Distance-aware Calibration for Pre-trained Language Models

Alberto Gasparin
Amazon
Berlin, Germany
albgas@amazon.de

Gianluca Detommaso[†]
Helsing
Berlin, Germany
detommaso.gianluca@gmail.com

Abstract

Language Models for text classification often produce overconfident predictions for both in-distribution and out-of-distribution samples, i.e. the model's output probabilities do not match their accuracy. Prior work showed that simple post-hoc approaches are effective for mitigating this issue, but are not robust in noisy settings, e.g., when the distribution shift is caused by spelling mistakes. In this work, we propose Distance Aware Calibration (DAC), a post-hoc approach that changes the confidence scores of a Language Model leveraging the distance between new samples been evaluated and the in-domain training set. We show that using DAC on top of a Language Model can improve in-domain calibration, robustness to different kind of distribution shift and also the model's ability to detect out-of-distribution samples. We provide an extensive evaluation on common text classification benchmark for both calibration and out-of-distribution detection tasks.

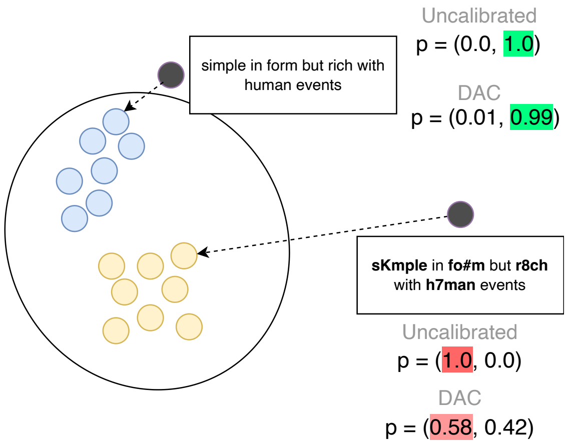

Figure 1: An example of DAC's calibrated confidence. The clean test sample on top is correctly classified with high confidence (green background), but once it's corrupted the model's prediction flips but the confidence is unchanged (red background). DAC leverages the samples' distances from the training set to decrease the confidence on the predictions, leaving the correctly classified one almost unchanged while yielding low confidence for the sample being far from the training dataset (white circle).

1 Introduction

Fine-tuning language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) has been established as the de-facto standard approach for text classification whenever a large amount of annotated training examples is available. Even in the era of Large Language Models (LLMs), smaller (and more efficient) fine-tuned models have been shown to still be able to outperform in context learning (ICL) methods for much larger LLMs (Edwards and Camacho-Collados, 2024).

One of the main open challenges with Language Models (LMs) of any size is how to obtain reliable uncertainty estimates for their predictions (Xiao et al., 2022); this is particularly important in safety critical applications where LMs are deployed to assist humans in their decision making. These applications span from medicine (Zhang et al., 2021) to finance (Liu et al., 2020) and legal (Xiao et al., 2021). Unfortunately, like most deep neural networks, language models have been shown to be miscalibrated and overconfident (Desai and Durrett, 2020; Dan and Roth, 2021), i.e., the model's predictive confidence over its predictions does not match the empirical accuracy, but usually exceeds it. This behaviour gets worse in the presence of distribution shift and in general whenever the training set and test set are not i.i.d (Xiao et al., 2022; Zhang et al., 2023).

To mitigate this problem, prior work has focused on either designing new training strategies to achieve better calibration (He et al., 2022; Park and Caragea, 2022), or on the use of post-hoc methods

[†] Work done while working at Amazon Web Services, Berlin, Germany

Findings of the Association for Computational Linguistics: EACL 2024, pages 12434–12447
November 12-16, 2024 ©2024 Association for Computational Linguistics

(Desai and Durrett, 2020), such as Temperature Scaling (Guo et al., 2017). The latter can be applied on top of any model, independently from its architecture and its training setup, which makes them highly appealing for practitioners.

Despite their simplicity, post-hoc methods like Temperature Scaling (TS) have been shown to be extremely effective for both in-domain (IND) and out-of-domain (OOD)[†] calibration in text classification (Xiao et al., 2022), outperforming more complex (and time-consuming) bayesian methods that rely on some kind of (multiple) models average in order to estimate the uncertainty around a prediction. Nevertheless, TS is not reliable for all kinds of distribution-shift, as it suffers in the presence of grammatical errors, typos and other common mistakes which are easily found in real world applications (Zhang et al., 2023).

In this work, we leverage recent findings on distance awareness (Liu et al., 2023; Mukhoti et al., 2023; Van Amersfoort et al., 2020) and its connections with high quality uncertainty quantification and propose *Distance-Aware Calibration* (DAC), a post-hoc method to improve model's calibration and its robustness to distribution shifts.

Our method is built on the belief that the confidence over a model's prediction should decrease as the inputs move away from the training domain. Thus, we propose to adjust a model's confidence using sample-specific information, namely, the distance between the embeddings of a new sample and the representation of the samples in the training set. We hypothesised that this would mitigate LMs's overconfidence, benefiting OOD calibration and, to a smaller extent, in-domain calibration. Figure 1 shows an illustrative example of DAC's adjusted confidence for test points located at various distances from the training set.

In this work we first introduce our calibration method DAC (§3.2.1) and detail the procedure used to optimize it for both IND and OOD calibration (§3.2.2). Then, we show the efficacy of our method in both traditional settings - where model calibration is measured on datasets that share the same task (e.g., sentiment analysis) but differ for domain or writing style - and noisy settings, where the OOD datatset is a corrupted version of the in-domain one (§5.1). Moreover, we will show that a LM equipped with DAC not only achieves better calibration but it's confidence scores can also be used to distinguish between in-domain and out-of-domain samples, boosting performance in the OOD detection task (§5.2).

