DO NOT TRAIN AWAY UNCERTAINTY: EARLY UNCERTAINTY ANCHORED CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks, including large language models, have achieved remarkable performance across various tasks. However, they are prone to overconfidence during training or fine-tuning. In this work, we observe a consistent phenomenon across different models that the early model is better calibrated, while later training or fine-tuning yields marginal accuracy gains but substantially increases calibration errors. Our analysis suggests that the early model retains uncertainty awareness in both its predictions and features, which is gradually lost with continued training. To avoid training away this uncertainty awareness, we propose **EUA-Cal**, a novel method that exploits the **E**arly model as an **U**ncertainty **A**nchor for **Cal**ibration. EUA-Cal introduces early prediction regularization to preserve early predictive uncertainty and prototype structure regularization to exploit uncertainty reflected in the early feature space, jointly mitigating overconfidence. Extensive experiments on image classification and multiple-choice question answering across eight diverse models demonstrate that EUA-Cal outperforms state-of-the-art calibration methods. ***Code is available in the supplementary materials.***

1 INTRODUCTION

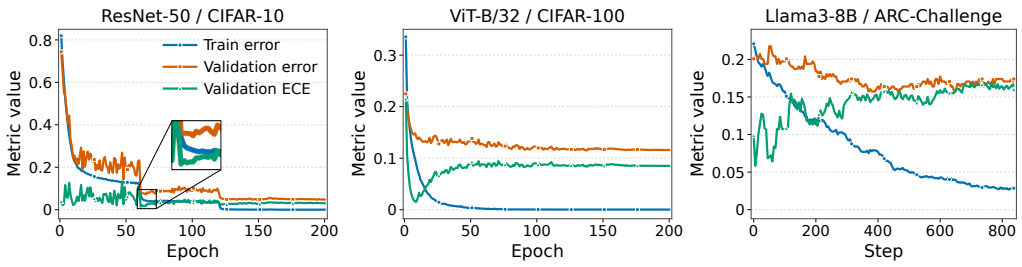
In recent years, deep neural networks (DNNs), including large language models (LLMs), have achieved significant progress in image classification (Zagoruyko & Komodakis, 2016), object segmentation (Kirillov et al., 2023), and natural language processing (Vaswani et al., 2017). Despite their remarkable performance, these models are prone to overconfidence during training or fine-tuning, with the predicted confidence exceeding the actual accuracy (Guo et al., 2017; Jiang et al., 2021). This overconfidence reflects an underestimate of the uncertainty in their predictions and can reduce their reliability, particularly in high-risk applications such as autonomous driving (Zhang et al., 2025) and medical diagnosis (Huang et al., 2025b).

To calibrate model confidence, various calibration techniques have been proposed, including post-hoc calibration (Guo et al., 2017; Pan et al., 2020), loss function and regularization designs (Szegedy et al., 2016; Mukhoti et al., 2020; Liu et al., 2022; Ni et al., 2025), mixup-based methods (Thulasidasan et al., 2019; Bouniot et al., 2025), uncertainty-aware methods (Sensoy et al., 2018; Li et al., 2025), etc. Expected Calibration Error (ECE) (Guo et al., 2017) is a commonly used metric for evaluating the calibration performance of these methods, which measures the discrepancy between predicted confidence and empirical accuracy across confidence bins. A lower value of ECE indicates better calibration performance, i.e., the model is aware of how uncertain its predictions are.

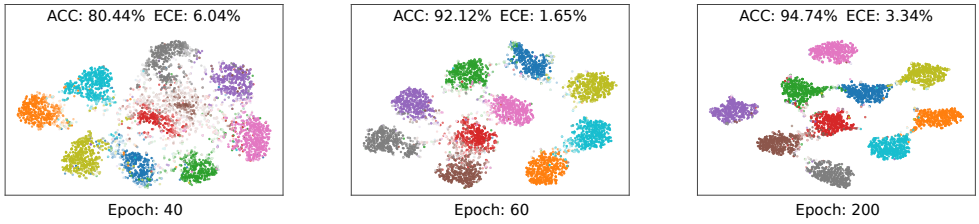
Some studies (Guo et al., 2017; Mukhoti et al., 2020) have shown that DNNs tend to produce overconfident predictions with continued training. We observe a similar pattern across various models, where the early model is better calibrated but its calibration worsens as training continues. Figure 1(a) illustrates the classification errors of three different models on the train and validation sets, along with the ECE on the validation set. On all three models, we observe that ECE reaches its minimum at an early stage of training or fine-tuning, but then increases in later stages, while accuracy improves only marginally. This indicates that there exists an early model that achieves the lowest ECE and exhibits the greatest uncertainty awareness in its predictions. However, this awareness is gradually lost during subsequent training, resulting in worse calibration performance.

Beyond predictions of the early model, we observe that it also exhibits uncertainty awareness in feature structure. Figure 1(b) visualizes the feature structure and the corresponding calibration per-

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.



(a)



(b)

Figure 1: **(a)** Training and validation classification errors and validation ECE for ResNet-50 on CIFAR-10, ViT-B/32 on CIFAR-100, and Llama3-8B on ARC-Challenge. **(b)**: t-SNE visualizations of features from ResNet-50 on CIFAR-10 at epochs 40, 60, and 200. Lighter shades indicate features associated with lower confidence. More examples are provided in Appendix B.

formance of ResNet-50 at different training epochs. We find that features with low confidence (shown in lighter shades) are crucial for calibration. At epoch 40, many features have low confidence, but their cluster structure remains poorly defined. The model exhibits the lowest accuracy and the highest ECE. At epoch 60, corresponding to the early model in Figure 1(a), clearer clusters emerge while features with low confidence remain near their boundaries. At this stage, accuracy improves and ECE reaches its minimum. By epoch 200, the model achieves the highest accuracy with the most compact clusters. However, features with low confidence almost disappear, while the ECE increases. These observations suggest that the early model with the best ECE forms a clear feature structure while retaining some low confidence features, demonstrating the best uncertainty awareness in its features. However, this uncertainty awareness also fades as training continues.

To avoid training away this uncertainty awareness in predictions and features, we propose **EUA-Cal**, a novel method that exploits the **E**arly model as an **U**ncertainty **A**nchor for **Cal**ibration. Specifically, in the pre-learning stage, we obtain the early model by stopping training at a specific epoch, which is selected on validation sets. We then introduce *Early Prediction Regularization* (EPR) to preserve early predictive uncertainty and *Prototype Structure Regularization* (PSR) to exploit uncertainty awareness in features of the early model. Finally, we combine EPR and PSR to formulate the overall EUA-Cal objective, aiming to preserve uncertainty and mitigate overconfidence.

Our contributions are summarized as follows:

- We identify a general phenomenon in training or fine-tuning where the early model is better calibrated. Our analysis further shows that the early model preserves uncertainty awareness in both the predictions and the feature structure. However, this uncertainty is gradually lost as training continues, resulting in overconfidence.
- We propose EUA-Cal, a novel method that exploits the uncertainty awareness retained by the early model as an anchor. Our method is achieved through regularization terms that EPR aligns current predictions with early predictions and PSR derives probability distributions from early feature structures, jointly mitigating overconfidence.
- Our method achieves superior performance in calibration across diverse network architectures, including deep neural networks such as ResNet and DenseNet, as well as

Transformer-based models such as ViT, Llama, and Qwen. Extensive experiments on eight models and six datasets demonstrate that EUA-Cal outperforms state-of-the-art methods.

2 Related Work

Calibration for DNNs. Calibration aims to ensure that predicted confidence is indicative of the actual likelihood of correctness (Zhang et al., 2023). Post-hoc methods learn a calibration function on a validation set (Kull et al., 2017; Guo et al., 2017; Pan et al., 2020). Among them, Temperature Scaling (Guo et al., 2017) is a widely used method that learns a single temperature to adjust the output logits. Despite their simplicity and effectiveness, post-hoc methods perform poorly under distribution shifts (Ovadia et al., 2019). Loss function and regularization designs reduce overconfidence by implicitly or explicitly maximizing the entropy of predictive distributions. Label Smoothing (LS) (Szegedy et al., 2016) softens one-hot labels, while Focal Loss (FL) (Mukhoti et al., 2020) implicitly increases the entropy of the predicted distribution. MbLS (Liu et al., 2022) controls the margin on logit distances to prevent the model from non-informative predictions. BalCAL (Ni et al., 2025) combines a fixed ETF classifier with a standard linear classifier to adjust confidence. Mixup-based methods were originally proposed to improve generalization and have been proven effective for calibration (Thulasidasan et al., 2019; Noh et al., 2023; Bouniot et al., 2025). Wang et al. (2023) note that the confidence penalty of Mixup would harm calibration and propose MIT to avoid the negative impact of the confidence penalty. Several methods propose calibration methods for specific network structures (Jordahn & Olmos, 2024; Wang & Zhang, 2024; Ni et al., 2026), but their generalizability across different structures remains unclear.

