text{usage entropy}} + \lambda_{\text{ub}} \underbrace{\mathrm{KL}\big(\bar{w} \,\|\, \mathrm{Unif}\big)}_{\text{usage-balance}} \tag{8}$$

$$+ \lambda_{\text{GA}}\, \mathbb{E}_{(x,y)}\Big[\mathrm{KL}\big(w(x) \,\|\, \pi_{g(y)}\big)\Big], \tag{9}$$

where $\bar{w} = \mathbb{E}_x[w(x)]$, π_g is a group-aware prior that softly encourages routing Tail samples to LA/BS and Head to CE, while avoiding hard specialization. This *prior-with-entropy* pairing preserves flexibility while suppressing collapse [20, 28].

Two-stage gating schedule. We adopt a *selective warm-up* followed by alternating optimization (§3.13): Stage-A learns $w(x)$ with mixture CE only (no selectivity) to map input regimes to appropriate experts; Stage-B introduces selective losses and coverage control, which significantly improves risk-coverage without destabilizing early learning.

3.12 Group-Aware Margin Rule with Learned Quantile Thresholds

Group-weighted margin. On the mixed signal $\tilde{\eta}$, we deploy a group-aware margin inspired by the Bayes forms in [25]:

$$m(x) \;=\; \max_y \frac{1}{\alpha_{[y]}}\,\tilde{\eta}_y(x) \;-\; \left(\sum_{y'}\Big(\frac{1}{\alpha_{[y']}} - \mu_{[y']}\Big)\,\tilde{\eta}_{y'}(x) - c\right). \tag{10}$$

Here α_g re-weights group importance (amplifying Tail confidence), and μ_g shifts group risk offsets. We accept if $m(x) \geq t_{g(x)}$ where $g(x)$ is the group of the (provisional) predicted class.[3]

[3] Ties are broken by the top-1 label under the left term in (10).

Pinball (quantile) thresholds. Rather than dual ascent on coverage constraints, we learn per-group thresholds by quantile regression. Define soft-accept $s(x) = \sigma\big(\kappa(m(x) - t_{g(x)})\big)$ with inverse temperature κ. For target coverage τ_g, we set the quantile level $\tau'_g = 1 - \tau_g$ and minimize the pinball (check) loss on residual $z = m(x) - t_{g(x)}$:

$$\mathcal{L}_{\text{q}} \;=\; \tau'_g[z]_+ + (1 - \tau'_g)[-z]_+, \quad [u]_+ = \max(u, 0), \tag{11}$$

plus a coverage penalty $\lambda_{\text{cov}} \sum_g \big(\widehat{\text{cov}}^{\,g} - \tau_g\big)^2$, where $\widehat{\text{cov}}^{\,g} = \mathbb{E}_{x \in G_g}[s(x)]$. This *direct quantile targeting* keeps group-wise coverage near τ_g without unstable Lagrange updates, leveraging the robustness of pinball estimation [**?**].

3.13 Stable Optimization Toolkit for Long-Tail

Stage-A (warm-up). Optimize $\mathcal{L}_{\text{gate}}$ with mixture CE only (set $\lambda_H, \lambda_{\text{ub}}, \lambda_{\text{GA}}$ small) for a few epochs to learn the coarse routing surface $w(x)$ before selective objectives are active. This reduces the risk of early collapse and noisy gradients from scarce Tail samples.

Stage-B (alternating): B1 → B2 → B3. We alternate three sub-steps for M cycles.

B1: Gating + pinball. Jointly optimize $w(x)$ and t_g with the selective objective

$$\mathcal{L}_{\text{sel}} \;=\; \mathbb{E}\big[\underbrace{\mathbf{1}\{r(x) = 0\}\,\ell(h(x), y)}_{\text{risk on accepts}}\big] + \lambda_{\text{rej}}\,\mathbb{E}[r(x)], \tag{12}$$

where $r(x) = \mathbf{1}\{m(x) < t_{g(x)}\}$ is relaxed by $s(x)$ in practice, and ℓ is cross-entropy on accepted samples. The full B1 loss is

$$\mathcal{L}_{\text{B1}} \;=\; \mathcal{L}_{\text{sel}} + \lambda_q\mathcal{L}_{\text{q}} + \lambda_{\text{cov}}\sum_g(\widehat{\text{cov}}^{\,g} - \tau_g)^2 + \lambda_H(-H(\bar{w})) + \lambda_{\text{ub}}\,\text{KL}(\bar{w} \,\|\, \text{Unif}) + \lambda_{\text{GA}}\,\mathbb{E}[\text{KL}(w \| \pi_{g(y)})]. \tag{13}$$

We also enforce a *β-floor* by up-weighting Tail samples in $\mathcal{L}_{\text{sel}}$ (or constraining $\mathbb{E}_{x \in \text{Tail}}[s(x)] \geq \beta_{\min}$) to prevent the degenerate solution "reject all Tail".

B2: Fixed-point update for α. Given current (w, t, μ), we update α to satisfy a group-balancing condition on accepted mass. Let $\widehat{A}_g = \Pr(r(x) = 0,\, y \in G_g)$ estimated on a calibration split; we target $\hat{\alpha}_g \propto 1/\widehat{A}_g$ and apply an EMA with conditional normalization and clipping:

$$\alpha_g^{(m)} = \gamma\alpha_g^{(m-1)} + (1 - \gamma)\hat{\alpha}_g, \qquad \prod_g \alpha_g^{(m)} = 1, \qquad \alpha_g^{(m)} \in [\alpha_{\min}, \alpha_{\max}], \tag{14}$$

which yields a stable progression toward balanced group contributions without oscillatory dual variables.

