EVIDENCE-AWARE ROBUST ENSEMBLING FOR CALIBRATED UNCERTAINTY IN FINE-TUNED FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As foundation models are increasingly deployed in high-stakes domains, their tendency to generate incorrect or low-quality content has become a major reliability concern. This risk is often amplified after fine-tuning for downstream tasks, where domain shift and limited task-specific data can degrade calibration and output quality. A common approach to uncertainty estimation is to sample multiple candidate responses through probabilistic decoding and use their disagreement as an uncertainty signal. From a Bayesian perspective, aggregating multiple generations can be interpreted as approximating marginalization over a posterior distribution over model parameters, whereas selecting a single maximum-probability output corresponds to a point estimate. However, fine-tuning on small or heterogeneous datasets can induce complex, flat, and potentially multimodal posteriors over model parameters, for which decoding-based sampling provides only limited and locally biased coverage, leading to miscalibrated uncertainty estimates. To address this limitation, we propose a diversity-inducing ensemble framework guided by Distributionally Robust Optimization (DRO) and evidential theory. It promotes complementary hypotheses that better cover disparate modes of the fine-tuned posterior, while inducing structured perturbations in the loss landscape to approximate Bayesian marginalization. Meanwhile, evidential modeling enables a principled decomposition of predictive uncertainty. We establish theoretical guarantees linking coverage of complementary posterior modes to improved uncertainty calibration. Experiments on benchmark datasets and large vision-language models show that our method consistently improves calibration and more accurately estimates content quality (correctness of generated outputs through confidence-based filtering), thereby enhancing the reliability of fine-tuned foundation models.

1 INTRODUCTION

As foundation models are increasingly deployed in real-world applications, concerns about the quality and reliability of their generated content have continued to grow, particularly in high-stakes domains. Recent work has explored incorrect output detection by quantifying model uncertainty or confidence. One common strategy estimates confidence from the conditional probabilities of generated tokens (Ren et al., 2022). However, semantically equivalent paraphrases can receive different token-level probabilities, limiting the reliability of such estimates. To mitigate this, several methods prompt models to generate explicit self-assessments of confidence, referred to as verbalized confidence (Xiong et al., 2023; Wang et al., 2024; Lin et al., 2022a). Empirical studies show that self-reported confidence estimates often exhibit systematic overconfidence, resulting in unreliable error detection (Xiong et al., 2023). Another line of work trains linear probes on intermediate embeddings to predict uncertainty scores using labeled correct and incorrect samples (Li et al., 2023a; Azaria & Mitchell, 2023; Marks & Tegmark, 2023). Yet probe-based methods are sensitive to the choice of embedding layer, depend heavily on the size and quality of the supervision data, and may overfit, thereby compromising the reliability of confidence estimation.

All of the above approaches quantify uncertainty based on a single generated output. In contrast, multi-sample methods generate multiple outputs for the same query, encouraging diverse response pathways and enabling uncertainty estimation from their variation (Bakman et al., 2024; Nikitin

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

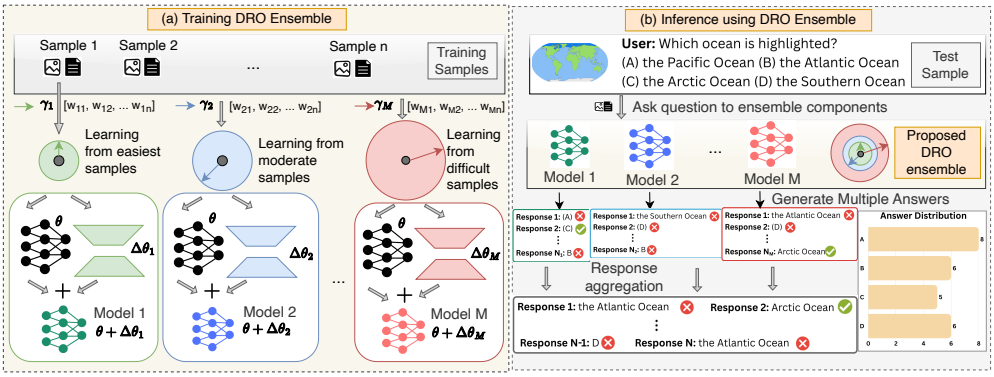


Figure 1: a) DRO ensemble training that balances easy and difficult samples based on uncertainty estimation (b) Inference that generates multiple responses to reduce overconfidence.

et al., 2024; Farquhar et al., 2024; Kuhn et al., 2023b). From this perspective, deriving a final answer can be viewed as marginalizing over multiple plausible answer paths (Wang et al., 2022). The inconsistency among these generated responses then provides a natural signal for quantifying overall uncertainty (Chen et al., 2024; Kuhn et al., 2023a; Lin et al., 2023; Bakman et al., 2024; Nikitin et al., 2024; Farquhar et al., 2024). Multiple outputs can be generated through a variety of decoding strategies, including top-k sampling, nucleus sampling, multinomial beam search, and temperature scaling. Recent studies further improve uncertainty estimation by merging semantically equivalent generations into shared answer clusters, thereby distinguishing meaningful semantic disagreement from superficial lexical variation (Farquhar et al., 2024; Kuhn et al., 2023b; Nikitin et al., 2024; Bakman et al., 2024).

Multi-output approaches have been shown to improve empirical performance in prior studies (Farquhar et al., 2024; Kuhn et al., 2023b; Xiong et al., 2023). However, their ability to quantify uncertainty can degrade substantially after fine-tuning. When a pre-trained foundation model is adapted to a downstream task, domain shift, limited task-specific data, and noisy supervision can lead to overfitting and poor calibration. As a result, uncertainty estimates produced by standard multi-output methods may become unreliable. This can be understood from a Bayesian perspective. Ideally, uncertainty estimation should account for both predictive variability and uncertainty over model parameters, yielding $p(y \mid x, D) = \int p(y \mid x, \theta) p(\theta \mid D)\, d\theta$. Standard multi-output methods approximate this marginalization only indirectly: they sample diverse response paths from the predictive distribution induced by a single fine-tuned model, rather than explicitly exploring the posterior distribution over plausible model parameters. This approximation is effective when the model is well calibrated and the posterior is highly concentrated. However, after fine-tuning, the posterior may become flatter or potentially multimodal, and a single fine-tuned model may represent only one local solution. As a result, decoder-based sampling can provide limited and locally biased coverage of the predictive space, leading to miscalibrated uncertainty estimates. Ensemble-based methods have been explored to better capture model uncertainty (Malinin & Gales, 2020), typically by initializing multiple models independently to obtain diverse parameter configurations. While this improves over sampling from a single model, random initialization alone often provides limited diversity and fails to adequately explore the space of plausible solutions after fine-tuning.

To address this limitation, we propose an *evidence-aware, diversity-inducing ensemble framework* that provides a coverage-aware surrogate to approximate Bayesian marginalization and posterior uncertainty, leading to improved uncertainty calibration for detecting incorrect outputs. Our approach is guided by principles from Distributionally Robust Optimization (DRO) and evidential theory. Specifically, we introduce a supervised fine-tuning objective that maximizes the joint likelihood of supervision tokens while encouraging ensemble members to specialize on different regions of the data distribution, as illustrated in Figure 1. More specifically, we design a DRO-based loss that adaptively reweights training samples according to their difficulty, enabling different ensemble components to focus on distinct subsets of the data. This promotes complementary hypotheses and improves coverage of diverse solution modes. From an evidential perspective, we show that the resulting objective admits a decomposition into two sources of uncertainty: prediction ambiguity and epistemic

uncertainty due to limited knowledge. This decomposition enables the ensemble to capture varying levels of difficulty reflected in the data. Further, for each ensemble member, the proposed DRO-based loss acts as a structured objective perturbation, yielding Maximum A Posteriori (MAP) estimates under a perturbed posterior that together with other ensemble components approximates Bayesian marginalization via Randomized MAP Sampling. Our contributions are threefold:

- a novel DRO-evidential training objective that promotes diversity among ensemble members, improves coverage of plausible solution modes and provides a structured surrogate for approximated Bayesian marginalization.
- a thorough theoretical analysis linking proposed formulation to improved uncertainty calibration and posterior coverage.
- consistent empirical gains in calibration and error detection across compared settings.

2 Related Work

With the increasing deployment of large language and vision-language models, understanding and improving uncertainty calibration has become a critical problem (Ling et al., 2024; Xiao & Wang, 2021; 2018; Lin et al., 2024). In classification settings, calibration is commonly measured using Expected Calibration Error (ECE) and its variants (Naeini et al., 2014; Nixon et al., 2019; Naeini et al., 2015; Kull et al., 2017). Existing approaches improve calibration either through post-hoc methods that adjust logits or probabilities (Platt et al., 1999; Guo et al., 2017; Mozafari et al., 2018; Pandey et al., 2024), or through training-time regularization techniques such as mixup, label smoothing, and focal loss (Thulasidasan et al., 2019; Müller et al., 2019; Lin et al., 2017). Ensemble methods further enhance calibration by capturing predictive uncertainty via diverse model outputs (Lakshminarayanan et al., 2017; Sapkota et al., 2023; Trinh et al., 2023).

Extending these ideas to generative models introduces additional challenges due to the open-ended nature of outputs. Existing uncertainty estimation methods for LLMs include verbalized confidence (Yang et al., 2024; Tao et al., 2024; Lin et al., 2022b; Wang et al., 2024), token-level probability-based confidence (Ren et al., 2022; Bakman et al., 2024), and supervised approaches using probes or auxiliary models (Kapoor et al., 2024; Azaria & Mitchell, 2023; Liu et al., 2024; Ulmer et al., 2024). However, these methods are often limited by overconfidence, sensitivity to prompt or representation choices, or reliance on labeled data. More recent work estimates uncertainty from multiple sampled outputs by measuring agreement or consistency across generations (Qiu & Miikkulainen, 2024; Lin et al., 2023; Farquhar et al., 2024; Kuhn et al., 2023a). For example, semantic entropy groups semantically equivalent outputs before computing entropy (Kuhn et al., 2023a), while kernel-based approaches quantify ambiguity via similarity-based distributions (Nikitin et al., 2024). While effective, these methods rely primarily on stochastic decoding to induce diversity, which may provide limited coverage of the underlying predictive uncertainty.

Our work builds on this multi-output paradigm by introducing a diversity-inducing ensemble framework guided by DRO and evidential uncertainty. Unlike prior approaches that rely on random initialization or sampling-based diversity (Malinin & Gales, 2020), our method explicitly promotes complementary hypotheses across ensemble components, leading to improved coverage of the predictive space and better-calibrated uncertainty estimates for autoregressive generation.

3 Methodology

Notation: For the generative setting, we consider a general case and define an input query containing both text and image $x_{\text{query}} = [x_{\text{image}}, x_{\text{text}}]$. Given x_{query}, the tokens generated by the model is $\boldsymbol{y} = [y_1, y_2, ...y_T]$. The selection of the tth token is based on the probability $p(y_t|x_{\text{query}}, \boldsymbol{y}_{<t})$ conditioned on the input query and previously generated tokens. The inference process results in predicted output $\hat{\boldsymbol{y}}$, along with confidence score s.

3.1 Diversity-Inducing Ensemble via Distributionally Robust Optimization

To address miscalibration caused by overfitting during fine-tuning, a common approach is to construct a vanilla ensemble by training multiple models with different random initializations. While this introduces some diversity, the resulting models are often highly correlated and may fail to adequately explore the space of plausible solutions. To systematically enhance diversity, we adopt a Distributionally Robust Optimization (DRO) framework (Duchi & Namkoong, 2019; Sapkota et al., 2023). Instead of minimizing the empirical risk under a fixed data distribution, DRO considers an adversarial

reweighting of training samples within an uncertainty set, leading to the robust objective $\mathcal{L}^{\texttt{robust}}(\theta)$.

$$\mathcal{L}^{\texttt{robust}}(\theta) = \max_{\boldsymbol{w}\in\mathcal{U}} \sum_{i=1}^{n} w_i \mathcal{L}_i(\theta), \tag{1}$$

where $\mathcal{L}_i(\theta) = -\sum_{t=1}^{T} \log p(y_t|x_{\text{query}}, \boldsymbol{y}_{<t})$ denotes the auto-regressive loss for the i-th training sample, and $\boldsymbol{w} = (w_1, \ldots, w_n)$ is a probability distribution over n training samples. The uncertainty set $\mathcal{U}$ is defined as:

$$\mathcal{U} = \left\{ \boldsymbol{w} \in \mathbb{R}^n \,\middle|\, \boldsymbol{w}^\top \mathbf{1} = 1,\ \boldsymbol{w} \geq 0,\ D_f\big(\boldsymbol{w} \,\|\, \tfrac{1}{n}\mathbf{1}\big) \leq \eta \right\}, \tag{2}$$

where $D_f(\cdot\|\cdot)$ denotes an f-divergence and η controls the size of the uncertainty set. This formulation allows the model to adaptively focus on different regions of the training distribution. When $\eta \to 0$, the uncertainty set collapses to the uniform distribution and the objective reduces to standard empirical risk minimization (ERM). As η increases, the optimization increasingly emphasizes high-loss (*i.e.,* difficult) samples, effectively shifting the training distribution toward harder regions.

