AdUE: Improving uncertainty estimation head for LoRA adapters in LLMs

Anonymous ACL submission

Abstract

Uncertainty estimation remains a critical challenge in adapting pre-trained language models to classification tasks, particularly under parameter-efficient fine-tuning approaches such as adapters. We introduce AdUE[1], an efficient post-hoc uncertainty estimation (UE) method, to enhance softmax-based estimates. Our approach (1) uses a differentiable approximation of the maximum function and (2) applies additional regularization through L2-SP, anchoring the fine-tuned head weights and regularizing the model. Evaluations on five NLP classification datasets across four language models (RoBERTa, ELECTRA, LLaMA-2, Qwen) demonstrate that our method consistently outperforms established baselines such as Mahalanobis distance and softmax response. Our approach is lightweight (no base-model changes) and produces better-calibrated confidence.

1 Introduction

Large-scale pretrained language models (LLMs) have become foundational tools in natural language processing (NLP), offering strong performance on a wide range of tasks through transfer learning. To adapt such models to downstream classification problems, *parameter-efficient fine-tuning* methods—such as adapters (Houlsby et al., 2019) and LoRA (Hu et al., 2022)—have emerged as scalable alternatives to full model fine-tuning. These methods update only a small subset of parameters while retaining the majority of the pre-trained backbone, enabling efficient adaptation with reduced computational and storage costs.

Despite their advantages, adapter-based methods are prone to overfitting, especially on low-resource tasks, and often produce overconfident predictions even when incorrect. This limits their reliability in risk-sensitive applications such as medical triage, toxic content filtering, or automated moderation. A principled estimate of predictive *uncertainty*—the model's confidence in its own outputs—is therefore essential. Common uncertainty estimation techniques include softmax response (Geifman and El-Yaniv, 2017), Mahalanobis distance (Lee et al., 2018)and deep ensembles (Lakshminarayanan et al., 2017). However, these approaches face trade-offs between computational cost, scalability, and generalization.

We revisit the softmax response as a lightweight yet effective uncertainty measure. Although softmax-based confidence is often unreliable due to model miscalibration (Guo et al., 2017), we show that it can be improved *post-hoc* by fine-tuning a small classification head using a *smooth approximation of the max function*. Our method **AdUE** builds on frozen, fine-tuned adapter models and optimizes a *multi-task objective* that combines binary classification loss, regularization towards the original softmax output, and L2-SP parameter anchoring (Xuhong et al., 2018). This approach enhances uncertainty estimation without degrading task performance or inducing forgetting.

We evaluate our approach across five diverse text classification datasets—SST-2, SST-5, CoLA, 20 Newsgroups, and ToxiGen— and across four Transformer-based architectures representing both encoders (RoBERTa, ELECTRA) and decoders (LLaMA-2, Qwen). Our experiments demonstrate that the proposed softmax response fine-tuning (AdUE) outperforms baselines Mahalanobis-based and robust distance-based uncertainty estimation methods in terms of AUC-ROC between predicted confidence and classification error.

Our main contributions are as follows:

- We propose AdUE — a lightweight post-hoc fine-tuning method based on softmax response uncertainty estimation in adapter-based LLMs. Figure 1 illustrates the approach.

