Similarity-Distance-Magnitude Activations

Allen Schmaltz
Reexpress AI
allen@re.express

Abstract

We introduce the SIMILARITY-DISTANCE-MAGNITUDE (SDM) activation function, a more robust and interpretable formulation of the standard softmax activation function, adding SIMILARITY (i.e., correctly predicted depth-matches into training) awareness and DISTANCE-to-training-distribution awareness to the existing output MAGNITUDE (i.e., decision-boundary) awareness, and enabling interpretability-by-exemplar via dense matching. We further introduce the SDM estimator, based on a data-driven partitioning of the class-wise empirical CDFs via the SDM activation, to control the class- and prediction-conditional accuracy among selective classifications. When used as the final-layer activation over pre-trained language models for selective classification, the SDM estimator is more robust to covariate shifts and out-of-distribution inputs than existing calibration methods using softmax activations, while remaining informative over in-distribution data.

1 Introduction

Neural-network-based language models (LMs) pose a challenge for interpretable and reliable deployment given the non-identifiability of their parameters (Hwang and Ding, 1997, inter alia)[1], which can number in the billions or more. Instead of directly interpreting parameters, one option is to move the focus of interpretation to auditing predictions as a form of interpretability by example, or *exemplar*, over the representation space of such models via dense matching (Schmaltz, 2021). However, for real-world deployments, robust approaches for predictive uncertainty are also needed, both for human decision-making and for constructing sequentially dependent LM pipelines.

[1]Informally, this means that two or more distinct sets of values for the parameters can result in identical output distributions. As a consequence, interpreting the parameters of such models is typically more complicated than with a simple linear regression model, for example.

Known theoretical results limit the statistical quantities that can be derived over LMs. Statistical guarantees in the distribution-free setting are limited to approximately conditional quantities (Valiant, 1984; Lei and Wasserman, 2014; Foygel Barber et al., 2020, inter alia). Further, even typical approximately conditional quantities can be difficult to obtain in practice, since the minimal assumption of exchangeability with a known held-out dataset is itself often violated with covariate and label shifts, which can be difficult to foresee with existing methods. Epistemologically, the prevalence of hallucinations and highly-confident wrong answers with widely deployed LMs suggests a technical impasse in effectively modeling the predictive uncertainty, despite significant work from Bayesian, Frequentist, and empirically motivated perspectives (Gal and Ghahramani, 2016; Angelopoulos et al., 2021; Guo et al., 2017; Lakshminarayanan et al., 2017; Ovadia et al., 2019, inter alia). A foundational piece is evidently missing from the picture.

Given these intrinsic challenges, we approach the problem of uncertainty quantification over LMs from a new angle and ask: *Can we leverage the metric learning and dense matching capabilities of neural networks over high-dimensional inputs to at least aim to decompose signals of epistemic (reducible) uncertainty in a manner that is interpretable and actionable?*

We address this question with a conceptually parsimonious, data-driven partitioning of the data to decompose sources of epistemic uncertainty: Correctly predicted depth-matches into the training set (SIMILARITY), the DISTANCE to the training set, and the distance to the decision-boundary (MAGNITUDE). We use these signals to construct a new activation function, the SDM activation, which can be used as a replacement for the standard softmax activation, as, for example, the final-layer activation. The SDM activation enables more re-

Findings of the Association for Computational Linguistics: ACL 2026, pages 22037–22057

liable estimates of the predictive uncertainty for selective classification (Chow, 1957; Geifman and El-Yaniv, 2017, inter alia), which addresses the need for uncertainty quantification with LMs used in multi-stage decision pipelines, in settings subject to covariate shifts and out-of-distribution inputs.

In summary, in this work:

- We introduce the SIMILARITY-DISTANCE-MAGNITUDE (SDM) activation function.
- We introduce the SDM estimator for use in controlling class- and prediction-conditional accuracy among selective classifications, based on a data-driven partitioning of the class-wise empirical cumulative distribution functions (eCDFs) over the SDM activation output.
- We examine the behavior of the SDM activation as a final-layer activation over pre-trained language models, using the SDM estimator for selective classification. We demonstrate empirically that the SDM estimator is more robust to covariate-shifts and out-of-distribution inputs than existing classes of post-hoc calibration methods, while remaining informative over in-distribution data.

2 Preliminaries

2.1 Setting

We consider the standard multi-class classification setting. We are given a training dataset, $\mathcal{D}_{\text{tr}} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ of inputs, $\boldsymbol{x} \in \mathcal{X}$, paired with their corresponding ground-truth discrete labels, $y \in \mathcal{Y} = \{1, \ldots, C\}$, and a labeled calibration dataset, $\mathcal{D}_{\text{ca}}$, drawn from the same distribution as $\mathcal{D}_{\text{tr}}$. We are then given a new test instance, $\boldsymbol{x}$, from an unlabeled test set, $\mathcal{D}_{\text{te}}$, and seek to estimate the label with a prediction, $\hat{y}$, via the unnormalized log probabilities ("logits", informally) of a final linear layer: $\boldsymbol{z} = \boldsymbol{W}^T\boldsymbol{h} + \boldsymbol{b}$, where $\boldsymbol{h} = \text{network}(\boldsymbol{x}; \theta)$ is the final hidden state of a network parameterized by θ. The discrete prediction is taken as $\hat{y} = \arg\max \boldsymbol{z}$; however, for learning θ, $\boldsymbol{W}$, and $\boldsymbol{b}$, and for human decision-making, we also seek an estimate of the predictive uncertainty, $p(y \mid \boldsymbol{x})$, which is typically obtained by normalizing $\boldsymbol{z}$ via the softmax activation described next. We will make a distinction between models, $\mathcal{M}$ (defined by θ, $\boldsymbol{W}$, and $\boldsymbol{b}$, and when applicable, the exemplar adaptor, described below), which produce the prediction, $\hat{y}$, and estimators, $\mathcal{E}$, which provide an estimate of $p(y \mid \boldsymbol{x})$, because different estimators can be used over the same model.

2.2 Softmax and the Cross-Entropy loss

The softmax activation is commonly used in neural network architectures, including, for example, as a router in self-attention mechanisms (Vaswani et al., 2017) and mixture-of-experts models (Shazeer et al., 2017), and forming the basis of the cross-entropy loss used for next-token training of LMs. It is the typical final output layer of LMs, converting the un-normalized model logits to a normalized probability distribution:

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{\tau \cdot z_i}}{\sum_{c=1}^{C} e^{\tau \cdot z_c}}, 1 \leq i \leq C, \tau \geq 0 \tag{1}$$