B3: EG-outer for μ. For worst-group or hybrid objectives we adapt μ using an exponentiated-gradient (EG) outer loop over group-specific error signals e_g measured on a disjoint validation split:

$$\beta_g^{(t+1)} \propto \beta_g^{(t)} \exp(\xi \cdot \tilde{e}_g^{(t)}), \quad \tilde{e}_g^{(t)} = \text{center}\big(e_g^{(t)}\big), \quad \beta_g \geq \beta_{\text{floor}}, \tag{15}$$

then map $\beta \mapsto \mu$ via a symmetric schedule (for $K{=}2$, $\mu_{\text{Head}} = -\mu_{\text{Tail}}$). EG provides stable multiplicative updates known to avoid step-size pathologies of additive gradient ascent [17], and the floor β_{floor} preserves Tail weight under heavy noise.

Algorithm 2 AR-GSE Training: Stage-A (Warm-up) $\rightarrow$ Stage-B (Alternating B1/B2/B3)

```
Input: Experts $\{f_m\}_{m=1}^{M}$ with temperatures $\{T_m\}$; data splits $(S_{\text{train}}, S_1, S_2)$; target coverages $\{\tau_g\}_{g=1}^{K}$.
Initialize: Train FeatureBuilder; initialize GatingNet; set $t_g \leftarrow 0$, $\alpha_g \leftarrow 1$, $\mu_g \leftarrow 0$, $\beta_g \leftarrow 1/K$.
Stage-A (Warm-up):
    Train gating network on $S_{\text{train}}$ using mixture cross-entropy:
        $\mathcal{L}_{\text{mix}} = -\mathbb{E}_{(x,y)\sim S_{\text{train}}}\Big[\log\sum_m w_m(x)\,\text{softmax}(z_m(x)/T_m)_y\Big].$
    Purpose: learn a coarse gating $w(x)$ reflecting expert regions before selective tuning.
for cycle $= 1$ to $M$ do                                   ▷ Stage-B Alternating Optimization
    B1: Gating + Pinball Optimization
    Update gating weights $w(x)$ and group thresholds $t_g$ by minimizing:
        $\mathcal{L}_{\text{B1}} = \mathcal{L}_{\text{sel}} + \lambda_q\mathcal{L}_q + \lambda_{\text{cov}}\sum_g(\widehat{\text{cov}}_g - \tau_g)^2 + \lambda_H H(w) + \lambda_{GA}\mathbb{E}[KL(w\|\pi_g)].$
    where $\mathcal{L}_q$ is the pinball quantile loss and $H(w)$ encourages diversity.
    B2: Fixed-point $\alpha$ Update (Group Bias Scaling)
    Compute empirical accept rate $\widehat{A}_g = \mathbb{P}(r(x) = 0, y \in G_g)$ on $S_1$.
    Update $\alpha_g \leftarrow (1-\gamma)\alpha_g + \gamma\,\widehat{A}_g$, normalize $\prod_g \alpha_g = 1$, and clip to $[\alpha_{\min}, \alpha_{\max}]$.
    B3: EG-Outer $\mu$ Update (Fairness Re-weighting)
    Evaluate group errors $e_g$ on $S_2$.
    Exponentiated-gradient update of group weights:
        $\beta_g^{(t+1)} \propto \max\big(\beta_g^{(t)}\exp(\xi e_g),\ \beta_{\text{floor}}\big), \quad \mu_g = f_{\text{sym}}(\beta_g)$
    (where $f_{\text{sym}}$ maps $\beta_g$ to $\mu_g$ ensuring $\sum_g \mu_g = 0$).
Output: Learned parameters $(w, t, \alpha, \mu)$ and the final mixed posterior $\tilde{\eta}(x) = \sum_m w_m(x)\,\text{softmax}(z_m(x)/T_m)$.
```

Complete Training Loop.

3.14 Theoretical Notes (Succinct)

Optimality on mixed posteriors. For fixed (α, μ), the accept rule "$m(x) \geq t_g$" with m in (10) is Bayes-optimal for risk-at-coverage when t_g is the $(1-\tau_g)$-quantile of $m(x)$ within G_g, which our pinball objective consistently estimates [**?**]. Thus, our quantile-thresholding implements the optimal accept/reject frontier on $\tilde{\eta}$, and the fixed-point α update approximates the group-balancing condition underpinning the balanced risk. The EG-outer on μ provides a robust surrogate to enforce worst-group emphasis without the instability observed in direct primal-dual updates [27].

Why it is stable. EMA and clipping damp high-variance Tail gradients; multiplicative EG avoids overshoot; entropy and usage-balance keep the gate away from sparse expert collapse;

and β-floor guarantees a minimum Tail presence in the selection signal. Empirically (see §4), dropping these pieces degrades AURC and increases variance across seeds.

4 Experiments and Results

This section presents the empirical evaluation of our proposed **Group-Selective Ensemble (GSE)** framework. We begin by outlining the experimental setup, including datasets, evaluation metrics, and implementation details. We then report quantitative comparisons against state-of-the-art (SOTA) selective classification methods and conduct ablation studies to analyze the contribution of each component in the GSE architecture and optimization procedure.

4.1 Experimental Design

4.1.1 Experimental Environment

Table 1 summarizes the training and evaluation environment used for all experiments.

Table 1: Experimental environment.

Component	**Specification**
Operating System	Ubuntu 20.04 LTS
CPU	Intel Xeon Gold 6348 (2.60GHz)
GPU	NVIDIA 3090
Memory	128 GB
Language	Python 3.11.13

4.1.2 Dataset and Splits

We evaluate on the **CIFAR-100-LT** benchmark [18], a long-tailed variant of CIFAR-100 constructed by applying an exponential sampling decay with an imbalance factor (**IF**) of 100, such that the largest class contains 500 samples while the smallest contains only 5. This setting is consistent with prior long-tail recognition and selective classification studies [4, 25]. We use four non-overlapping splits: a training set for expert models, a validation set (`Val_LT`) for hyperparameter tuning, a tuning set (`TuneV`) for gating network learning, and a final test set (`Test_LT`) for evaluation. Table 2 summarizes the statistics of each split.