To construct a diverse ensemble, we train multiple models using different values of η. Models trained with small η primarily capture dominant patterns from easy and representative samples, while those trained with larger η focus on rarer and more challenging examples. This mechanism yields a set of complementary models that specialize in different regions of the data distribution. Compared to standard ensembles based on random initialization, the proposed DRO-based ensemble explicitly controls the diversity of ensemble members through the uncertainty set. This results in complementary hypotheses that better cover different regions of the solution space, which is critical for improving calibration after fine-tuning.

3.2 Uncertainty-Aware DRO Ensemble

The DRO formulation above controls sample reweighting through the uncertainty-set parameter η. While this provides a principled way to emphasize harder samples, the induced weights are primarily determined by training loss and do not explicitly distinguish or calibrate different sources of predictive uncertainty. To bridge this gap, we introduce an evidential uncertainty-guided weighting scheme, where sample difficulty is measured through calibrated evidential uncertainty rather than loss alone. This allows the ensemble components to specialize across different uncertainty regimes, ultimately leading to more diverse and better-calibrated predictions. We introduce the theory of evidence next.

Theory of evidence: For a K-class classification problem, the theory of evidence allows the assignment of belief mass $b_k, \forall k \in \{1, 2...K\}, b_k \geq 0$, along with the uncertainty mass $u, u \geq 0$ (Jøsang, 2016; Sensoy et al., 2018), satisfying $\sum_{k=1}^{K} b_k + u = 1$. Belief mass b_k represents the portion of total certainty assigned to class k, while uncertainty mass u represents the remaining belief, capturing ignorance due to lack of evidence. The model expresses a *lack of knowledge* (also referred to as vacuity) when belief masses are uniformly low. In contrast, confusion between multiple labels is reflected by the total belief assigned to incorrect classes. Therefore, evidential uncertainty naturally decomposes into two components: lack of knowledge (u) and predictive confusion (b^{inc}). For a model $f(\theta)$, these quantities are computed as:

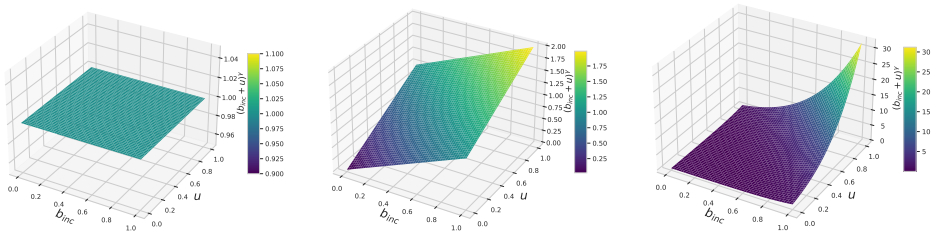
$$b_k = \frac{e_k}{\sum_{k=1}^{K} e_k + K}, \quad u = \frac{K}{\sum_{k=1}^{K} e_k + K}, \quad e_k = \mathcal{E}(f_\theta(y_k|x)), \quad b^{inc} = \sum_{k\neq gt} b_k. \tag{3}$$

$\mathcal{E}$ is a non-negative transformation (e.g., ReLU) to convert logit of a class $f_\theta(y_k|x)$ to evidence.

Evidential uncertainty-guided weight assignment: To compute uncertainty for a generated output $\boldsymbol{y}$, we denote the logits at the tth generation step as $f_\theta(\boldsymbol{y}_t|x_{\text{query}}, \boldsymbol{y}_{<t}) \in \mathbb{R}^V$. We leverage fine-grained evidential uncertainties (u and b^{inc}) to guide the training of ensemble components. In the generative setting, we define token-level evidence $\boldsymbol{e}_{tv} = \mathcal{E}(f_\theta(\boldsymbol{y}_{tv}|x_{\text{query}}, \boldsymbol{y}_{<t}))$, and compute:

$$b_{tv} = \frac{\boldsymbol{e}_{tv}}{V' + \sum_{v=1}^{V'} \boldsymbol{e}_{tv}}, \quad u_t = \frac{V'}{V' + \sum_{v=1}^{V'} \boldsymbol{e}_{tv}}, \quad b_t^{inc} = \sum_{v\neq y_t} b_{tv}, \tag{4}$$

Because evaluating over the full vocabulary ($V \approx 32,000$) causes uncertainty to saturate near maximum ($u_t \approx 1$), we dynamically set $V' < V$ to the number of tokens with positive logits at each step. This ensures the computed uncertainty reflects actual prediction difficulty rather than vocabulary



(a) $\gamma = 0$ (b) $\gamma = 1$ (c) $\gamma = 5$

Figure 2: Visualization of weight of the loss with change of b_{inc} and u for different values of γ.

size. To obtain a sample-level difficulty measure, we aggregate uncertainty across generation steps by focusing on the dominant uncertainty signal:

$$u = \max_{t=1}^{T} u_t, \quad b^{inc} = \max_{t=1}^{T} b_t^{inc}. \tag{5}$$

We use the max operator because the overall correctness of a reasoning chain is often bottlenecked by its most uncertain logical step. We then define the sample difficulty weight and robust loss:

$$w_i = (u + b^{inc})^{\gamma}, \quad \mathcal{L}^{\texttt{robust}} = \sum_{i=1}^{n} w_i \mathcal{L}_i(\theta) \tag{6}$$

where γ controls the emphasis on uncertain samples. This formulation serves as a structured, uncertainty-aware surrogate for the sample weighting in DRO, where samples with high and well-calibrated uncertainty are emphasized during training. Specifically, the exponent γ in Equation 6 provides an explicit, closed-form parameterization of the uncertainty set radius η defined in Section 3.1. Rather than dynamically solving the implicit dual optimization problem at each training step, γ directly governs the sensitivity to uncertainty: $\gamma = 0$ corresponds to standard empirical risk minimization ($\eta \to 0$), while higher values of γ enforce stronger worst-case robustness by concentrating weight on high-uncertainty instances. The effect of different γ values on uncertainty aware weight w_i is illustrated in Figure 2: Figure 2(a) $\gamma = 0$ allocates equal weight w_i across all uncertainty values of b_{inc} and u, Figure 2(b-c): as γ increases, weight w_i increases with increase in uncertainties: b_{inc} and u. By training ensemble components with $\{\gamma_m\}_{m=1}^{M}$, we obtain M models that focus on different uncertainty regimes, yielding a diverse and complementary ensemble. To ensure computational efficiency, all ensemble components share a single frozen base model θ and are fine-tuned using parameter-efficient Low-Rank Adapters (LoRA) $\{\Delta\theta^{(m)}\}_{m=1}^{M}$, incurring negligible parameter and memory overhead compared to full fine-tuning.

Inference: During inference, we draw N/M outputs per model, yielding a set of N total generations $\mathcal{Y}$. To distinguish true semantic disagreement, we group semantically equivalent outputs ($\boldsymbol{y} \equiv \tilde{\boldsymbol{y}}_k$) into K unique candidate answers $\tilde{\mathcal{Y}} = \{\tilde{\boldsymbol{y}}_k\}_{k=1}^{K}$. The marginal probability for each unique answer $p(\tilde{\boldsymbol{y}}_k \mid x_{\text{query}})$ (fraction of the counts of answers that are semantically equivalent to the $\tilde{\mathbf{y}}_k$ out of all the answers in $\mathcal{Y}$), along with the final ensemble prediction $\hat{\boldsymbol{y}}$ and confidence score s, are derived as:

$$p(\tilde{\boldsymbol{y}}_k \mid x_{\text{query}}) \approx \sum_{\boldsymbol{y} \in \mathcal{Y}} \frac{|\boldsymbol{y} \equiv \tilde{\boldsymbol{y}}_k|}{|\mathcal{Y}|}, \quad \hat{\boldsymbol{y}} = \arg\max_{k \in K} p(\tilde{\boldsymbol{y}}_k \mid x_{\text{query}}), \quad s = \max_{k \in K} p(\tilde{\boldsymbol{y}}_k \mid x_{\text{query}}). \tag{7}$$

3.3 Theoretical Analysis

In this section, we theoretically analyze how the proposed evidential uncertainty-guided ensemble improves calibration and enforces mode separation among ensemble members. The use of DRO to improve calibration through ensemble diversity has precedent beyond our setting. DRE (Sapkota et al., 2023) constructs a DRO-guided ensemble of sparse sub-networks using uncertainty sets to make members specialize from easy to difficult samples, and shows this improves calibration in the sparse-network classification setting. We adopt the same underlying principle, DRO-guided ensemble diversity as a route to calibration, and make a novel extension to the generative setting, guided by evidential uncertainty. Our calibration analysis builds on the connection between sample reweighting and uncertainty-aware learning. We begin by noting that focal loss is a widely used mechanism for emphasizing difficult samples by assigning weights of the form $(1 - \hat{p})^{\gamma}$, where $\hat{p}$ is the predicted probability for the ground-truth label (Mukhoti et al., 2020). While effective, focal loss defines

difficulty purely in terms of prediction confidence, without distinguishing the underlying sources of uncertainty. In contrast, our approach defines sample difficulty through evidential uncertainty, which decomposes predictive uncertainty into two components: lack of knowledge (vacuity) and predictive confusion (incorrect belief). We show that this formulation naturally generalizes focal loss by providing a principled decomposition of its weighting scheme. The following lemma establishes that the focal loss weight can be expressed as a function of evidential uncertainty components.

Lemma 3.1. *Let $w_i = (1 - \hat{p}_{iv})^\gamma$ be the focal loss weight for the ith sample, where $\hat{p}_{iv}$ is the predicted probability for the ground-truth token v. Under the evidential formulation, this weighting can be expressed as a function of vacuity and incorrect belief:*

$$w_i = (\hat{u}_i + b_i^{inc})^\gamma \tag{8}$$

where $\hat{u}_i = u_i - \frac{1}{S_i}$ is the adjusted vacuity, u_i is the uncertainty mass, $S_i = \sum_{v=1}^{V} e_{iv} + 1$ is the total evidence strength, and b_i^{inc} is the total belief assigned to non-ground-truth tokens.

This lemma reveals that focal loss implicitly combines two distinct sources of uncertainty. The first component, vacuity, captures lack of knowledge due to insufficient evidence, while the second component, incorrect belief, reflects confusion among competing predictions. Unlike focal loss, which treats all low-confidence predictions uniformly, the evidential formulation explicitly distinguishes these two cases. As a result, the proposed weighting scheme provides a more structured and interpretable notion of sample difficulty. In particular, increasing γ emphasizes samples with high vacuity, encouraging the model to learn from under-represented or uncertain regions of the data. Simultaneously, samples with high incorrect belief are also emphasized, promoting correction of confidently wrong predictions. This dual emphasis prevents the model from becoming overconfident on incorrect samples, leading to improved calibration.

Theorem 3.2. *Let $\mathcal{L}^{EU\text{-}Ensemble}(\Theta)$ denote the evidential uncertainty-guided ensemble loss and $\mathcal{L}^{ERM}(\theta)$ the standard autoregressive loss. The proposed loss admits the following lower bound:*

$$\mathcal{L}^{EU\text{-}Ensemble}(\Theta) \geq \sum_{m=1}^{M} \left[\mathcal{L}_m^{ERM}(\theta) - \gamma_m \mathcal{H}^m[\hat{p}] + \gamma_m C\right], \tag{9}$$

where $\mathcal{H}^m[\hat{p}]$ denotes the entropy of the predictive distribution, C is a normalization term, and γ_m controls the strength of uncertainty weighting.

Theorem 3.2 demonstrates that minimizing our ensemble loss drives down a Right-Hand Side (RHS) composed of cross-entropy ERM loss and a negative-entropy term $-\gamma_m \mathcal{H}^m[\hat{p}]$. This formulation exposes a direct trade-off between minimizing empirical risk and maximizing predictive entropy. For samples where the ERM loss cannot be further reduced (e.g., difficult or ambiguous examples), the only way to decrease the RHS is to increase the entropy $\mathcal{H}^m[\hat{p}]$, discouraging the model from making highly confident wrong predictions. Conversely, on easy samples where the ERM loss can be driven close to zero, the bound imposes little pressure on entropy, so the model remains confident.

The parameter γ_m dictates this balance. For confidently wrong samples, Eq. 8 assigns larger weights when γ_m is high, thereby strengthening the entropy regularization effect and reducing overconfident predictions. In contrast, smaller values of γ_m place less emphasis on uncertain or difficult samples, making the model behave more similarly to standard ERM and allowing higher-confidence predictions. The confidence trend for small and large values of γ_m is empirically illustrated in Figure 3. Ensemble members trained with different γ_m values exhibit complementary accuracy-confidence behaviors, resulting in accurate and calibrated predictions.

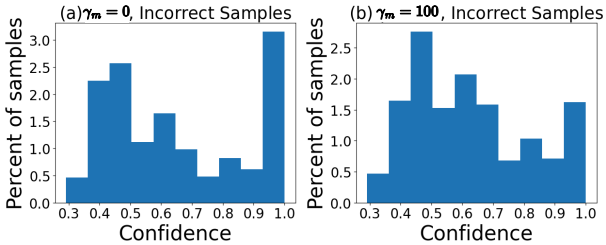


Figure 3: Effect of γ_m. As shown for the lower γ_m value i.e., $\gamma = 0$ value the confidences shifted right compared to the one with relatively higher γ_m value i.e., $\gamma_m = 100$.