Calibration for Fine-tuned LLMs. Recent studies have focused on fine-tuned LLM calibration (Jiang et al., 2021; Yang et al., 2024). Traditional methods, such as MC-Dropout (Gal & Ghahramani, 2016) and Deep Ensemble (Lakshminarayanan et al., 2017) provide uncertainty information but require multiple inferences when applied to fine-tuned LLMs. Several methods originally developed for the calibration of DNNs, such as Mixup and LS, have also been shown to effectively calibrate LLMs during fine-tuning (Park & Caragea, 2022; Huang et al., 2025a). Compared with these methods, our method is also applicable to fine-tuned LLMs and achieves superior performance. Distillation (Guo et al., 2021) improves calibration by knowledge distillation from calibrated model ensembles. However, this approach relies on an ensemble of multiple models and an additional validation set for calibration. IB-EDL (Li et al., 2025) introduces an information bottleneck into the EDL (Sensoy et al., 2018) framework to suppress spurious and excessive evidence, thus mitigating overconfidence in fine-tuned LLMs.

3 Methodology

In this section, we first provide the problem statement of model calibration in Sec. 3.1 and then introduce the proposed method, regarding details of early prediction regularization, prototype structure regularization, and the overall EUA-Cal objective in Sec. 3.2. The overview of our method is shown in Figure 2.

3.1 Problem Statement

We consider a K-class classification problem with a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a training sample, $y_i \in \mathcal{Y}$ is its corresponding label and N denotes the total number of samples. Let $f_{\boldsymbol{\theta}}$ be a model parameterized by $\boldsymbol{\theta}$ that produces a logit $\boldsymbol{z}_i = f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \in \mathbb{R}^K$, and the class probabilities are defined as $\boldsymbol{p}_i = \{p_{i,k}\}_{k=1}^{K} = softmax(\boldsymbol{z}_i)$, where $p_{i,k}$ is the predicted probability for class k. Based on these probabilities, the predicted label and confidence are defined as $\hat{y}_i = \arg\max_k p_{i,k}$ and $\hat{p}_i = \max_k p_{i,k}$, respectively. We further separate the model $f_{\boldsymbol{\theta}}$ into a feature extractor $h_{\boldsymbol{\varphi}} : \mathcal{X} \to \mathbb{R}^d$ and a linear classifier $\boldsymbol{W} \in \mathbb{R}^{d\times K}$, such that $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = h_{\boldsymbol{\varphi}}(\boldsymbol{x}_i)\boldsymbol{W}$. During training, we denote the parameters obtained at the t-th epoch by $\boldsymbol{\theta}^t = \{\boldsymbol{\varphi}^t, \boldsymbol{W}^t\}$, and the corresponding model is $f_{\boldsymbol{\theta}^t}(\cdot) = h_{\boldsymbol{\varphi}^t}(\cdot)\boldsymbol{W}^t$.

A model is well-calibrated if the predicted confidence $\hat{p}_i$ reflects the actual probability of correctness, i.e., $\mathbb{P}(\hat{y}_i = y_i | \hat{p}_i = p) = p$ for all $p \in [0, 1]$. However, DNNs trained or LLMs fine-tuned with CE

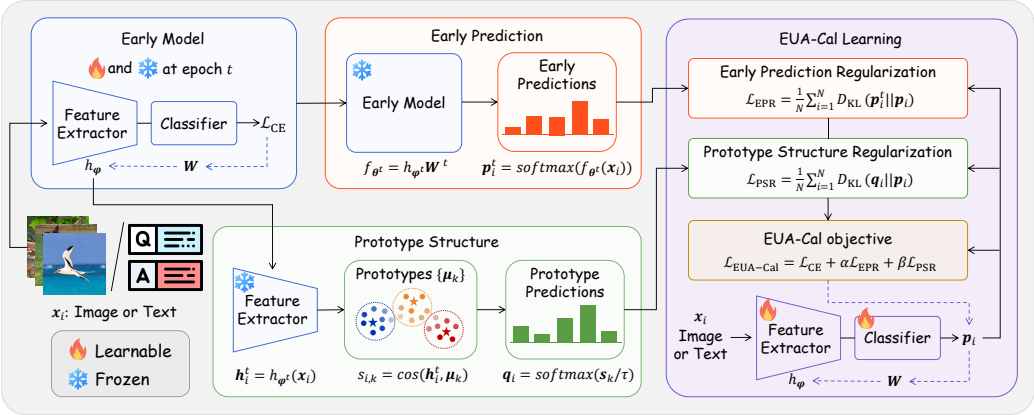


Figure 2: An illustration of the proposed EUA-Cal approach for calibration. In the pre-learning stage, we freeze the early model as an uncertainty anchor at the selected training epoch. Its predictions are used for EPR, while confidence-weighted class prototypes constructed from early features provide structural targets for PSR. The overall objective guides training or fine-tuning to preserve uncertainty and mitigate overconfidence.

loss tend to be overconfident, producing higher confidence than its accuracy. Confidence calibration aims to align predicted confidence with actual accuracy, thereby producing more reliable predictions.

3.2 Early Uncertainty Anchored Calibration

Based on previous analysis, we propose EUA-Cal, a novel method that exploits uncertainty awareness retained by the early model for calibration.

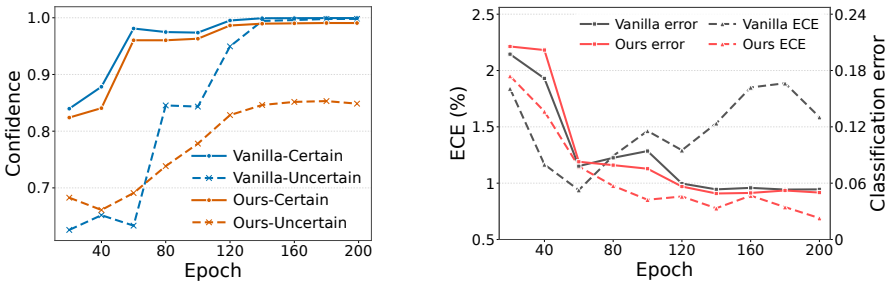
Early Prediction Regularization. As analyzed in Sec. 1, prediction of the early model is better calibrated and retains reliable uncertainty awareness for samples. Motivated by this insight, we introduce *Early Prediction Regularization* (EPR) to preserve the specific uncertainty of samples captured at the early training stage. Formally, given the parameters $\boldsymbol{\theta}^t$ from the selected training iteration t, we freeze the corresponding model $f_{\boldsymbol{\theta}^t}$. The EPR loss is defined as the Kullback-Leibler (KL) divergence between the early and current prediction distributions, i.e.,

$$\mathcal{L}_{\text{EPR}} = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}(\boldsymbol{p}_i^t || \boldsymbol{p}_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} p_{i,k}^t \log\left(\frac{p_{i,k}^t}{p_{i,k}}\right), \tag{1}$$

where $\boldsymbol{p}_i^t = \{p_{i,k}^t\}_{k=1}^K = \text{softmax}(f_{\boldsymbol{\theta}^t}(\boldsymbol{x}_i))$ is the early prediction distribution of sample $\boldsymbol{x}_i$ from the frozen early model. The traditional CE loss with hard label continuously pushes all samples toward one-hot prediction, regardless of the sample specific uncertainty. In contrast, early predictions retain sample uncertainty before the overconfidence emerges. EPR utilizes them as reliable references to constrain excessive confidence growth on uncertain samples, while preserving confident predictions for certain samples.

To verify the role of uncertainty awareness in early predictions, we divide the samples into certain and uncertain groups according to the confidence of the selected early model (e.g., epoch 60 on ResNet-50). Figure 3(a) illustrates the average confidence of CIFAR-10 samples. For certain samples, confidence remains high throughout training. However, samples identified as uncertain at the early stage gradually become highly confident under vanilla CE training, and their confidence eventually approaches that of certain samples. In contrast, our method leverages the uncertainty awareness captured by early predictions to prevent overconfident predictions on uncertain samples.