2 Related Work

Calibration in Language Models Several methods have been developed to mitigate language model's overconfidence, thus providing more calibrated predictions. The methods in the literature can be broadly classified in two groups: i) post-hoc methods and ii) methods that require model re-training.

To the first class of methods belongs temperature scaling (TS) (Guo et al., 2017), which has been shown to be beneficial for model calibration in text classification settings, both IND and OOD (Desai and Durrett, 2020). More evidence in favour of TS is presented in Xiao et al. (2022) where TS is shown to be superior to other uncertainty quantification methods such as MC Dropout (Gal and Ghahramani, 2016), (last layer) Variational Inference (Blundell et al., 2015) and Deep Ensembles (Lakshminarayanan et al., 2017).

More recent works focused on a second class of model calibration techniques, those that require retraining. Park and Caragea (2022) uses Mixup to improve the calibration of fine-tuned language models, showing improvement in expected calibration error (ECE) w.r.t. base TS and label smoothing. Li et al. (2022) leverage model explanations to enhance calibration. The proposed method is used in conjunction with TS to improve model performance for IND and OOD calibration. Recently, He et al. (2022) showed that Pretrained-Language Models (PLMs) are better calibrated on the masked language modelling (MLM) task than their fine-tuned counterpart. The authors hypothesised that LM's overconfidence is due to catastrophic forgetting and showed that preserving the pre-trained features while fine-tuning LMs can improve IND and OOD calibration. Finally, Zhang et al. (2023) discovered that TS performs poorly when the in-domain samples are perturbed with noise. Thus, they propose an extension of TS based on the measured distribution-shift between the in-domain dataset and the noisy one. This work closely relates to ours as they also used some notion of "distance" to the in-domain dataset. However, unlike their work, DAC's confidence is calibrated sample-wise and does not require access to samples coming

[†] In this work we will use the terms in-domain (out-of-domain) and in-distribution (out-of-distribution) interchangeably

from the same OOD distribution that will be later use for evaluation. Also, we do not focus exclusively on noisy-settings like Zhang et al. (2023).

Distance Awareness The concept of distant-aware uncertainty is native in Gaussian Processes (GPs) (Williams and Rasmussen, 2006), where a kernel captures a measure of distance between pairs of inputs. Modern approaches combine Radial Basis Function (RBF) kernels with deep feature extractors, i.e. deep neural networks that transform the input space in order to obtain a better fit of the data (Chen et al., 2016). This approach is commonly referred to as Deep Kernel Learning (DKL) (Wilson et al., 2016b,a), and has recently inspired a variety of methods for deterministic uncertainty estimation. SNGP (Liu et al., 2023) is one such method, which builds upon DKL proposing to approximate a GP via Random Fourier Features (RFFs) (Rahimi and Recht, 2007). A Similar approach is proposed by Van Amersfoort et al. (2020). These methods are competitive with Deep Ensembles on multiple OOD benchmarks, but still require i) changes to the model architecture and ii) model re-training, making them quite expensive and not suitable to be applied post-hoc. Recently, in the context of computer vision, Tomani et al. (2023) proposed a post-hoc calibration method that shares with ours the idea of assigning higher uncertainty to samples that are far away from the training data. However, the actual methodology used to change the underlying model's confidence is different from our proposal and so is the tuning procedure for the calibration parameters.

3 Methodology

3.1 Preliminary on Calibration

Consider an input $\mathbf{x} \in \mathcal{X}$, a target variable $y \in \{1, \ldots, K\}$, and a score function $f(\mathbf{x})$ estimating the confidence of a classification model in each class. Standard top-label calibration (Guo et al., 2017) requires:

$$\mathbb{P}(y = \hat{y} | f_{\hat{y}}(\mathbf{x}) = p) = p, \tag{1}$$

for all $p \in [0, 1]$ such that $\mathbb{P}(f_{\hat{y}}(\mathbf{x}) = p) > 0$. In other words, we say that a model is top-label calibrated (or just calibrated, in short) if for every level set of the model with positive probability, the fraction of correct predictions matches the confidence of the model in the predicted label. To evaluate a model's calibration performance, the most popular metric is the Expected Calibration Error (ECE), where the model's confidence scores are sorted and organized into M bins of equal length and then the metric is computed as the absolute weighted average between confidence and accuracy in each bin:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{2}$$

where $|B_m|$ is the number of samples in the m-th bin, acc(B_m) is the model's accuracy in the m-th bin while conf(B_m) is the average model's confidence score in the m-th bin.

3.2 Distance Aware Calibration (DAC)

In this section, we present our proposed method DAC. We first introduce our method and it's underlying assumption in §3.2.1. Then, in §3.2.2, we describe how DAC's parameter is optimized to maintain high quality in-domain calibration while also being able to assign low confidence scores to out-of-distribution samples.

3.2.1 Post-hoc approach

In this work we leverage the idea that a model should be able to express high confidence when evaluated on new samples that are close to the domain it was originally trained on, while it should yield low confidence predictions for samples that are far-away from it. This is particularly relevant in scenarios where a distribution-shift has happened as we would like our model's confidence to be correlated with the magnitude of the shift (see Figure 1). To formalize this concept, we introduce a distance function $d_{\mathcal{X}}(\mathbf{x}) > 0$ which measures the distance of an arbitrary input $\mathbf{x}$ from the training set $\mathcal{X}$. We can now define a sample $\mathbf{x}$ as OOD if $d_{\mathcal{X}}(\mathbf{x})) > T$, with T a random variable whose distribution we do not know and we need to assume a priori. In this way, the probability that a sample $\mathbf{x}$ do not belong to the in-domain increases with its distance from the training set. We argue that, whenever an input $\mathbf{x}$ is OOD, the model's predictive distribution should reflect maximum uncertainty (Liu et al., 2023), that is

$$p(\mathbf{y}|\mathbf{x}, \mathbf{x} \notin \mathcal{X}_{\text{IND}}) = p_{\text{uniform}}(\mathbf{y}|\mathbf{x}) \tag{3}$$

where p_{uniform} is a discrete uniform distribution over the output space and $\mathcal{X}_{\text{IND}}$ is the full (in-domain) input space, of which the training set is a subset, i.e., $\mathcal{X} \subset \mathcal{X}_{\text{IND}}$. Then, by conditioning on