Posterior Mode Coverage: Assume a pre-trained LVLM with parameters θ. Performing fine-tuning of the model with parameter-efficient fine-tuning approaches results in updated parameters

$\theta' = \theta + \Delta\theta$. Following Bayes' theorem (where *posterior* $\propto$ *prior* $\times$ *likelihood*), using $P(\theta)$ as a prior results in a posterior distribution $p(\theta|D)$ by finetuning with Dataset D. Ideally, obtaining the predictive distribution requires marginalizing over this parameter posterior: $p(\mathbf{y}|x_{\text{query}}, D) = \int p(\mathbf{y}|x_{\text{query}}, \theta), p(\theta|D), d\theta$.Because $p(\theta|D)$ is intractable, standard deep ensembles approximate this integral using an ensemble $\{\theta^{(m)}\}_{m=1}^{M} \sim q(\theta)$ initialized randomly. While this provides some coverage of the distribution, solving the same optimization ($\mathcal{L}^{ERM}$) limits the ensemble models from exploring the space of plausible solutions. In addition to random initialization, we increase the coverage of the posterior distribution by providing each member more freedom to explore plausible solutions via γ_m under $\mathcal{L}^{\text{robust}}$. Next, we formalize how varying γ_m across ensemble components enforces mode separation of ensemble members in parameter space relative to the standard ERM baseline $\theta_0 = \arg\min_\theta \mathcal{L}^{\text{ERM}}(\theta)$.

Lemma 3.3. *Let $\hat{\theta}^{(m)} = \arg\min_\theta \mathcal{L}_m^{robust}(\theta)$ denote the local Maximum A Posteriori (MAP) estimate for ensemble member m under exponent γ_m, and let $\theta_0 = \arg\min_\theta \mathcal{L}^{ERM}(\theta)$ denote the unperturbed ERM baseline optimum ($\gamma = 0$). Under a local second-order expansion around θ_0, the parameter distance between any two ensemble members m and m' is bounded below by:*

$$\left\|\hat{\theta}^{(m)} - \hat{\theta}^{(m')}\right\|_2 \;\geq\; \frac{|\gamma_m - \gamma_{m'}|}{\lambda_{\max}(\mathbf{H}_{ERM})} \left\| \sum_{i=1}^{n} f(u_i, b_i^{inc}) \nabla_\theta \mathcal{L}_i(\theta_0) \right\|_2 \tag{10}$$

where $\mathbf{H}_{ERM}$ is the Hessian evaluated at the ERM baseline θ_0, $\lambda_{\max}(\cdot)$ is its maximum eigenvalue, and $f(u_i, b_i^{inc})$ is the non-negative evidential difficulty factor.

Lemma 3.3 demonstrates that training under $\mathcal{L}^{\text{robust}}$ overcomes the limited coverage of plausible solutions by applying a structured objective perturbation via γ_m that guarantees a lower bound on pairwise parameter distance. Crucially, this mode separation depends not only on the gap $|\gamma_m - \gamma_{m'}|$, but is explicitly scaled by the gradient magnitude of an evidential difficulty function. Data points characterized by high vacuity or incorrect belief thus actively push parameter trajectories further apart along task-critical gradient directions $\nabla_\theta \mathcal{L}_i(\theta_0)$. By leveraging this structured perturbation, our framework connects to Randomized MAP Sampling (RMS) (Pearce et al., 2020), where training ensemble members over perturbed optimization landscapes yields MAP estimates under an implicit, perturbed posterior $\hat{\theta}^{(m)}$. Rather than relying solely on random initialization, allowing each member a distinct slice of sample difficulty via γ_m provides the ensemble with freedom to explore complementary, plausible posterior modes. This serves as a principled surrogate for approximated Bayesian marginalization, ultimately ensuring broader posterior coverage and superior predictive calibration.

3.4 Adaptive Sampling for Cost Reduction

Estimating confidence with multi-output methods requires generating N answer samples for the same query, which can substantially increase inference cost. Empirically, we observe that calibration generally improves as N increases, but many queries reach stable confidence estimates before all N samples are generated. To reduce unnecessary computation, we propose an adaptive sampling strategy that uses a query-specific number of samples $N_{\text{adaptive}} < N$ while preserving calibration performance. For each query, we generate answer samples sequentially and update the confidence score after each new sample. At iteration j, let s_j denote the current confidence estimate computed from the first j samples. We monitor the confidence change between consecutive iterations:

$$\Delta s_j = |s_j - s_{j-1}|. \tag{11}$$

Sampling stops when either the maximum sampling budget N is reached or the confidence estimate becomes stable, i.e., $\Delta s_j \leq \epsilon$, where ϵ controls the stopping criterion. The final number of generated samples is denoted as N_{adaptive}. Intuitively, early stopping occurs when additional samples no longer substantially change the estimated answer distribution, indicating that the model has reached a stable confidence estimate for the query. If the confidence remains unstable, the method continues sampling until the full budget N is exhausted. The proposed adaptive sampling strategy is agnostic to the underlying multi-output generation method and can therefore be applied to both standard decoding-based sampling and our ensemble-based sampling framework.

4 Experiments

Datasets, models, and evaluation metrics. We conduct experiments on three Multiple Choice Questions (MCQ) datasets: ScienceQA (Lu et al., 2022), CommonsenseQA (Talmor et al., 2019) and

Table 1: Performance Comparison on three benchmark datasets and two LVLMs

Baseline	CommonsenseQA			Commonsense Reasoning			ScienceQA			TriviaQA		
	ECE (↓)	AUROC (↑)	Acc. (↑)	ECE (↓)	AUROC (↑)	Acc. (↑)	ECE (↓)	AUROC (↑)	Acc. (↑)	ECE (↓)	AUROC (↑)	Acc. (↑)
Model: gemma-3-4b-it												
Least Probability	0.262	60.294	81.980	0.500	52.430	73.540	0.285	63.880	83.447	0.087	72.387	84.840
Product Probability	0.464	62.240	81.980	0.690	54.980	73.540	0.366	62.178	83.447	0.083	72.516	84.840
Perplexity	-	63.060	81.980	-	54.930	73.540	-	62.178	83.447	-	72.294	84.840
Lexical Similarity	-	61.010	82.470	-	53.380	81.320	-	67.480	82.100	-	73.830	83.150
Semantic Entropy (beam)	0.108	69.220	82.470	0.117	71.510	82.470	0.091	85.690	83.560	0.180	73.485	82.010
Multi-output (top-k)	0.132	70.737	82.473	0.046	79.561	82.473	0.059	87.410	83.490	0.079	76.043	85.133
Multi-output (temp)	0.151	63.578	82.391	0.057	80.953	82.801	0.082	84.368	83.046	0.102	74.429	84.775
Multi-output (nucleus)	0.146	67.884	81.572	0.031	80.965	82.228	0.063	85.414	83.565	0.092	73.477	84.775
KLE	0.146	67.020	81.572	0.031	68.870	82.228	0.063	70.220	83.565	0.092	71.170	84.775
Degree	0.146	67.010	81.572	0.031	71.060	82.228	0.063	69.460	83.565	0.092	76.140	84.775
Semantic Density	0.146	67.880	81.572	0.031	67.020	82.228	0.063	68.440	83.565	0.092	76.110	84.775
MARS	0.146	65.900	81.572	0.031	63.570	82.228	0.063	65.860	83.565	0.092	56.770	84.775
Vanilla Ensemble	0.099	77.929	82.555	0.048	81.945	82.801	0.054	89.850	83.540	0.070	73.780	86.750
Proposed Ensemble	**0.051**	**82.445**	83.047	**0.017**	**82.568**	82.801	**0.040**	**90.375**	84.485	**0.063**	**76.856**	87.085
Model: LLaVA-OneVision-Qwen2-0.5B												
Least Probability	0.168	65.503	63.390	0.420	52.710	58.390	0.108	75.980	75.147	0.141	74.866	65.419
Product Probability	0.168	67.020	63.390	0.569	53.027	58.390	0.108	75.980	75.147	0.204	76.490	65.419
Perplexity	-	67.029	63.390	-	53.024	58.390	-	75.223	75.147	-	76.424	65.419
Lexical Similarity	-	65.940	63.390	-	53.350	58.720	-	68.183	76.161	-	68.130	61.290
Semantic Entropy (beam)	0.317	65.336	63.390	0.359	61.460	58.720	0.187	68.194	76.160	0.120	64.448	59.466
Multi-output (top-k)	0.195	71.307	62.817	0.117	72.107	57.330	0.103	75.810	74.794	0.083	78.000	63.500
Multi-output (temp)	0.241	68.006	63.718	0.190	71.541	58.313	0.121	72.610	75.925	0.087	77.544	64.281
Multi-output (nucleus)	0.156	71.521	63.800	0.085	72.158	56.921	0.089	76.465	75.218	0.084	76.966	64.769
KLE	0.156	69.770	63.800	0.085	58.558	56.920	0.089	63.080	75.218	0.084	71.660	64.769
Degree	0.156	72.060	63.800	0.085	68.630	56.920	0.089	72.410	75.218	0.084	76.840	64.769
Semantic Density	0.156	69.800	63.800	0.085	59.550	56.920	0.089	63.390	75.218	0.084	74.840	64.769
MARS	0.156	68.240	63.800	0.085	61.580	56.920	0.089	74.050	75.218	0.084	71.210	64.769
Vanilla Ensemble	0.170	72.881	62.735	0.065	71.165	57.821	0.078	76.757	75.595	0.077	77.299	64.443
Proposed Ensemble	**0.040**	**73.189**	63.145	**0.048**	**72.338**	57.903	**0.051**	**78.921**	74.534	**0.067**	**78.287**	64.573

Note: KLE, Degree, Semantic Density, and MARS share identical Accuracy and ECE because both metrics depend solely on the shared self-consistency predictions, whereas AUROC evaluates their distinct uncertainty-scoring methods.

Commonsense Reasoning (Rajani et al., 2019) and open-ended QA dataset: TriviaQA (Joshi et al., 2017). Although we consider 3 MCQ datasets, the number of tokens in output varies (ScienceQA: $\approx$ 200 tokens, CommonsenseQA: $\approx$ 1 tokens, and Commonsense reasoning: $\approx$ 10 tokens on average), and answer retrieval is a key part to obtain the most consistent answer. Details about datasets and answer retrieval are in the Appendix.

We finetune two LVLMs for our experiments: LLaVA-OneVision-Qwen2-0.5B (Li et al., 2024) and gemma-3-4b-it (Google, 2024) with a train split of each of the datasets. For the evaluation of calibration, we leverage Expected Calibration Error (ECE) as a metric. For fault tolerance performance, we leverage AUROC, which measures the area under the curve for a binary classification problem of detecting correct/incorrect test generation using confidence score. We also report the accuracy of each comparison. More details on experiments are presented in the Appendix.

Comparison baselines. We compare with both single and multi-output baselines. For multi-output methods, we include semantic entropy (Kuhn et al., 2023a), lexical similarity (Fomicheva et al., 2020), KLE (Nikitin et al., 2024), degree (Lin et al., 2023), semantic density (Qiu & Miikkulainen, 2024) and MARS (Bakman et al., 2024) for comparison. We also leverage decoding strategies to generate N samples (top-k sampling, nucleus sampling (Holtzman et al., 2019), and temperature sampling) as baselines. We include vanilla ensemble as a baseline. Similarly, for a single-output method, we include the least probability, product probability (Ren et al., 2022), and perplexity as baselines. Details about the baselines are included in the Appendix. As perplexity and lexical similarity do not provide probability outputs, we do not report ECE for these baselines.

4.1 Calibration and Fault Tolerance Performance

We present the performance of the proposed DRO ensemble and baseline methods in Table 1. Across three different datasets and 2 models of varying range, we observe that the calibration (ECE) and fault tolerance performance (AUROC) of the proposed method are superior in comparison to baselines while maintaining accuracy and inference compute ($\mathcal{O}(N)$). In general, it is seen that multi-output-based methods demonstrate competitive performance. An ensemble of models trained with different seeds (vanilla ensemble baseline) shows improvement over other baselines across multiple settings. The proposed ensemble further enhances the diversity among ensemble components and outperforms the vanilla ensemble method.

Adaptive Sampling Performance: In this section, we present the ECE results of the proposed adaptive function in two datasets using the Gemma3-4b-it model. Along with the proposed DRO Ensemble, adaptive sampling can be leveraged with other multi-output baselines as well. The results for gemma-3-4b-it are presented in Table 2.