Prototype Structure Regularization. Beyond early predictions, the feature structure of the early model also preserves uncertainty awareness. We further introduce *Prototype Structure Regularization* (PSR) to exploit this uncertainty awareness of the early feature structure. Specifically, given the feature extractor $h_{\boldsymbol{\varphi}^t}$ of the frozen early model, we first construct the prototype for each class. For a sample $\boldsymbol{x}_i$, we extract its feature $\boldsymbol{h}_i^t$ from the frozen early model, i.e., $\boldsymbol{h}_i^t = h_{\boldsymbol{\varphi}^t}(\boldsymbol{x}_i) \in \mathbb{R}^d$. Given



(a) (b)

Figure 3: **(a)** Confidence of certain and uncertain samples under the vanilla CE model and our method. Samples are divided according to the confidence of the selected early model. **(b)** Classification errors and ECE of the prototype predictions on the validation set using prototypes constructed from different training epochs.

the uncertainty awareness in early predictions, we utilize the probability assigned by the early model to the true class of each sample as a reliability weight, i.e., $w_i = p^t_{i,y_i}$. Accordingly, the prototype $\boldsymbol{\mu}_k$ is constructed using normalized confidence-weighted averaging over the early features of class k, i.e.,

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} \mathbb{I}(y_i = k) w_i \overline{\boldsymbol{h}}^t_i}{\sum_{i=1}^{N} \mathbb{I}(y_i = k) w_i}, \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function and $\overline{\boldsymbol{h}}^t_i = \boldsymbol{h}^t_i / ||\boldsymbol{h}^t_i||_2$ is the normalized feature. Given the set of prototypes $\mathcal{M} = \{\boldsymbol{\mu}_k\}_{k=1}^K$, the similarity between the feature of the sample $\boldsymbol{x}_i$ in the early model and the prototype of class k is

$$s_{i,k} = \cos(\boldsymbol{h}^t_i, \boldsymbol{\mu}_k) = \frac{(\boldsymbol{h}^t_i)^\top \boldsymbol{\mu}_k}{||\boldsymbol{h}^t_i||_2 ||\boldsymbol{\mu}_k||_2}. \tag{3}$$

As observed in Figure 1(b), features near cluster boundaries tend to have lower confidence, whereas those closer to cluster centers tend to have higher confidence. This suggests that similarities to cluster centers, i.e., class prototypes, can reflect the early model's uncertainty awareness in its features. Accordingly, we convert the similarity into a probability distribution over classes for the sample $\boldsymbol{x}_i$ using the softmax function, i.e.,

$$\boldsymbol{q}_i = \{q_{i,k}\}_{k=1}^K = \text{softmax}\left(\frac{\boldsymbol{s}_i}{\tau}\right), \quad q_{i,k} = \frac{\exp(s_{i,k}/\tau)}{\sum_{j=1}^K \exp(s_{i,j}/\tau)}, \tag{4}$$

where $\boldsymbol{q}_i$ is defined as the prototype prediction of the sample $\boldsymbol{x}_i$, $\boldsymbol{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,K}]$ is the similarity vector between the feature of the sample $\boldsymbol{x}_i$ and the prototypes, and $\tau > 0$ is the scaling factor. The PSR loss is defined as the KL divergence between the prototype prediction distributions and the current prediction distributions, i.e.,

$$\mathcal{L}_{\text{PSR}} = \frac{1}{N} \sum_{i=1}^{N} D_{\text{KL}}(\boldsymbol{q}_i || \boldsymbol{p}_i) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} q_{i,k} \log\left(\frac{q_{i,k}}{p_{i,k}}\right). \tag{5}$$

Note that the prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^K$ are constructed from the frozen early model, preventing them from drifting during training. Figure 3(b) shows the classification errors and ECE of the prototype predictions for ResNet-50 on CIFAR-10. The ECE of prototype predictions from the vanilla model reaches its minimum at epoch 60 and increases in later stages. In contrast, the ECE of our prototype predictions continues to decrease, demonstrating better uncertainty awareness of the feature structure. The PSR loss leverages this awareness of uncertainty to further improve the calibration performance.

The overall objective of our method is:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{EUA-Cal}} = \min_{\boldsymbol{\theta}} \mathcal{L}_{\text{CE}} + \alpha \mathcal{L}_{\text{EPR}} + \beta \mathcal{L}_{\text{PSR}}, \tag{6}$$

Algorithm 1 EUA-Cal

Input: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, initial model $f_{\boldsymbol{\theta}} = h_{\boldsymbol{\varphi}}\boldsymbol{W}$, selected epoch t, parameters α and β, scaling factor τ, and total number of epochs T.
Output: Calibrated model $f_{\boldsymbol{\theta}^T}$.

```
Pre-learn f_θ with L_CE until epoch t to obtain the early model f_{θ^t}
Compute p_i^t = softmax(f_{θ^t}(x_i)), h̄_i^t = h_i^t / ||h_i^t||_2, and w_i = p_{i,y_i}^t for all (x_i, y_i) ∈ D.
Construct the class prototypes {μ_k}_{k=1}^K according to Eq. (2).
Compute the prototype similarities and distributions q_i according to Eq. (3) and Eq. (4).
for epoch from 1 to T do
    for each training batch sampled from D do
        Compute L_EPR and L_PSR according to Eq. (1) and Eq. (5).
        Update θ by minimizing L_EUA-Cal in Eq. (6).
    end for
end for
```

where α and β are hyperparameters that control the strengths of EPR loss and PSR loss, respectively. The pseudo-code of our EUA-Cal method is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Datasets and Models. We evaluate our method on both image classification and multiple-choice question answering. For image classification, we conduct experiments on CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), and Tiny-ImageNet (Deng et al., 2009) datasets with different network architectures, including ResNet-50, ResNet-101 (He et al., 2016), Wide-ResNet-26-10 (Zagoruyko & Komodakis, 2016), DenseNet-121 (Huang et al., 2017), and ViT-B/32 (Dosovitskiy et al., 2021). For multiple-choice question answering, we consider SciQ (Welbl et al., 2017), ARC-Challenge (ARC-C) and ARC-Easy (ARC-E) (Clark et al., 2018), and CommonsenseQA (CSQA) (Talmor et al., 2019) datasets. We fine-tune Llama3-8B, Llama3.1-8B (Grattafiori et al., 2024), and Qwen3-8B-Base (Yang et al., 2025) using LoRA (Hu et al., 2022).

Baselines and Metrics. For image classification, we compare EUA-Cal with various calibration methods: (i) the vanilla model trained with CE; (ii) implicit or explicit regularization, including LS (Szegedy et al., 2016), FL (Mukhoti et al., 2020), CRL (Moon et al., 2020), Distillation (Guo et al., 2021), MbLS (Liu et al., 2022), BalCAL (Ni et al., 2025), and CSAM (Tan et al., 2026); (iii) data augmentation, including Mixup (Thulasidasan et al., 2019) and MIT (Wang et al., 2023); (iv) architecture-specific methods, including PLP (Wang & Zhang, 2024) and MaC-Cal (Ni et al., 2026). For multiple-choice question answering with fine-tuned LLMs, we evaluate all methods above except BalCAL, PLP, and MaC-Cal since they rely on specific components. We further include (v) uncertainty-aware methods, including EDL (Sensoy et al., 2018) and IB-EDL (Li et al., 2025). Calibration is evaluated using ACC, ECE (Guo et al., 2017), and adaptive ECE (AECE) (Nixon et al., 2019). For out-of-distribution (OOD) detection, we adopt AUROC and FPR95 as metrics.

More implementation details including dataset splitting, training policies, hyperparameters, and evaluation protocols are provided in Appendix A.

4.2 Calibration Results

We provide comprehensive comparisons of our method against state-of-the-art methods for model calibration on both image classification and multiple-choice question answering. Table 1 reports the calibration performance of our method with ResNet-50 on CIFAR-10, CIFAR-100, and Tiny-ImageNet. Compared with other calibration methods, our method significantly improves calibration performance across all datasets and achieves the best average rank among all methods. Several methods improve calibration at the cost of accuracy, such as the FL and PLP methods. In contrast, our EUA-Cal method obtains the lowest ECE and AECE on all three datasets while maintaining

Table 1: Performance comparison between different methods with ResNet-50 on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	CIFAR-10			CIFAR-100			Tiny-ImageNet			Avg. Rank
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	
Vanilla	3.36±0.17	3.31±0.19	94.57±0.15	9.19±0.16	9.07±0.26	76.75±0.62	5.54±0.44	5.52±0.43	64.45±0.45	10.11
LS	3.74±0.24	3.93±0.33	94.78±0.13	3.96±0.22	3.99±0.24	76.77±0.13	3.78±0.27	3.70±0.36	63.58±0.67	8.78
FL	1.81±0.31	1.58±0.27	94.40±0.28	1.77±0.31	1.71±0.33	75.88±0.35	1.57±0.11	1.48±0.18	62.94±0.33	6.67
CRL	0.86±0.09	0.61±0.03	93.94±0.22	5.81±0.30	5.74±0.27	76.87±0.39	2.99±0.31	3.04±0.33	63.43±0.33	7.67
Distillation	3.22±0.52	3.27±0.49	94.10±0.67	7.51±0.20	7.40±0.16	76.64±0.12	6.04±0.24	6.01±0.19	65.21±0.20	10.22
MbLS	3.51±0.23	3.48±0.21	94.54±0.35	8.62±0.38	8.52±0.37	77.41±0.49	1.42±0.05	1.34±0.22	64.44±0.48	8.00
BalCAL	2.46±0.12	2.39±0.11	94.38±0.13	4.03±0.27	3.93±0.28	77.33±0.31	1.48±0.14	1.52±0.13	65.26±0.12	6.56
CSAM	1.04±0.15	0.90±0.19	95.36±0.20	1.40±0.39	1.30±0.50	**78.82±0.88**	6.10±0.13	6.12±0.13	**66.54±0.40**	4.67
Mixup	1.74±0.38	2.25±0.29	95.01±0.37	5.40±0.62	5.34±0.55	77.78±0.58	3.30±0.23	3.20±0.15	63.58±0.74	6.78
MIT-A	1.64±0.06	1.54±0.05	**95.57±0.03**	1.95±0.11	1.87±0.30	78.46±0.48	3.66±0.82	3.77±0.76	64.34±0.89	5.22
PLP	2.35±0.12	2.31±0.13	94.24±0.19	4.66±0.38	4.64±0.36	74.60±0.68	2.07±0.51	2.02±0.51	60.85±1.68	9.00
MaC-Cal	0.83±0.15	0.82±0.11	94.81±0.05	3.32±0.14	3.14±0.21	76.07±0.10	1.44±0.12	1.33±0.23	61.34±0.20	5.22
EUA-Cal	**0.54±0.12**	**0.50±0.09**	94.63±0.05	**1.28±0.11**	**1.18±0.21**	77.66±0.32	**1.38±0.17**	**1.29±0.11**	65.31±0.47	**2.00**