[1] https://anonymous.4open.science/r/AdUE-BB68

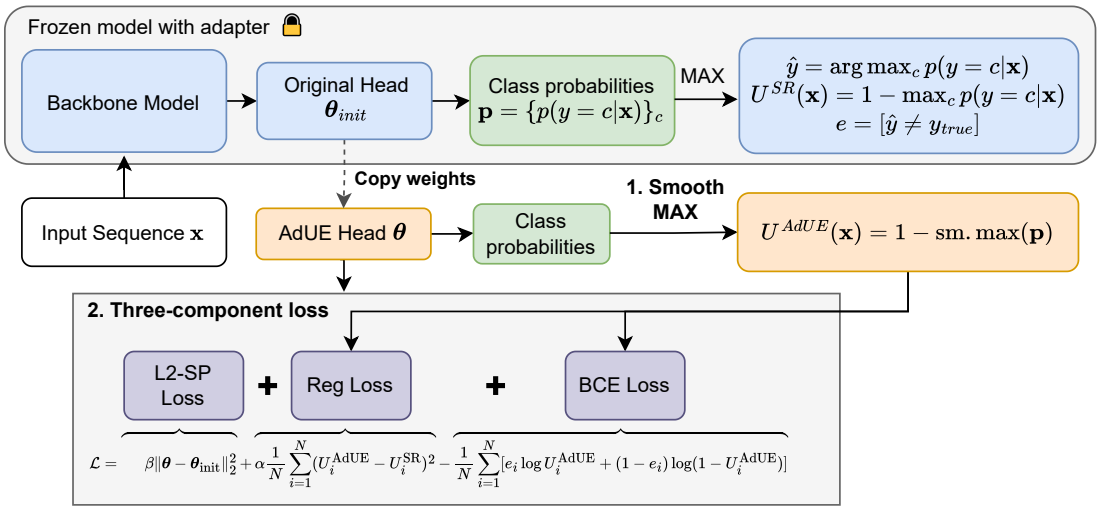

Figure 1: U^{AdUE} head training scheme. We initialize the new uncertainty head with the original classifier's weights θ_{init} and fine-tune it with a three-term loss (binary CE, softmax-regularization, L2-SP). The hard max is replaced by a differentiable SmoothMax during training

- We introduce a loss that balances classification accuracy, softmax regularization, and parameter anchoring for improved robustness.
- We extensively evaluate our approach on five NLP benchmarks and four different LoRA fine-tuned LLMs, demonstrating performance gains in uncertainty estimation.

2 Background Uncertainty Estimation Methods

Let $\mathbf{z} = h(\mathbf{x}) \in \mathbb{R}^d$ denote the latent representation of an input instance $\mathbf{x}$, typically extracted from the penultimate layer of a neural network. For a classification problem with classes $c \in \mathcal{C}$, we aim to provide an uncertain score $U(\mathbf{x})$ that reflects the possibility of an error for an example $\mathbf{x}$. We focus on methods that require low computational cost.

2.1 Softmax Response

The Softmax Response (SR) (Geifman and El-Yaniv, 2017) provides a computationally efficient measure of aleatoric uncertainty by examining the model's maximum class probability:

$$U^{\text{SR}}(\mathbf{x}) = 1 - \hat{y}, \qquad (1)$$

with $\hat{y} = \max_{c \in \mathcal{C}} p(y = c|\mathbf{x})$, where $p(y = c|\mathbf{x})$ represents the softmax probability for class $c \in \mathcal{C}$.

2.2 Mahalanobis Distance (MD) Method

The MD-based uncertainty estimation method (Lee et al., 2018) models each class as a Gaussian distribution characterized by class centroids $\boldsymbol{\mu}_c = \mathbb{E}[h(\mathbf{x})|y = c]$, a shared covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu}_c)(\mathbf{z} - \boldsymbol{\mu}_c)^T]$.

The uncertainty score for an input $\mathbf{x}$ is given by the minimum Mahalanobis distance to any class centroid:

$$U_{\text{MD}}(\mathbf{x}) = \min_{c \in \mathcal{C}} (\mathbf{z} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}_c). \qquad (2)$$

where $\mathbf{z}$ is a representation from the penultimate layer.

Two methods extend MD to make it more robust: the Relative Mahalanobis Distance (RMD) (Ren et al., 2021) and the Robust Distance Estimation (RDE) (Yoo et al., 2022). These modifications enhance robustness to outliers while preserving the discriminative power of the original representations.

Further details on distance-based methods are available in Appendix B.

3 Methodology

For experimentation, we chose Low-Rank Adaptation methods because of their computational efficiency. After fine-tuning the LoRA model, we extract embedding and fine-tuning the classification head with smooth max and regularization to obtain the probability of classification error.

3.1 Efficient Fine-tuning with LoRA

We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022), which is still one of the most popu-

lar fine-tuning approaches (Tuggener et al., 2024) from the peft library (Mangrulkar et al., 2022), to efficiently fine-tune the attention weights. This approach reduces the number of trainable parameters compared to full fine-tuning, while maintaining competitive quality.

3.2 Softmax Response Fine-Tuning

Softmax response (SR) (1) performs better in various uncertainty estimation scenarios without additional training (Vazhentsev et al., 2023; Holm et al., 2023).