Table 2: Performance of adaptive sampling

Baseline	ϵ	CommonsenseQA % Samples	CommonsenseQA ECE (↓)	Commonsense Reasoning % Samples	Commonsense Reasoning ECE (↓)
Top-k	-	100.00	0.132	100.00	0.046
Top-k + Adaptive	0.1	42.162	0.137	50.041	0.097
Top-p	-	100.00	0.146	100.00	0.031
Top-p + Adaptive	0.1	42.129	0.150	48.010	0.089
Vanilla Ensemble	-	100.00	0.099	100.00	0.048
Vanilla Ensemble + Adaptive	0.1	44.341	0.112	48.485	0.087
DRO Ensemble	-	100.00	0.051	100.00	0.017
DRO Ensemble + Adaptive	0.1	46.716	0.067	50.762	0.029
DRO Ensemble + Adaptive	0.01	61.556	0.066	74.758	0.024

The row with 100% samples is the non-adaptive version of each method. For the proposed DRO ensemble, we show the calibration performance with two epsilon values 0.1 and 0.01. When we use $\epsilon = 0.1$, with only 46% of the total samples, we can achieve competitive calibration performance. Similarly, with $\epsilon = 0.01$, with 61% of the total samples, we can further improve calibration performance. The adaptive sampling method can be applied to other sampling and vanilla ensemble methods. For the other baselines, we observe similar competitive calibration performance while reducing the samples.

4.2 Ablation Study

Table 3: Ablation Study

Setting	ECE (↓)
Multi-output (nucleus)	0.085
Vanilla Ensemble	0.065
DRO w/o Uncertainty	0.057
Proposed Ensemble	**0.048**

Impact of individual components: We conduct an ablation to evaluate the performance gain from individual components. In particular, we compare the calibration performance of (i) vanilla ensemble, (ii) DRO ensemble without evidential uncertainty, and (iii) the proposed method (DRO + Evidential) for Commonsense Reasoning with LLaVA-OneVision-Qwen2-0.5B in Table 3. We observe that using a vanilla ensemble of models obtains improved calibration performance of the multi-output method. The standard DRO-based ensemble shows further improvement. DRO guided by evidential uncertainty demonstrates the best calibration performance (0.048).

Table 4: Impact of γ

Setting	ECE (↓)	Acc. (↑)
$\gamma = 1$	0.160	63.309
$\gamma = 2$	0.165	63.554
$\gamma = 5$	0.156	63.391
$\gamma = 50$	0.097	61.998
$\gamma = 100$	0.044	50.942
$\gamma \in \{1, 2, 5\}$	0.161	63.145
$\gamma \in \{1, 2, 100\}$	0.056	63.473
$\gamma \in \{1, 50, 100\}$	0.043	62.817
$\gamma \in \{2, 50, 100\}$	0.043	62.244
$\gamma \in \{5, 50, 100\}$	0.040	63.145

Impact of γ: We study the impact of γ on the accuracy and calibration performance of individual models and the proposed DRO ensemble. The study conducted on the CommonsenseQA dataset and Llavaonevision model is presented in Table 4. When we use different values of γ, we observe that higher values of gamma result in a well-calibrated lower accuracy model, as evidenced by the result of $\gamma = 100$. Increasing the value of γ allows the model to focus on more difficult samples, resulting in inaccurate results for easy samples. Our theoretical analysis suggests that the ensemble approach that focuses on a diverse range of difficulty results in more calibrated models. This is also observed from experiments. If we ensemble similar values of $\gamma \in \{1, 2, 5\}$, the calibration performance is still poor. However, as we include $\gamma = 100$ in the ensemble, the performance significantly improves while maintaining the accuracy. The γ values of $\{5, 50, 100\}$ result in the most optimal performance. The inclusion of diverse values of γ also shows competitive performance. This demonstrates the importance of γ for the proposed DRO ensemble.

4.3 Additional Results

Due to space constraints, we present additional results in the Appendix. Impact study of sample size per query N, hyperparameters: V', ϵ, calibration error metric, external model and alternative difficulty scores for proposed ensemble formulation are presented in Appendix F.1. To demonstrate the robustness properties of the proposed method, we conduct experiments by fine-tuning the model on limited and long-tailed data, which are presented in Appendix F.2. Experiments on additional vision-language models are presented in Appendix F.3. The comparison of adaptive sampling in the Llava model is presented in Appendix F.4. Qualitative analysis of demonstrating the suppression of confidence on incorrect samples from the proposed method is presented in Appendix F.7.

5 Conclusion

In this work, we study the reliability of fine-tuned foundation models from the perspective of uncertainty calibration. We propose a Bayesian-inspired, diversity-inducing ensemble framework guided by Distributionally Robust Optimization and evidential theory. By using evidential uncertainty to guide sample reweighting, our method encourages ensemble components to specialize across different uncertainty regimes, leading to more diverse and better-calibrated predictions. We further provide theoretical analysis showing how the proposed objective promotes calibration through an entropy-regularized formulation and how proposed formulation enforces mode separation of ensemble members. Finally, we introduce an adaptive sampling strategy to reduce inference cost while maintaining competitive calibration performance. Extensive experiments demonstrate that our method consistently improves calibration and incorrect-output detection over strong baselines.

AI USAGE

In this work, the use of generative AI tools for synthetic data generation, cleaning and reformatting dataset, qualitative and thematic data analysis are not applicable. LLM tools have been utilized to polish writing, not for generating ideas, or methods or theoretical contribution. Tools were used to help debug Python errors while implementing methodolology, and latex errors while creating the plots, and tables for the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our experimental results, we provide details across the paper and Appendix. Experimental details, hyperparameters, decoding settings, and hardware specifications are described in Section 4 and Appendix. Descriptions of the datasets, prompt templates, and evaluation metrics used in our experiments are provided in Section 4 and Appendix. Furthermore, source code containing our training pipelines, and evaluation code is linked anonymously in Appendix.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7752–7767, 2024.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Sebastian Farquhar, Jannik Kossen, Sebastian Farquhar, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 2024.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Google. Gemma: Introducing new state-of-the-art open models. *https://blog.google/technology/developers/gemma-open-models/*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QzvWyggrYB.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, abs/2302.09664, 2023a. URL https://api.semanticscholar.org/CorpusID:257039062.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023b.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Feng Li. Repository to fine tune llava-one vision models. https://github.com/LLaVA-VL/LLaVA-NeXT, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words, 2022b. URL https://arxiv.org/abs/2205.14334.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL https://arxiv.org/abs/2305.19187.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. Uncertainty quantification for in-context learning of large language models, 2024. URL https://arxiv.org/abs/2402.10189.

Xin Liu, Farima Fatahi Bayat, and Lu Wang. Enhancing language model factuality via activation-based confidence calibration and guided decoding, 2024. URL https://arxiv.org/abs/2406.13230.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*, 2018.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Binary classifier calibration: Non-parametric approach. *arXiv preprint arXiv:1401.3390*, 2014.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Deep Shankar Pandey, Spandan Pyakurel, and Qi Yu. Be confident in what you know: Bayesian parameter efficient fine-tuning of vision foundation models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, pp. 234–244. PMLR, 2020.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Hitesh Sapkota, Dingrong Wang, Zhiqiang Tao, and Qi Yu. Distributionally robust ensemble of lottery tickets towards calibrated sparse network training. *Advances in Neural Information Processing Systems*, 36:62657–62681, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to trust llms: Aligning confidence with response quality, 2024. URL https://arxiv.org/abs/2404.17287.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Trung Q Trinh, Markus Heinonen, Luigi Acerbi, and Samuel Kaski. Input gradient diversity for neural network ensembles. *CoRR*, 2023.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only, 2024. URL https://arxiv.org/abs/2403.05973.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yu won Lee. Repository to fine tune gemma models. https://github.com/2U1/Gemma3-Finetune, 2024.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. *CoRR*, abs/1811.07253, 2018. URL http://arxiv.org/abs/1811.07253.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. *CoRR*, abs/2103.15025, 2021. URL https://arxiv.org/abs/2103.15025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms, 2024. URL https://arxiv.org/abs/2412.14737.

Supplementary Material

TABLE OF CONTENTS

A Organization of Appendix **15**

B Summary of Notations **15**

C Connection between DRO, evidential uncertainty, posterior mode coverage, and calibration **15**

C.1 Proposed formulation: DRO + evidential uncertainty 15

C.2 Proposed Formulation and Posterior Mode Coverage 16

C.3 Proposed Formulation and Predictive Calibration 16

C.4 Summary 17

D Mathematical Proofs **17**

D.1 Proof of Lemma 3.1 17

D.2 Proof of Theorem 3.2 18

D.3 Proof of Lemma 3.3 18

E Details of Experiments **20**

F Additional Experiments **21**

F.1 Ablation Studies 21

F.2 Handling limited and long-tail data distribution 24

F.3 Results on additional models 25

F.4 Additional Results On Adaptive Sampling 25

F.5 Results on POPE dataset 26

F.6 Comparison with post-hoc calibration methods 26

F.7 Qualitative Analysis 28

F.8 Analysis of lower accuracy 28

G Limitations **30**

H Impact Statement **30**

I Source Code **30**

J Computational Cost Analysis **30**

A Organization of Appendix

In this Appendix, we first summarize the major notations used in our paper in Section B. We present connection of proposed components (DRO, evidential uncertainty, posterior mode coverage and calibration) in Section C. We provide the detailed mathematical proofs for our theoretical contributions in Section D. Next, we provide the details of the experiment in Section E and additional experiments in Section F. We discuss limitations in Section G. We provide the link to the code in Section I.

B Summary of Notations

Table 5 summarizes the major notations used in our paper.

Table 5: Symbols with Descriptions

Symbol Group	Notation	Description
Autoregressive data generation	$\{x_{\text{query}}, \mathbf{y}\}$	Input output pair
	x_{image}	Image input
	x_{text}	Text input
	y_t	t-th token
	T	total tokens in $\mathbf{y}$
	N	Total number of sampled output
	$\mathcal{Y}$	set of N outputs
	K	Total classes
	$\tilde{\mathcal{Y}}$	Set of K answers obtained by clustering
	$\boldsymbol{y}_k$	k^{th} answer
	$p(\tilde{\boldsymbol{y}}_k \mid x_{\text{query}})$	Probability of class k
	s	Maximum probability among K classes
Proposed Ensemble	$\mathcal{L}_i(\Theta)$	Auto-regressive loss of ith training sample
	$\mathcal{L}^{robust}$	Proposed DRO Loss
	w_i	Difficulty weight of ith training sample
	$\boldsymbol{w}$	weight distribution for n training samples
	$\mathcal{E}$	Non-negative transformation function
	V	Vocabulary size
	$\boldsymbol{e}_{tv}$	Evidence of vth in t^{th} generation
	b_{tv}	Belief of vth in t^{th} generation
	u_t	Uncertainty mass of tth generation
	b_t^{inc}	Incorrect belief of tth generation
	u	Uncertainty mass of generation
	b^{inc}	Incorrect belief of generation
	γ	Hyperparameter to control w_i
Adaptive Sampling	$N_{adaptive}$	Number of adaptive sampled output
	ϵ	Parameter to control convergence
	Δs_j	Confidence change between between consecutive iterations

C Connection between DRO, evidential uncertainty, posterior mode coverage, and calibration

The following demonstrates the precise connection between the four pieces of the contribution. In C.1, we summarize the proposed formulation from the perspective of DRO and evidential uncertainty. In C.2, we provide the connection of the proposed formulation with posterior mode coverage. In C.3, we provide the coneection of the proposed formulation with calibration.

C.1 Proposed formulation: DRO + evidential uncertainty

Equations 1–2 define the DRO objective as a worst-case reweighting of training samples within an f-divergence uncertainty set. The difficulty weights assigned to training samples are defined by w,

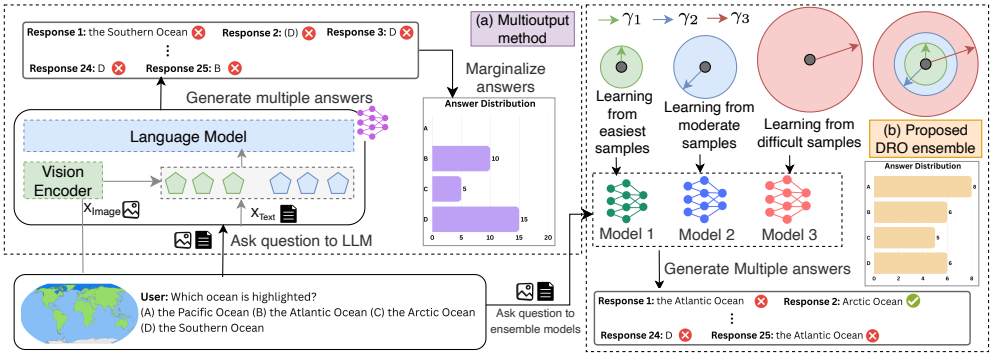


Figure 4: Comparative approaches for generating answer distributions. (a) A multi-output architecture utilizing a single fine-tuned model. (b) An ensemble-based approach utilizing Distributionally Robust Optimization (DRO).

which is primarily defined using training loss. Sample loss is a measure of the difficulty of the model in learning the specific training sample. In our proposed method, evidential uncertainty quantifies the training difficulty in a fine-grained manner through lack of knowledge and predictive confusion. In Lemma 3.1, we show how a difficulty score only dependent on probability $(1 - \hat{p})$ can be expressed through fine-grained evidential uncertainties, making the connection more precise.