Noise Intensity	Text
0	involved in creating the layered richness of the imagery in this chiaroscuro of madness and light
1	involved in creating the layered richne(ss of the imagery in this chiaroscuro of madness and ight
2	involved in creating the layered richness of the imKazgery in thixs chViarNoscuro of madfnesss and light
3	nivovled in Rneatinq the layered richness of the FimTager@y in this chariscoruo of madness and lig1o

Table 1: Example of a sentence coming from the SST2 dataset and its transformations under increasing noise intensity (Random).

whether the input is OOD or not, we can write the predictive distribution of the model as:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= p(\mathbf{y}|\mathbf{x}, \mathbf{x} \in \mathcal{X}_{\text{IND}})p(\mathbf{x} \in \mathcal{X}_{\text{IND}}) \\ &\quad + p(\mathbf{y}|\mathbf{x}, \mathbf{x} \notin \mathcal{X}_{\text{IND}})p(\mathbf{x} \notin \mathcal{X}_{\text{IND}}) \end{aligned} \tag{4}$$

where $p(\mathbf{x} \notin \mathcal{X}_{\text{IND}}) = \mathbb{P}(T < d_{\mathcal{X}}(x))$ and $p(\mathbf{x} \in \mathcal{X}_{\text{IND}}) = 1 - p(\mathbf{x} \notin \mathcal{X}_{\text{IND}})$.

The model has learned $p(\mathbf{y}|\mathbf{x}, \mathbf{x} \in \mathcal{X}_{\text{IND}})$ as part of its training process, and we defined $p(\mathbf{y}|\mathbf{x}, \mathbf{x} \notin \mathcal{X}_{\text{IND}})$ in Eq (3). In order to define $p(\mathbf{x} \notin \mathcal{X}_{\text{IND}})$, we observe that the distance score $d_{\mathcal{X}}(\mathbf{x})$ is positive, thus modelling T with an exponential distribution with rate parameter ϕ seems a natural choice, (i.e., $T \sim \text{Exp}(\phi)$). We highlight that when the distance being used is bounded (e.g. between $[0, 1]$) a different distribution should be considered. In practice, with this formulation, the domain membership decreases exponentially as the distance increases and so does the model's confidence.

We can finally rewrite Eq. (4) as:

$$\begin{aligned} &p(\mathbf{y}|\mathbf{x}, d_{\mathcal{X}}(\mathbf{x})) = d) = \\ &\qquad p(\mathbf{y}|\mathbf{x}, \mathbf{x} \in \mathcal{X}_{\text{IND}})e^{-\phi d} + \frac{1}{K}(1 - e^{-\phi d}) \end{aligned} \tag{5}$$

From a calibration perspective, as the distance d grows, the model's confidence converges to $\frac{1}{K}$ (K being the number of classes), on the other hand, standard calibration requirements are maintained for in-distribution data, where the values for d are small. For all values of d between 0 and ∞, the model's confidence over its (top-label) prediction is lower compared to the uncalibrated counterpart, and the magnitude of this confidence decrease is related to the sample's distance from the training set. In this work $d_{\mathcal{X}}(\mathbf{x})$ is the K-nearest neighbors (KNN) (Sun et al., 2022) distance between the feature embedding of $\mathbf{x}$ and the ones of the training samples; we considered alternatives in §5.3. The embedding vector for each sample $\mathbf{x}$ is the hidden representation of the [CLS] token.

3.2.2 Optimization of $p(\mathbf{x} \notin \mathcal{X}_{\text{IND}})$

The simple parametric form introduced in Eq. (5) (i.e., $p(\mathbf{x} \notin \mathcal{X}_{\text{IND}}) = 1 - e^{-\phi d}$) has one free parameter $\phi \in \mathbb{R}$, which controls the impact of the distance d on the model's confidence. In this section we first introduce the calibration sets that will be used to fit DAC, and later detail the actual optimization procedure.

Let $\mathbb{D}^{(c)}_{\text{IND}} = \{(x_i, y_i)\}_{i=1}^{L}$ be a set of samples not overlapping with the training set but coming from the same in-domain distribution, i.e., $x_i \in \mathcal{X}_{\text{IND}}$, $y_i \in \{1, ..., K\}$ and $L \leq N$, with N being the number of training samples. Let's also have $\mathbb{D}^{(c)}_{\text{OOD}} = \{x_j\}_{j=1}^{T}$ be a set of samples from a different distribution w.r.t in-domain one, i.e., $x_j \notin \mathcal{X}_{\text{IND}}$. Finally, we introduce a binary classification dataset $\mathbb{D}^{(c)}_{\text{IND+OOD}} = \{(x_i, y_i)\}_{i=1}^{L+T}$ where x_i may come from either $\mathbb{D}^{(c)}_{\text{IND}}$ or $\mathbb{D}^{(c)}_{\text{OOD}}$ and $y_i = \mathbf{1}_{x_i \in \mathcal{X}_{\text{IND}}}$. In Appendix A we detail the composition and size of the calibration datasets used in the experiments.