Table 2: Performance comparison between different methods with Llama3-8B on ARC-C, ARC-E, and SciQ datasets. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	ARC-C			ARC-E			SciQ			Avg. Rank
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	
Vanilla	17.24±0.25	17.12±0.31	81.31±0.64	6.54±0.03	6.48±0.03	92.81±0.12	6.42±0.38	6.10±0.41	93.40±0.52	8.56
LS	15.23±0.96	15.54±1.26	79.81±0.96	3.83±0.17	5.44±0.78	91.91±0.16	3.16±0.35	5.61±0.44	92.83±0.21	8.11
FL	12.97±1.31	12.85±0.93	80.94±0.81	4.34±0.24	4.91±0.24	91.93±0.10	4.49±0.12	4.78±0.22	93.63±0.06	5.89
CRL	18.17±0.62	18.04±0.55	80.46±1.19	6.98±0.84	6.88±0.86	91.83±0.87	6.21±0.22	5.99±0.27	93.27±0.38	10.67
Distillation	11.99±0.76	12.26±0.69	81.77±0.50	3.79±0.27	4.69±0.45	93.00±0.45	3.90±0.13	5.03±0.22	93.57±0.12	4.00
MbLS	12.17±10.50	13.39±8.18	80.97±0.82	7.62±0.40	7.60±0.40	91.98±0.32	6.73±0.09	6.71±0.10	93.07±0.12	9.89
CSAM	13.17±0.34	13.08±0.45	**82.19±0.69**	5.13±0.63	5.22±0.43	**93.23±0.65**	5.26±0.07	5.03±0.05	93.63±0.06	5.00
Mixup	10.53±0.68	10.86±0.29	79.61±0.48	6.16±0.24	6.32±0.25	91.88±0.28	5.84±0.34	5.76±0.30	93.67±0.25	7.56
MIT-A	6.52±0.39	7.10±0.66	81.31±0.53	5.87±0.07	5.80±0.08	92.75±0.21	4.47±0.14	4.33±0.27	**94.73±0.35**	4.00
EDL	11.25±0.59	13.25±0.66	80.23±0.34	2.60±0.17	5.75±0.16	92.51±0.15	2.86±0.05	6.22±0.19	93.47±0.15	6.67
IB-EDL	3.98±0.50	9.98±1.20	81.03±0.47	2.53±0.54	5.82±0.57	92.65±0.19	2.31±0.26	5.91±0.59	93.30±0.30	5.00
EUA-Cal	**3.60±0.34**	**4.17±0.60**	80.94±0.47	**1.67±0.19**	**1.79±0.29**	92.72±0.19	**2.19±0.24**	**2.57±0.49**	93.90±0.28	**2.22**

competitive accuracy. These results support the use of uncertainty information retained by the early model for calibration.

Table 2 reports the calibration results of Llama3-8B on SciQ, ARC-C, and ARC-E. Our method also achieves the best average rank and the lowest ECE and AECE on all three datasets. Compared with the second best method on each dataset, our method reduces ECE and AECE by 16.24% and 47.92% on average, respectively. Additional results on ResNet-101, Wide-ResNet-26-10, DenseNet-121, ViT-B/32, Llama3.1-8B, and Qwen3-8B-Base are provided in Appendix C.1. The consistent improvements across these models further demonstrate the effectiveness and generalizability of EUA-Cal across diverse architectures.

4.3 Calibration under Distribution Shift

Maintaining reliable confidence estimates under distribution shift is particularly important in safety-critical applications such as autonomous driving, where systems must make safe decisions even in adverse weather. To evaluate the calibration robustness of our EUA-Cal method in this setting, we first train ResNet-50 on CIFAR-10 and CIFAR-100 and then evaluate on their corrupted versions, CIFAR-10-C and CIFAR-100-C (Hendrycks & Dietterich, 2019), respectively. These corruptions include noise, blur, compression artifacts, and changes in brightness. Figure 4 illustrates the calibration performance of various methods with different shift levels. Across both datasets and all shift levels, our EUA-Cal method achieves lower ECE than most methods. This indicates that our EUA-Cal method maintains robust calibration performance under distribution shifts. In addition, we also present the OOD detection results in Table 4. Across all settings, our EUA-Cal method consistently outperforms the vanilla model, improving AUROC by an average of 1.31% and reducing FPR95

Table 3: Calibration performance of various methods under distribution shift. Llama3-8B is fine-tuned on ARC-E and evaluated on ARC-C, SciQ, and CSQA. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	ARC-C			SciQ			CSQA			Avg.
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	Rank
Vanilla	18.25±0.24	18.03±0.12	**79.81±0.05**	6.72±0.39	6.64±0.42	92.43±0.35	25.42±0.94	25.39±0.99	72.07±0.98	7.56
LS	18.69±0.26	18.80±0.43	76.17±0.32	4.98±0.82	7.65±0.20	90.90±0.90	21.82±1.55	21.66±1.60	72.24±1.44	8.11
FL	16.26±1.39	16.29±1.33	76.99±1.45	5.47±1.19	6.16±1.10	90.97±1.32	19.49±1.33	19.38±1.33	72.89±0.57	5.67
CRL	21.47±1.05	21.41±0.99	75.81±0.91	8.07±1.00	7.77±1.18	91.20±1.13	25.01±0.26	25.01±0.26	72.48±0.12	10.00
Distillation	16.55±0.78	16.53±0.65	77.87±1.04	5.17±0.62	6.03±0.78	92.00±0.95	21.84±0.49	21.78±0.56	72.67±0.58	5.78
MbLS	23.50±1.08	23.33±1.11	75.65±1.41	8.32±0.79	8.18±0.84	91.37±1.03	27.86±1.63	27.82±1.67	70.98±1.77	11.67
CSAM	20.80±0.43	20.63±0.53	77.93±0.75	7.21±0.32	6.98±0.29	92.37±0.40	25.33±1.06	25.20±1.09	**73.46±0.90**	7.78
Mixup	17.85±0.76	17.72±0.69	78.50±0.70	6.14±0.26	6.20±0.39	91.87±0.47	23.78±0.65	23.72±0.64	71.14±0.34	7.33
MIT-A	16.81±0.31	16.52±0.68	79.72±0.50	6.64±0.28	6.50±0.25	91.57±0.06	21.83±0.54	21.60±0.69	73.41±0.29	5.56
EDL	13.27±0.87	13.69±0.77	79.66±0.21	**2.32±0.83**	5.48±0.10	**92.73±0.42**	18.40±0.63	18.07±0.72	72.54±0.71	2.89
IB-EDL	11.14±0.12	12.02±0.35	79.80±0.49	3.17±0.94	5.79±0.59	92.43±0.15	16.05±1.22	16.55±1.02	72.18±1.02	3.00
EUA-Cal	**9.25±0.50**	**9.11±0.42**	77.59±0.47	2.50±0.36	**3.12±0.34**	92.17±0.45	**12.48±0.96**	**12.26±0.88**	73.11±0.45	**2.56**

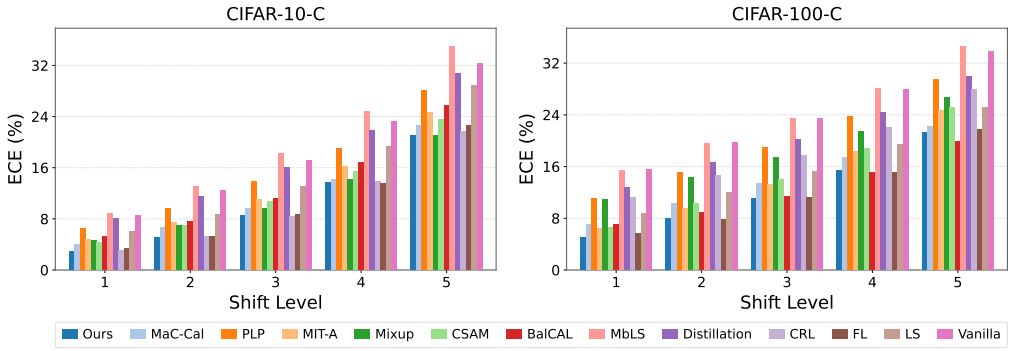


Figure 4: Calibration performance of various methods under distribution shift on CIFAR-10-C and CIFAR-100-C. The horizontal axis denotes the shift level from weakest to strongest and the vertical axis reports ECE (lower is better).

by 1.79%. These results demonstrate that EUA-Cal maintains reliable uncertainty awareness under distribution shift.