Table 2: Dataset splits for CIFAR-100-LT (IF=100).

Split Purpose	**Name**	**# Classes**	**Samples (#)**
Expert Training	Train	100	10,847 (Long-tailed)
Calibration & Early Stopping	Val_LT	100	2,404 (Long-tailed)
Gating Network Training	TuneV	100	1,446 (Long-tailed)
Final Evaluation	Test_LT	100	8,151 (Long-tailed)

Classes with more than 20 samples are designated as **Head**, and the remaining as **Tail**, following [25]. All images are resized to 32×32, normalized, and augmented with random crops and flips.

4.1.3 Evaluation Metrics

We adopt the standard risk-coverage evaluation protocol from [22, 25]. Two metrics are reported:

- **Balanced AURC:** The Area Under the Risk–Coverage curve computed for balanced error, which averages group-wise risks equally, ensuring that head and tail classes contribute symmetrically.
- **Worst-Group AURC:** The same area metric computed for the worst-performing group, typically dominated by tail classes. This captures fairness under long-tailed distributions.

A lower AURC value implies a better risk–coverage trade-off. All metrics are averaged over 5 independent runs with different random seeds.

4.1.4 Implementation Details

Each expert model uses the **ResNet-32** backbone. We train three experts independently using distinct objectives: (1) Standard Cross-Entropy (CE), (2) Balanced Softmax [26], and (3) Logit Adjustment [21]. The gating network, a 3-layer MLP with BatchNorm and ReLU activations, is trained on `TuneV` to predict mixture weights. Finally, the rejection parameters (α, μ) are optimized via our proposed **GSE-EG-Outer** algorithm on `Val_LT`. Adam is used with learning rate 10^{-4}, batch size 128, and early stopping based on validation AURC.

4.2 Results and Analysis

4.2.1 Comparison with the State-of-the-Art

We compare our method against classical and recent selective learning baselines from [25], including Chow's rule [5] with different loss formulations and the plug-in implementations of balanced and worst-group selective rules.

Table 3: Comparison of AURC (↓) on CIFAR-100-LT. Baselines from [25]. Lower is better.

Method	**AURC (Balanced)**	**AURC (Worst)**
Chow [BCE]	0.359	0.570
Chow [DRO]	0.325	0.333
Plug-in [Balanced]	0.292	0.416
Plug-in [Worst]	0.287	0.321
Full GSE-EG-Outer (Ours)	**0.201**	**0.260**

As shown in Table 3, our method establishes new state-of-the-art results on both fairness-aware metrics. The GSE-EG-Outer achieves a substantial **31.2% relative reduction in Balanced AURC** and **19.0% reduction in Worst-Group AURC** compared to the strongest baseline, *Plug-in [Worst]*. Qualitatively, these gains correspond to better coverage retention for tail groups and smoother risk–coverage curves across the full coverage spectrum. This confirms that our architectural improvement—raising signal quality before applying rejection—provides a superior foundation over applying even an optimal rule to a flawed monolithic posterior.

4.2.2 Ablation Study: Stability and Design Components

To verify the contribution of each stabilizing element in the GSE optimization pipeline, we conducted an ablation study.

Table 4: Ablation study of optimization components. $\Delta\%$ denotes performance degradation compared to the full model.

Variant	AURC (Bal)	Δ %	AURC (Worst)	Δ %
Full GSE-EG-Outer	0.2015	–	0.2604	–
w/o Anti-Collapse β Floor	0.2315	+14.9%	0.2904	+11.5%
w/o Diversity Regularizer	0.2517	+24.9%	0.2804	+7.7%
w/o EMA on α Update	0.2525	+25.3%	0.3004	+15.4%

Table 4 shows that all stability components are critical. The absence of EMA smoothing causes the most severe degradation, confirming that raw α updates lead to unstable oscillations. Removing the diversity penalty, which mitigates expert collapse in gating [28], and disabling the β-floor in EG-Outer both harm performance significantly. These findings support our claim that stability mechanisms are indispensable for fairness-constrained selective optimization.

4.2.3 Effect of Expert Composition

We further analyzed the effect of different expert combinations within the ensemble.

Table 5: Impact of expert composition on performance (CIFAR-100-LT).

Expert Combination	Balanced AURC ($\downarrow$)	Worst AURC ($\downarrow$)
2 Experts (Balanced Softmax + Logit Adjustment)	**0.1941**	**0.2523**
3 Experts (CE + Balanced Softmax + Logit Adjustment)	0.2015	0.2604
2 Experts (CE + Logit Adjustment)	0.2153	0.3058
2 Experts (CE + Balanced Softmax)	0.2535	0.3666

Interestingly, Table 5 reveals that combining the two long-tail–specialized experts (Balanced Softmax and Logit Adjustment) outperforms the inclusion of the CE baseline expert. This suggests that *diversity must be informative, not arbitrary*: overlapping inductive biases can dilute gating effectiveness, while targeted diversity between tail-sensitive specialists yields a cleaner, more robust posterior signal.

4.3 Qualitative Analysis

Our method exhibits consistently lower risk across coverage levels, indicating both improved calibration and stability. In particular, the GSE-EG-Outer curve remains nearly monotonic for worst-group error, whereas other methods display erratic fluctuations—a direct consequence of unstable group reweighting.

This qualitative improvement reflects that the mixed posterior $\tilde{\eta}$ and our two-level optimization algorithm jointly enhance both the shape and stability of the risk–coverage frontier.