Our goal is to obtain a model whose confidence score are high for in-domain samples and low for out-of domain samples, while also enhancing its in-domain calibration. To achieve this, DAC is fit over the previously defined calibration sets to solve the following multi-objective problem:

$$\min_{\phi \in \Phi}(1 - c_1(\mathbb{D}^{(c)}_{\text{IND+OOD}}; \phi, \theta), c_2(\mathbb{D}^{(c)}_{\text{IND}}; \phi, \theta)) \tag{6}$$

where c_1 is the Area under the Precision-Recall Curve (PRAUC), c_2 is the Expected Calibration Error (ECE), Φ is a grid of calibration parameters and θ are the model's parameters which are kept fixed. The choice of the two objectives is a natural one, as the ECE is the golden standard for measuring model calibration, while PRAUC is known to be a good proxy for evaluating model's uncertainty over out-of-domain samples, as it requires a model to assign different uncertainty scores to IND and OOD samples in order to distinguish them. For PRAUC, we use the (top-label) confidence scores as uncertainty quantifier.

Given that in multi-objective optimization we usually do not have a unique solution simultaneously minimizing all objectives, within this paper we rely on the Kneedle algorithm developed by (Satopaa et al., 2011) to obtain a single solution. The Kneedle algorithm is a simple and general approach to on line and off line knee detection, where a knee is defined as the point of maximum curvature for a function, or, in more practical terms, it is the point where the cost (e.g., an increase in $1 - c1$) of changing the parameter ϕ is no longer worth the performance improvement (e.g., a decrease in $c2$). We highlight that the additional latency required to compute $p(\mathbf{x} \notin \mathcal{X}_{\text{IND}})$ scales with the logarithm of the number of training samples when using libraries that supports efficient Approximate Nearest Neighbour search such as FAISS (Johnson et al., 2019).

Finally, one may argue that since we are interested in improving the model's calibration we should jointly optimize for in-domain ECE and out-of-domain ECE instead of using PRAUC in Eq. 6. The reason why we did not do that is a practical one: in order to optimize for OOD ECE one should be able to collect samples coming from a different domain w.r.t. the training one, but still sharing the same output space (i.e., it should be the same task). This is significantly more restrictive then our requirement, where we do not impose any constraint on the out-of-distribution samples one needs to collect. It's important to highlight that DAC is not sensitive to the source of out-of-distribution data as we will show in §5.3.

4 Experimental Setup

In this work we aim at improving the model's uncertainty estimates, thus we designed different experiments aimed at assessing i) the model's calibration under different distribution shifts and ii) the model's OOD detection capabilities by means of its confidence scores. In this section we will introduce the datasets we used for evaluation, as well as the metrics and the baselines.

4.1 Datasets

4.1.1 Model calibration under distribution shift

Synthetic noise We study the robustness of our method under different types of synthetic noise injection. In particular we focus on character level perturbations as they have been shown to be more challenging for transformer models compared to other types of synthetic noise injections (Zhang et al., 2023). In particular, we will focus on two types of synthetic noise:

Random, the perturbation consists in a random operation on a character within a word. The allowed operation are i) the insertion of a new character, ii) the deletion of the character, iii) the swap of two contiguous characters. All the operations (insert/delete/swap) have equal probability of being selected. An example of one such perturbation can be found in Table 1.

Keyboard Typo (Belinkov and Bisk, 2018), the perturbation consists in the replacement of a character within a word with an adjacent key, assuming a QWERTY keyboard. An example of one such perturbation can be found in Appendix A.1 in Table 6. Each word within a sentence have a probability p_w of being perturbed, and, each character within a word has a probability p_c of being replaced.

We will study the performance of our model and the baselines with increasing noise intensity levels on two different datasets: the Stanford Sentiment Treebank (SST2) (Socher et al., 2013), a binary sentiment analysis dataset consisting of movie reviews, and the 20 News Group dataset. The models will be trained on the clean datasets and evaluated both in-distribution (clean test set) and out-of-distribution (noisy test set).

Content shift To further test the calibration of our model under distribution shift we experimented with pairs of datasets from different domains: natural language inference and paraphrase detection. For each task we have an in-distribution dataset and an out-of-distribution dataset that share the same label space but differ in content. For the natural language inference task the Stanford Natural Language Inference (SNLI) dataset (Young et al., 2014) will be used as in-distribution dataset while the Multi-Genre Natural Language Inference (MNLI) dataset (Williams et al., 2018) will serve as the OOD one, while for paraphrase detection we'll use Quora Question Pair (QQP) and TwitterPPDB (Lan et al., 2017) as IND and OOD datasets,

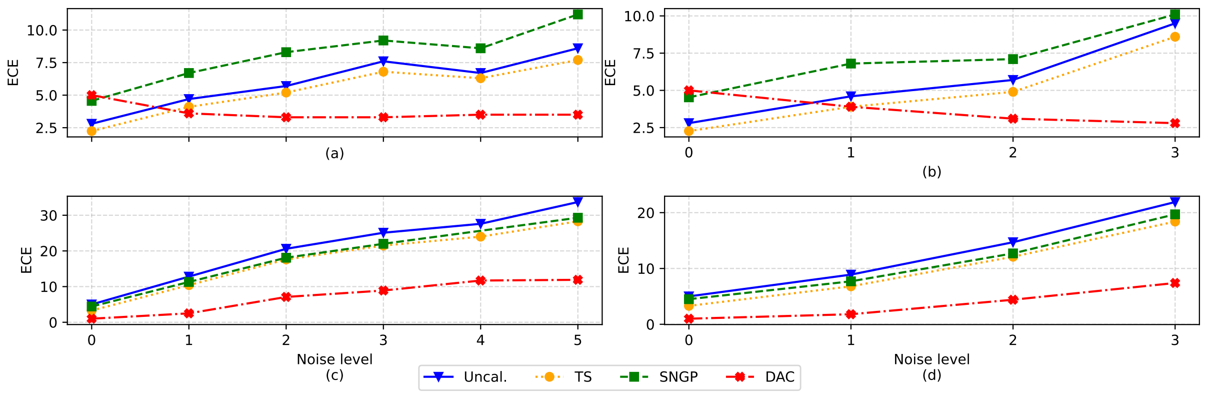

Figure 2: Overall calibration performance (ECE) of baselines and our method under increasing levels of synthetic noise using roberta-base as the PLM. Level 0 is the clean, in-distribution test set, while level 1-5 are to be considered OOD. **a)** In-distribution dataset is 20 News Group and the noise injection is of type Keyboard Typo, **b)** In-distribution dataset is 20 News Group and the noise injection is of type Random, **c)** In-distribution dataset is SST-2 and the noise injection is of type Keyboard Typo, **d)** In-distribution dataset is SST-2 and the noise injection is of type Random.

respectively.