C.2 Proposed Formulation and Posterior Mode Coverage

In generative LVLM tasks, obtaining a predictive distribution $p(\tilde{\mathbf{y}}_k \mid x_{\text{query}})$ requires aggregating N output responses $\mathcal{Y}$ into unique semantic clusters $\tilde{\mathcal{Y}} = \{\tilde{\mathbf{y}}_k\}_{k=1}^{K}$ via semantic equivalence marginalization:

$$p(\tilde{\boldsymbol{y}}_k|x_{\text{query}}) \approx \sum_{\boldsymbol{y}\in\mathcal{Y}} \frac{|\boldsymbol{y} \equiv \tilde{\boldsymbol{y}}_k|}{|\mathcal{Y}|}, \quad \hat{\mathbf{y}} = \arg\max_{k\in K} p(\tilde{\mathbf{y}}_k \mid x_{\text{query}}). \tag{12}$$

While single-model multi-output methods draw $\mathcal{Y}$ from a single parameter state θ', true Bayesian inference requires marginalizing over the parameter posterior $p(\theta \mid D)$. However, the posterior distribution is intractable. It can be approximated using an ensemble of models $\{\theta^{(m)}\}_{m=1}^{M}$, such that $\theta^{(m)} \sim q(\theta)$, where $q(\theta)$ is an approximation of the posterior distribution. In deep ensemble, approximation is obtained using random initialization of model weights. However, when all members optimize same loss $\mathcal{L}^{\text{ERM}}$, random initializations alone fail to prevent models from collapsing into the same local mode. Our framework resolves this by leveraging $\mathcal{L}_m^{\text{robust}}$. In Lemma 3.3, it can be seen that assigning distinct exponents γ_m introduces a structured objective perturbation, that guarantees a strict lower bound on pairwise parameter distance:

$$\left\|\hat{\theta}^{(m)} - \hat{\theta}^{(m')}\right\|_2 \geq \frac{|\gamma_m - \gamma_{m'}|}{\lambda_{\max}(\mathbf{H}_{\text{ERM}})} \left\|\sum_{i=1}^{n} \ln(u_i + b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0)\right\|_2. \tag{13}$$

This formulation connects directly to Randomized MAP Sampling (RMS) (Pearce et al., 2020). Rather than relying solely on initialization, each member m solves a perturbed MAP objective, $\hat{\theta}^{(m)} \approx \arg\max_\theta p(\theta \mid D, \gamma_m)$, driving parameter trajectories along gradient directions $(\nabla_\theta \mathcal{L}_i(\theta_0))$ scaled by sample vacuity and incorrect belief. Instead of generating multiple answers by repeatedly sampling from a single model, gathering candidate predictions across distinct ensemble members $\{\hat{\theta}^{(m)}\}_{m=1}^{M}$ ensures our semantic aggregation (Eq. 12) combines viewpoints from fundamentally different, plausible parameter modes. The illustration of obtaining multiple answers from (a) single finetuned model θ' with multi-output method and (b) finetuned ensembles guided by the proposed formulation is presented in Figure 4.

C.3 Proposed Formulation and Predictive Calibration

As shown in Theorem 3.2, minimizing the ensemble loss $\mathcal{L}^{\texttt{EU-Ensemble}}$ implicitly maximizes a lower bound governed by an explicit trade-off between standard empirical risk $\mathcal{L}_m^{\text{ERM}}$ and predictive entropy

$-\gamma_m \mathcal{H}^m[\hat{p}]$. For difficult or ambigious samples where empirical risk cannot be easily driven to zero, increasing γ_m strengthens the negative-entropy penalty, forcing the model to lower its confidence and suppress overconfident wrong predictions. On easy samples, ERM loss dominates and entropy regularization relaxes, allowing high-confidence predictions. By constructing an ensemble across different values of γ_m values, we combine member models with complementary accuracy-confidence dynamics, leading to calibrated predictions.

C.4 Summary

DRO provides a framework for a model to focus on different slices of the training data distribution. By leveraging evidential uncertainty, DRO is augmented with the capability to focus on samples based on fine-grained uncertainties. By leveraging multiple values of γ_m, we obtain multiple models each focusing on different data regions, resulting in increased coverage of the posterior distribution. Ultimately, this multi-region focus yields comparable predictive accuracy while significantly improving calibration.

D Mathematical Proofs

In this section, we provide the mathematical proof for Lemma 3.1, Theorem 3.2 and Lemma 3.3.

D.1 Proof of Lemma 3.1

In this section, we show the that focal loss weight can be decomposed as combination of the vacuity and incorrect below. Focal loss weight is defined as

$$w_i = (1 - \hat{p}_{iv})^{\gamma} \tag{14}$$

From the Subjective logic, we leverage the following relationship.

$$\hat{p}_{iv} = \frac{\hat{\alpha}_{iv}}{S_i} = \frac{\hat{b}_{iv} S_i + 1}{S_i} \tag{15}$$

Using Equation 15, we update the Equation 14 as follows:

$$w_i = \left(1 - \frac{\hat{b}_{iv} S_i + 1}{S_i}\right)^{\gamma} \tag{16}$$

$$= \left(1 - \hat{b}_{iv} - \frac{1}{S_i}\right)^{\gamma} \tag{17}$$

Where $\hat{b}_{iv}$ is the belief associated with the ground truth token v and $S_i = \sum_{j=1}^{V} e_{ij} + 1$. Next, using the standard evidential learning equation:

$$u_i + \sum_{j=1}^{V} b_{ij} = 1 \tag{18}$$

$$u_i + \sum_{j=1, j\neq k}^{V} b_{ij} + \hat{b}_{iv} = 1 \tag{19}$$

$$u_i + \sum_{j=1, j\neq k}^{V} b_{ij} = 1 - \hat{b}_{iv} \tag{20}$$

Now, using the value of $1 - \hat{b}_{iv}$ from Equation 20 to simplify Equation 17, we have:

$$w_i = \left(u_i + \sum_{j=1, j\neq v}^{V} b_{ij} - \frac{1}{S_i} \right)^{\gamma} \tag{21}$$

In Equation 21, the second term refers to the sum of the beliefs associated with the incorrect classes. Denoting $b_i^{inc} = \sum_{j=1, j\neq v}^{V} b_{ij}$, we have the following

$$w_i = \left(u_i + b_i^{inc} - \frac{1}{S_i} \right)^{\gamma} \tag{22}$$

Representing $\hat{u}_i = u_i - \frac{1}{S_i}$ we have the following

$$w_i = (\hat{u}_i + b_i^{inc})^{\gamma} \tag{23}$$

This completes our Lemma 3.1

D.2 Proof of Theorem 3.2

In the Lemma 3.1 we have shown that focal loss weight being combination of the vacuity and incorrect belief. In this Theorem, we show that our ensemble model loss is upper bounded by the autoregressive loss and entropy. Specifically, following (Mukhoti et al., 2020) Appendix, for each model we can write the following:

$$\mathcal{L}^{robust} \geq \mathcal{L}^{ERM} - \gamma \mathcal{H}^m[\hat{p}_y] + \gamma C \tag{24}$$

where C is a constant for which the above inequality holds true for all ensemble base models. If $a > b$, $c > d$ then we can write $a + b > c + d$. Applying this to the ensemble by summing $\mathcal{L}^{robust}$ across M models, we can write the following

$$\mathcal{L}^{EU-Ensemble}(\Theta) \geq \sum_{m=1}^{M} [\mathcal{L}_m^{ERM}(\theta) - \gamma_m \mathcal{H}^m[\hat{p}] + \gamma_m C] \tag{25}$$

This completes the Theorem 3.2.

D.3 Proof of Lemma 3.3

Let the sample-weighted distributionally robust loss for component m be defined as:

$$\mathcal{L}_m^{\text{robust}}(\theta) = \sum_{i=1}^{n} w_{i,m} \mathcal{L}_i(\theta), \quad \text{where } w_{i,m} = (u_i + b_i^{\text{inc}})^{\gamma_m} \tag{26}$$

Let $\hat{\theta}^{(m)} = \arg\min_\theta \mathcal{L}_m^{\text{robust}}(\theta)$ denote the local Maximum A Posteriori (MAP) estimate for component m at convergence. Setting the gradient of the loss to zero yields the first-order stationary condition:

$$\nabla_\theta \mathcal{L}_m^{\text{robust}}(\hat{\theta}^{(m)}) = \sum_{i=1}^{n} w_{i,m} \nabla_\theta \mathcal{L}_i(\hat{\theta}^{(m)}) = 0 \tag{27}$$

$$\sum_{i=1}^{n} w_i(\gamma) \nabla_\theta \mathcal{L}_i(\hat{\theta}(\gamma)) = 0 \tag{28}$$

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \left(\sum_{i=1}^{n} w_i(\gamma) \nabla_\theta \mathcal{L}_i(\hat{\theta}(\gamma)) \right) = 0 \tag{29}$$

Using product and chain rule,

$$\sum_{i=1}^{n} \left(\frac{\mathrm{d}w_i(\gamma)}{\mathrm{d}\gamma} \nabla_\theta \mathcal{L}_i(\hat{\theta}(\gamma)) + w_i(\gamma) \nabla_\theta^2 \mathcal{L}_i(\hat{\theta}(\gamma)) \cdot \frac{\partial \hat{\theta}}{\partial \gamma} \right) = 0 \tag{30}$$

Note that $\frac{\mathrm{d}w_i(\gamma)}{\mathrm{d}\gamma} = \frac{\mathrm{d}}{\mathrm{d}\gamma}\left((u_i + b_i^{\text{inc}})^\gamma\right) = \ln(u_i + b_i^{\text{inc}}) \cdot (u_i + b_i^{\text{inc}})^\gamma = -\ln\left(\frac{1}{u_i + b_i^{\text{inc}}}\right) \cdot (u_i + b_i^{\text{inc}})^\gamma$.

We define the non-negative evidential difficulty function $f(u_i, b_i^{\text{inc}}) = \ln\left(\frac{1}{u_i + b_i^{\text{inc}}}\right) \geq 0$ for $u_i + b_i^{\text{inc}} \in (0, 1]$, such that $\left.\frac{\mathrm{d}w_i}{\mathrm{d}\gamma}\right|_{\gamma=0} = -f(u_i, b_i^{\text{inc}})$. Evaluating the stationary condition at the unperturbed cross-entropy baseline θ_0 (where $\gamma = 0$, $w_i(0) = 1$, and $\mathbf{H}_{\text{ERM}}(\theta_0) = \sum_{i=1}^n \nabla_\theta^2 \mathcal{L}_i(\theta_0)$):

$$-\sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) + \left(\sum_{i=1}^n \nabla_\theta^2 \mathcal{L}_i(\theta_0)\right) \cdot \left.\frac{\partial \hat{\theta}}{\partial \gamma}\right|_{\gamma=0} = 0 \; [w_i(\gamma) = 0 \text{ for } \gamma = 0] \tag{31}$$

$$-\sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) + \mathbf{H}_{\text{ERM}}(\theta_0) \cdot \left.\frac{\partial \hat{\theta}}{\partial \gamma}\right|_{\gamma=0} = 0 \tag{32}$$

$$\left.\frac{\partial \hat{\theta}}{\partial \gamma}\right|_{\gamma=0} = \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \cdot \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) \tag{33}$$

Let $\theta_0 = \arg\min_\theta \mathcal{L}^{\text{ERM}}(\theta)$ be the unperturbed baseline ($\gamma = 0$). Now, for two models with parameters m and k, using Taylor's expansion $f(x) \approx f(x_0) + (x - x_0) \cdot f'(x_0) + \mathcal{O}((x - x_0)^2)$:

$$\hat{\theta}^{(m)} = \theta_0 + \gamma_m \left.\frac{\partial \hat{\theta}}{\partial \gamma}\right|_{\gamma=0} + \mathcal{O}(\gamma_m^2) \tag{34}$$

$$\hat{\theta}^{(m)} \approx \theta_0 + \gamma_m \cdot \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) \tag{35}$$

$$\hat{\theta}^{(m')} = \theta_0 + \gamma_{m'} \left.\frac{\partial \hat{\theta}}{\partial \gamma}\right|_{\gamma=0} + \mathcal{O}(\gamma_{m'}^2) \tag{36}$$

$$\hat{\theta}^{(m')} \approx \theta_0 + \gamma_{m'} \cdot \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) \tag{37}$$

Subtracting the parameter expansion of Model m' from Model m:

$$\begin{aligned} \hat{\theta}^{(m)} - \hat{\theta}^{(m'k)} &= \left[\theta_0 + \gamma_m \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0)\right] \\ &\quad - \left[\theta_0 + \gamma_{m'} \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0)\right] + \mathcal{O}\left((\gamma_m - \gamma_{m'})^2\right) \qquad (38) \\ &= (\gamma_m - \gamma_{m'}) \cdot \mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0) + \mathcal{O}\left((\gamma_m - \gamma_{m'})^2\right) \qquad (39) \end{aligned}$$

(40)

Taking the Euclidean norm

$$\left\|\hat{\theta}^{(m)} - \hat{\theta}^{(m')}\right\|_2 = |\gamma_m - \gamma_{m'}| \cdot \left\|\mathbf{H}_{\text{ERM}}^{-1}(\theta_0) \sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0)\right\|_2 \tag{41}$$

For any positive-definite matrix $\mathbf{A}$ and vector $\mathbf{v}$, $\|\mathbf{A}^{-1}\mathbf{v}\|_2 \geq \lambda_{\min}(\mathbf{A}^{-1})\|\mathbf{v}\|_2 = \frac{1}{\lambda_{\max}(\mathbf{A})}\|\mathbf{v}\|_2$.

$$\left\|\hat{\theta}^{(m)} - \hat{\theta}^{(m')}\right\|_2 \geq \frac{|\gamma_m - \gamma_{m'}|}{\lambda_{\max}(\mathbf{H}_{\text{ERM}})} \cdot \left\|\sum_{i=1}^n f(u_i, b_i^{\text{inc}}) \nabla_\theta \mathcal{L}_i(\theta_0)\right\|_2 \tag{42}$$

This completes the proof of Lemma 3.3.