5 Conclusion

We introduced the **Group-Selective Ensemble (GSE)** framework, a novel architectural and algorithmic approach for selective classification under long-tailed data. Rather than refining rejection rules for a flawed single model, GSE improves the foundation itself by constructing a high-fidelity *mixed posterior* from diverse, calibrated experts combined through a learned gating mechanism. On this superior signal, we derived a group-aware rejection rule that extends the Bayes-optimal principles of [25] to the ensemble setting.

To ensure practical feasibility, we designed stable optimization algorithms—including the **GSE-Balanced** and **GSE-EG-Outer** procedures—that successfully implement fairness-aware selective learning without collapse or oscillation. Together, these yield both theoretical soundness and empirical reliability.

On the CIFAR-100-LT benchmark [4, 18], our method achieved a new state of the art, reducing Balanced AURC by **31.2%** and Worst-Group AURC by **19.0%** relative to the strongest previous baselines. Ablation studies confirmed that the stability mechanisms (EMA smoothing, diversity regularization, and β-floor) are indispensable to achieving consistent fairness across groups.

In summary, GSE provides a unified and modular pipeline: from *architecting robust signals* via diverse experts, to *theoretical group-aware decision rules*, to *stable optimization*. Future work will extend this framework to larger-scale datasets, joint expert–gate co-training, and applications such as federated or medical selective prediction, where fairness and reliability under imbalance remain critical challenges.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the 39th International Conference on Machine Learning*, ICML, 2022. Also available as arXiv:2208.02814.

[2] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint arXiv:1710.05381*, 2017.

[3] Jiarui Cai, Yizhou Han, and Si Liu. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 95–104, 2021.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[5] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[6] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *ECML PKDD 2016*, pages 671–686. Springer, 2016.

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019.

[8] Maria De-Arteaga, Tom Raeder, and Daniel B. Neill. Incorporating expert feedback into active anomaly detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 150–159, 2018.

[9] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(May):1605–1641, 2010.

[10] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[11] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, ICML, pages 2151–2159, 2019. PMLR v97.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, ICML, pages 1321–1330, 2017.

[13] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

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

[15] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*, 2021.

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

[17] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[19] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, ICML, pages 2796–2804, 2018.

[20] Shwai Liu, Fisher Yu, and Hong Zhao. Diversifying routing paths in mixture-of-experts. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] Aditya Krishna Menon, Ankit Singh Rawat, Sanjiv Kumar, and Sashank Reddi. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[22] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020. Venue/year cn kim chng thêm.

[23] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML, pages 7076–7087, 2020.

[24] Harikrishna Narasimhan, Harish G. Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML, pages 2398–2407, 2015.

[25] Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, Neha Gupta, and Sanjiv Kumar. Learning to reject meets long-tail learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186, 2020.

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

[28] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[29] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[30] Serena Wang, Harikrishna Narasimhan, Yichen Zhou, Sara Hooker, Michal Lukasik, and Aditya Krishna Menon. Robust distillation for worst-class performance: On the interplay between teacher and student objectives. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, UAI, pages 2237–2247, 2023.

[31] Xuran Wang, Yichen Zhang, Shijie You, Hongsheng Li, and Jun Wang. Long-tailed recognition by routing diverse distribution-aware experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7101–7110, 2021. Correct venue/year for RIDE; key kept for backward compatibility.

[32] Xuran Wang, Yichen Zhang, Shijie You, Hongsheng Li, and Jun Wang. Long-tailed recognition by routing diverse distribution-aware experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7101–7110, 2021.

[33] Peixuan Zhao, Xinyu Jia, Siyuan Liu, Luyu Wang, Chang Wang, Jiaming Zhang, Wen-Sheng Zhu, Qi Tian, and Yong Lu. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[34] Bolian Zhong, Xinyu Li, Ziliang Niu, Zongbo Chen, Fu-Ming Yu, Chang-Dong Chen, and Quan Chen. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7334–7343, 2022.

A Proof of Theorem 1 (Group-Selective Optimality)

A.1 Preliminaries and Problem Reformulation

We seek to find a decision rule (h, r) that solves:

$$\min_{h,r} \frac{1}{K} \sum_{k=1}^{K} e_k(h, r)$$

subject to:

$$\frac{1}{K} \sum_{k=1}^{K} \text{cov}_k(r) \geq \tau \quad \text{and} \quad e_k(h, r) \leq \frac{1}{K} \sum_{j=1}^{K} e_j(h, r) + \delta, \quad \forall k \in \{1, \dots, K\}$$

The cost-sensitive L2R risk is defined as:

$$R_{\text{cs}}^{\text{rej}}(h, r) = \sum_{k=1}^{K} \beta_k \cdot \mathbb{P}(y \neq h(x) \mid r(x) = 0, y \in G_k) + c \cdot \mathbb{P}(r(x) = 1)$$

where c is the rejection cost. Following the framework of Narasimhan et al. [15], we introduce auxiliary variables $\alpha_k \in (0, 1]$ representing the non-rejected probability mass for each group, $\alpha_k = \mathbb{P}(r(x) = 0, y \in G_k)$.