Practical deployment of LLMs requires them to generalize beyond the domains covered by their fine-tuning data. To evaluate calibration under such distribution shifts, we fine-tune Llama3-8B on ARC-E and then evaluate on ARC-C, SciQ, and CSQA, as reported in Table 3. Our method achieves the best average rank and the lowest AECE on all three datasets. It also achieves the lowest ECE on ARC-C and CSQA and ranks second on SciQ. Compared with the second best method on each dataset, our method reduces AECE by 30.94% on average. These results demonstrate that EUA-Cal maintains effective calibration performance when fine-tuning LLMs under distribution shift.

4.4 Further Analysis

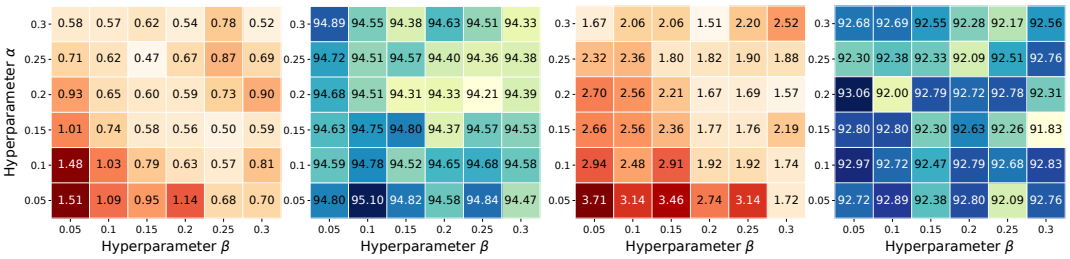
Ablation Study. As shown in Table 5, we evaluate several variants of our EUA-Cal method on ResNet-50 and Llama3-8B to investigate the effectiveness of EPR, PSR, and the selection of the anchor model. We find that the EPR variant reduces the ECE across all six datasets, while the PSR variant also improves calibration on most datasets. The complete EUA-Cal, which combines the two regularization terms, achieved the lowest ECE on each dataset, suggesting that predictions and feature structures of the early model provide complementary uncertainty awareness for calibration. For anchor model selection, we compare the early model with a late model that has higher validation accuracy but also higher validation ECE. Although the late model leads to slightly higher accuracy in most settings, its calibration is substantially worse. More analysis on the impact of model selection is provided in Appendix C.2.

Table 4: OOD detection results (%). The best and second best results are shown in bold and underlined, respectively.

ID DATASET	CIFAR-10				Tiny-ImageNet			
OOD DATASET	Tiny-ImageNet		CIFAR-100		CIFAR-10		CIFAR-100	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Vanilla	88.41	58.30	87.41	60.91	<u>85.01</u>	<u>67.25</u>	<u>83.61</u>	<u>68.40</u>
MbLS	88.97	58.91	87.37	63.08	81.57	74.82	80.16	75.47
BalCAL	87.43	56.97	86.67	60.71	83.34	71.71	82.32	71.60
PLP	88.95	60.63	88.24	63.56	80.94	81.21	80.37	79.44
MaC-Cal	<u>89.02</u>	**55.78**	<u>88.56</u>	**57.30**	83.27	74.60	82.29	74.24
EUA-Cal	**89.32**	<u>56.09</u>	**89.16**	<u>59.85</u>	**86.45**	**64.83**	**84.75**	**66.94**

Table 5: Ablation study of EPR, PSR, and the anchor model. The best and second best results are shown in **bold** and <u>underlined</u>, respectively.

EPR	PSR	Anchor Model	ResNet-50						Llama3-8B					
			CIFAR-10		CIFAR-100		Tiny-ImageNet		ARC-C		ARC-E		SciQ	
			ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑	ECE ↓	ACC ↑
×	×	–	3.36	94.57	9.19	76.75	5.54	64.45	17.24	**81.31**	6.54	<u>92.81</u>	6.42	93.40
✓	×	Early	<u>0.75</u>	<u>94.69</u>	3.30	77.57	<u>2.50</u>	<u>65.48</u>	<u>7.47</u>	<u>81.10</u>	<u>2.67</u>	91.25	4.21	93.30
×	✓	Early	0.83	94.67	<u>2.23</u>	77.37	6.62	64.41	7.79	81.07	4.67	91.33	<u>2.44</u>	93.10
✓	✓	Late	2.14	**94.78**	6.24	**78.11**	3.63	**66.15**	9.71	80.12	6.64	**92.85**	5.11	**94.30**
✓	✓	Early	**0.54**	94.63	**1.28**	<u>77.66</u>	**1.38**	65.31	**3.60**	80.94	**1.67**	92.72	**2.19**	<u>93.90</u>



(a) ResNet-50 on CIFAR-10 (b) Llama3-8B on ARC-E

Figure 5: Sensitivity to α and β of EUA-Cal. ECE (↓) is shown in warmer colors, and ACC (↑) is shown in cooler colors.

Impact of Hyperparameters. We explore the impact of hyperparameters α, β and τ on ResNet-50 and Llama3-8B. α and β control the strengths of EPR loss and PSR loss, respectively. We recommend setting them within the range of $[0.1, 0.3]$. As shown in Figure 5, our EUA-Cal method consistently improves the calibration performance within this range. Figure 9 shows the performance of EUA-Cal under different values of τ, where τ^* is obtained by minimizing NLL on the training set. Our EUA-Cal method maintains relatively stable accuracy and competitive calibration performance under different τ settings. We recommend determining τ through minimizing NLL on the training set, avoiding a separate hyperparameter search. For more detailed analyses, see Appendix C.3.

5 Conclusion

In this work, we observe a consistent phenomenon across different models that the early model is better calibrated, while continued training often yields modest accuracy gains while worsening calibration. Motivated by this observation, we proposed EUA-Cal, a novel method that exploits the early model as an uncertainty anchor for calibration. Our method is achieved through the regularization terms EPR and PSR to leverage the uncertainty awareness of the early model for calibration. Extensive experiments on image classification and multiple-choice question answering on eight models indicate that our EUA-Cal method consistently improves calibration performance while maintaining competitive accuracy across different models and tasks.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with grammar, spelling, and sentence clarity. We have reviewed all AI-assisted work and assume full responsibility for all content generated by the LLM. The generative AI contributed in no other way to this work. The research idea, experiment design, and all other content were developed and completed by ourselves.

ETHICS STATEMENT

We confirm that our work adheres to the ICLR Code of Ethics. The research utilizes publicly available datasets that contain no personally identifiable information. As a strictly technical contribution, we foresee no ethical issues or potential misuse.

REPRODUCIBILITY STATEMENT

The complete code for our proposed method is provided in the Supplementary Material to facilitate the reproduction of our results. Our experimental implementation is also described in Appendix A.

REFERENCES

Quentin Bouniot, Pavlo Mozharovskyi, and Florence d'Alché-Buc. Tailoring mixup to data for calibration. In *The Thirteenth International Conference on Learning Representations*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *The Ninth International Conference on Learning Representations*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pp. 1050–1059, 2016.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. An overview of uncertainty calibration for text classification and the role of distillation. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, pp. 289–306, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *The Seventh International Conference on Learning Representations*, 2019.

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

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, 2017.

Jerry Huang, Peng Lu, and Qiuhao Zeng. Calibrated language models and how to find them with label smoothing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 25593–25612, 2025a.

Xijie Huang, Xinyuan Wang, Hantao Zhang, Yinghao Zhu, Jiawen Xi, Jingkun An, Hao Wang, Hao Liang, and Chengwei Pan. Medical mllm is vulnerable: cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, pp. 3797–3805, 2025b.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Mikkel Jordahn and Pablo M. Olmos. Decoupling feature extraction and classification layers for calibrated neural networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 22530–22550, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003, 2023.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

Meelis Kull, Telmo de Menezes e Silva Filho, and Peter A. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 623–631, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Yawei Li, David Rügamer, Bernd Bischl, and Mina Rezaei. Calibrating llms with information-theoretic evidential deep learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 80–88, 2022.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7034–7044, 2020.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15288–15299, 2020.

Jiani Ni, He Zhao, Jintong Gao, Dandan Guo, and Hongyuan Zha. Balancing two classifiers via a simplex etf structure for model calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30712–30721, 2025.

Jiani Ni, He Zhao, Yibo Yang, and Dandan Guo. Deep neural network calibration by reducing classifier shift with stochastic masking. *Pattern Recognit.*, 177:113217, 2026.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–41, 2019.

Jongyoun Noh, Hyekang Park, Junghyup Lee, and Bumsub Ham. Rankmixup: Ranking-based mixup training for network calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1358–1368, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13969–13980, 2019.

Feiyang Pan, Xiang Ao, Pingzhong Tang, Min Lu, Dapeng Liu, Lei Xiao, and Qing He. Field-aware calibration: A simple and empirically strong method for reliable probabilistic predictions. In *Proceedings of The Web Conference*, pp. 729–739, 2020.