E DETAILS OF EXPERIMENTS

Datasets: We conduct experiments on three datasets. ScienceQA is a multimodal dataset across 26 topics of science where answer contains multiple-choice option, along with explanations (≈ 200 tokens on average). CommonsenseQA is a text-only dataset built on a large commonsense knowledge graph where answer contains multiple choice options (≈ 1 tokens on average). Commonsense reasoning is a text-only dataset where the answer contains multiple choice options, along with a natural language explanation (≈ 10 tokens on average). TriviaQA is a reading comprehension and open-domain question-answering dataset, where targets in this dataset consist of short factual entities or entity names without multiple-choice constraints.

Models and Implementation Details: We use two models, gemma-3-4b-it and LLaVA-OneVision-Qwen2-0.5B-ov. We use pretrained models from Huggingface with repository names: google/gemma-3-4b-it and lmms-lab/llava-onevision-qwen2-0.5b-ov. For experiments, we use the device with a GPU: NVIDIA RTX A6000 and memory: 48 GB. We perform Lora fine-tuning of the models using the training split as provided in the original splits, and present the results for the test split. We follow the fine-tuning instructions from repositories (Li., 2024; won Lee., 2024). For additional models (Qwen/Qwen3.5-4B and microsoft/Phi-4-mini-instruct(3.8B)), we leveraged HuggingFace SFTTrainer to finetune. For each ensemble component, we use nucleus sampling with values: $top_p = 0.9$ and temperature $= 0.7$. We provide the detailed code to reproduce the experiments in section I.

Answer Retrieval: To check whether the answer corresponds to one of the options in the multiple choice question, we use the text containing the obtained answer and the multiple choice option as an input to a natural language inference model. Given x_{query} with options, and the generated answer $\mathbf{y}$, we use the following text:

$$\text{TEXT} = < x_{query} > + < \mathbf{y} > + \text{[SEP]} + < x_{query} > + <\text{Multiple Choice Option}> \tag{43}$$

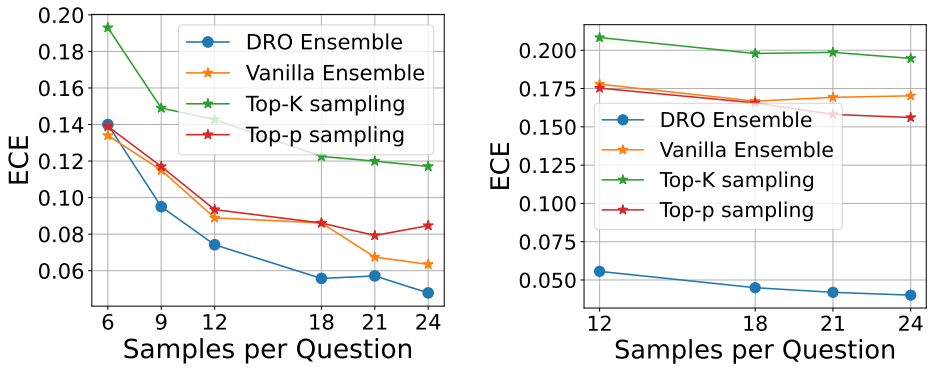
NLI model returns the logits for 3 classes. If the obtained answer $\mathbf{y}$ entails the multiple choice option, the logit is highest for 3rd class. Based on the logit, we can retrieve the answer. To accurately retrieve the answer, it is recommended to use both the character and actual answer in the <Multiple Choice Option>. For the generated answers in Triviaqa, we aggregated them into K unique answers by leveraging the semantic equivalence decision from an NLI model as described in clustering of Semantic entropy (Kuhn et al., 2023b).

Baselines: For all the baselines, first, we finetune the model with the training split of the dataset. For top-p sampling, we use the value of $top_p = 0.9$. For beam sampling, we use $num_beams = 5$. For temperature sampling, we use the value of temperature as 0.7. For top-k sampling, we use the value of $top_k = 50$. For all multi-output baselines except temperature sampling, we use the value of temperature as 1.0. For the least probability, we obtain confidence as the minimum of the probability of all generated tokens, as $\min_{t=1}^{T} p(y_t|y_{<t}, x)$. For the product probability, we multiply the probability of all generated tokens as $\prod_{t=1}^{T} p(y_t|y_{<t}, x)$.

Expected Calibration Error: Given N answers from decoding methods, we have an answer set $\mathcal{Y}$. We place each answer into the respective multiple-choice option, aggregating them to K unique answers $\tilde{\mathcal{Y}} = \{\tilde{\mathbf{y}}_k\}_{k=1}^{K}$. The probability distribution over k^{th} unique label is represented by $p(\tilde{\mathbf{y}}_k|x_{\text{query}})$. Confidence score and predicted label are given by:

$$p(\tilde{\mathbf{y}}_k|x_{\text{query}}) \approx \sum_{\mathbf{y}\in\mathcal{Y}} \frac{|\mathbf{y} \equiv \tilde{\mathbf{y}}_k|}{|\mathcal{Y}|}, \quad \hat{\mathbf{y}} = \arg\max_{k\in K}(p(\tilde{\mathbf{y}}_k|x_{\text{query}})), \quad s = \max_{k\in K}(p(\tilde{\mathbf{y}}_k|x_{\text{query}})) \tag{44}$$

For a test dataset $\{x^i, y^i, \hat{\mathbf{y}}^i\}_{i=1}^{N'}$, where x^i is i^{th} dataset input, y^i is the corresponding ground truth label, and $\hat{\mathbf{y}}^i$ is the prediction from the model. First, we bin the dataset into B equally spaced confidence intervals (Naeini et al., 2014; Nixon et al., 2019). For each bin B_b accuracy(B_b)



(a) Impact of N in Commonsense Reasoning

(b) Impact of N in CommonsenseQA

Figure 5: Impact of N for (a) Commonsense Reasoning dataset (b) CommonsenseQA datset

denotes the accuracy of the dataset belonging to the bin B_b, and confidence(B_b) denotes the average confidence of the dataset belonging to the bin B_m. The expected calibration error is calculated as the mean of the difference between accuracy and confidence across all the bins.

$$ECE = \sum_{b=1}^{B} \frac{|B_b|}{N'} |\text{accuracy}(B_b) - \text{confidence}(B_b)| \tag{45}$$

For ECE calculation, we use the bin size of 10.

F Additional Experiments

F.1 Ablation Studies

F.1.1 Impact of N

In this section, we investigate the effect of N (the number of samples per question) on calibration performance for the proposed methods, along with competitive multi-output baselines. The comparison result on the commonsense reasoning dataset and commonsense QA dataset by sampling $N = \{24, 21, 18, 12, 9, 6\}$ samples per query are presented in Figure 5a and 5b. For all methods, as we increase the value of N, we have comparable or better calibration performance (decreasing trend of ECE). The results demonstrate an improvement in confidence capability as the value of N increases. Next, we compare performance among baselines. For all the values of N, we achieve performance improvement of the proposed ensemble method over the compared baselines. This comparison further demonstrates the effectiveness of the proposed DRO-based ensemble in calibration performance across different values of N.

F.1.2 Impact of V'

In equation 4, we use V' to evaluate the vacuity and belief. In evidential theory, V' refers to the number of classes in the standard classification setting. For token generation, the total number of classes is V, which is normally a large value (e.g., 32000 for LLaVA-OneVision-Qwen2-0.5B). As a result, the two uncertainty values become $u_t \approx 1, b_t^{inc} \approx 0$ when $V' = V$, resulting in uniform weights for all the samples. As a practical solution, we used $V' < V$ for the experiments. To be specific, we set V' as the total number of tokens with a positive logit, instead of a constant value for all tokens. In this way, the resulting uncertainties reflect the actual difficulty of learning and result in a calibrated output from the ensemble. To demonstrate the impact of using different values of V', we experiment with OneVision-Qwen2-0.5B model and the Commonsense reasoning dataset. Calibration comparison is presented in Table 6. Using

Table 6: Impact of V'

Setting	ECE
Vanilla Ensemble	0.065
$V' = V$	0.061
$V' = 100$	0.057
$V' = 50$	0.052
Ours	**0.048**

$V' = V(32000)$ demonstrates approximately similar performance to vanilla ensemble. Lowering the value of V' to 100 and 50 gives an improvement in calibration. Setting V' as the number of tokens with positive logits for every token generation demonstrates the superior performance.

Table 7: Impact of calibration error

Method	Commonsense QA		Commonsense Reasnoning		ScienceQA	
	Gemma	LLaVa	Gemma	LLaVa	Gemma	LLaVa
Beam Search	0.069	0.133	0.059	0.150	0.046	0.084
Top-p	0.060	0.113	0.050	0.114	0.045	0.073
Top-k	0.062	0.108	0.053	0.111	0.046	0.070
Vanilla Ensemble	0.054	0.109	0.049	0.112	0.044	0.070
DRO Ensemble	**0.049**	**0.102**	**0.049**	**0.111**	**0.041**	**0.069**

F.1.3 Impact of calibration error

As an ablation study, we leverage the Brier score (Brier, 1950) as a metric to measure calibration error. The Brier score for a generative test dataset can be calculated as, Brier Score $= \sum_{k=1}^{K}(p_k - y_k)^2$. Here, y_k is the ground truth label of whether the sample belongs to class K or not, and p_k is the probability of the sample belonging to class k. The comparison of the proposed method is presented along with competitive baselines in the Table 7. Using the Brier score, we observe that the proposed ensemble sampling achieves competitive performance over baselines, consistent with the ECE-based calibration error metric.

Table 8: Impact of ϵ

ϵ	% of samples	ECE
0.5	31.581	0.118
0.1	50.713	0.091
0.05	68.436	0.064
0.01	87.224	0.068
0.005	92.580	0.053
0.001	93.120	0.053
–	100.00	0.040

F.1.4 Impact of ϵ

The hyperparameter ϵ is set to stop the iteration in adaptive sampling. A higher value of ϵ results in early stopping of iteration and generates a lower percentage of samples. The comparison of different values of $\epsilon = \{0.5, 0.1, 0.05, 0.01, 0.005, 0.001\}$, along with calibration performance and generated percentage of samples, is presented in the Table 8. As we lower the value of ϵ, we generate more samples, resulting in a higher calibration performance.

F.1.5 Impact of external model

Our method follows standard multi-output uncertainty quantification frameworks: (i) samples multiple candidate responses, (ii) clusters them into semantically equivalent groups, and (iii) computes confidence over these groups. Relying on an external model for semantic equivalence is established practice in multi-output literature. In this section, we study the impact of using different external models in an open-ended setting (TriviaQA). The results are presented in Table 9. While errors in semantic grouping naturally affect all multi-output methods, our approach consistently handles this noise better than existing baselines, demonstrated empirically in an open-ended setting (TriviaQA) below with external models such as an NLI model (DeBERTa-MNLI) and a sentence transformer (all-mpnet-base-v2). We also included Rouge-L-based clustering that does not require an external model. While NLI is the better choice, the proposed method outperforms the baselines in each external model experiment.

Table 9: Impact of external model.

Clustering Backend	Method	ECE (↓)	AUROC (↑)
NLI (DeBERTa-MNLI)	Nucleus	0.092	73.47
	Vanilla ensemble	0.070	73.78
	Ours	**0.063**	**76.85**
Sentence Transformers	Nucleus	0.139	55.82
	Vanilla ensemble	0.125	59.06
	Ours	**0.097**	**65.00**
Rouge-L	Nucleus	0.151	50.85
	Vanilla ensemble	0.147	53.53
	Ours	**0.107**	**59.95**

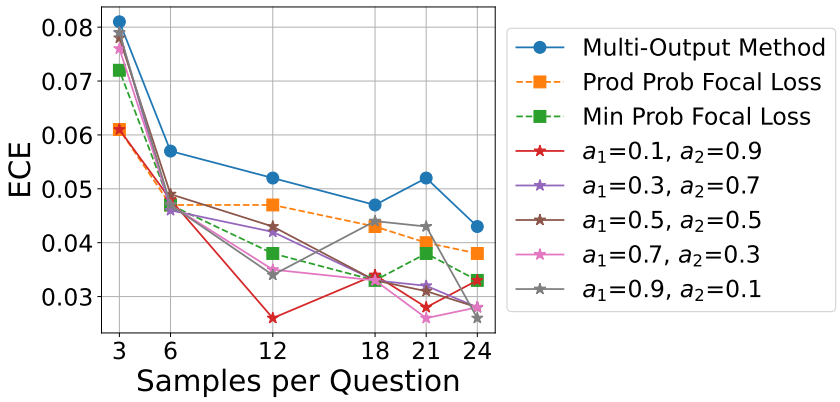


Figure 6: Ablation of the proposed ensemble: Impact of putting weights a_1 and a_2 on uncertainty u and incorrect belief b^{inc} on w_j of the loss function, and impact of using a different focal loss function

F.1.6 Proposed Ensemble Formulation

In this section, we conduct the ablation study on the proposed ensemble. Each ensemble component is trained using the loss function guided by uncertainty. First, we study the impact of different uncertainties: u and b^{inc} in Equation 6. We re-formulate the weight as $w_i = (a_1 u + (1 - a_1) b^{inc})^{\gamma}$. We vary the value of $a_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for $\gamma = 2$. For each setting, we sample N samples per query from two ensemble components. The first component is trained with $w_j = 1$ and the second component is trained with $w_i = (a_1 u + (1 - a_1) b^{inc})^2$. For a multi-output method, all the N answers are sampled from the model trained with $w_i = 1$. Next, we study the impact of using a loss function guided by probability. For product probability-based focal loss, we use the value of $w_i = (1 - \prod_{t=1}^{T} p(y_t|y_{<t}, x))^2$, and for min probability-based focal loss, we use the value of $w_i = (1 - \min_{t=1}^{T} p(y_t|y_{<t}, x))^2$.