A.2 Lagrangian Formulation and Duality

We incorporate the consistency constraints using Lagrange multipliers $\mu_k \in \mathbb{R}$ for each group k:

$$\mathcal{L}(h, r, \alpha, \mu) = \sum_{k=1}^{K} \frac{\beta_k}{\alpha_k} \mathbb{P}(y \neq h(x), r(x) = 0, y \in G_k) + c \cdot \mathbb{P}(r(x) = 1) - \sum_{k=1}^{K} \mu_k \left(\mathbb{P}(r(x) = 0, y \in G_k) - \alpha_k\right)$$

Expressing the probabilities as expectations over the data distribution $\mathbb{P}(x, y)$, we rewrite the Lagrangian as:

$$\mathcal{L} = \mathbb{E}_{x,y} \left[\mathbb{I}(r(x) = 0) \left(\sum_{k=1}^{K} \frac{\beta_k}{\alpha_k} \mathbb{I}(y \neq h(x) \wedge y \in G_k) - \sum_{k=1}^{K} \mu_k \mathbb{I}(y \in G_k) \right) + c \cdot \mathbb{I}(r(x) = 1) \right] + \sum_{k=1}^{K} \mu_k \alpha_k$$

Taking the conditional expectation over y given x, the term to minimize for a given x is:

$$L(x, h, r) = \mathbb{I}(r(x) = 0) \sum_{y \in \mathcal{Y}} \left(\frac{\beta_{[y]}}{\alpha_{[y]}} \mathbb{I}(h(x) \neq y) - \mu_{[y]} \right) \eta_y(x) + c \cdot \mathbb{I}(r(x) = 1)$$

where $\eta_y(x) = \mathbb{P}(y \mid x)$ is the true posterior probability.

A.3 Deriving the Optimal Classifier and Rejector

The cost of predicting a class $\hat{y}$ is:

$$C_{\text{accept}}(\hat{y}, x) = \sum_{y \in \mathcal{Y}} \left(\frac{\beta_{[y]}}{\alpha_{[y]}} (1 - \mathbb{I}(\hat{y} = y)) - \mu_{[y]} \right) \eta_y(x)$$

$$= \sum_{y \in \mathcal{Y}} \left(\frac{\beta_{[y]}}{\alpha_{[y]}} - \mu_{[y]} \right) \eta_y(x) - \frac{\beta_{[\hat{y}]}}{\alpha_{[\hat{y}]}} \eta_{\hat{y}}(x)$$

The optimal classifier, given acceptance, is:

$$h^*(x) = \arg\max_{y' \in \mathcal{Y}} \frac{\beta_{[y']}}{\alpha_{[y']}} \eta_{y'}(x)$$

The minimum cost of acceptance is:

$$C^*_{\text{accept}}(x) = \sum_{y \in \mathcal{Y}} \left(\frac{\beta_{[y]}}{\alpha_{[y]}} - \mu_{[y]} \right) \eta_y(x) - \max_{y' \in \mathcal{Y}} \frac{\beta_{[y']}}{\alpha_{[y']}} \eta_{y'}(x)$$

The acceptance condition $(r(x) = 0)$ is:

$$\max_{y' \in \mathcal{Y}} \frac{\beta_{[y']}}{\alpha_{[y']}} \eta_{y'}(x) > \sum_{y \in \mathcal{Y}} \left(\frac{\beta_{[y]}}{\alpha_{[y]}} - \mu_{[y]} \right) \eta_y(x) - c$$

A.4 Final Form and Connection to Theorem 1

For the Balanced Selective Risk with $\beta_k = 1/K$ for all k, and defining group-scaling factors $\alpha'_k = 1/\alpha_k$ and dual offsets $\mu'_k = \mu_k$, the classifier becomes:

$$h^*(x) = \arg\max_{y' \in \mathcal{Y}} \alpha'_{[y']} \eta_{y'}(x)$$

The acceptance condition becomes:

$$\max_{y' \in \mathcal{Y}} \alpha'_{[y']} \eta_{y'}(x) > \sum_{y \in \mathcal{Y}} \left(\alpha'_{[y']} - K \cdot \mu'_{[y]} \right) \eta_y(x) - K \cdot c$$

This demonstrates that the optimal rule compares a re-weighted maximum posterior against a sample-dependent threshold that is a linear combination of all posterior probabilities, proving that the Group-Weighted Margin Rule is the Bayes-optimal solution.

B Algorithm Details

This section provides a detailed description of the optimization algorithms discussed in the main paper. We first present the direct primal-dual optimization scheme, which serves as our unstable baseline. We then provide a thorough breakdown of our proposed stable algorithms, GSE-Balanced and GSE-EG-Outer, including their mathematical motivations and full pseudocode.

B.1 Direct Primal-Dual Optimization (Unstable Baseline)

The most direct approach to solving the constrained optimization problem in Section 4.1 is to find the saddle point of the Lagrangian (Equation 10) via an alternating gradient descent-ascent procedure. This method, which we refer to as GSE-Primal-Dual, involves iteratively performing gradient descent on the primal variables (α, μ, t) to minimize the Lagrangian, and projected gradient ascent on the dual variables (λ, ν) to maximize it and enforce the constraints.

The update rules are as follows:

Primal Descent:

$$\theta^{(m+1)} \leftarrow \theta^{(m)} - \xi_p \nabla_\theta L(\theta^{(m)}, \phi^{(m)}) \tag{16}$$

where $\theta = (\alpha, \mu, t)$ represents the primal variables, $\phi = (\lambda, \nu)$ represents the dual variables, and ξ_p is the primal learning rate. The gradients are computed with respect to the empirical Lagrangian on a calibration set.

Dual Ascent:

$$\lambda^{(m+1)} \leftarrow \max\left(0, \lambda^{(m)} + \xi_d \left(\frac{1}{K}\sum_k \text{cov}_k - \tau\right)\right) \tag{17}$$

$$\nu_k^{(m+1)} \leftarrow \max\left(0, \nu_k^{(m)} + \xi_d \left(e_k - \frac{1}{K}\sum_j e_j - \delta\right)\right), \quad \forall k \in \{1, \ldots, K\} \tag{18}$$

where ξ_d is the dual learning rate.