The inverse-temperature parameter, τ, controls the sharpness of the distribution. As $\tau \to 0$, the output of $\text{softmax}(\boldsymbol{z})$ converges to a uniform distribution where each class has probability $\frac{1}{C}$; as $\tau \to \infty$, the output converges to a distribution in which all of the mass is assigned to a single class. In deep learning, τ is treated as a learnable, *global* hyperparameter; *instance-wise* variation in the distance to the decision-boundary is thus determined by the relative MAGNITUDE of $z_{\hat{y}}$. This model is learned by minimizing the cross-entropy loss between $\boldsymbol{z}$ and the index of the true labels over $\mathcal{D}_{\text{tr}}$. The *natural* logarithm of the loss is the counterpart to the base e of the softmax:

$$\mathcal{L}(\theta, \boldsymbol{W}, \boldsymbol{b}; \mathcal{D}_{\text{tr}}) = -\frac{1}{N} \sum_{n}^{N} \log_e \left(\frac{e^{\tau \cdot z_{y_n}}}{\sum_{c=1}^{C} e^{\tau \cdot z_c}} \right) \tag{2}$$

3 Methods

In this work, we revisit Eq. 1 and 2. We seek to decouple the sources of epistemic uncertainty via a new activation function that is conceptually:

$$\text{SDM}(\boldsymbol{z})_i = \frac{\text{SIMILARITY}^{\text{DISTANCE} \cdot \text{MAGNITUDE}_i}}{\sum_{c=1}^{C} \text{SIMILARITY}^{\text{DISTANCE} \cdot \text{MAGNITUDE}_c}} \tag{3}$$

with a corresponding negative log likelihood loss that takes into account the change of base (§ 3.1). Unique to this setting, a modification to label-conditional conformal prediction (Vovk et al., 2005) then follows via a parsimonious partitioning of the class-wise empirical CDFs, providing a principled basis for controlling the class-conditional

accuracy among selective classifications, combined with empirically-robust prediction-conditional estimates.

3.1 Similarity-Distance-Magnitude Activation Functions

Calculating the SDM activation involves training an exemplar adaptor, a 1-D CNN adaptor (with a final linear layer) over the frozen hidden states of a network, to induce distilled, compressed representations of the underlying network's representation space conditional on its predictions. The resulting representations provide a probabilistic mapping to the training, or support, set. In this way, neural networks, including large pre-trained networks, can be viewed as *hidden* instance-based metric learners (Schmaltz, 2021), from which we can then derive signals of the epistemic uncertainty.

3.1.1 Exemplar Adaptor

We take as the CNN of our exemplar adaptor $g : \boldsymbol{h} \in \mathbb{R}^D \mapsto \boldsymbol{h}' \in \mathbb{R}^M$, a 1-D CNN that takes as input $\boldsymbol{h}$ of the underlying network. The CNN has M filters, the filter applications of which produce $\boldsymbol{h}'$, the distilled representation of the underlying network. A final linear layer, $\boldsymbol{z}' = \boldsymbol{W}'^T\boldsymbol{h}' + \boldsymbol{b}', \boldsymbol{z}' \in \mathbb{R}^C$, then replaces the underlying network's final linear layer, with the discrete prediction taken as $\hat{y} = \arg\max \boldsymbol{z}'$. This exemplar adaptor will then enable us to derive the SIMILARITY, DISTANCE, and MAGNITUDE values, as defined next.

3.1.2 SIMILARITY

We define the SIMILARITY (q) of an instance to the training set as the count of consecutive nearest matches in $\mathcal{D}_{\text{tr}}$ that are correctly predicted *and* match $\hat{y}$ of the test instance.[2] Concretely, we first sort $\mathcal{D}_{\text{tr}}$ (for which we have both model predictions and ground-truth labels) based on the L^2 distance from $\boldsymbol{h}'$, $\left[(\boldsymbol{x}^{tr}_{(1)}, \hat{y}^{tr}_{(1)}, y^{tr}_{(1)}), \ldots, (\boldsymbol{x}^{tr}_{(N)}, \hat{y}^{tr}_{(N)}, y^{tr}_{(N)})\right]$, such that $||\boldsymbol{h}' - \boldsymbol{h}'^{tr}_{(1)}||_2 \leq \ldots \leq ||\boldsymbol{h}' - \boldsymbol{h}'^{tr}_{(N)}||_2$, and then calculate $q \in \{0, \ldots, |\mathcal{D}_{\text{tr}}|\}$ as:

$$q = \sum_{i=1}^{|\mathcal{D}_{\text{tr}}|} \mathbf{1}_{\hat{y}=\hat{y}^{\text{tr}}_{(i)}} \cdot \mathbf{1}_{\hat{y}^{\text{tr}}_{(i)}=y^{\text{tr}}_{(i)}} \cdot \mathbf{1}_{i-1=\sum_{j=1}^{i-1} \mathbf{1}_{\hat{y}=\hat{y}^{\text{tr}}_{(j)}} \cdot \mathbf{1}_{\hat{y}^{\text{tr}}_{(j)}=y^{\text{tr}}_{(j)}}} \quad (4)$$

where the rightmost indicator function, $\mathbf{1} \in \{0, 1\}$, ensures consecutive (depth-wise) matches.[3] By definition, q cannot exceed the count of the most prevalent class label in $\mathcal{D}_{\text{tr}}$, and since we assume an approximately equal number of points for each class, $q \ll |\mathcal{D}_{\text{tr}}|$ is typical. For the special case of calculating q for $\boldsymbol{x} \in \mathcal{D}_{\text{tr}}$, which only occurs during learning, we exclude the self-match.

3.1.3 DISTANCE

The L^2 distance to the nearest match in $\mathcal{D}_{\text{tr}}$ follows from above: $d_{\text{nearest}} = ||\boldsymbol{h}' - \boldsymbol{h}'^{tr}_{(1)}||_2$. We normalize these values by defining the DISTANCE, $d \in [0, 1]$, in terms of the class-wise empirical CDFs of d_{nearest} over $\mathcal{D}_{\text{ca}}$, as the most conservative quantile relative to the distance to the nearest matches observed in the labeled, held-out set:

$$d = \min\left[1 - \text{eCDF}^{y_1}_{\text{ca}}(d_{\text{nearest}}), \ldots, 1 - \text{eCDF}^{y_C}_{\text{ca}}(d_{\text{nearest}})\right] \quad (5)$$

The empirical CDFs are determined by the labeled points in $\mathcal{D}_{\text{ca}}$ for which $q > 0$, where, as indicated by the superscripts, the stratification of points is by the true labels, y. For example, $\text{eCDF}^{y_1}_{\text{ca}}(d_{\text{nearest}})$ is the empirical CDF of d_{nearest} values in $\mathcal{D}_{\text{ca}}$ for which $y = 1$, a notation convention we will use throughout. (Points with $q = 0$ are effectively out-of-distribution points and treated as such in downstream decision-making, so they are excluded to avoid biasing the estimates.) At test time, we do not see y; instead, the minimum is calculated over the quantiles of each of the class-conditional eCDFs, regardless of $\hat{y}$. As with q, for the special case of calculating d for $\boldsymbol{x} \in \mathcal{D}_{\text{tr}}$, we replace $\text{eCDF}^{y_c}_{\text{ca}}$ with the analogous $\text{eCDF}^{y_c}_{\text{tr}}$, the class-wise empirical CDFs of d_{nearest} over $\mathcal{D}_{\text{tr}}$ excluding self-matches.

Appendix A.5 provides a method for accounting for error in the estimation of the eCDFs given the effective sample size.