4.1.2 Out-of-Distribution Detection

We decided to test the extent to which our method can be used for out-of-distribution detection. For this task, we will rely only on the confidence score obtained after the calibration process. We will test the model in the presence of background and semantic shifts. Background shift usually refers to changes in the input space (e.g., domain differences, style differences, etc) and not in the the output space (i.e., the labels are unchanged). On the other hand, semantic shift refers to changes in the input space that cannot be captured within the output space a model was trained on. To test the impact of background shift we will rely on common datasets for this task (Baran et al., 2023). We will use IMDB (Maas et al., 2011) as the IND dataset and SST-2 as the OOD one, while for semantic shift we experimented with CLINC OOS (Larson et al., 2019), a dataset for task-oriented dialog systems that includes queries that are out-of-scope, i.e., queries that do not fall into any of the system's supported intents. This dataset has been specifically designed to test the robustness of dialog systems in real-world settings.

4.2 Evaluation Metrics

Calibration For measuring model's calibration we will rely on Expected Calibration Error (ECE). For all the calibration experiments we'll report calibration metrics for both the in-distribution and out-of-distribution dataset.

OOD Detection To assess the model's ability to distinguish between in-distribution and out-of-distribution samples we will report: 1) the area under the receiver operating characteristic curve (AUROC), 2) the area under the precision-recall curve (PRAUC) 3) the false positive rate at 95% true positive rate (FPR95), i.e., the probability that an OOD (negative) example is classified as a positive when the Recall is above 95%.

4.3 Baselines and Models

Uncalibrated - The fine-tuned language model, without any calibration

TS (Guo et al., 2017) - the most widely known post-hoc calibration method. Temperature scaling relies on a single calibration parameter to scale the logit of the uncalibrated model, preserving its accuracy.

SNGP (Liu et al., 2023) - a distance-aware method that showed significant improvements w.r.t. base model both in calibration and out-of-domain detection for NLP tasks. We highlight that this is not a post-hoc method as it requires i) changes to the architecture of the model and also ii) to retrain it.

All experiments are carried out using the base versions of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The training details are in Appendix B. For DAC, we relied on deep k-nearest neighbour distance (Sun et al., 2022) but alternative options are considered in the ablation study in §5.3.

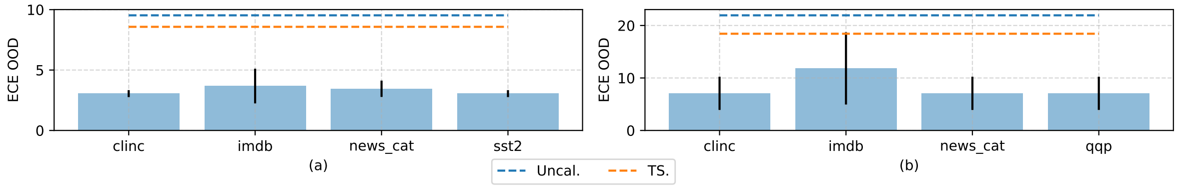

Figure 3: DAC's out-of-distribution ECE on 20 News Group (a) and SST-2 (b) with noise level 3 of type Random using different out-of-distribution sources within the calibration set. Performance for each calibration OOD source is averaged over four runs, each with a different size of the calibrations set (100, 500, 1000, 5000).

5 Results

5.1 DAC improves OOD calibration under different domain-shifts

In Figure 2 we report the overall calibration performance in the synthetic noise setting using RoBERTa as the LM.. Since the focus of this work is on post-hoc methods (with the exception of SNGP) that do not affect the overall accuracy of the model we focus on calibration metrics only. We can observe that our method consistently improves over the baselines when evaluated on OOD samples independently of the noise type. In particular, we observe larger gains when significant amount of noise is present in the text input (noise intensity > 1). Besides improving calibration on OOD samples, our method proves to be competitive for in-distribution calibration as well, leading to large improvements over the uncalibrated model and, outperforming TS for SST2. Results for the BERT based models are available in the Appendix in Figure 5; DAC still outperforms all baselines by a large margin. We noticed that overall, the out-of-the-box calibration of RoBERTa appears to be better then the one of BERT, but the gap reduces significantly once a post-hoc calibration method is used. Overall, we can conclude that Distance Aware Calibration is extremely effective in challenging noisy settings as calibration on OOD samples greatly improves for all noise types and intensities.

Next, we study whether our method is effective in a different setting, when the OOD samples are not noisy but they either have different writing style w.r.t in-distribution samples or they come from different domains. In Table 2 we show both the in-domain and out-of-domain ECE on all the studied datasets. DAC is still the most effective method for OOD calibration in this settings, but, compared to the noisy settings, TS proves to be a much stronger baseline. In fact, in the Natural Language Inference task when using RoBERTa as the PLM, it proves better then DAC at OOD calibration. For in-distribution calibration our method consistently improves over the Uncalibrated model but TS is generally more effective.