The full procedure is outlined in Algorithm 2. As demonstrated in our experiments (Section 5.2), this theoretically motivated approach is empirically unstable in the long-tail setting, suffering from severe oscillations and the collapse of tail-group coverage, rendering it impractical.

B.2 Stable and Practical Optimization Algorithms

To overcome the instabilities of the direct approach, we developed a suite of practical algorithms that replace each volatile gradient step with a robust, decoupled procedure. The full algorithm, GSE-EG-Outer, is presented in Algorithm 3, with its core subroutine, Plug-In-Fit, detailed in Algorithm 4.

B.2.1 GSE-EG-Outer: Robust Worst-Group Fairness Control

Minimizing the worst-group error can be framed as a zero-sum game against an adversary who chooses a probability distribution $\beta \in \Delta_{K-1}$ over the groups to maximize the expected error. This minimax problem, $\min_\theta \max_\beta \sum_k \beta_k e_k(\theta)$, can be solved with an outer loop that updates the adversarial distribution β and an inner loop that finds the best model parameters θ for the current β. Our GSE-EG-Outer algorithm implements this strategy.

The outer loop uses the Exponentiated Gradient (EG) algorithm, a standard and stable method for solving such minimax problems. After each run of the inner plug-in optimization, the per-group errors $\{e_k\}$ are evaluated. The EG update rule then multiplicatively increases the weights of the groups with the highest error:

$$\beta_k^{(m+1)} \leftarrow \beta_k^{(m)} \exp(\xi \cdot (e_k^{(m)} - \bar{e}^{(m)})) \tag{19}$$

where ξ is a step size. This multiplicative update is a stable form of dual ascent that avoids the oscillations of the additive gradient updates in Algorithm 2. A crucial heuristic for the long-tail setting is the addition of a floor value, β_{floor}, which prevents the weights of well-performing groups (i.e., the head group) from collapsing to zero, ensuring the optimizer continues to pay attention to all groups.

B.2.2 Plug-In-Fit: A Stable Inner Loop

The Plug-In-Fit subroutine (Algorithm 4) is the core of our stable optimization. It replaces each of the unstable gradient steps from Algorithm 2 with a robust, decoupled counterpart.

Quantile Threshold Fitting for t: To break the oscillation-inducing cycle between the threshold t and the coverage dual variable λ, we eliminate the gradient-based update for t. Instead, t is computed deterministically by taking the $(1-\tau)$-quantile of the raw margin scores on the calibration set S_1. This non-parametric "plug-in" approach directly enforces the target coverage τ by construction, completely removing the need for the unstable dual variable λ.

Fixed-Point Update for α: The parameter α_k is theoretically tied to the group's coverage: $\alpha_k^* = K \cdot P(r^*(x) = 0, y \in G_k)$. Instead of using gradient descent, we leverage this relationship directly in a fixed-point iteration. We estimate the target $\hat{\alpha}_k$ from the current coverage and update α towards it. To prevent the collapse of α_k to zero for noisy tail groups, this update is smoothed using an Exponential Moving Average (EMA).

Grid Search for μ: The dual offset parameters μ_k are decoupled from the volatile dual ascent process. We instead perform a grid search over a pre-defined set of candidate values for μ. For each candidate, the performance is evaluated on the validation set S_2, and the best-performing value is selected. This guarantees stability by replacing a dynamic optimization with an exhaustive search.

C Additional Implementation and Experimental Details

This section provides a comprehensive overview of the implementation details for our experiments to ensure full reproducibility. We detail the training protocols for the expert models, the architecture of the gating network, and the specific hyperparameter settings used for our proposed GSE algorithms.

C.1 Expert Model Training

To ensure a fair comparison, the training of our base expert models closely follows the experimental setup of recent works in long-tail learning and selective classification [21, 25].

Network Architecture. All expert models use a ResNet-32 architecture [**?**], a standard choice for the CIFAR-100 dataset.

Training Protocol. The models are trained from scratch on the CIFAR-100-LT training set (80% of the full data). We use a Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. The models are trained for 200 epochs with a mini-batch size of 128.

Learning Rate Schedule. We use an initial learning rate of 0.1. Following the schedule in [21], the learning rate is decayed by a factor of 0.1 at the 160th and 180th epochs to ensure stable convergence.

Expert Ensemble. As stated in the main text, our ensemble consists of five experts, each trained with a different, representative long-tail learning strategy to ensure diversity:

1. Standard Cross-Entropy (CE)

2. Balanced Softmax (BS) [26]

3. Logit Adjustment (LA) [21]

Table 6: Hyperparameter settings for GSE algorithms.

Parameter	**Value**	**Description**
GSE-EG-Outer (Outer Loop)		
Outer Iterations T_{outer}	25	Number of Exponentiated Gradient outer loops [25].
EG Step Size ξ	1.0	Learning rate for updating the group weights $\boldsymbol{\beta}$ [25].
EG Momentum ρ	0.9	Momentum applied to the $\boldsymbol{\beta}$ update for smoothing.
Weight Floor β_{floor}	0.1	Minimum weight for any group to prevent collapse of head-group
Plug-In-Fit (Inner Loop)		
Inner Iterations M_{inner}	10	Number of fixed-point iterations for $\boldsymbol{\alpha}$ [25].
EMA Decay γ	0.9	Smoothing factor for the fixed-point updates of $\boldsymbol{\alpha}$.
$\boldsymbol{\mu}$ Grid Search	$\lambda \in [-2.0, 2.0]$	Range for the scalar λ used to generate $\boldsymbol{\mu}$ candidates. For $K = 2$ groups, candidates are of the form $\boldsymbol{\mu} = [\lambda/2, -\lambda/2]$.