3.1.4 MAGNITUDE

We take as the MAGNITUDE, or distance to the decision boundary, $z'_{\hat{y}}$, as in the standard softmax case but via $\boldsymbol{z}'$ from the linear layer of the exemplar adaptor.

[2] We use the letter q, as this value **q**uantizes the closeness of a point to the training set with a discrete estimate.

[3] This seemingly simple rule differs from traditional KNN rules (Cover and Hart, 1967; Devroye et al., 1996, inter alia) in two critical respects: The neural network serves as a semi-supervised metric learner of the distances between the dense representations that identify the instances, and there is a model prediction (in addition to the ground-truth label) for each instance in the support set. The former enables effective partitioning despite the curse of high dimensions; the latter provides an additional indicator of reliability for each instance.

3.1.5 SDM Activation: Formulation

We use the above quantities to define the SDM activation function:

$$\text{SDM}(\boldsymbol{z}')_i = \frac{(2+q)^{d \cdot z'_i}}{\sum_{c=1}^{C} (2+q)^{d \cdot z'_c}}, 1 \leq i \leq C \quad (6)$$

The output distribution becomes sharper with larger values of q and d, as well as with larger relative values of $z'_{\hat{y}}$, as with the standard $\mathrm{softmax}$. When d_{nearest} exceeds the largest distance observed in the labeled data, $d = 0$ and the output distribution is uniform, reflecting maximally high uncertainty. The standard $\mathrm{softmax}$ with $\tau = 1$ is recovered by setting $q = e - 2, d = 1$. Provided $d \neq 0$, $\arg\max \text{SDM}(\boldsymbol{z}') = \arg\max \boldsymbol{z}'$, as with the $\mathrm{softmax}$ activation.

3.1.6 SDM Activation: Loss and Training

A loss analogous to Eq. 2 then follows with the applicable change of base. We use this loss to train the weights of the exemplar adaptor, which includes the parameters of the linear layer ($\boldsymbol{W}'$ and $\boldsymbol{b}'$), as well as the convolution weights and biases, which we collectively represent with $\boldsymbol{G}$. The weights of the underlying $\mathrm{network}$ remain fixed.

$$\mathcal{L}(\boldsymbol{G}, \boldsymbol{W}', \boldsymbol{b}'; \mathcal{D}_{\text{tr}}) = \\ -\frac{1}{N} \sum_{n}^{N} \log_{(2+q)} \left(\frac{(2+q)^{d \cdot z'_{y_n}}}{\sum_{c=1}^{C} (2+q)^{d \cdot z'_c}} \right) \quad (7)$$

The first epoch of training is initialized with a standard $\mathrm{softmax}$ (i.e., setting $q = e-2, d = 1$). Training then proceeds by re-calculating q and d for each $\boldsymbol{x} \in \mathcal{D}_{\text{tr}}$ after each epoch. We take as the stopping criteria for one learning round as the epoch with the lowest balanced (across classes) average loss over $\mathcal{D}_{\text{ca}}$. We repeat this process for J iterations of random shuffles and splits of $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$ and parameter initializations, choosing the final model as that with the globally lowest balanced (across classes) average loss over $\mathcal{D}_{\text{ca}}$.

3.2 Evaluating Selective Classification

As a common, unambiguous baseline quantity for comparing selective classifiers over a held-out test set, $\mathcal{D}_{\text{te}}$, we seek an easy-to-interpret and easy-to-evaluate metric, reflecting real-world applications. Among the selective classifications from an estimator, we seek (Quantity I) prediction-conditional accuracy at or above a given threshold, $\alpha \in (\frac{1}{C}, 1]$, and (Quantity II) class-conditional accuracy at or above that same threshold, α.

To evaluate this metric, we only consider the points for which the given estimator assigns a high-probability of at least α, which is typically near 1, such as $\alpha = 0.95$ in our experiments. We refer to this set of points as the *admitted*, or *non-rejected*, set. Then, given ground-truth values for $\mathcal{D}_{\text{te}}$, we assess whether the conditional accuracies of the admitted set are at least α when (Quantity I) stratifying by the predicted labels, $\hat{y}$, and when (Quantity II) stratifying by the true labels, y.

The estimator that rejects all points would meet these conditions. However, given two estimators that meet these conditions, we prefer that which rejects fewer points, ceteris paribus. In other words, we seek estimators that meet our reliability condition and are informative (i.e., maximize the number of points that are properly admitted), but when the estimator is uncertain, we prefer rejection over unexpectedly falling under the desired α probability threshold.

Quantity I corresponds to top-label calibration (Gupta and Ramdas, 2022), but with a single bin for evaluation, $[\alpha, 1]$, removing ambiguity with regard to the choice of binning the probabilities. Quantity II does not directly correspond to quantities typically examined in the calibration literature (Brier, 1950; Dawid, 1982; Guo et al., 2017; Vaicenavicius et al., 2019; Kull et al., 2019, inter alia), but it approximates label-conditional conformal coverage in the special case of class-wise thresholds that only admit prediction sets of cardinality 1. We introduce a straightforward procedure to estimate this quantity next.

3.2.1 Controlling the Class-conditional Accuracy among Selective Classifications with SDM Estimators

In general, the statistical coverage guarantee of marginal split-conformal estimators is not directly informative for selection, since the coverage guarantee is not conditional on the set size. We may instead seek one of various approximately conditional notions of coverage (Romano et al., 2020; Angelopoulos et al., 2021, inter alia); however, there is no guarantee that when we stratify by sets of cardinality 1, coverage will be maintained. However, there is a *special case* in which label-conditional conformal estimators *do* provide a meaningful notion of class-conditional coverage for selection. Assuming $\mathcal{D}_{\text{ca}}$ and $\mathcal{D}_{\text{te}}$ are exchangeable, if the conformity score for each label is from a categorical distribution and the resulting thresh-

olding of the class-wise empirical CDFs results in class-wise thresholds that are all greater than $\frac{1}{C}$, then the cardinality 1 sets will, on average, obtain class-conditional coverage, by definition.[4] Unfortunately, it may be rare to encounter this restricted setting over the full data distributions of real-world tasks. Instead, we will use the SDM activation to estimate a partitioning of the distribution into a region that approximately fulfills these assumptions.