Model	PLM	PD	NLI
Uncal	bert	5.6 / 11.3	5. / 13.6
	roberta	6.4 / 12.4	4.9 / 11.6
TS	bert	**0.8** / 7.5	**1.5** / 6.2
	roberta	2.6 / 8.7	**1. / 4.5**
SNGP	bert	8.55 / 12.8	4.9 / 14.7
	roberta	6.4 / 12.5	5.8 / 12.7
DAC	bert	2.1 / **3.8**	2.7 / **2.8**
	roberta	**1.6 / 4.4**	2.6 / 6.6

Table 2: Calibration performance (ECE) for DAC and the other baselines on Paraphrase Detection (PD) and Natural language Inference (NLI). Each cell contains the in-domain ECE and the out-of-domain ECE.

5.2 DAC makes language models better OOD detectors

If a model is well calibrated its confidence scores should allow to easily discriminate between in-domain and out-of-domain samples. Nevertheless, classical post-hoc approaches such as TS provide almost no improvement to the underlying model performance on OOD detection tasks. On the contrary, DAC is tailored for this kind of tasks, as it has been designed specifically to assign different confidence levels to samples that are far from the training domain, i.e., OOD samples. Thus, complementary to the previous section where we studied DAC's effectiveness on typical calibration metrics, in this section we investigate whether a fine-tuned language model, once equipped with DAC is able to improve on different OOD detection tasks. Table 3 summarizes the results on the pairs of in-domain and out-of-distribution datasets presented in §4.1.2. All the detection metrics are computed using the top label confidence score for all baselines with the

Model	**PLM**	**IMDB**			**CLINC**		
		ROCAUC	PRAUC	FPR95	ROCAUC	PRAUC	FPR95
Uncal.	bert	81.65	97.56	62.48	96.23	99.00	13.64
	roberta	81.34	97.49	55.90	95.01	98.42	24.31
TS	bert	81.65	97.56	62.48	96.68	99.10	12.50
	roberta	81.34	97.49	55.90	94.34	98.26	27.66
SNGP	bert	84.52	98.18	**35.88**	97.01	99.14	13.71
	roberta	87.85	98.55	34.64	**96.87**	**99.16**	**13.42**
DAC	bert	**85.48**	**98.18**	46.53	**97.19**	**99.30**	**11.51**
	roberta	**90.18**	**98.87**	**29.02**	95.01	98.42	24.31

Table 3: Overall performance on the OOD detection task.

exception of SNGP where we used the predictive entropy. The results show that in general, DAC is always able to boost the performance of the underlying model, topping all baselines in all cases but one. We can conclude that besides providing calibration improvements, our method is also effective in OOD detection tasks.

5.3 Ablation Study

To use DAC on top of a fine-tuned language model there are two main design choices that need to be made: i) what distance function should be used and ii) what data should be used in the calibration procedure outlined in §3.2.2. We conduct an ablation study to assess the extent to which this two design decision affect DAC's overall performance. In Figure 3 we look at the OOD ECE for both 20 News Group and SST-2 when using different sources for the OOD samples needed to build the calibration sets described in §3.2.2. To further investigate whether size, and not just content, matters, DAC's performance - for each OOD dataset - is averaged over four runs, each with a different size for the calibrations set (100, 500, 1000, 5000). As it can be observed in Figure 3, DAC's performance exhibit some variance due the calibration set size and the OOD source being used. In particular, the reader can notice that when IMDB is the source of OOD samples DAC exhibit the worst performance and the highest variance across all OOD sources. Since the performance gap between IMDB and the other OOD sources is wider when SST-2 is the in-domain dataset we hypothesised that this may be due to the fact that IMDB and SST-2 come from the same domain (movie reviews). However, this hypothesis only partially explain the behaviour observed in Figure 3 as DAC's performance - despite having a smaller performance gap w.r.t the other OOD sources - still have the highest variance when 20 News Group is the in-domain dataset and IMDB is the OOD source. Nevertheless, despite the observed variance, our method still proves to be the better option as its performance are consistently improving over the baselines for all sources of OOD data.

Next, we study alternatives to the distance function we have considered in this work so far, i.e., KNN distance. In Figure 4 we show the overall performance of DAC in the noisy settings for three different distance function: i) KNN, ii) Mahalnaobis distance and iii) a Gaussian Mixture Model (GMM), with a single Gaussian component per class, where we use the the negative marginal likelihood of the new sample representation as a replacement for the distance score. For both Mahalnaobis and GMM, the estimators are fit on the feature embeddings of the training samples. The results clearly show that a distribution-free distance such as KNN provides better performance independently from the dataset being used.

The ablation has been performed using RoBERTa as the PLM, results for BERT are available in the Appendix C.2 in Figure 6 and 7.

6 Conclusions

In this work we introduce DAC, a new post-hoc calibration approach that leverages per-sample information in order to compute new calibrated confidence scores. Unlike previous work, our approach improves the performance of the underlying language models on a variety of settings, including noisy ones, which have been shown to be particularly challenging for transformers based models. Additionally, when coupled with DAC, the underlying language model shows significant gains in pure OOD settings, where the new calibrated con-

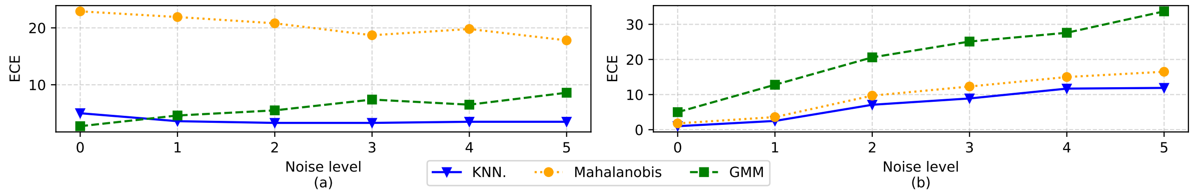

Figure 4: Overall calibration performance (ECE) of DAC using different distance functions for 20 News Group (a) and SST-2 (b) under noisy settings

fidence allows to better distinguish between in-domain and out-of-domain samples. Finally, we performed an ablation study to further understand whether DAC's performance is sensitive to some core design choices, concluding that, the use of a distribution-free distance such as KNN is the one yielding the best overall performance.