Seo Yeon Park and Cornelia Caragea. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5364–5374, May 2022.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pp. 3183–3193, 2018.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *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, 2019.

Chengli Tan, Yubo Zhou, Haishan Ye, Guang Dai, Junmin Liu, Zengjie Song, Jiangshe Zhang, Zixiang Zhao, Yunda Hao, and Yong Xu. Towards understanding the calibration benefits of sharpness-aware minimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13888–13899, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Deng-Bao Wang and Min-Ling Zhang. Calibration bottleneck: Over-compressed representations are less calibratable. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 52156–52170, 2024.

Deng-Bao Wang, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Min-Ling Zhang. On the pitfall of mixup for uncertainty calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7609–7618, 2023.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

Ruifei Zhang, Wei Zhang, Xiao Tan, Sibei Yang, Xiang Wan, Xiaonan Luo, and Guanbin Li. Vldrive: Vision-augmented lightweight mllms for efficient language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5923–5933, 2025.

Xu-Yao Zhang, Guo-Sen Xie, Xiuli Li, Tao Mei, and Cheng-Lin Liu. A survey on learning to reject. *Proceedings of the IEEE*, 111(2):185–215, 2023.

A Implementation Details

A.1 Datasets

To comprehensively evaluate the effectiveness of our method, we adopt CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), and Tiny-ImageNet (Deng et al., 2009) for image classification, and SciQ (Welbl et al., 2017), ARC-C and ARC-E (Clark et al., 2018), and CommonsenseQA (CSQA) (Talmor et al., 2019) for multiple-choice question answering.

CIFAR-10 and CIFAR-100 each contain 60,000 RGB images of size 32×32, covering 10 and 100 classes, respectively. For each dataset, we divide the original 50,000 training images into 45,000 training and 5,000 validation images, and retain the 10,000 test images.

Tiny-ImageNet contains 200 classes of natural images at 64×64 resolution. We divide its 100,000 training images into 90,000 for training and 10,000 for validation, and evaluate on the labeled test set of 10,000 images.

SciQ contains 13,679 four-choice science questions that cover physics, chemistry, biology, and related subjects. We use the official training/validation/test splits of 11,679/1,000/1,000 examples. During fine-tuning, each question and its answer choices are provided as input, while the correct answer serves as the supervision label.

ARC-C and ARC-E consist of grade-school science examination questions. Specifically, ARC-C contains challenging questions that retrieval-based and word co-occurrence baselines fail to answer, while ARC-E contains the remaining questions. We follow the official training/validation/test splits of 1,119/299/1,172 for ARC-C and 2,251/570/2,376 for ARC-E.

CommonsenseQA contains 12,102 five-choice questions requiring common sense knowledge about everyday situations. Its official training/development/test splits contain 9,741/1,221/1,140 examples. Since test labels are unavailable, we use the labeled development set for evaluation.

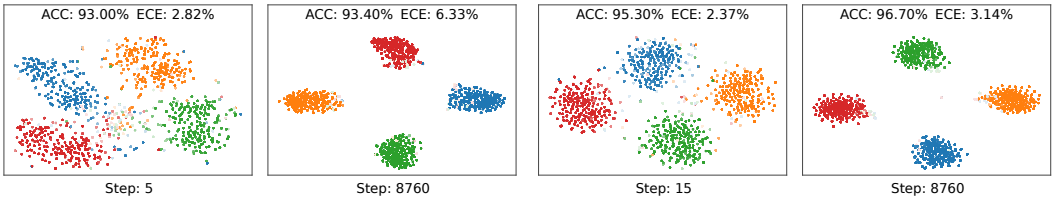
A.2 Training Details

For image classification, we conduct experiments on five models including ResNet-50, ResNet-101, Wide-ResNet-26-10, DenseNet-121, and ViT-B/32. To learn these models, we use SGD as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4} and an initial learning rate of 0.1 unless otherwise specified. On CIFAR-10 and CIFAR-100, models are trained for 200 epochs with a batch size of 128. The learning rate is divided by 5 at epochs 60, 120, and 160. On Tiny-ImageNet, models are trained for 100 epochs with a batch size of 64. The learning rate is divided by 10 at epochs 40 and 60. We initialize ViT-B/32 from ImageNet-1K pretrained weights and fine-tune all its parameters using 224×224 inputs. For ViT-B/32 on CIFAR, we use a batch size of 64, an initial learning rate of 0.01 with cosine decay, no weight decay, and a gradient clipping norm of 1.

For Multiple-choice question answering, we fine-tune Llama3-8B, Llama3.1-8B, and Qwen3-8B-Base using LoRA. Following (Li et al., 2025), we define the label space $\mathcal{Y}$ as the tokens corresponding to the possible options (A/B/C/D). The adapters are applied to the query, key, value, and output projections, with rank 8, scaling factor 16, and dropout 0.1. The learning rate is set to 5×10^{-5} and annealed using a cosine schedule. The maximum token length is limited to 256 for all other datasets. The training is conducted with bfloat16 precision.

A.3 Hyperparameter Settings

We compare our EUA-Cal method with various calibration methods and provide hyperparameter settings of these baseline methods as follows. (a) LS (Szegedy et al., 2016): We use a smoothing factor of $\alpha = 0.05$. (b) FL (Mukhoti et al., 2020): We set the focusing parameter to $\gamma = 2$. (c) CRL (Moon et al., 2020): We set the coefficient of the confidence ranking loss to 1.0. (d) Distillation (Guo et al., 2021): We use two teachers and set the distillation loss weight to 1.0. The distillation temperature is determined by minimizing the validation Brier score. (e) MbLS (Liu et al., 2022): We use a margin of $m = 10$ and a regularization coefficient of 0.1. (f) BalCAL (Ni et al., 2025): We set $\beta = 1, \delta = 0.95$ with a dynamic γ. (g) CSAM (Tan et al., 2026): We use a perturbation radius $\rho \in [0.05, 0.2]$ and $\gamma \in [0.5, 2.0]$. (h) Mixup (Thulasidasan et al., 2019): We use $\alpha = 0.2$. (i) MIT



(a) Llama3-8B on SciQ (b) Qwen3-8B-Base on SciQ

Figure 6: t-SNE visualizations of features from Llama3-8B and Qwen3-8B-Base on SciQ at different fine-tuning steps. Lighter shades indicate features associated with lower confidence.

Table 6: Performance comparison between different methods with ResNet-101 on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	CIFAR-10			CIFAR-100			Tiny-ImageNet			Avg. Rank
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	
Vanilla	3.26±0.13	3.21±0.12	94.81±0.16	8.97±0.28	8.94±0.27	78.21±0.67	5.95±0.39	5.99±0.36	65.74±0.26	9.78
LS	4.05±0.27	4.21±0.18	94.72±0.38	3.81±0.47	4.01±0.40	77.57±0.48	1.90±0.05	1.69±0.17	66.22±0.38	8.33
FL	1.63±0.14	1.42±0.09	94.68±0.26	1.57±0.26	1.61±0.19	77.03±0.61	1.38±0.08	1.29±0.19	63.68±0.29	6.11
CRL	0.83±0.13	0.67±0.05	94.10±0.15	5.91±0.24	5.78±0.06	77.36±0.50	3.57±0.19	3.49±0.24	64.73±0.28	8.33
Distillation	2.87±0.15	2.95±0.14	94.86±0.04	7.17±0.39	7.12±0.35	77.93±0.26	7.36±0.14	7.36±0.13	66.04±0.15	9.33
MbLS	3.57±0.08	3.58±0.10	94.71±0.02	9.21±0.53	9.18±0.48	77.84±0.53	1.49±0.30	1.34±0.16	65.68±0.56	9.44
BalCAL	2.23±0.20	2.08±0.18	94.75±0.07	3.56±0.13	3.40±0.09	78.17±0.35	1.32±0.22	1.44±0.23	66.59±0.50	5.44
CSAM	1.36±0.08	1.28±0.07	95.53±0.14	1.64±0.26	1.70±0.25	**79.60±0.06**	5.72±0.26	5.70±0.30	**67.12±0.42**	4.44
Mixup	1.61±0.41	2.21±0.06	95.47±0.20	4.93±0.83	4.88±0.85	78.52±0.49	1.89±0.54	1.79±0.70	65.19±0.84	6.44
MIT-A	1.66±0.19	1.60±0.18	**95.87±0.19**	2.07±0.57	2.05±0.54	79.13±0.46	2.84±0.73	2.76±0.76	65.31±0.42	5.56
PLP	2.25±0.17	2.22±0.17	94.04±0.37	8.76±0.34	8.75±0.34	75.74±0.45	2.23±0.36	2.06±0.47	59.49±2.05	10.56
MaC-Cal	1.18±0.11	1.25±0.14	94.94±0.03	3.97±0.18	3.80±0.20	78.31±0.07	1.27±0.26	1.30±0.16	63.22±0.27	5.00
EUA-Cal	**0.52±0.10**	**0.49±0.09**	94.74±0.14	**1.07±0.27**	**1.03±0.28**	78.58±0.19	**1.07±0.10**	**1.11±0.14**	66.24±0.08	**2.22**

(Wang et al., 2023): We adopt MIT-A with $\alpha = 0.5$ and $\Delta\lambda > 1/2$. (j) PLP(Wang & Zhang, 2024): We set the hyperparameter γ to 1.0.(k) MaC-Cal (Ni et al., 2026): We set the weighting factor γ to 0.7. (l) EDL (Sensoy et al., 2018): We limit the annealing coefficient to 1.0. (m) IB-EDL (Li et al., 2025): We use $K = 20$ samples and $\beta \in [5 \times 10^{-7}, 9 \times 10^{-5}]$, depending on the dataset.