First, we rescale the SIMILARITY estimate to take into account the DISTANCE and MAGNITUDE, given the predicted class. The resulting value[5] will be the basis for partitioning the distribution:

$$q' = \min\left(q, (2+q)^{\text{SDM}(\boldsymbol{z}')_{\hat{y}}}\right) \tag{8}$$

Next, we estimate label-conditional conformal thresholds, $[\psi_1, \ldots \psi_C]$, over the output from the SDM activation among a subset of the distribution constrained by progressively larger values of q' (among $q' > 0$) until all thresholds are at least α. By setting the stopping criteria at α rather than $\frac{1}{C}$, we also restrict the region to the empirically-motivated prediction-conditional quantity, $\text{SDM}(\boldsymbol{z}')_{\hat{y}}$. The procedure appears in Alg. 1. If we find a finite $q'_{\min}$ that obtains such thresholds, we refer to the resulting region as the HIGH-RELIABILITY (SDM_{HR}) region, taking membership in this region as our selection criteria:

$$\text{SDM}_{\text{HR}} := \begin{cases} \hat{y} & \text{if } q' \geq q'_{\min} \wedge \text{SDM}(\boldsymbol{z}')_{\hat{y}} \geq \psi_{\hat{y}} \\ \perp & \text{otherwise} \end{cases} \tag{9}$$

where $\perp$ indicates a rejected (non-admitted) point and $\hat{y} = \arg\max \boldsymbol{z}'$.[6]

To calculate this quantity for new, unseen test points $\boldsymbol{x} \in \mathcal{D}_{\text{te}}$, we require $\mathcal{D}_{\text{tr}}$ to calculate q and d_{nearest}; the cached class-wise empirical CDFs of the distances over $\mathcal{D}_{\text{ca}}$ of Eq. 5; and $q'_{\min}$ and the thresholds, $[\psi_1, \ldots \psi_C]$. Evaluation of the SDM_{HR} selection criteria is straightforward: We simply assess the conditional accuracies for the admitted points after stratifying by the predictions and the true labels, each in turn. When Alg. 1 returns $q'_{\min} = \infty$, we obtain a useful empirical indicator that the model is too weak, or the data insufficient, to reliably obtain class- and prediction-conditional estimates at the specified α value.

[4] Although the formal interpretations are not identical, the evaluation of class-conditional coverage for a single estimate of such cardinality 1 prediction sets in this restricted setting over $\mathcal{D}_{\text{te}}$ is numerically equivalent to assessing the class-conditional accuracy, when not considering the sample-size dependent error term. See Appendix A.5 for a further discussion of analyzing the effective sample size.

[5] The min ensures that q' remains 0 for points with $q = 0$, which are effectively out-of-distribution points.

[6] The convention is to refer to the basic architecture of Eq. 6 as the SDM activation function, and using the activation with the selection criteria of Eq. 9 as the SDM estimator.

4 Experiments

We provide controlled comparisons of our proposed methods over representative LMs and tasks, systematically ablating relevant components, holding the data and underlying LM constant, ceteris paribus. We consider in-distribution, covariate shifted, and out-of-distribution test sets. We consider representative estimators over the existing LM architecture (i.e., without additional parameters); with CNN adaptors; and with the SDM activation layer. Additional details are provided in the Appendix.

4.1 Task: Binary Sentiment Classification

SENTIMENT: $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$. Our first task is predicting the sentiment of movie reviews. We use the commonly used benchmark data of Maas et al. (2011). This is a binary classification task with $y \in \{0 = \text{negative}, 1 = \text{positive}\}$. $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$ are constructed from a total of 18k instances. The held-out set for evaluation (SENTIMENT), $|\mathcal{D}_{\text{te}}| = 1583$, is from the same distribution as $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$.

SENTIMENTOOD. To evaluate the behavior of the estimators over out-of-distribution (OOD) data, we consider an additional evaluation set, SENTIMENTOOD, $|\mathcal{D}_{\text{te}}| = 4750$. We use the SemEval-2017 Task 4a test set (Rosenthal et al., 2017), which consists of short-form social media posts that differ in the distribution of topics, language styles, and lengths relative to the movie reviews. We balance the test set, dropping the third class (neutral), setting the semantics of the true labels to be the same as that of the movie reviews: $y \in \{0 = \text{negative}, 1 = \text{positive}\}$.

Far OOD Challenge Sets. In Appendix A.3, we consider two additional out-of-distribution challenge test sets, SENTIMENTSHUFFLED and SENTIMENTOODSHUFFLED, constructed by randomly shuffling the input documents for each of SENTIMENT and SENTIMENTOOD, respectively.

4.2 Task: FACTCHECK

FACTCHECK. As a more challenging binary classification task for LMs, we consider the fact

Algorithm 1 Search Algorithm to Find $q'_{\min}$ and $[\psi_1, \dots \psi_C]$ to Estimate the HIGH-RELIABILITY Region

Input: cached $(q', \text{SDM}(\boldsymbol{z}'))\ \forall\ \boldsymbol{x} \in \mathcal{D}_{\text{ca}}, \alpha \in (\frac{1}{C}, 1]$

procedure ESTIMATE-HIGH-RELIABILITY-REGION(cached $(q', \text{SDM}(\boldsymbol{z}'))\ \forall\ \boldsymbol{x} \in \mathcal{D}_{\text{ca}}, \alpha \in (\frac{1}{C}, 1]$)
  $q'_{\min} \leftarrow \infty$ ▷ A finite $q'_{\min}$ may not be found.
  $[\psi_1, \dots \psi_C] \leftarrow [\infty, \dots, \infty]$ ▷ Class-wise output thresholds
  sortedList $\leftarrow$ sorted $[q' \in \mathcal{D}_{\text{ca}}$ s.t. $q' > 0]$ ▷ Restricted to $q' > 0$ to exclude OOD
  for $q^* \in$ sortedList **do**
    Construct $\text{eCDF}^{y_1}_{\text{ca}}, \dots, \text{eCDF}^{y_C}_{\text{ca}}$ for all $q' \geq q^*$ in $\mathcal{D}_{\text{ca}}$ ▷ eCDFs for $\text{SDM}(\boldsymbol{z}')$ (Eq. 6), stratified by y
    Calculate $\psi_c = \text{inverseCDF}^{y_c}_{\text{ca}}(1-\alpha)\ \forall\ c \in \{1, \dots, C\}$ ▷ Quantile functions are inverses of L. 6
    if all($[\psi_1, \dots \psi_C] \geq \alpha$) **then** ▷ Element-wise comparison
      $q'_{\min} \leftarrow q^*$
      break
  return $q'_{\min}, [\psi_1, \dots \psi_C]$

Output: $q'_{\min}, [\psi_1, \dots \psi_C]$

check data of Azaria and Mitchell (2023). The training and calibration sets, a combined total of 6k instances, consist of single sentence statements that have been semi-automatically generated via templates and a knowledge base. The task is to determine whether the statement is true or false, $y \in \{0 = \text{false}, 1 = \text{true}\}$. The held-out eval set (FACTCHECK), $|\,\mathcal{D}_{\text{te}}\,| = 245$, the focus of our analysis, has been constructed by having an LM generate a statement continued from a true statement not otherwise in the dataset. These evaluation statements are checked manually and assigned labels by human annotators. In addition to being a relatively challenging task that evaluates—at least in principle—the latent knowledge stored within an LM's parameters, the test set is representative of the types of covariate shifts over high-dimensional inputs that can be problematic for real applications, and challenging to characterize without model assistance and ground-truth labels. It was observed in Azaria and Mitchell (2023) that the accuracy of existing LM classifiers is lower on this generated, held-out test set compared to the calibration set. However, these test sentences would seem to also be simple true-false statements, reflecting that it is not necessarily straightforward for a human user to detect distribution shifts over high-dimensional inputs. As such, we seek for our models and estimators to reflect such shifts via the predictive uncertainty.