After training LoRA adapters for the task, we attach a small trainable head—the *SmoothMax* head—initialized with the classification head weights, to predict an uncertainty score (error probability). This head approximates the softmax response and is fine-tuned independently, using only the frozen model representations.

Naive fine-tuning this head often underperforms the original softmax response due to two key challenges:

- **Overfitting**: A fully connected layer in traditional classification heads contains orders of magnitude more trainable parameters than available training samples, leading to severe overfitting.
- **Optimization Difficulty**: Direct optimization of the maximum function in uncertainty estimation proves computationally challenging due to its non-smooth nature.

To address these limitations, we introduce the *SmoothMax Classifier Nonlinearity*, which extends a conventional classification head with an additional smooth operation to produce similar output to *softmax response*. Given class probabilities $\{p_i\}_{i=1}^{C}$ from the base classifier, our head computes the smooth maximum as:

$$U^{\text{AdUE}} = 1 - \text{SmoothMax}(\mathbf{p}),$$

$$\text{SmoothMax}(\mathbf{p}) = \frac{1}{\lambda} \log \sum_{i=1}^{C} e^{\lambda p_i}, \qquad (3)$$

where $\lambda > 0$ is a temperature parameter controlling the smoothing intensity. This function is differentiable, allowing backpropagation.

3.3 Training Objective

The smooth head with parameters $\boldsymbol{\theta}$ is trained using a multi-task objective that combines three components, that takes into account model error $e_i = [\hat{y} \neq y_{true}]$, uncertainty estimates $U^{\text{SR}}(\mathbf{x})$ and parameters $\boldsymbol{\theta}_{\text{init}}$ from the original head:

The first is classic Binary Cross-Entropy Loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \Big[e_i \log U_i^{\text{AdUE}} + (1 - e_i) \log(1 - U_i^{\text{AdUE}}) \Big].$$

The next now is a regularization loss:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} (U_i^{\text{AdUE}}(\mathbf{x_i}) - U^{\text{SR}}(\mathbf{x_i}))^2.$$

This encourages the new uncertainty scores to stay close to the original softmax-based confidence on average, preventing drastic shifts.

Finally, we add L2-SP loss adopted from the transfer learning domain (Xuhong et al., 2018), this keeps the fine-tuned weights near their initialization, avoiding forgetting and overfitting:

$$\mathcal{L}_{\text{L2SP}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{init}}\|_2^2.$$

The combined training objective, where α and β control the regularization strengths:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \alpha \mathcal{L}_{\text{reg}} + \beta \mathcal{L}_{\text{L2SP}}, \qquad (4)$$

4 Experiments

4.1 Models

We evaluate AdUE on four pretrained models: `roberta-base` (Liu et al., 2019) (125M, masked language modeling), `electra-base-discriminator` (Clark et al., 2020) (110M, replaced token detection), and two 7B autoregressive models—`Qwen2.5-7B` (Yang et al., 2024) and `LLaMA-2-7b` (Touvron et al., 2023).

4.2 Training algorithm

After training a LoRA-adapter model on the task and obtain representation z:

1. Build a UE head initialized with θ_{init}.
2. Compute U^{SR} on each training example.
3. Train the uncertainty head to predict probability of error using the combined loss (4) replacing the hard max with the differentiable `SmoothMax` (3).