We experiment on the minitest split of ScienceQA. The results are presented in Figure 6. All the settings exhibit better calibration performance than the multi-output baseline. Among the values of a_1, the settings with the lower value of a_1 demonstrate better ECE values. Specifically, for $N < 18$, $a_1 = 0.1$ has the best ECE, and for $N \geq 18$, $a_1 = 0.7$ has the best ECE. Lower values of a_1 are the settings where we put more weight on incorrect beliefs than on uncertainty. Incorrect beliefs refer to the model's beliefs about the incorrect labels. Penalizing those beliefs is essential for the model to be accurate. It should be noted that too much penalty on incorrect belief does not always result in best calibration performance(as demonstrated by best performance $a_1 = 0.7, N \geq 18$).

Among min-prob and prod-prob based focal loss, we have superior performance of the min-prob based focal loss. Minimum probability among tokens helps the model to focus on the most uncertain tokens. This study further justifies the design of the proposed loss function for guiding the model to focus on the key token while fine-tuning. When compared with other settings, min-probability-based

focal loss performs comparably to the setting that puts equal weight on uncertainty and incorrect belief. High incorrect beliefs and high uncertainty are two contributing factors for a low probability value. Loss function based on probability score does not allow fine-grained control on the source of low probability in the training ensemble component.

F.2 Handling limited and long-tail data distribution

F.2.1 Results on limited data setting

In this section, we study the robustness of the proposed method in a limited data setting. Instead of fine-tuning LLaVA-OneVision and Gamma models with the full dataset as done for experiments in Table 1, we experimented on a limited data setting by using only 5% and 8% of training samples of the Commonsense Reasoning dataset. The results (ECE, Accuracy) are presented in Table 10. The proposed method outperforms vanilla ensemble and other baselines, demonstrating the robustness in a limited data scenario.

Table 10: Fine-tuning on Limited data

	LLama model				Gemma model			
	5%		8%		5%		8%	
Method	ECE	Acc.	ECE	Acc.	ECE	Acc.	ECE	Acc.
Least Probability	0.201	44.307	0.216	44.635	0.601	67.731	0.668	68.386
Product Probability	0.436	44.307	0.441	44.635	0.396	67.731	0.473	68.386
Multi-output (nucleus)	0.115	52.007	0.099	51.925	0.114	73.790	0.075	74.201
Vanilla Ensemble	0.106	52.088	0.092	52.580	0.094	74.770	0.056	74.939
Ours	**0.100**	**52.416**	**0.083**	**53.317**	**0.086**	**74.770**	**0.035**	**75.758**

F.2.2 Results on Long-tailed data setting

In our setting, the weighting term w_i in Equation 6 enables the model to adaptively focus more on difficult samples while assigning less emphasis to easier ones. For challenging samples, w_i becomes large either due to high vacuity or because more belief mass is placed on incorrect classes, compared to easier cases. Consequently, during backpropagation, the model prioritizes reducing the loss associated with these difficult samples.

Importantly, DRO does not encourage overfitting; rather, it helps prevent it through the implicit regularization induced by the entropy term. As established in Theorem 3.2, our loss decomposes into the cross-entropy minus the entropy, where the negative entropy term discourages the model from producing incorrect, overconfident predictions.

Table 11: Fine-tuning on long-tailed data

Setting	**ECE** (↓)	**Acc.** (↑)
Multi-output (nucleus)	0.116	76.415
Vanilla Ensemble	0.106	74.764
Ours	**0.059**	**77.358**

To empirically verify the effectiveness of the proposed method over baselines for handling long-tail data distribution, we conducted experiment on the ScienceQA dataset with the Gemma3-4b-it model. We fine-tune the model with the minitrain split of the dataset. While the original minitrain split contains 115 samples from the "magnet" category, we include only 4 samples from the same category, making it a tail class. We present the performance comparison between the proposed method and baselines in the Table 11.

The results contain both the calibration and accuracy performance for the minitest split. We observe a clear performance gain of the proposed method over baselines. Further, we compare performance over only the tail class, where we leverage a total of 110 samples from "magnet" category. The results are presented in the Table 12.

The accuracy performance of all methods is lower compared to the whole minitest split. However, this is expected, as we leverage very few samples in training. When we compare the calibration performance among the methods, our method shows significant improvement over the baselines. These results provide important empirical evidence to demonstrate the robustness of the proposed method for long-tail data.

Finally, we present a qualitative analysis in the Figure 7. We have a training sample where the question is about whether magnets "repel" or "attract" each other. In this specific sample, we have

Table 12: Results on tail class "magnet"

Setting	ECE (↓)	Acc. (↑)
Multi-output (nucleus)	0.333	37.273
Vanilla Ensemble	0.305	31.818
Ours	**0.164**	**44.545**

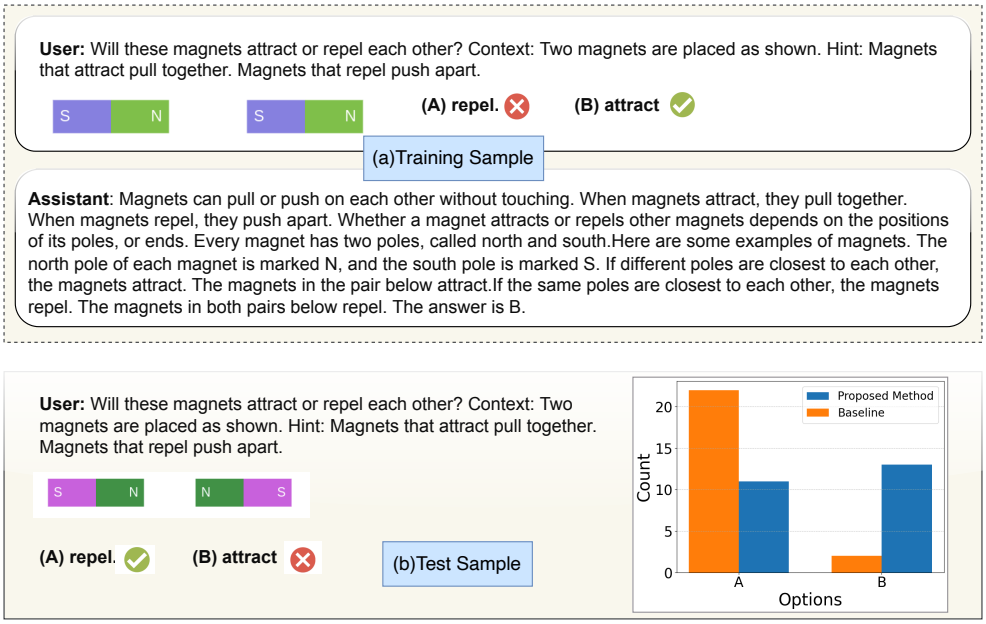


Figure 7: Qualitative analysis of tail class result in long tailed data study.

the case of "attract" (north and south poles facing each other). In testing, we present a similar question, but with north poles facing each other (i.e., the correct answer being "repel"). The baseline method (multi-output with nucleus sampling) outputs answer of "attract" 22 out of 24 times, making a confidently wrong prediction. In contrast, proposed method outputs answer of "repel" 13 out of 24 times. When only few samples of "magnet" category is shown in training, the baseline method suffers from overfitting and results in overconfident predictions. The proposed method demonstrates superior calibration over the baseline.

F.3 Results on additional models

In this section, we study the effectiveness of the proposed method for additional models: google/gemma-3-12b-it, lmms-lab/llava-onevision-qwen2-7b-ov, Qwen/Qwen3.5-4B, and microsoft/Phi-4-mini-instruct(3.8B). These models are implemented using the Huggingface library: transformers. We conduct experiments on Commonsense Reasoning with these models. The experiment results of our method, along with vanilla ensemble and multi-output nucleus-based sampling baselines, are presented in Table 13. The experiments demonstrate the effectiveness of the proposed method using these LLMs as well (including larger-sized LLMs).

F.4 Additional Results On Adaptive Sampling

In the main paper, we demonstrate the effectiveness of adaptive sampling for gemma-3-4b-it model. In this section, we study the performance of adaptive sampling for the proposed DRO ensemble, along with other competitive baselines for lava-onevision-qwen2-0.5b-ov model. The results are presented in Table 14. For the proposed ensemble, when $\epsilon = 0.1$ is used, with only 50% of the samples, the

Table 13: Performance comparison across 7B and 12B models.

Model	Setting	ECE (↓)	Acc. (↑)
Gemma-12B	Multi-output	0.047	86.077
	Vanilla Ensemble	0.036	86.159
	Ours	**0.026**	**86.568**
LLaVA-7B	Multi-output	0.042	84.029
	Vanilla Ensemble	0.039	84.357
	Ours	**0.033**	**84.439**
Qwen3.5-4B	Multi-output	0.060	83.374
	Vanilla Ensemble	0.057	83.702
	Ours	**0.040**	**84.029**
Phi-4-mini-instruct	Multi-output	0.113	65.766
	Vanilla Ensemble	0.104	66.093
	Ours	**0.087**	**66.749**

calibration performance is competitive compared to the non-adaptive setting. When we use $\epsilon = 0.01$, the calibration performance further improves. Other competitive baselines, in comparison to the proposed DRO ensemble, perform poorly. When compared with its own non-adaptive version, these methods also demonstrate reasonable calibration performance upon leveraging adaptive sampling.

Table 14: Calibration performance comparison of adaptive sampling for lava-onevision-qwen2-0.5b-ov model

Baseline	ϵ	CommonsenseQA		Commonsense Reasoning	
		% Samples	ECE	% Samples	ECE
Top-k	-	100.00	0.195	100.00	0.117
Top-k + Adaptive	0.1	46.47	0.209	50.172	0.140
Top-p	-	100.00	0.156	100.00	0.085
Top-p + Adaptive	0.1	47.486	0.174	50.925	0.104
Vanilla Ensemble	-	100.00	0.170	100.00	0.065
Vanilla Ensemble + Adaptive	0.1	47.715	0.172	50.631	0.102
DRO Ensemble	-	100.00	0.040	100.00	0.048
DRO Ensemble + Adaptive	0.1	50.106	0.077	50.713	0.091
DRO Ensemble + Adaptive	0.01	87.109	0.067	87.224	0.068

F.5 Results on POPE dataset

In this section, we present the comparison of calibration performance in POPE dataset (Li et al., 2023b) for the proposed ensemble sampling and multi-output method using nucleus sampling. For fine-tuning, we use "popular" split from the dataset to train the Lora component for gemma3-4b-it model. For testing, we use "random" split. The comparison is presented in Figure 8 for $N \in \{24, 21, 18\}$. Using the proposed ensemble sampling, the calibration performance improves over the multi-output method. This study further confirms the superior calibration performance of the proposed method over the multi-output method.

F.6 Comparison with post-hoc calibration methods

We performed a post-hoc temperature scaling to improve calibration performance on Commonsense reasoning with the Google/Gemma-3-4b-it model. For the held-out dataset, we searched for the temperature parameter that obtains the optimal Negative Log Likelihood loss of the output generation. For a generated output $\mathbf{y} = [y_1, y_2, ...y_T]$, the dimension of the logits is $T \times V$. In case of the model considered, the size of the vocabulary is $V = 262,208$. The results of the experiment are presented in the Table 15.