FACTCHECKSHUFFLED. As with the sentiment task, in Appendix A.3 we also consider an additional out-of-distribution challenge test set, FACTCHECKSHUFFLED, constructed by randomly shuffling the input documents of the FACTCHECK task.

4.3 Models

We consider two representative, open-weight decoder-only Transformer-based LMs, the parameters of which stay fixed: The 3.8 billion-parameter `Phi-3.5-mini-instruct` model (PHI3.5) (Abdin et al., 2024), and the 47 billion-parameter `Mixtral 8x7B Instruct v0.1` mixture-of-experts model (MIXTRAL8X7B) (Jiang et al., 2024).

Hidden states. We take as $\boldsymbol{h}$ the concatenation of the final-layer hidden state of the final sequence position (i.e., the hidden state that is the input to the linear-layer over the output vocabulary for the `Yes` or `No` generation) with the mean over all final-layer hidden states. For PHI3.5, this results in $\boldsymbol{h} \in \mathbb{R}^{6144}$, and for MIXTRAL8X7B, $\boldsymbol{h} \in \mathbb{R}^{8192}$.

4.4 Estimators

We examine representative Frequentist, Bayesian, and empirically-motivated classes of estimators used with neural networks, setting $\alpha = 0.95$ for all experiments. At the most basic, but perhaps the most commonly used in practice, representing the absence of a post-hoc calibration method, we simply threshold the output, $\text{softmax}(\boldsymbol{z}) \geq \alpha$, where the temperature $\tau = 1$. For this, we use the label softmax. As an established empirical approach for calibrating neural networks, we provide a comparison to temperature scaling (Guo et al., 2017), a single parameter version of post-hoc Platt-scaling (Platt, 1999), with the label TEMPSCALING. In this case, the estimator is the thresholding of the output $\text{softmax}(\boldsymbol{z}; \tau) \geq \alpha$ after learning a value for τ over $\mathcal{D}_{\text{ca}}$. We also provide a comparison to two representative conformal predictors, the APS method of Romano et al. (2020) and the adaptiveness-optimized RAPS algorithm of Angelopoulos et al. (2021). The admission criteria for the APS and RAPS estimators is prediction sets

of size 1, at the 0.05 level (i.e., $1 - \alpha$, as defined here). We consider these estimators over the logits corresponding to the `Yes` and `No` indexes of the output linear-layer of the underlying LM (PHI3.5 and MIXTRAL8X7B), which provides a reference point without introducing additional adaptor layers. We also consider these baselines over 1-D CNN adaptors over $\boldsymbol{h}$ of each LM (PHI3.5+ADAPTOR and MIXTRAL8X7B+ADAPTOR). These baseline adaptors are identical to those used in the corresponding SDM activation layers, with $M = 1000$, and are similarly trained for $J = 10$ iterations of 200 epochs, but unlike the SDM activation layers, the stopping criteria is the minimum balanced (across classes) average cross-entropy loss.

We further compare to variational Bayesian last-layer neural networks (VBLL), a computationally efficient Bayesian approach (Harrison et al., 2024). We consider the discriminative versions over `Phi-3.5-mini-instruct` and `Mixtral 8x7B Instruct v0.1` (PHI3.5+DISCVBLLMLP and MIXTRAL8X7B+DISCVBLLMLP, respectively) in the main text, with additional comparisons in Appendix A.8.

We then compare to the final-layer SDM activations over $\boldsymbol{h}$ of each LM (PHI3.5+SDM and MIXTRAL8X7B+SDM). For reference, we provide the result of thresholding a $\mathrm{softmax}$ over these adaptors at α, as above, as well as a thresholded $\mathrm{softmax}$ that simply treats d as the inverse-temperature, $\textsc{softmax}(d \cdot \boldsymbol{z}')$, which is equivalent to setting $q = e - 2$ in the SDM activation. We consider an analogous threshold over the activation output, $\textsc{sdm}(\boldsymbol{z}') \geq \alpha$, for which we use $\textsc{sdm}_\alpha$ as the estimator label. Finally, for the eponymous "SDM estimator" using the selection criteria of Eq. 9, we use the label $\textsc{sdm}_{\textsc{hr}}$ in the results tables.

As a common point of reference, the label NO-REJECT refers to predictions without any selective filtering (i.e., the raw output accuracies, derived from the $\arg\max$ over the final linear-layer).

5 Results

In-distribution data. Representative results are provided in Table 1, with additional results in the Appendix. Even on the in-distribution SENTIMENT dataset, the estimators over the underlying LMs without adaptor layers exhibit over-confidence, which is reflected in conditional accuracies that fall below the expected α. The estimators over the adaptor layers all obtain the desired conditional accuracies, with the class-wise accuracies of the models themselves $\geq \alpha$ (see the NO-REJECT rows in Table 3), with differences arising in the proportion of admitted points. Here and elsewhere, $\textsc{softmax}(d \cdot \boldsymbol{z}')$ tends to be overly conservative in rejecting points. To be expected, the $\textsc{sdm}_{\textsc{hr}}$ estimator tends to be more conservative than simply thresholding the SDM activation at α ($\textsc{sdm}_\alpha$), but the latter lacks the assurances on the class-conditional accuracy obtained by the constraints on the HIGH-RELIABILITY region. In practice, this behavior can be used as a basis to triage selective classifications: Documents in the HIGH-RELIABILITY region might be treated as automated, or semi-automated, predictions in the decision pipeline, whereas other documents might be triaged by $\textsc{sdm}(\boldsymbol{z}')_{\hat{y}}$ for calling more resource-intensive LM tools, or human adjudication. Importantly, as we discuss below with the covariate-shifted and out-of-distribution datasets, the non-SDM-based estimators provide a less reliable substrate for basing such conditional branching decisions.

Covariate-shifted and Out-of-distribution data. With SENTIMENTOOD, the distinctions among the estimators become clear, with the non-SDM-based estimators performing poorly, even in terms of marginal calibration. That would come as a surprise to end-users, whereas with the $\textsc{sdm}_{\textsc{hr}}$ estimator, the out-of-distribution documents are more reliably rejected, with the few admitted predictions generally obtaining high conditional accuracies, despite the relatively low accuracies over the test set without selection (see NO-REJECT in Table 3). A similar pattern is observed over the FACTCHECK dataset in Table 1, and Table 4 in the Appendix.