Dataset	Cola	News	SST2	SST5	Toxigen	Cola	News	SST2	SST5	Toxigen	Rank ↓
Model	Electra					Roberta					
RMD	70.0±2.6	83.7±0.6	79.2±4.1	57.2±1.3	67.9±1.5	65.3±1.7	84.7±0.7	73.0±2.8	58.1±1.1	69.1±2.2	3.7
MD	**80.9±1.1**	74.1±0.7	**88.2±1.0**	55.5±1.0	74.0±1.6	73.7±0.6	79.5±0.4	**82.8±2.2**	55.6±1.9	65.1±9.2	3.2
RDE	80.0±1.3	70.8±1.8	68.8±24.1	56.2±0.9	74.3±1.3	73.4±1.4	77.3±0.5	67.3±25.4	55.9±1.5	65.1±8.9	4.2
SR	79.2±1.5	83.4±0.6	86.6±2.1	60.8±1.5	74.5±1.4	75.1±2.5	84.8±0.2	80.8±2.0	60.9±1.8	74.3±0.6	2.4
AdUE	79.4±1.3	**84.9±0.6**	86.8±2.5	**62.4±0.9**	**75.2±1.1**	**75.4±3.0**	**86.0±0.3**	80.8±2.0	**62.1±1.5**	**74.6±1.0**	**1.5**
Model	LLaMA					Qwen					
RMD	67.3±6.0	80.7±1.1	78.7±2.5	53.4±0.4	67.5±2.5	65.2±4.1	81.2±0.7	78.6±4.1	54.3±1.4	66.4±1.9	3.0
MD	60.5±4.3	56.3±2.7	59.9±10.0	51.4±1.0	56.3±1.8	56.7±2.6	53.9±0.7	54.2±6.4	52.1±1.2	58.1±5.0	4.3
RDE	62.1±5.8	47.6±3.3	72.2±8.7	49.2±1.8	53.6±2.0	47.2±2.4	48.2±2.8	63.8±16.2	48.0±0.9	54.9±5.5	4.7
SR	77.8±0.7	86.3±0.4	88.3±1.8	59.5±2.4	76.7±1.6	78.1±2.1	85.6±0.4	86.3±1.7	60.8±1.1	72.1±2.4	2.0
AdUE	**77.8±0.5**	**88.4±0.4**	**88.4±1.7**	**61.7±2.7**	**77.2±2.5**	**79.7±2.1**	**87.4±0.1**	**86.3±2.1**	**62.5±1.4**	**72.8±2.4**	**1.0**

Table 1: ROC AUC mean and standard deviation values of five runs for error classification problem, our AdUE improves performance for almost all datasets and language models.

For all models, we train the LoRA adapter five times with different seeds and apply the AdUE method to compute the mean and standard deviation for each dataset (see Appendix D). Hyperparameter search details are provided in Appendix C.

4.3 Results

AdUE improves results for both generative and encoder-based transformer architectures across most datasets. The results are presented in Table 1. We also performed an ablation study to see what would happen if we trained the linear layer on $\mathbf{z}$ or initialized the classification layer with random parameters. The results of disabling each loss term are shown in Table 2 with details provided in the Appendix F.

5 Related work

Predictive uncertainty in LLMs can be estimated using various methods. Among these, *information-based* methods are widely utilized, which analyze token probability distributions by accessing logits or outputs from the internal layers of LLMs (Takayama and Arase, 2019; Fomicheva et al., 2020; van der Poel et al., 2022; Colombo et al., 2023), or by relying on the generated text (Tian et al., 2023). However, these methods are often outperformed by *sample diversity* methods, which involve generating multiple outputs for LLM and either aggregating their confidence scores or assessing their diversity (Kuhn et al., 2023; Malinin and Gales, 2021; Duan et al., 2024).

In contrast, **density-based** methods approximate the distribution of training data using embeddings of training instances (Lee et al., 2018; Yoo et al., 2022; Ren et al., 2023; Vazhentsev et al., 2023). Additionally, predictive uncertainty can be estimated using **reflexive** methods, in which the model is asked directly to provide confidence levels for its responses (Kadavath et al., 2022; Tian et al., 2023).

Method	Rank ↓
AdUE (Ours)	**2.3**
Loss: BCE	2.95
Loss: BCE+L2SP	2.75
Loss: BCE+reg	2.9
Full loss + rand cls	4.75
Full loss + random linear	5.35

Table 2: Ablation of loss components and initialization strategies. Using all loss terms with classification head initialization yields the best performance.

6 Conclusion

This work introduces AdUE Softmax Response Fine-Tuning, a lightweight post-hoc method to enhance uncertainty estimation in parameter-efficient fine-tuned LLMs. Empirical results demonstrate that AdUE improves the correlation between predictive uncertainty and classification error across a diverse set of tasks and model architectures, outperforming both traditional and robust distance-based uncertainty estimation techniques. Looking ahead, we aim to explore whether combining AdUE with existing calibration techniques or uncertainty-aware training objectives may further enhance robustness.