For our EUA-Cal experiments, we choose α and β from $[0.1, 0.3]$ and determine τ by minimizing the NLL of prototype predictions on training samples. The anchor is selected at epoch 60 for CIFAR-10 and CIFAR-100 and epoch 40 for Tiny-ImageNet, except for ViT-B/32, which uses epoch 5. In LLM experiments, we use either the pretrained model or a model obtained after 5-15 fine-tuning steps as the anchor, depending on the model and dataset.

A.4 Evaluation Protocols

We report expected calibration error (ECE) and adaptive expected calibration error (AECE) to evaluate confidence calibration. ECE measures the difference between accuracy and confidence in expectation, i.e., $\mathbb{E}[|P(\hat{y}_i = y_i|\hat{p}_i) - \hat{p}_i|]$. Dividing the confidence interval $[0, 1]$ into M equally spaced bins $\{B_m\}_{m=1}^{M}$, ECE is defined as the weighted average of the difference between accuracy and confidence in each bin, i.e.,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{7}$$

where N is the total number of samples, $|B_m|$ is the number of samples in bin m, $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denote the accuracy and the average confidence of the samples in this bin, respectively. Unlike ECE, AECE adaptively divides the confidence interval with the same sample size. Lower ECE and AECE indicate better calibration performance. In our experiments, we set the number of bins as 20 for both ECE and AECE. The results are reported as the mean and standard deviation over three random seeds.

Table 7: Performance comparison between different methods with Wide-ResNet-26-10 on CIFAR-10 and CIFAR-100. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	CIFAR-10			CIFAR-100		
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑
Vanilla	2.52±0.07	2.43±0.05	95.84±0.08	5.41±0.12	5.26±0.17	79.64±0.19
MbLS	2.79±0.10	2.75±0.12	95.64±0.11	7.48±0.11	7.32±0.09	79.66±0.15
BalCAL	2.57±0.24	2.40±0.12	95.69±0.07	4.76±0.20	4.53±0.33	78.39±0.58
CSAM	0.70±0.07	**0.54±0.07**	96.52±0.08	1.95±0.16	1.84±0.16	**81.70±0.13**
MIT-A	1.30±0.11	1.28±0.10	**96.70±0.12**	2.98±0.15	2.89±0.20	81.30±0.12
MaC-Cal	1.18±0.16	1.25±0.03	95.93±0.04	2.70±0.10	2.55±0.03	78.87±0.11
EUA-Cal	**0.59±0.12**	0.58±0.14	95.79±0.07	**1.72±0.12**	**1.61±0.15**	80.95±0.04

Table 8: Performance comparison between different methods with DenseNet-121 on CIFAR-10 and CIFAR-100. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	CIFAR-10			CIFAR-100		
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑
Vanilla	2.92±0.18	2.87±0.18	94.69±0.19	8.38±0.26	8.34±0.25	76.11±0.25
MbLS	2.96±0.04	2.92±0.07	94.25±0.20	2.38±0.33	2.12±0.33	75.95±0.44
BalCAL	1.07±0.02	0.99±0.08	93.96±0.41	7.93±0.49	7.93±0.49	74.41±0.52
CSAM	4.33±0.12	4.32±0.10	94.50±0.07	4.42±0.31	4.40±0.31	77.26±0.21
MIT-A	0.90±0.04	0.78±0.01	**95.87±0.12**	2.77±0.10	2.40±0.22	**78.57±0.31**
MaC-Cal	0.83±0.07	0.70±0.03	94.28±0.07	1.41±0.07	1.25±0.06	70.92±0.62
EUA-Cal	**0.55±0.11**	**0.40±0.08**	94.34±0.09	**1.13±0.21**	**1.03±0.26**	76.27±0.18

B Visualizations

Figure 6 illustrates a similar pattern during LLM fine-tuning. For both models, the later features form more distinct class clusters, while the number of the low-confidence samples visible at the early steps decreases. Accuracy increases from 93.00% to 93.40% for Llama3-8B and from 95.30% to 96.70% for Qwen3-8B-Base, but ECE also rises from 2.82% to 6.33% and from 2.37% to 3.14%, respectively. This indicates that the uncertainty awareness of the features preserved by the early model is lost in the later training stage.

C More Experimental Results

C.1 Calibration on Additional Model Architecture

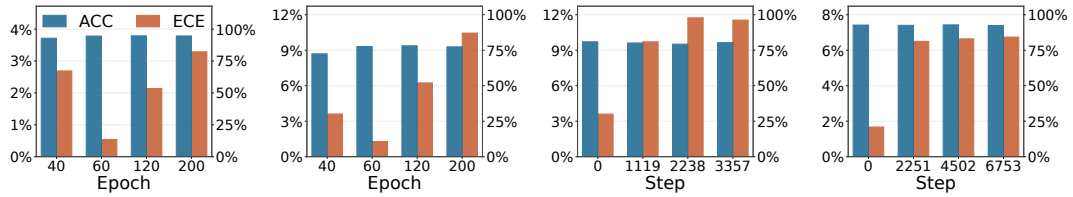
Tables 6–9 report additional image classification results on ResNet-101, Wide-ResNet-26-10, DenseNet-121, and ViT-B/32. Our EUA-Cal method achieves the lowest ECE on every dataset for all four models and the lowest AECE in eight of the nine cases while maintaining competitive accuracy. In particular, on ViT-B/32, it reduces ECE from 1.92% to 0.48% on CIFAR-10 and from 9.56% to 0.89% on CIFAR-100 compared with the vanilla model. Table 10 and Table 11 present multiple-choice question answering results on Llama3.1-8B and Qwen3-8B-Base, respectively. Our EUA-Cal method achieves the best average rank on both models, with the lowest AECE in five of six settings. These results further demonstrate its effectiveness across diverse models and tasks.

C.2 Impact of Anchor Model Selection

Figure 7 illustrates how the selection of anchor affects EUA-Cal. On CIFAR-10 and CIFAR-100, the model at epoch 60 yields the lowest ECE. Selecting the model at epoch 40 results in lower accuracy, whereas later models substantially increase the ECE with marginal accuracy gain. For ARC-C and ARC-E, using the pretrained Llama3-8B as the anchor model yields the lowest ECE. Later models

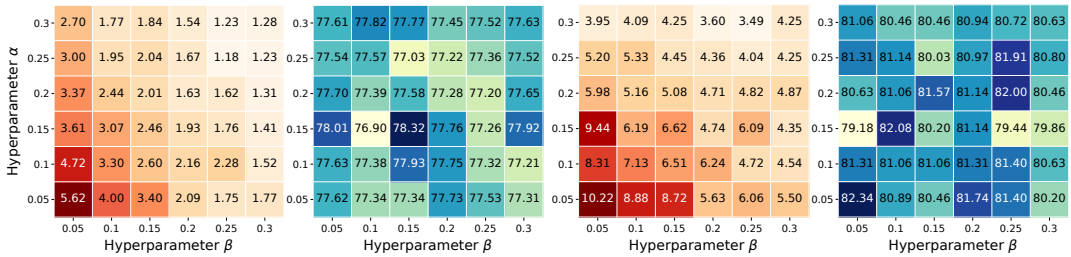
Table 9: Performance comparison between different methods with ViT-B/32 on CIFAR-10 and CIFAR-100. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	CIFAR-10			CIFAR-100		
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑
Vanilla	1.92±0.10	1.90±0.10	97.78±0.10	9.56±0.06	9.54±0.07	87.61±0.07
MbLS	1.80±0.09	1.77±0.11	97.90±0.10	8.65±0.19	8.58±0.19	87.71±0.16
BalCAL	3.88±0.08	4.05±0.14	97.67±0.07	4.83±0.16	5.60±0.21	87.49±0.19
CSAM	1.52±0.06	1.49±0.05	97.56±0.02	5.32±0.03	5.28±0.04	87.26±0.01
MIT-L	1.20±0.02	1.13±0.04	**98.13±0.05**	6.67±0.06	6.65±0.05	**87.94±0.13**
MaC-Cal	0.60±0.03	0.81±0.01	97.80±0.01	3.85±0.26	3.77±0.34	87.03±0.13
EUA-Cal	**0.48±0.08**	**0.20±0.03**	97.43±0.01	**0.89±0.07**	**0.84±0.13**	87.54±0.12



(a) CIFAR-10 (b) CIFAR-100 (c) ARC-C (d) ARC-E

Figure 7: Impact of anchor model selection. The horizontal axis indicates the epoch at which the anchor model is selected for ResNet-50 on CIFAR-10 and CIFAR-100, or the optimization step for Llama3-8B on ARC-C and ARC-E.