Understanding $q'_{\min}$. For reference, Table 2 provides the results over $\mathcal{D}_{\text{ca}}$ for the SDM-based estimators. The value of $q'_{\min}$ tends to increase as the accuracy over $\mathcal{D}_{\text{ca}}$ decreases, reflecting a more conservative HIGH-RELIABILITY region that admits fewer points. Alg. 1 failed to find a finite $q'_{\min}$ for MIXTRAL8X7B+SDM over the FACTCHECK calibration set at $\alpha = 0.95$, so for reference, we also show the HIGH-RELIABILITY region at $\alpha = 0.94$, as well as in Tables 1 and 4. In this way, $q'_{\min}$ provides a principled, data-driven indicator of the reliability of the estimates, which is interpretable as a simple indicator as to whether the conditional accuracies are, or are not, obtainable over $\mathcal{D}_{\text{ca}}$ at the specified α. Similar to conformal estimators, and unlike TEMPSCALING and VBLL estimators, Alg. 1

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert D_{te} \rvert}$	ACC.	$\frac{n}{\lvert D_{te} \rvert}$	ACC.	$\frac{n}{\lvert D_{te} \rvert}$	ACC.	$\frac{n}{\lvert D_{te} \rvert}$	ACC.	$\frac{n}{\lvert D_{te} \rvert}$
SENTIMENT	PHI3.5	SOFTMAX	0.98	0.50	0.86	0.48	0.88	0.56	0.98	0.42	0.93	0.98
SENTIMENT	PHI3.5	TEMPSCALING	0.99	0.49	0.91	0.41	0.93	0.52	0.99	0.38	0.95	0.90
SENTIMENT	PHI3.5	APS	0.99	0.49	0.92	0.40	0.94	0.51	0.99	0.37	0.96	0.89
SENTIMENT	PHI3.5	RAPS	0.99	0.48	0.91	0.41	0.93	0.51	0.99	0.38	0.95	0.90
SENTIMENT	PHI3.5+ADAPTOR	SOFTMAX	0.99	0.42	1.00	0.42	1.00	0.42	0.99	0.42	0.99	0.84
SENTIMENT	PHI3.5+ADAPTOR	TEMPSCALING	0.99	0.42	1.00	0.41	1.00	0.42	0.99	0.41	0.99	0.83
SENTIMENT	PHI3.5+ADAPTOR	APS	0.98	0.45	0.98	0.45	0.98	0.45	0.98	0.45	0.98	0.90
SENTIMENT	PHI3.5+ADAPTOR	RAPS	0.98	0.45	0.98	0.44	0.98	0.45	0.98	0.44	0.98	0.89
SENTIMENT	PHI3.5+DISCVBLLMLP	VBLL	0.99	0.42	1.00	0.40	1.00	0.42	0.99	0.40	0.99	0.82
SENTIMENT	PHI3.5+SDM	SOFTMAX$(d \cdot \mathbf{z}')$	0.99	0.30	0.99	0.24	1.00	0.30	0.99	0.24	0.99	0.54
SENTIMENT	PHI3.5+SDM	SDM_α	0.99	0.43	0.99	0.38	0.99	0.43	0.99	0.38	0.99	0.81
SENTIMENT	PHI3.5+SDM	SDM_{HR}	1.00	0.37	0.99	0.30	0.99	0.38	1.00	0.30	0.99	0.68
SENTIMENT	MIXTRAL8X7B	SOFTMAX	0.98	0.50	0.88	0.50	0.89	0.55	0.98	0.45	0.93	1.00
SENTIMENT	MIXTRAL8X7B	TEMPSCALING	0.99	0.50	0.90	0.48	0.91	0.54	0.98	0.44	0.94	0.98
SENTIMENT	MIXTRAL8X7B	APS	0.98	0.49	0.91	0.47	0.92	0.52	0.98	0.44	0.95	0.96
SENTIMENT	MIXTRAL8X7B	RAPS	0.99	0.49	0.92	0.47	0.93	0.52	0.98	0.44	0.95	0.96
SENTIMENT	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.99	0.45	0.99	0.43	0.99	0.45	0.99	0.43	0.99	0.87
SENTIMENT	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.99	0.43	0.99	0.41	0.99	0.43	0.99	0.41	0.99	0.84
SENTIMENT	MIXTRAL8X7B+ADAPTOR	APS	0.99	0.46	0.98	0.45	0.98	0.46	0.99	0.44	0.99	0.91
SENTIMENT	MIXTRAL8X7B+ADAPTOR	RAPS	0.99	0.46	0.98	0.45	0.98	0.47	0.98	0.45	0.98	0.92
SENTIMENT	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.87
SENTIMENT	MIXTRAL8X7B+SDM	SDM_{HR}	0.99	0.41	0.98	0.33	0.99	0.41	0.98	0.33	0.99	0.74
SENTIMENTOOD	PHI3.5	SOFTMAX	1.00	0.50	0.54	0.46	0.70	0.71	0.99	0.25	0.78	0.96
SENTIMENTOOD	PHI3.5	TEMPSCALING	1.00	0.49	0.58	0.30	0.80	0.62	0.99	0.17	0.84	0.79
SENTIMENTOOD	PHI3.5	APS	1.00	0.49	0.59	0.28	0.81	0.60	0.99	0.17	0.85	0.77
SENTIMENTOOD	PHI3.5	RAPS	1.00	0.49	0.59	0.28	0.81	0.60	0.99	0.17	0.85	0.77
SENTIMENTOOD	PHI3.5+ADAPTOR	SOFTMAX	0.57	0.03	0.96	0.07	0.84	0.02	0.85	0.07	0.85	0.09
SENTIMENTOOD	PHI3.5+ADAPTOR	TEMPSCALING	0.60	0.02	0.97	0.05	0.86	0.01	0.87	0.06	0.87	0.07
SENTIMENTOOD	PHI3.5+ADAPTOR	APS	0.46	0.14	0.83	0.18	0.67	0.09	0.68	0.22	0.68	0.32
SENTIMENTOOD	PHI3.5+ADAPTOR	RAPS	0.48	0.13	0.82	0.18	0.66	0.10	0.68	0.22	0.68	0.32
SENTIMENTOOD	PHI3.5+DISCVBLLMLP	VBLL	0.89	0.03	0.96	0.04	0.93	0.02	0.94	0.05	0.94	0.07
SENTIMENTOOD	PHI3.5+SDM	SDM_{HR}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	0.01
SENTIMENTOOD	MIXTRAL8X7B	SOFTMAX	1.00	0.50	0.35	0.49	0.61	0.82	1.00	0.17	0.68	0.99
SENTIMENTOOD	MIXTRAL8X7B	TEMPSCALING	1.00	0.49	0.37	0.41	0.66	0.75	0.99	0.15	0.71	0.90
SENTIMENTOOD	MIXTRAL8X7B	APS	1.00	0.45	0.44	0.32	0.71	0.63	0.99	0.14	0.77	0.77
SENTIMENTOOD	MIXTRAL8X7B	RAPS	1.00	0.45	0.44	0.32	0.72	0.63	0.99	0.14	0.77	0.77
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.98	0.02	0.83	0.07	0.66	0.04	0.99	0.06	0.87	0.10
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.98	0.01	0.90	0.05	0.67	0.02	1.00	0.05	0.91	0.06
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	APS	0.94	0.14	0.63	0.18	0.67	0.20	0.93	0.12	0.77	0.32
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	RAPS	0.94	0.14	0.63	0.18	0.67	0.20	0.93	0.12	0.76	0.32
SENTIMENTOOD	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.94	0.01	0.97	0.11	0.79	0.02	0.99	0.10	0.96	0.12
SENTIMENTOOD	MIXTRAL8X7B+SDM	SDM_{HR}	0.9487	0.01	0.96	0.01	0.9487	0.01	0.96	0.01	0.95	0.02
FACTCHECK	PHI3.5	SOFTMAX	0.94	0.51	0.73	0.46	0.79	0.60	0.92	0.36	0.84	0.97
FACTCHECK	PHI3.5	TEMPSCALING	0.97	0.38	0.79	0.37	0.83	0.45	0.96	0.31	0.88	0.76
FACTCHECK	PHI3.5	APS	0.98	0.22	0.82	0.27	0.82	0.27	0.98	0.23	0.89	0.50
FACTCHECK	PHI3.5	RAPS	0.98	0.20	0.84	0.28	0.81	0.24	0.98	0.24	0.90	0.47
FACTCHECK	PHI3.5+ADAPTOR	SOFTMAX	0.40	0.08	0.99	0.33	0.89	0.04	0.87	0.37	0.87	0.41
FACTCHECK	PHI3.5+ADAPTOR	TEMPSCALING	0.38	0.07	0.99	0.29	0.86	0.03	0.88	0.33	0.88	0.36
FACTCHECK	PHI3.5+ADAPTOR	APS	0.26	0.14	0.99	0.38	0.90	0.04	0.78	0.48	0.79	0.52
FACTCHECK	PHI3.5+ADAPTOR	RAPS	0.36	0.18	0.98	0.35	0.89	0.07	0.74	0.46	0.76	0.53
FACTCHECK	PHI3.5+DISCVBLLMLP	VBLL	0.35	0.07	1.	0.31	1.	0.02	0.88	0.36	0.88	0.38
FACTCHECK	PHI3.5+SDM	SDM_{HR}	R	0.	1.	0.12	R	0.	1.	0.12	1.	0.12
FACTCHECK	MIXTRAL8X7B	SOFTMAX	0.98	0.51	0.48	0.49	0.66	0.76	0.95	0.24	0.73	1.
FACTCHECK	MIXTRAL8X7B	TEMPSCALING	0.99	0.50	0.46	0.43	0.68	0.73	0.98	0.20	0.75	0.93
FACTCHECK	MIXTRAL8X7B	APS	1.	0.18	0.80	0.16	0.84	0.21	1.	0.13	0.90	0.34
FACTCHECK	MIXTRAL8X7B	RAPS	1.	0.14	0.66	0.20	0.67	0.21	1.	0.13	0.80	0.35
FACTCHECK	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.68	0.11	0.97	0.31	0.90	0.09	0.89	0.34	0.89	0.42
FACTCHECK	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.70	0.09	0.97	0.29	0.89	0.07	0.91	0.31	0.91	0.39
FACTCHECK	MIXTRAL8X7B+ADAPTOR	APS	0.62	0.22	0.96	0.37	0.89	0.16	0.80	0.44	0.83	0.59
FACTCHECK	MIXTRAL8X7B+ADAPTOR	RAPS	0.65	0.22	0.96	0.35	0.92	0.16	0.81	0.41	0.84	0.57
FACTCHECK	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.85	0.08	0.97	0.27	0.89	0.08	0.95	0.27	0.94	0.35
FACTCHECK	MIXTRAL8X7B+SDM	SDM_{HR}	R	0.	R	0.	R	0.	R	0.	R	0.
FACTCHECK	MIXTRAL8X7B+SDM	$\text{SDM}_{\text{HR}}, \alpha = 0.94$	1.	0.03	0.95	0.16	0.80	0.04	1.	0.15	0.96	0.19