7 Limitations

While our method demonstrates promising improvements in uncertainty estimation, several limitations remain. On certain datasets, notably SST-2—where encoder models were originally trained

on substantially larger samples—AdUE underperforms relative to alternative methods. This indicates that the advantage of our approach may depend significantly on training set characteristics and traditional methods may be more effective in data-rich scenarios.

Additionally, we recognize that the evaluation scope could be expanded. First, our experiments focus exclusively on **adapter-based fine-tuning**, specifically LoRA, and do not explore other parameter-efficient methods such as prompt or prefix tuning, which might exhibit different uncertainty characteristics. Second, our assessment is restricted to classification tasks. The proposed SmoothMax head has not yet been evaluated in **generation tasks** or other contexts, such as sequence labeling, where uncertainty dynamics differ, and factors like beam diversity or exposure bias become relevant. Third, training and evaluating the uncertainty head using pseudo-error labels derived from the model's own predictions could introduce bias, particularly when the base model tends toward overconfidence. Lastly, our primary evaluation metric, uncertainty–error AUC, does not directly measure **calibration metrics** such as Expected Calibration Error (ECE). Future research should address these limitations, extending the analysis to additional architectures, tasks, and evaluation frameworks.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Pierre Colombo, Maxime Darrin, and Pablo Panitainada. 2023. Rainproof: An umbrella to shield text generators from out-of-distribution data. In *EMNLP*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

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

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

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

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Andreas Nugaard Holm, Dustin Wright, and Isabelle Augenstein. 2023. Revisiting softmax for uncertainty approximation in text classification. *Information*, 14(7).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Peter J. Rousseeuw. 1984. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lukas Tuggener, Pascal Sager, Yassine Taoudi-Benchekroun, Benjamin F Grewe, and Thilo Stadelmann. 2024. So you want your private LLM at home? a survey and benchmark of methods for efficient GPTs. In *2024 11th IEEE Swiss Conference on Data Science (SDS)*, pages 205–212. IEEE.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, Manvel Avetisian, and Leonid Zhukov. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252, Dublin, Ireland. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Li Xuhong, Yves Grandvalet, and Franck Davoine. 2018. Explicit inductive bias for transfer learning with convolutional networks. In *International conference on machine learning*, pages 2825–2834. PMLR.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

A Hardware

The experiments utilized a high-performance computing cluster with the above specifications. Training and evaluating 100 model instances across all datasets and architectures required substantial computational resources to ensure reliable estimation of means and standard deviations.

CPU Cores	128
CPU Memory	2 TB
GPU	8 × NVIDIA A100 80GB
Total GPU Hours	340

B Details on distance-based UE methods

Here we described implementation as well as extensions of MD — a Mahalanobis distance-based uncertainty estimation method.

B.1 Implementation details

For Electra and Roberta models incorporate spectral normalization (Miyato et al., 2018) in the penultimate layer of their classification heads, which has been shown to stabilize density-based uncertainty estimation (Vazhentsev et al., 2022) and also test L2 normalization for representation vector $\mathbf{z}$, results shown in Tables 7, 8.

B.2 Relative Mahalanobis Distance (RMD)

The Relative Mahalanobis Distance (RMD) (Ren et al., 2021) improves upon standard MD for near-OOD detection by comparing class-conditional and background distributions:

$$\begin{aligned} \mathrm{RMD}_k(\mathbf{z}) &= \mathrm{MD}_k(\mathbf{z}) - \mathrm{MD}_0(\mathbf{z}), \\ \text{where} \quad \mathrm{MD}_k(\mathbf{z}) &= (\mathbf{z}-\mu_k)^T \Sigma_k^{-1} (\mathbf{z}-\mu_k), \\ \mathrm{MD}_0(\mathbf{z}) &= (\mathbf{z}-\mu_0)^T \Sigma_0^{-1} (\mathbf{z}-\mu_0), \end{aligned}$$

with $\mu_0 = \mathbb{E}[\mathbf{z}]$, $\Sigma_0 = \mathrm{Cov}(\mathbf{z})$ computed across all classes.

B.3 Robust Distance Estimation (RDE)

The RDE (Yoo et al., 2022) extends MD with two key improvements:

1. Class-specific covariance matrices $\mathbf{\Sigma}_c$ estimated via Minimum Covariance Determinant (MCD), achieving higher robustness of the computation (Rousseeuw, 1984).