(a) ResNet-50 on CIFAR-100 (b) Llama3-8B on ARC-C

Figure 8: Sensitivity to α and β of EUA-Cal. ECE (↓) is shown in warmer colors, and ACC (↑) is shown in cooler colors.

increase the ECE, while the accuracy changes only slightly. These results indicate that the selection of the anchor model should consider the balance between classification performance and calibration performance, rather than simply selecting the earliest model.

C.3 Impact of Hyperparameters

Sensitivity to α and β. Figure 8 illustrates that our EUA-Cal method improves calibration performance for most settings of α and β on both CIFAR-100 and ARC-C, while ACC remains relatively stable. This indicates that the calibration improvement of our method is not restricted to a particular choice of α and β, suggesting that EUA-Cal is robust to these hyperparameters.

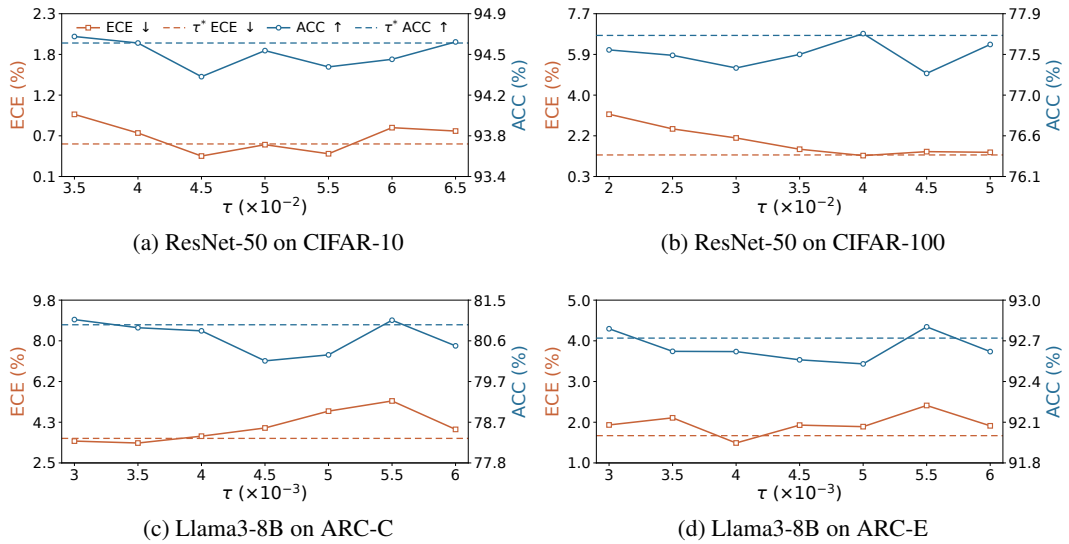
Sensitivity to τ. Figure 9 reports the ACC and ECE of our EUA-Cal method under different values of τ on CIFAR-10, CIFAR-100, ARC-C, and ARC-E. As τ varies, ACC remains relatively stable, while ECE varies more. Despite this variation, selecting τ^* by minimizing training-set NLL achieves competitive ECE on all four datasets without a separate search over τ.

Table 10: Performance comparison between different methods with Llama3.1-8B on ARC-C, ARC-E, and SciQ datasets. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	ARC-C			ARC-E			SciQ			Avg. Rank
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	
Vanilla	17.36±0.59	17.29±0.53	81.54±0.38	6.75±0.11	6.73±0.07	92.76±0.11	5.83±0.10	5.72±0.13	94.03±0.12	10.00
LS	14.30±0.27	14.67±0.35	80.66±0.34	**3.01±0.08**	5.11±0.23	92.55±0.11	**1.79±0.32**	4.57±0.29	94.43±0.31	5.56
FL	12.40±2.18	12.74±1.64	80.97±0.12	4.05±0.25	4.03±0.08	92.92±0.09	3.50±0.17	4.21±0.17	94.27±0.31	4.56
CRL	16.26±1.29	16.01±1.38	82.30±0.18	5.80±0.09	5.71±0.03	**93.10±0.11**	5.23±0.15	4.94±0.22	94.37±0.15	6.67
Distillation	9.12±0.90	10.41±0.58	81.48±0.64	3.03±0.10	4.06±0.14	92.80±0.15	3.81±0.18	4.97±0.22	94.10±0.10	5.00
MbLS	18.41±1.20	18.33±1.17	80.69±1.16	6.73±0.03	6.64±0.08	92.87±0.09	5.63±0.28	5.51±0.33	94.23±0.29	9.56
CSAM	15.25±0.42	15.24±0.42	**83.22±0.43**	4.85±0.21	4.77±0.15	92.42±0.11	3.87±0.10	3.73±0.08	94.53±0.15	5.56
Mixup	12.76±0.57	12.96±0.57	80.20±0.30	6.07±0.28	6.07±0.26	92.17±0.15	4.52±0.38	4.54±0.40	**95.00±0.46**	7.78
MIT-A	7.87±0.45	7.79±0.45	80.57±0.30	5.80±0.22	5.73±0.15	92.66±0.28	4.55±0.30	4.40±0.42	94.77±0.31	5.78
EDL	11.15±0.19	12.07±0.63	80.69±0.40	4.26±0.53	5.29±0.48	92.45±0.25	1.88±0.36	5.42±0.46	94.37±0.38	6.00
IB-EDL	4.86±0.29	9.82±0.32	80.77±0.21	6.30±0.26	6.34±0.20	92.35±0.06	1.97±0.18	5.82±0.20	94.10±0.10	7.44
EUA-Cal	**3.82±0.49**	**3.95±0.30**	82.42±0.39	3.38±0.05	**3.34±0.19**	92.35±0.11	1.86±0.14	**1.83±0.25**	94.10±0.17	**3.33**

Table 11: Performance comparison between different methods with Qwen3-8B-Base on ARC-C, ARC-E, and SciQ datasets. Results are reported as mean ± standard deviation (%). The best and second best results are shown in **bold** and underlined, respectively.

Method	ARC-C			ARC-E			SciQ			Avg. Rank
	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	ECE ↓	AECE ↓	ACC ↑	
Vanilla	8.10±0.71	7.89±0.70	91.50±0.56	2.41±0.02	2.32±0.07	97.31±0.08	3.64±0.15	3.51±0.15	96.27±0.12	10.22
LS	3.53±0.03	5.62±0.25	91.75±0.27	3.86±0.18	3.94±0.10	97.40±0.10	2.75±0.39	4.42±0.09	96.53±0.23	8.22
FL	3.33±0.23	3.60±0.25	91.70±0.05	**1.21±0.14**	**1.01±0.04**	97.45±0.07	1.67±0.39	2.31±0.44	96.20±0.40	4.44
CRL	5.00±0.11	5.14±0.06	91.70±0.05	1.36±0.08	1.37±0.10	97.46±0.09	3.07±0.13	2.87±0.42	96.10±0.36	7.33
Distillation	3.77±0.11	4.57±0.26	91.87±0.20	1.47±0.03	1.46±0.02	96.97±0.00	2.60±0.31	3.74±0.19	96.07±0.06	7.44
MbLS	6.93±0.19	6.67±0.08	92.18±0.13	2.25±0.04	2.25±0.08	97.38±0.09	3.28±0.07	2.96±0.08	96.77±0.12	7.78
CSAM	4.02±0.20	3.62±0.08	92.28±0.06	1.27±0.07	1.15±0.05	97.66±0.05	2.01±0.22	1.52±0.09	**97.13±0.06**	3.00
Mixup	4.41±0.23	4.93±0.18	92.09±0.10	1.33±0.15	1.76±0.13	97.53±0.13	2.91±0.03	2.89±0.03	96.63±0.06	5.89
MIT-A	3.71±0.64	3.48±0.28	**92.89±0.21**	1.56±0.11	1.44±0.05	**97.97±0.02**	2.61±0.14	2.30±0.23	97.07±0.21	3.44
EDL	5.48±0.30	5.91±0.63	92.35±0.32	3.02±0.19	3.69±0.32	97.35±0.18	1.80±0.37	3.97±0.10	96.67±0.06	7.56
IB-EDL	8.32±0.14	10.97±0.09	92.18±0.05	3.87±0.24	4.67±0.16	97.57±0.16	2.96±0.67	4.80±0.16	96.43±0.25	9.33
EUA-Cal	**2.04±0.46**	**2.68±0.10**	91.72±0.26	1.33±0.08	1.11±0.06	97.49±0.13	**1.55±0.42**	**1.25±0.12**	96.77±0.15	**2.89**



(a) ResNet-50 on CIFAR-10

(b) ResNet-50 on CIFAR-100

(c) Llama3-8B on ARC-C

(d) Llama3-8B on ARC-E

Figure 9: Sensitivity to τ of EUA-Cal. ECE (↓) is shown in warmer colors, and ACC (↑) is shown in cooler colors.