Table 1: Comparison of estimators for the sentiment and factcheck datasets, with $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents. The rows corresponding to the proposed SDM estimator (Eq. 9) are highlighted.

runs after the parameters of the adaptor layer have been fixed, so this indicator also in effect serves as a check on the optimization process (Eq. 7) itself.

Understanding SDM_{HR}. The relative proportion of points in the HIGH-RELIABILITY region reflects the data and model, and the uncertainty therein. A high abstention rate at a given α over in-distribution data suggests a need for more (or higher quality) data, a stronger model, and/or a reduction in α. For certain applications, it may be informative to construct estimators across a range of α values, assigning a point to the HIGH-RELIABILITY region of the most conservative (i.e., closer to 1) α in which it is a member.

Far out-of-distribution data. The Appendix (Tables 5 and 6) provides results for the shuffled variants. The selection criteria of Eq. 9 reliably rejects the challenging predictions, whereas the non-SDM-based estimators fare poorly, in general. In this way, the SDM estimator serves as an effective out-of-distribution detection method. With existing methods, defining an out-of-distribution point has been task-specific, and generally challenging over high-dimensional inputs, typically requiring additional modeling beyond that of the calibration or selection method. In contrast, Eq. 9 provides a principled approach for determining such cut-offs in a data- and model-driven manner, with mini-

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{ca}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{ca}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{ca}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{ca}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{ca}}\rvert}$
SENTIMENT $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	NO-REJECT	0.95	0.50	0.96	0.50	0.96	0.50	0.95	0.50	0.96	1.
SENTIMENT $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SOFTMAX	0.96	0.48	0.97	0.48	0.97	0.47	0.96	0.49	0.97	0.96
SENTIMENT $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	0.99	0.31	0.99	0.24	1.00	0.31	0.99	0.24	0.99	0.55
SENTIMENT $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SDM$_\alpha$	0.99	0.42	0.99	0.39	0.99	0.42	0.99	0.39	0.99	0.81
SENTIMENT $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SDM$_{\text{HR}}, \alpha = 0.95, q'_{\min} = 52.2$	0.99	0.38	0.99	0.32	0.99	0.38	0.99	0.31	0.99	0.69
SENTIMENT $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	NO-REJECT	0.96	0.50	0.96	0.50	0.96	0.50	0.96	0.50	0.96	1.
SENTIMENT $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SOFTMAX	0.96	0.49	0.97	0.49	0.97	0.49	0.96	0.49	0.97	0.98
SENTIMENT $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	1.00	0.43	0.99	0.35	0.99	0.43	1.00	0.34	0.99	0.78
SENTIMENT $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_\alpha$	0.99	0.47	0.98	0.43	0.98	0.47	0.98	0.43	0.98	0.90
SENTIMENT $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_{\text{HR}}, \alpha = 0.95, q'_{\min} = 63.0$	0.99	0.41	0.99	0.34	0.99	0.41	0.99	0.34	0.99	0.74
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	NO-REJECT	0.90	0.50	0.91	0.50	0.91	0.49	0.90	0.51	0.90	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SOFTMAX	0.94	0.32	0.96	0.41	0.95	0.31	0.96	0.41	0.96	0.72
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	0.98	0.08	1.00	0.07	1.00	0.08	0.98	0.07	0.99	0.15
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SDM$_\alpha$	0.98	0.33	0.99	0.27	0.99	0.33	0.98	0.27	0.98	0.60
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+SDM	SDM$_{\text{HR}}, \alpha = 0.95, q'_{\min} = 95.0$	1.00	0.19	0.99	0.12	1.00	0.19	0.99	0.12	1.00	0.31
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	NO-REJECT	0.90	0.50	0.92	0.50	0.92	0.49	0.90	0.51	0.91	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SOFTMAX	0.94	0.44	0.96	0.45	0.96	0.43	0.94	0.46	0.95	0.89
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	0.98	0.28	0.98	0.17	0.99	0.27	0.96	0.17	0.98	0.44
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_\alpha$	0.97	0.41	0.95	0.32	0.96	0.41	0.95	0.32	0.96	0.73
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_{\text{HR}}, \alpha = 0.95, q'_{\min} = \infty$	R	0.	R	0.	R	0.	R	0.	R	0.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_\alpha, \alpha = 0.94$	0.96	0.42	0.95	0.33	0.96	0.42	0.95	0.32	0.96	0.74
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+SDM	SDM$_{\text{HR}}, \alpha = 0.94, q'_{\min} = 134.0$	1.00	0.33	0.96	0.14	0.99	0.33	0.99	0.13	0.99	0.46