4. Decoupling (Decouple) [16]

5. RIDE [32]

C.2 Gating Network and Posterior Mixture

The gating network is a simple Multi-Layer Perceptron (MLP) with one hidden layer of 128 units and a ReLU activation function. It takes the 512-dimensional feature vector from the penultimate layer of a ResNet-32 backbone (pre-trained on the training set) as input. The output layer has $M = 5$ units (one for each expert) followed by a Softmax function to produce the gating weights. The gating network is trained on the calibration set, $\mathcal{S}_1$, to minimize the cross-entropy loss of the final mixed posterior.

C.3 Hyperparameter Settings for GSE Algorithms

A complete list of hyperparameters used for our proposed `GSE-Balanced` and `GSE-EG-Outer` algorithms is provided in Table 6. All hyperparameters were selected based on performance on the validation set, $\mathcal{S}_2$.

The choice of these hyperparameters is critical for the stability and performance of our method. The EG parameters (ξ, ρ, β_{floor}) control the fairness optimization, while the inner loop parameters (M_{inner}, γ) ensure the stable estimation of the group-scaling factors $\boldsymbol{\alpha}$. The grid search over $\boldsymbol{\mu}$ is a key step that replaces the unstable dual ascent for the fairness multipliers, as discussed in Section 4.3.

D Theoretical Guarantees for GSE-EG-Outer

D.1 Overview

While the main paper demonstrates the empirical success of the stable algorithms, this section provides a theoretical justification for the convergence of the GSE-EG-Outer procedure. The algorithm is shown to be a principled method for solving the worst-group risk minimization problem by framing it as a zero-sum game and leveraging standard convergence guarantees from online learning theory.

D.2 Minimax Formulation of Worst-Group Risk

The problem of minimizing the worst-group selective error, subject to a coverage constraint, can be expressed as a minimax optimization problem. Let $\theta = (\alpha, \mu, t)$ denote the parameters of a rejector from the derived family of rules, and let $\mathcal{H}$ be the set of all such rejectors. The worst-group L2R risk for a given rejector (h_θ, r_θ) is defined as:

$$R_{\text{worst}}(h_\theta, r_\theta) = \max_{k \in \{1,\ldots,K\}} e_k(h_\theta, r_\theta)$$

where e_k is the selective error on group k. This max operator can be equivalently written as a maximization over the probability simplex Δ^{K-1}. Let $\beta \in \Delta^{K-1}$ be a vector of non-negative weights that sum to one. As shown in Narasimhan et al. [25], the worst-group risk can be expressed as:

$$R_{\text{worst}}(h_\theta, r_\theta) = \max_{\beta \in \Delta^{K-1}} \sum_{k=1}^{K} \beta_k e_k(h_\theta, r_\theta)$$

Therefore, the full optimization problem is to find the rejector parameters θ^* that solve the following minimax problem:

$$\min_{\theta \in \mathcal{H}} \max_{\beta \in \Delta^{K-1}} \sum_{k=1}^{K} \beta_k e_k(h_\theta, r_\theta)$$

This formulation can be interpreted as a zero-sum game between a "player" who chooses the rejector parameters θ to minimize the weighted-average error, and an "adversary" who chooses the group weights β to maximize it.

D.3 The Exponentiated Gradient Algorithm as a No-Regret Learner

The GSE-EG-Outer algorithm employs exponentiated-gradient updates in the outer loop to solve the minimax reweighting problem, a standard approach in regret minimization and online convex optimization [17]. It can be viewed as an instance of online learning where, at each iteration m, the algorithm: (1) receives a set of "experts," which are the K group-error functions $e_k(\cdot)$; (2) chooses a distribution over these experts, which is the group weight vector $\beta^{(m-1)}$; (3) plays this distribution by finding the best-response rejector $\theta^{(m)}$ that minimizes the $\beta^{(m-1)}$-weighted average of the group errors (handled by the Plug-In-Fit subroutine); (4) incurs a loss vector, which is the vector of empirical group errors $\hat{\mathrm{e}}^{(m)} = [\hat{e}_1(\theta^{(m)}), \ldots, \hat{e}_K(\theta^{(m)})]$; and (5) updates the distribution β for the next round using the Exponentiated Gradient (EG) update rule.

The EG algorithm is a "no-regret" algorithm, meaning that over time, the average loss it incurs is guaranteed to be not much worse than the loss of the best single expert (i.e., the best fixed group k) in hindsight. This property leads to convergence to the minimax value of the game.

D.4 Convergence Guarantee

Following the standard convergence analysis of the exponentiated gradient algorithm [17], we can state a formal guarantee for the algorithm, analogous to Theorem 4 in Narasimhan et al. [25].

Theorem 1 (Convergence of GSE-EG-Outer). *Let $(h^{(m)}, r^{(m)})$ for $m \in \{1, \dots, T\}$ be the sequence of rejectors generated by the GSE-EG-Outer algorithm. Let $\hat{e}_k(h, r)$ be the empirical selective error on the validation set S_2. Then, the average performance of the iterates satisfies:*

$$\max_{k \in \{1,\dots,K\}} \frac{1}{T} \sum_{m=1}^{T} \hat{e}_k(h^{(m)}, r^{(m)}) \leq \min_{(h,r)\in\mathcal{H}} \max_{k\in\{1,\dots,K\}} \hat{e}_k(h, r) + \sqrt{\frac{2\log K}{T}}$$

This theorem states that the worst-group error of the averaged predictor converges to the optimal empirical worst-group error at a rate of $O(1/\sqrt{T})$. In practice, the single best rejector found during the iterations is returned, whose performance is bounded by the average.