We observe a marginal improvement with temperature scaling. Our proposed method significantly outperforms the performance with and w/o temperature scaling. Similarly, we also performed temperature scaling for the multi-output-based baseline. With N generations of output, we aggregate

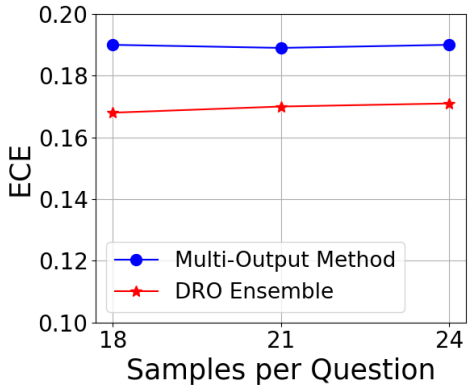


Figure 8: Calibration result on POPE dataset

Table 15: Comparison with post-hoc calibration

Setting	w/o Temperature Scaling	with Temperature Scaling
Single Output	0.690	0.679
Multi-Output (nucleus)	0.031	0.366

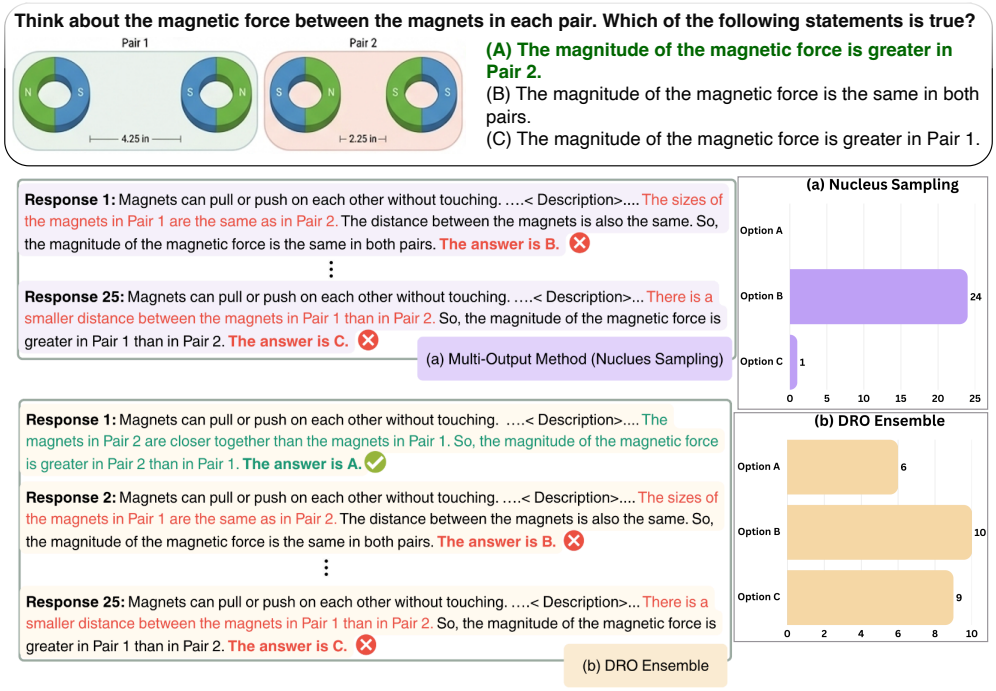


Figure 9: For the question, the prediction from both methods is incorrect. Using (a) Multi-output method (nucleus sampling): the model outputs the wrong answers with high confidence, (b) DRO Ensemble: Diversity components provide diverse answers with low confidence

them to K multiple-choice answers. We treat the logit for each option as the number of answers semantically equivalent to the option, $\sum_{\mathbf{y}\in\mathcal{Y}} |\mathbf{y} \equiv \tilde{\mathbf{y}}_k|$. The temperature scaling with the multi-output method significantly degrades the performance. In the case of MCQ datasets, although we have K options, the actual answer varies from one sample to another. Hence, searching for an optimum temperature for all the data samples does not lead to optimum calibration performance for the whole dataset.

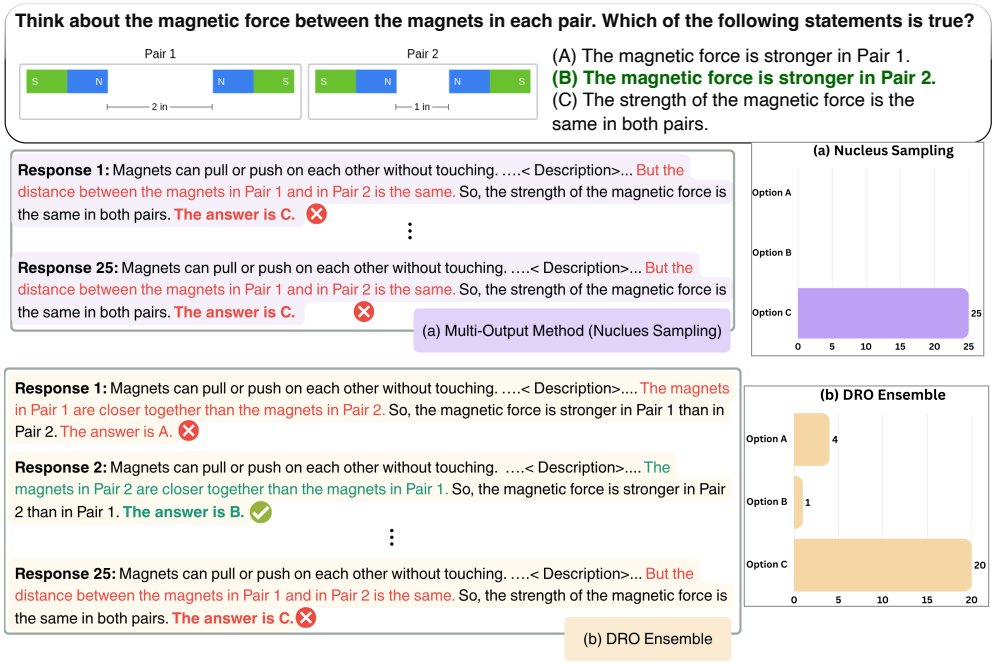


Figure 10: Qualitative analysis: For the question, the prediction from both methods is incorrect. Using (a) Multi-Output Method (Nucleus Sampling): the model answers the wrong answers with high confidence, (b) DRO Ensemble: Diversity components provide diverse answers with low confidence

F.7 Qualitative Analysis

First, we present cases of two qualitative analysis comparing the confidence estimates produced by the baseline and our proposed method on an incorrectly answered sample. For first case, as shown in Figure 9, the baseline multi-output method, nucleus sampling generates 25 responses for the same question, among which 24 responses select the same incorrect answer B and only one response differs. This yields a high confidence score of 0.96 despite the prediction being incorrect. In contrast, our method generates more diverse responses across different answer paths, reducing the confidence score from 0.96 to 0.40. Although the final prediction remains incorrect, the substantially lower confidence better reflects the model's uncertainty and therefore improves calibration. Notably, the baseline never produces the correct option A across all sampled decoding paths, whereas our method explores a broader set of plausible answers, demonstrating the benefit of diversity-inducing ensemble sampling. Next, in Figure 10, we observe that for the question asked, even after 25 samples, the multi-output method with nucleus sampling confidently provides the wrong output without leading to the diverse paths of A and B. In contrast, the proposed method outputs other, more diverse answers, lowering the confidence of the provided output from 1.0 to 0.80. In this way, the proposed method improves the calibration of generative content by suppressing confidence of incorrect samples.

F.8 Analysis of lower accuracy

We observe that we achieve lower accuracy in some cases in Table 1. We conducted an analysis of the method with highest accuracy baseline(baseline: beam sampling: 58.720) vs (our method: 57.903) for Commonsense reasoning and LLaVA-OneVision-Qwen2-0.5B. While the baseline achieves marginally higher accuracy, the probability assigned is mostly on the higher end, resulting in untrustworthy confidence when the answer is wrong. On contrary, our method does not concentrate probability on the higher end, resulting in lower confidence when the answer is wrong. This makes the proposed method more trustworthy. This is also demonstrated by ECE and AUROC (baseline: 0.359, 61.460) and (our method: 0.048, 72.338) in Table 1. We show two qualitative analyses of different cases. First, we present the test sample, where our method is wrong, but the baseline is

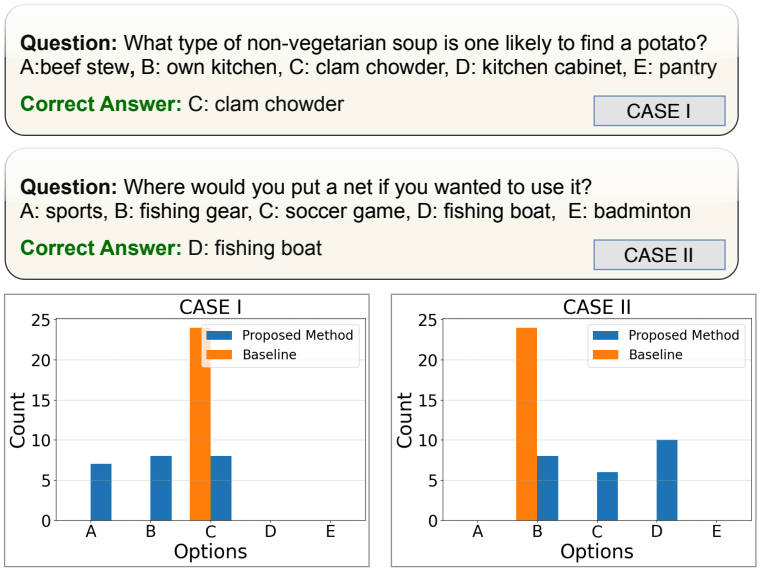


Figure 11: Qualitative analysis of the error case of the proposed method. Case I: The Proposed method is wrong, and the baseline is correct. Case II: The Proposed method is correct, and the baseline is wrong.

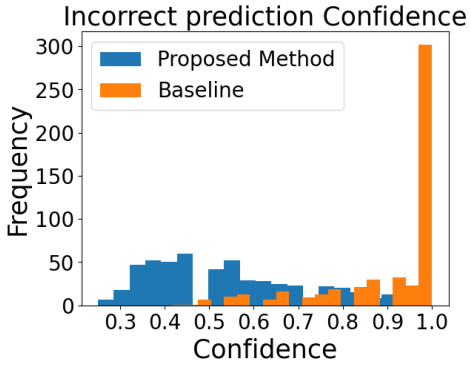


Figure 12: Confidence distribution of the proposed method and the baseline with the highest accuracy on incorrect predictions

correct. In the second case, we present the test sample, where our method is correct, and the baseline is wrong. It is presented in Figure 11. In case I, option C is correct. Baseline answers option C: 24 times. Our method answers option A: 7 times, B: 8 times and C: 8 times. Our method is wrong, but the confidence is low (0.3). In case II, option D is correct. Baseline answers option B: 24 times. Our method answers option B: 8 times, C: 6 times and D: 10 times. Baseline is confidently wrong.

We plot a histogram of confidence of incorrect predictions from the proposed method and the baseline in Figure 12. For the baseline, the confidence of wrong predictions is concentrated towards the higher end. Whereas, our method suppresses confidence of incorrect samples. We want to clarify that the proposed method does not suppress confidence uniformly for all the predictions; rather, it suppresses confidence on incorrect predictions far more aggressively than on correct ones. To demonstrate this empirically, we counted the incorrect and correct test samples from our method in the most confident bin. Out of a total of 707 correct and 514 incorrect test samples, our method places 156 correct predictions and only 16 incorrect predictions in the most confident bin, demonstrating a non-uniform confidence suppression. The confidence distribution reveals the trustworthiness of the proposed method over the most accurate baseline. Further, the performance comparison of our formulation, together with baselines from Table 1 reveals highest AUROC and ECE from our method, empirically

demonstrating that calibration gains come from true discrimination rather than uniform confidence suppression.

G LIMITATIONS

In this work, we propose a novel diversity-inducing ensemble with theoretical guarantees. We demonstrate the effectiveness of the proposed method in several datasets and models. The method requires sampling N answers in inference to estimate the confidence accurately, which is computationally expensive. To reduce the cost, we propose adaptive sampling and demonstrate competitive calibration performance from the proposed sampling. Adaptive sampling reduces the sampling cost from N to N_{adaptive}, which is still expensive compared to obtaining a single sample for a question. We leave the systematic study of calibration in LLMs from a single sample as future work.

H IMPACT STATEMENT

Large language models are becoming increasingly popular in a lot of applications. Although they can perform many tasks, the content generated is not always correct and trustworthy. The concern becomes more pronounced in critical applications (medical, education). The model can quantify a confidence or uncertainty score along with the generated content. The confidence can be used to decide whether we should trust the generated content. The research in this work is performed to ensure that the generated content has a calibrated confidence, such that quantified confidence correlates with the correctness of the output.

I SOURCE CODE

The source code is available at this link.

J COMPUTATIONAL COST ANALYSIS

The inference cost of single-output method is $\mathcal{O}(1)$ and multi-output methods(including our method and vanilla ensemble) is $\mathcal{O}(N)$. In our experiments, we generate a constant number of samples N for all multi-output methods, resulting in the same inference cost for our method and vanilla ensemble as the multi-output baselines. Training ensemble components requires an extra cost for the proposed method. We reduce the cost of training ensemble components by leveraging $M < N$ models for the proposed method. To ensure that M components cover the diversity of models to approximate the posterior, we propose a DRO-based ensemble guided by evidential uncertainty. Further, we leverage a parameter-efficient fine-tuning method to fine-tune the pretrained model to reduce the training cost. Fine-tuning of the ensemble components with downstream task is a one-time expense and the ensemble members can be run in parallel across multiple GPUs for computational efficiency. It should be noted that with the same training and inference cost as vanilla ensemble method, our method demonstrates superior performance.