Limitations

Given a new sample, DAC requires to compute its distance from the training set. In order to do this, the training data must be available, which is not always the case. Moreover, the embeddings of the training samples have to be pre-computed and their representation must be stored in an index for efficient retrieval. This operation may be quite expensive for models where the training set is particularly large. Another limitation of our method is that we require to have a small set of out-of-distribution samples in order to calibrate the model's parameter (see §3.2.2). Nevertheless, the ablation in §5.3 shows that our method is robust to the choice of the OOD dataset being picked for calibration, which partially mitigates this issue.

References

Mateusz Baran, Joanna Baran, Mateusz Wójcik, Maciej Zięba, and Adam Gonczarek. 2023. Classical out-of-distribution detection methods benchmark in text classification tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 119–129, Toronto, Canada. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Yushi Chen, Hanlu Jiang, Chunyang Li, Xiuping Jia, and Pedram Ghamisi. 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 54(10):6232–6251.

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Gianluca Detommaso, Alberto Gasparin, Michele Donini, Matthias Seeger, Andrew Gordon Wilson, and Cedric Archambeau. 2024. Fortuna: A library for uncertainty quantification in deep learning. *Journal of Machine Learning Research*, 25(238):1–7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? *Preprint*, arXiv:2403.17661.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

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.

Guande He, Jianfei Chen, and Jun Zhu. 2022. Preserving pre-trained features helps calibrate fine-tuned language models. In *The Eleventh International Conference on Learning Representations*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Calibration meets explanation: A simple and effective approach for model confidence estimates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2784, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. 2023. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42):1–63.

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. *arXiv preprint arXiv:1907.11692*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394.

Seo Yeon Park and Cornelia Caragea. 2022. 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 (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

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.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR.

Christian Tomani, Futa Kai Waseda, Yuesong Shen, and Daniel Cremers. 2023. Beyond in-domain scenarios: robust density-aware calibration. In *International Conference on Machine Learning*, pages 34344–34368. PMLR.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Christopher K Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Andrew G Wilson, Zhiting Hu, Russ R Salakhutdinov, and Eric P Xing. 2016a. Stochastic variational deep kernel learning. *Advances in Neural Information Processing Systems*, 29.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. 2016b. Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 370–378, Cadiz, Spain. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jun Zhang, Wen Yao, Xiaoqian Chen, and Ling Feng. 2023. Transferable post-hoc calibration on pretrained transformers in noisy text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13940–13948.

Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5882–5893, Online. Association for Computational Linguistics.

A Datasets

Table 4 presents the statistics of the different dataset used in this work organized by task. The calibration datasets was created by selecting up to 1000 samples from the validation split of the in-distribution dataset and an other 1000 samples from a different dataset (OOD). The statistics for the calibration sets are available in Table 5.

A.1 Synthetic Noise

In this section we provide more details on construction of the noisy datasets in 4.1. Both datasets are obtained using transformations from the nlpaug library [‡], and the noise intensity levels are obtained with the parameters reported below.

For noise of type Random: Level 1 ($p_w = 0.1, p_c = 0.1$), Level 2 ($p_w = 0.2, p_c = 0.1$), Level 3 ($p_w = 0.6, p_c = 0.4$).

For noise of type Keyboard Typo: Level 1 ($p_w = 0.1, p_c = 0.1$), Level 2 ($p_w = 0.3, p_c = 0.1$), Level 3 ($p_w = 0.3, p_c = 0.3$), Level 4 ($p_w = 0.5, p_c = 0.1$), Level 4 ($p_w = 0.5, p_c = 0.5$).

In the main body of the paper we have presented some examples for synthetic noise of type Random. In Table 6 we provide the same for Keyboard Typo kind of noise as well.

B Implementation Details

We used the base versions of BERT and RoBERTa models starting from the pretrained checkpoints available on HuggingFace (Wolf et al., 2019). All models were trained using Fortuna (Detommaso et al., 2024) for 3 epochs using AdamW (Loshchilov and Hutter, 2018) and the model gradients are clipped to a max norm of 1. For BERT models, the learning rate is $2e^{-5}$ the batch size is 32 and weight decay is set to 0. For RoBERTa models, the learning rate is $1e^{-5}$ the batch size is 16 and weight decay is 0.1. For TS and SNGP, we relied on the implementation provided in Fortuna. Regarding SNGP, for the Spectral normalization layer we set the number of power iterations to 1 and the upper bound for spectral norm to 6, for the output layer we used 1024 random features and a temperature of 30 (used to compute them mean-field approximation to the posterior predictive). Each model run on an NVIDIA A10G Tensor Core GPU.

C Results

C.1 Synthetic noise results (BERT)

In Figure 5 we show the overall calibration performance of BERT in the synthetic noise settings.

C.2 Ablation Study (BERT)

In this section we provide the ablation results for the BERT based models. We anticipate that the

[‡] `https://github.com/makcedward/nlpaug/tree/master`

Task	In-distribution				Out-of-distribution	
	Dataset	Train	Val	Test	Dataset	Test
Synthetic noise	SST2	53880	872	13469	SST2	13397
	20 News Group	11314	3766	3766	20 News Group	3649
Content shift	QQP	363178	20207	20215	TwiterPPDB	4949
	SNLI	549367	4921	4921	MNLI	4907
OOD Detection	IMDB	25000	8250	16750	SST-2	13469
	CLINC OOS	15000	3000	4500	CLINC OOS	1000

Table 4: Datasets statistics.