2. Dimensionality reduction of $\mathbf{z}$ through kernel PCA with Radial Basis Function (RBF).

C AdUE Hyperparameters search

We conducted extensive hyperparameter optimization across the following ranges:

- **Learning Rates**: 1×10^{-3}, 1×10^{-4}, 1×10^{-5}
- **Training Epochs**: 5, 10, 20
- **Softmax response fine tune parameters**:
  - λ: 100.0
  - Out regularization (α): 0.0, 0.1, 0.3, 1.0
  - L2SP (β): 0.01, 0.1, 0.5, 1.0

D Datasets

We evaluated our approach on five benchmark datasets spanning different text classification tasks. The sizes of the datasets and the number of classes represented are shown in the Table 5.

SST-2 & SST-5 The Stanford Sentiment Treebank provides sentence-level sentiment labels from movie reviews. SST-2 is a binary classification task (positive/negative), while SST-5 offers fine-grained sentiment analysis (very negative to very positive). Both datasets contain parse trees enabling full-sentence compositional analysis.

20Newsgroups A classic text categorization corpus comprising newsgroup posts across 20 topics. The dataset presents challenges in document-level understanding and contains significant class imbalance.

ToxiGen A large-scale dataset for detecting hate speech against 13 minority groups, with synthetic and human-annotated examples. It focuses on implicit toxicity detection in diverse linguistic constructions.

CoLA The Corpus of Linguistic Acceptability evaluates models' ability to judge grammatical correctness. It contains expert-labeled examples of English grammatical phenomena, testing linguistic competence.

E LoRA Adapter training

All models in our experiments employed parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA). This approach enables effective model adaptation while maintaining computational efficiency and preserving the original model knowledge. The LoRA implementation follows standard practices for transformer-based architectures,

Model			Electra					Roberta			
Dataset	Cola	News	SST2	SST5	Toxigen	Cola	News	SST2	SST5	Toxigen	Rank ↓
AdUE (Ours)	79.4±1.3	84.9±0.6	86.8±2.5	**62.4±0.9**	**75.2±1.1**	75.4±3.0	**86.0±0.3**	80.8±2.0	62.1±1.5	74.6±1.0	**2.3**
Loss: BCE	79.2±1.7	85.4±0.6	86.7±2.5	62.0±1.2	75.1±1.0	75.4±2.8	85.8±0.6	80.7±2.6	62.1±1.4	74.6±1.1	3.5
Loss: BCE+L2SP	79.3±1.4	84.8±0.5	86.8±2.3	62.3±0.9	75.1±1.0	75.3±2.9	86.0±0.3	80.8±1.9	62.1±1.6	**74.8±0.9**	3.1
Loss: BCE+reg	79.4±1.3	**85.5±0.6**	86.8±2.5	62.0±1.1	74.9±0.5	**75.6±2.9**	85.7±0.7	80.9±2.5	**62.2±1.5**	74.6±1.0	2.7
Full loss: rand cls	**80.6±1.4**	83.1±1.0	**88.0±1.4**	57.7±3.2	70.0±8.5	63.2±7.7	78.7±1.6	**81.4±3.2**	58.7±3.8	68.9±8.5	3.9
Full loss: random linear	77.4±7.5	78.0±1.9	87.9±1.6	56.7±2.0	72.8±3.4	59.4±6.4	77.8±1.3	77.4±11.5	57.5±1.6	59.2±10.6	5.5

Table 3: Ablation study on uncertainty estimation for encoder models. Mean ROC-AUC and standard deviation over 5 runs are reported. Results indicate that using all loss components combined with classification head initialization yields the best performance.