Proof Sketch. The proof relies on the standard regret bound for the exponentiated gradient algorithm [**?**]. The worst-group L2R risk can be written as a minimax problem over the empirical errors on the validation set S_2:

$$\min_{(h,r)\in\mathcal{H}} \max_{\beta\in\Delta^{K-1}} \sum_{k=1}^{K} \beta_k \hat{e}_k(h, r)$$

The EG algorithm guarantees that after T iterations, the following regret bound holds:

$$\frac{1}{T} \sum_{m=1}^{T} \sum_{k=1}^{K} \beta_k^{(m-1)} \hat{e}_k(h^{(m)}, r^{(m)}) - \min_{k^*} \frac{1}{T} \sum_{m=1}^{T} \hat{e}_{k^*}(h^{(m)}, r^{(m)}) \leq \sqrt{\frac{2\log K}{T}}$$

where k^* is the best single group in hindsight. By the design of the inner loop, the rejector $(h^{(m)}, r^{(m)})$ is the best response to the weights $\beta^{(m-1)}$, meaning it minimizes the weighted average error:

$$\sum_{k=1}^{K} \beta_k^{(m-1)} \hat{e}_k(h^{(m)}, r^{(m)}) = \min_{(h,r)\in\mathcal{H}} \sum_{k=1}^{K} \beta_k^{(m-1)} \hat{e}_k(h, r)$$

Combining these facts and applying properties of minimax optimization, it can be shown that the average performance converges to the minimax value. A full derivation can be found in the appendix of Narasimhan et al. [25].

Finally, the gap between the empirical risk on the validation set $\hat{e}_k$ and the true population risk e_k can be bounded using standard generalization analysis [**? ?**]. The rejectors are chosen from a class with finite capacity (as they are post-hoc adjustments to a fixed model), which ensures that convergence on the empirical risk implies convergence on the true risk, up to a generalization error term that diminishes with the size of the validation set.

E The Role of the Mixed Posterior Probability in GSE

The entire Group-Selective Ensemble (GSE) framework is predicated on a central hypothesis: that the quality of the underlying posterior probability distribution, $P(y|x)$, is the primary bottleneck for achieving fair and reliable selective classification in long-tailed settings. This section details the motivation for this hypothesis, defines the mixed posterior used in GSE, and provides a mathematical and intuitive justification for its superiority over the posteriors produced by single, monolithic models.

E.1 The Flawed Signal from a Single Generalist Model

In classical selective classification, the decision to reject is typically based on the posterior probability provided by the classifier [5]. For instance, Chow's rule abstains if the maximum posterior probability, $\max_y \eta_y(x)$, falls below a threshold. This approach implicitly assumes that the posterior is a reliable measure of the model's true confidence.

However, this assumption breaks down severely in the long-tail setting. A single neural network trained with a standard objective like cross-entropy on an imbalanced dataset learns a biased representation. The model becomes highly confident—often overconfident—on head-class samples while producing low-confidence, poorly calibrated, and uninformative posteriors for tail-class samples. This results in a flawed signal where:

Low Confidence for Tail Classes: The posterior for a correct tail-class sample is often low and spread across multiple classes, making it indistinguishable from a genuinely uncertain sample.

Poor Calibration: The model's reported confidence does not match its empirical accuracy. For tail classes, a model might report 60% confidence but only be correct 30% of the time.

Applying any rejection rule, even a theoretically optimal one like the rule derived in [25], to such a flawed posterior is fundamentally limiting. The rule is forced to make decisions based on an unreliable and biased signal, which is the core motivation for our architectural shift.

E.2 The GSE Solution: A High-Quality Mixed Posterior

The GSE framework addresses this issue at its source by replacing the single, flawed posterior with a high-quality mixed posterior probability, $\tilde{\eta}(y|x)$, generated by a committee of specialists.

Let $\{f_1, \ldots, f_M\}$ be the set of M expert models, each producing its own posterior probability distribution $p_m(y|x)$ for an input x. Let $g(x)$ be a gating network that, for the same input x, outputs a vector of weights $[g_1(x), \ldots, g_M(x)]$ such that $g_m(x) \geq 0$ and $\sum_{m=1}^{M} g_m(x) = 1$. The mixed posterior probability for class y is then defined as the weighted average of the expert posteriors:

$$\tilde{\eta}(y|x) = \sum_{m=1}^{M} g_m(x) \cdot p_m(y|x) \tag{20}$$

This mixed posterior, $\tilde{\eta}$, serves as the fundamental signal for our Group-Weighted Margin Rule. The gating network is trained to dynamically assign higher weights to the experts that are most likely to be correct for a given input, effectively creating a more robust and reliable probability distribution.

E.3 Analysis and Visualization of the Mixed Posterior

The superiority of the mixed posterior stems from its ability to leverage the diverse inductive biases of the specialist experts. An expert trained with Logit Adjustment may be better at separating tail classes, while another trained with standard Cross-Entropy may be more confident on head classes. The gating network learns to act as an intelligent router, combining these strengths.

Improved Calibration: By averaging the outputs of multiple, diverse models, the mixed posterior is often better calibrated than any single model. The overconfidence of one expert can be tempered by the uncertainty of another, leading to a more honest reflection of the true likelihood. This can be visualized with reliability diagrams.

Sharper, More Informative Distributions for Tail Classes: For a difficult tail-class sample, a single model might produce a flat, high-entropy posterior. The GSE framework, by routing the sample to the relevant expert(s), can produce a much sharper posterior with higher confidence concentrated on the correct class. This provides a much cleaner signal for the rejection rule.

In summary, the introduction of the mixed posterior is not merely an implementation detail; it is the central architectural contribution of the GSE framework. By constructing a higher-quality, better-calibrated, and more informative probability distribution, especially for challenging tail classes, it provides a solid foundation upon which our theoretically-derived, group-aware rejection rule can operate effectively. This addresses the fundamental limitation of prior work, which focused on refining the rejection rule without addressing the poor quality of the signal it was applied to.