Task	Calibration			
	$D^{(c)}_{\text{IND}}$	$\lvert\mathbb{D}^{(c)}_{\text{IND}}\rvert$	$\mathbb{D}^{(c)}_{\text{OOD}}$	$\lvert\mathbb{D}^{(c)}_{\text{OOD}}\rvert$
Synthetic noise	SST2	872	CLINC OOS	1000
	20 News Group	1000	SST2	1000
Content shift	QQP	1000	SNLI	1000
	SNLI	1000	QQP	1000
OOD Detection	IMDB	1000	CLINC OOS	1000
	CLINC OOS	1000	SST2	1000

Table 5: Calibration datasets statistics.

Noise Intensity	Text
0	she looks like the six-time winner of the miss hawaiian tropic pageant , so i do n't know what she 's doing in here
1	she lokks like the six - time winner of the <iss hawaiian tropic pageant, so i do n ' t know what she ' s doing in gere
2	she looks li>e the six - hime w&nner of the miss hawaiian Fropic pagean6, so i do n ' t knlw whst she ' s doung in heEe
3	she lKPks lLle the six - time winMeG of the miqd hwwwiKan tropic 9aheanR, so i do n ' t know Egat she ' s do7ny in hWrR
4	she lPoks li<e the six - tihe 2inner of the mise hzwaiian tropiD pageant, so i do n ' t knLw whag she ' s Coing in here
5	she lkoks lJke the six - tihe Sinner of the mids hawai*an tropic pag2ant, so i do n ' t kbow what she ' s dLing in Nere

Table 6: Example of a sentence coming from the SST2 dataset and its transformations under increasing noise intensity (Keyboard Typo).

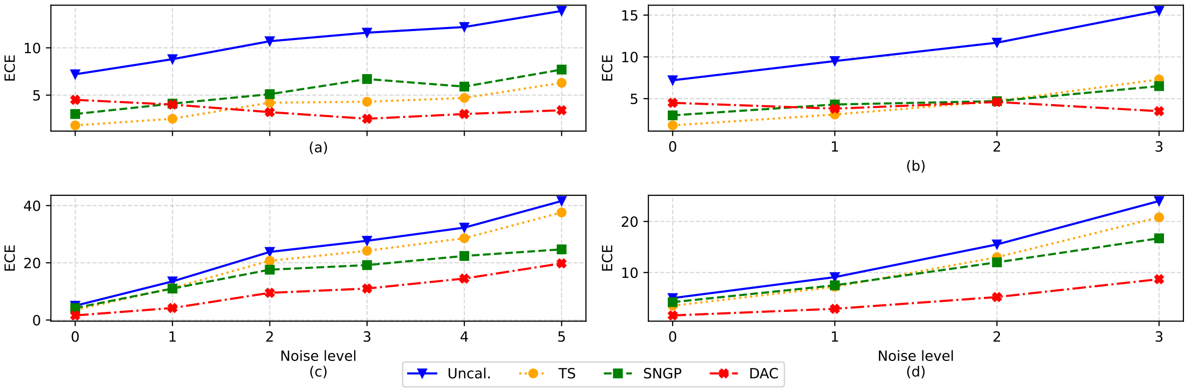

Figure 5: Overall calibration performance (ECE) of baselines and our method under synthetic noise injection using BERT as the PLM. **a)** In-distribution dataset is 20 NewsGroup and the noise injection is of type Keyboard Typo, **b)** In-distribution dataset is 20 NewsGroup and the noise injection is of type Random, **c)** In-distribution dataset is SST-2 and the noise injection is of type Keyboard Typo, **d)** In-distribution dataset is SST-2 and the noise injection is of type Random

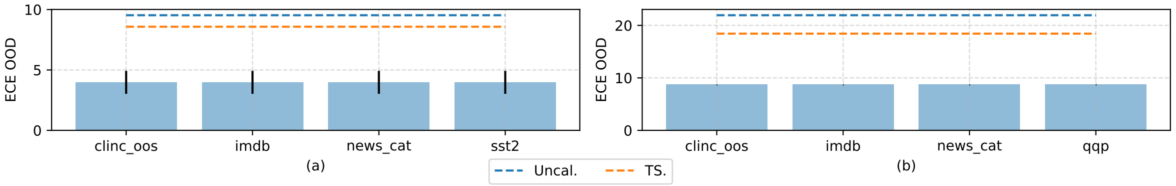

Figure 6: DAC's out-of-distribution ECE on 20 News Group (a) and SST-2 (b) with noise level 3 of type Random using different out-of-distribution sources within the calibration set. Performance for each calibration OOD source is averaged over four runs, each with a different size of the calibrations set (100, 500, 1000, 5000).

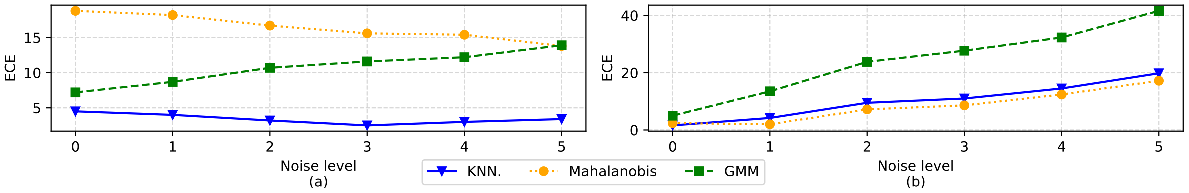

Figure 7: Overall calibration performance (ECE) of DAC using different distance functions for 20 News Group (a) and SST-2 (b) under noisy settings

conclusion that can be drawn from this results are in agreement with the results in §5.3 . In Figure 6 we look at the OOD ECE for both 20 News Group and SST-2 when using different sources for the OOD samples needed to build the calibration sets. In Figure 7 we show the overall performance of DAC in the noisy settings for the three distance functions presented in the main body of the paper (§5.3).