Model			LLaMA					Qwen			
Dataset	Cola	News	SST2	SST5	Toxigen	Cola	News	SST2	SST5	Toxigen	Rank ↓
AdUE (Ours)	77.8±0.5	88.4±0.4	88.4±1.7	61.7±2.7	**77.2±2.5**	**79.7±2.1**	87.4±0.1	**86.3±2.1**	62.5±1.4	**72.8±2.4**	**2.3**
Loss: BCE	**78.1±0.5**	**88.4±0.4**	88.5±1.4	62.3±1.2	76.9±2.3	79.7±2.1	**87.5±0.2**	86.2±2.2	62.4±1.4	72.6±2.7	2.4
Loss: BCE+L2SP	77.7±0.6	88.4±0.4	**88.6±1.4**	60.8±3.7	77.0±2.5	79.6±2.4	87.5±0.1	86.3±2.2	**62.5±1.4**	72.7±2.6	2.4
Loss: BCE+reg	77.7±0.5	88.4±0.4	88.4±2.3	**62.4±1.3**	76.9±2.3	79.7±2.1	87.5±0.2	85.9±1.8	62.5±1.4	72.6±2.7	3.1
Full loss: rand cls	68.7±6.7	71.8±3.1	88.0±1.0	54.7±2.1	66.4±10.3	78.9±2.9	69.1±3.2	85.7±1.8	54.1±1.8	65.3±7.9	5.6
Full loss: random linear	68.8±7.1	74.8±1.5	87.9±1.5	54.1±3.0	67.3±8.1	75.2±7.6	70.5±1.0	86.3±1.2	54.2±1.6	56.0±7.3	5.2

Table 4: Ablation study on uncertainty estimation for generative models. Mean ROC-AUC and standard deviation across 5 runs are presented. Similar to encoder models, the combination of all loss terms and initialization from the classification head provides the most effective configuration.

Dataset	**Description**	**Classes**	**Train size**	**Valid size**	**Test size**
SST-2	Stanford Sentiment Treebank (Socher et al., 2013)	2	53879	13470	872
SST-5	Stanford Sentiment Treebank (Socher et al., 2013)	5	6840	1711	1043
20Newsgroups	Topic classification (Lang, 1995)	20	9051	2263	7532
ToxiGen	Hate speech detection (Hartvigsen et al., 2022)	2	7168	1792	940
CoLA	Linguistic acceptability (Warstadt et al., 2019)	2	6840	1711	1043

Table 5: Dataset inforamtion, number of classes and split sizes.

with adaptations applied to attention mechanisms throughout the network. The specifications were as follows:

$$\text{LoRA config} = \begin{cases} \text{Weights} = \{W_q, W_k, W_v\} \\ \alpha = 16 \\ \text{rank} = 8 \\ \text{dropout} = 0.05 \end{cases}$$

Parameter	**Value**
Optimizer	AdamW
Learning rate	5×10^{-4}
Weight Decay	0.1
Batch Size	64
Learning Rate Schedule	Linear decay to 0
Warmup steps	0.1
Mixed-precision	bfloat16-mixed

E.1 Metrics

The final quality of each LoRA fine-tuned model is summarized in Table 6, which reports the mean and standard deviation of accuracy across multiple runs.

	News	SST2	SST5	Cola	Toxigen
LLaMA	77.5±2.2	**96.4±0.2**	**59.3±0.8**	**85.7±0.5**	**85.3±0.4**
Qwen	**79.3±0.7**	96.0±0.6	58.8±0.4	84.6±1.1	85.3±0.7
Electra	72.3±0.4	93.9±0.4	56.0±1.1	85.1±0.7	81.3±0.4
Roberta	74.2±0.7	93.2±0.5	55.2±1.2	83.3±0.9	79.4±1.0

Table 6: Classification accuracy across all datasets shows that the fine-tuned LLaMA model achieves the best performance among all the models we trained.

F Ablation

Also we do ablation study in different scenarios: first we disable each part of final loss, train our smooth head. We check what will happen if we do not init our head with trained classification head

Model			Electra					Roberta			
Dataset	Cola	News	SST2	SST5	Toxigen	Cola	News	SST2	SST5	Toxigen	Rank ↓
RMD	70.5±2.4	83.7±0.6	80.1±2.1	57.3±1.2	68.6±0.9	64.2±1.7	84.3±0.7	73.9±3.8	57.9±1.1	68.3±2.5	4.8
RMD L2 norm	70.0±2.6	83.7±0.6	79.2±4.1	57.2±1.3	67.9±1.5	65.3±1.7	84.7±0.7	73.0±2.8	58.1±1.1	69.1±2.2	4.8
MD	76.2±2.0	74.6±0.7	85.1±2.4	53.3±1.1	62.8±3.9	52.8±11.2	78.4±0.7	76.4±11.7	51.0±1.8	53.2±11.3	6.2
MD L2 norm	**80.9±1.1**	74.1±0.7	**88.2±1.0**	55.5±1.0	74.0±1.6	73.7±0.6	79.5±0.4	**82.8±2.2**	55.6±1.9	65.1±9.2	3.8
RDE	73.2±2.6	68.5±0.5	85.3±2.2	53.8±1.3	64.8±4.0	50.6±11.3	74.2±1.0	74.8±11.5	51.0±1.7	51.6±11.0	6.9
RDE L2 norm	80.0±1.3	70.8±1.8	68.8±24.1	56.2±0.9	74.3±1.3	73.4±1.4	77.3±0.5	67.3±25.4	55.9±1.5	65.1±8.9	5.5
SR	79.2±1.5	83.4±0.6	86.6±2.1	60.8±1.5	74.5±1.4	75.1±2.5	84.8±0.2	80.8±2.0	60.9±1.8	74.3±0.6	2.5
AdUE (Ours)	79.4±1.3	**84.9±0.6**	86.8±2.5	**62.4±0.9**	**75.2±1.1**	**75.4±3.0**	**86.0±0.3**	80.8±2.0	**62.1±1.5**	**74.6±1.0**	**1.5**

Table 7: L2 normalization on representation show better UE ROC-AUC for encoder models

Model			LLaMA					Qwen			
Dataset	Cola	News	SST2	SST5	Toxigen	Cola	News	SST2	SST5	Toxigen	Rank ↓
RMD	67.0±5.6	80.4±1.1	79.0±2.3	53.4±0.4	67.6±2.5	65.3±4.2	82.0±0.8	78.5±4.6	54.3±1.5	65.8±1.5	3.4
RMD L2 norm	67.3±6.0	80.7±1.1	78.7±2.5	53.4±0.4	67.5±2.5	65.2±4.1	81.2±0.7	78.6±4.1	54.3±1.4	66.4±1.9	3.6
MD	60.5±4.0	55.2±2.4	57.3±12.2	52.0±0.9	55.7±2.2	52.1±3.8	60.5±1.6	52.6±6.7	52.2±1.5	52.7±3.0	6.3
MD L2 norm	60.5±4.3	56.3±2.7	59.9±10.0	51.4±1.0	56.3±1.8	56.7±2.6	53.9±0.7	54.2±6.4	52.1±1.2	58.1±5.0	6.0
RDE	63.7±9.6	43.5±3.7	60.0±9.3	51.6±1.2	52.7±4.7	42.5±12.0	51.0±4.4	52.2±7.4	52.1±1.0	51.3±2.8	7.0
RDE L2 norm	62.1±5.8	47.6±3.3	72.2±8.7	49.2±1.8	53.6±2.0	47.2±2.4	48.2±2.8	63.8±16.2	48.0±0.9	54.9±5.5	6.7
SR	77.8±0.7	86.3±0.4	88.3±1.8	59.5±2.4	76.7±1.6	78.1±2.1	85.6±0.4	86.3±1.7	60.8±1.1	72.1±2.4	2.0
AdUE (Ours)	**77.8±0.5**	**88.4±0.4**	**88.4±1.7**	**61.7±2.7**	**77.2±2.5**	**79.7±2.1**	**87.4±0.1**	**86.3±2.1**	**62.5±1.4**	**72.8±2.4**	**1.0**

Table 8: For generative model L2 normalization on representation do not change significant ROC-AUC

and replace it with a random initialization of one linear layer.

- **Loss: BCE** - Uses only binary cross-entropy loss for training.
- **Loss: BCE+L2SP** - Combines binary cross-entropy with L2SP loss to prevent weights from deviating significantly from their initial values.
- **Loss: BCE+reg** - Extends binary cross-entropy with a regularization term that encourages model outputs to stay close to a softmax response distribution.
- **Full loss: rand linear** - Uses our complete proposed loss function and replaces the uncertainty prediction head with a single randomly initialized linear layer.
- **Full loss: rand cls** - Uses our full loss but initializes the uncertainty prediction layer with the same architecture as the classification head, though with random initialization.

Our results are presented in Tables 3, 4.

For all baselines, we compare two options for distance calculation: using the original representation or the L2-normalized representation, see Tables 7 and 8.