Table 2: Reference results over $\mathcal{D}_{\text{ca}}$ to illustrate the behavior of $q'_{\min}$. The value of $q'_{\min}$ tends to increase as the accuracy over $\mathcal{D}_{\text{ca}}$ decreases, reflecting a more conservative HIGH-RELIABILITY region. Alg. 1 failed to find a finite $q'_{\min}$ for MIXTRAL8X7B+SDM over the FACTCHECK calibration set at $\alpha = 0.95$, so for reference, we also show the HIGH-RELIABILITY region at $\alpha = 0.94$.

mal hyper-parameters, resulting in a separation of points over which the estimator tends to be reliable (namely, the admitted points) and those over which the estimates themselves tend to be unreliable.

6 Conclusion

We introduced SDM activation functions and SDM estimators, which are more robust and interpretable estimators of the predictive uncertainty than those based on the commonly used softmax function. In this way, SDM activations provide a principled, data-driven substrate for approaching selective classification, calibration, and out-of-distribution detection with language models.

Limitations

The SDM estimator requires the $\mathcal{D}_{\text{tr}}$ exemplar vectors at test-time, but the additional compute required for test-time matching is similar to that of commonly used dense retrieval mechanisms with LMs, so the additional overhead is achievable in typical use-cases.

For brevity of presentation, in the main text we omit consideration of the Beta-distributed error term that is a function of the effective sample size of split-conformal coverage (Vovk, 2012). In practice, this term is negligible in the present experiments given $|\mathcal{D}_{\text{ca}}|$ and the resolution of the comparisons. See Appendix A.5 for a further discussion of analyzing the effective sample size for both the class-conditional and prediction-conditional estimates.

Among points in the HIGH-RELIABILITY region, SDM estimators are relatively robust to covariate shifts that alter the relative proportion of points in the regions partitioned by the SIMILARITY, DISTANCE, and MAGNITUDE signals, as well as label shifts, the latter due to the class-wise thresholds of Alg. 1. For example, in the extreme case, the proportion of points for which $q = 0$ and/or $d = 0$ can be arbitrarily larger in the held-out $\mathcal{D}_{\text{te}}$ compared to that seen in training without altering the calibration of the HIGH-RELIABILITY region, ceteris paribus. Of course, it is not possible to maintain calibration over all possible distribution shifts.[7] However, in

[7] With label shifts, the marginal distribution over labels $P(Y)$ changes, while the conditional distribution $P(X \mid Y)$ remains unchanged. With covariate shifts, the marginal distribution over features $P(X)$ changes, while the conditional distribution $P(Y \mid X)$ remains unchanged. Label shifts are relatively straightforward to model in the selective classification setting via class-wise CDFs. Absent task-specific information (e.g., samples from the new distribution), controlling for covariate shifts is more complex and requires a partitioning of the high-dimensional space, as with SDM estimators. "Concept

contrast, the examined non-SDM-based calibration methods work quite poorly in maintaining calibration in high-probability regions in the presence of even modest distribution shifts, since they marginalize over those partitions, and they also marginalize over the labels. That limits the applicability of such calibration methods in practice. Importantly, due to the dense matching and partitioning of the calibration set, SDM estimators provide interpretability-by-exemplar: A user can examine the applicable documents in training and those in the applicable partition of the calibration set for further analysis and labeling, as needed.

For the experiments in the present work (§ 4.3), the input $\boldsymbol{h}$ to the adaptor layers is a mean-pool over the final-layer hidden states concatenated with the final-layer hidden state of the final sequence position. This achieves the desired behavior observed in the experiments, while enabling relatively efficient training since $\boldsymbol{h}$ can be calculated once and cached, with reasonable space requirements, for use across all epochs and J iterations of training. In this case, the notion of convolving over sequence positions is not used, since the kernel/filter-width of the CNN adaptor equals D, the dimension of $\boldsymbol{h}$. We hypothesize that a learned max-pooling over sequence positions as used in Schmaltz (2021) may be more sample efficient (i.e., higher proportions of points in the HIGH-RELIABILITY region, ceteris paribus) than mean-pooling, since the filters learn the aggregation rather than the marginalization over positions with mean-pooling. That would also enable the local-level feature decompositions examined in previous works. However, in that case, naive pre-caching of the frozen hidden states for each token position becomes space prohibitive for long sequences. A practical middle-ground may be to down-sample the position-wise hidden states via summary statistics at the token-level; at semantic boundaries (sentence, paragraph, etc.); and/or at conversation boundaries (prompt, answer, etc.), and to concatenate the resulting vectors with the $\boldsymbol{h}$ used in the present experiments. We leave to future higher-resourced experiments to examine the significance of those tradeoffs.

Contemporary commercial LMs served from cloud APIs may not expose the hidden states of the underlying network. In practice, the output from such models can be cross-encoded with an open-weight model, the composition of which then becomes the model over which to calibrate and by extension, the input to the SDM estimator. The present work examines the SDM-based calibration behavior with open-weight models to facilitate replication in controlled settings; we leave applied examples with commercial black-box APIs to future work.

shifts", where the conditional distribution $P(Y \mid X)$ changes, require information exogenous to the SDM estimator to model.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

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

C. K. Chow. 1957. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

A. P. Dawid. 1982. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Luc Devroye, László Györfi, and Gábor Lugosi. 1996. A Probabilistic Theory of Pattern Recognition. In *Stochastic Modelling and Applied Probability*.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. 1956. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. 2020. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

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

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

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

Chirag Gupta and Aaditya Ramdas. 2022. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*.

James Harrison, John Willes, and Jasper Snoek. 2024. Variational bayesian last layers. In *The Twelfth International Conference on Learning Representations*.

J. T. Gene Hwang and A. Adam Ding. 1997. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438):748–757.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Jing Lei and Larry Wasserman. 2014. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.

Ilya Loshchilov and Frank Hutter. 2019. 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.

P. Massart. 1990. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283.

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

John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Allen Schmaltz. 2021. Detecting local insights from global labels: Supervised and zero-shot sequence labeling via a convolutional decomposition. *Computational Linguistics*, 47(4):729–773.

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

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR.

L. G. Valiant. 1984. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vladimir Vovk. 2012. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore. PMLR.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Appendix

The prompts for the tasks appear in Appendix A.1. Expanded versions of the tables in the main text appear in Appendix A.2. We provide the results for the far out-of-distribution (OOD) shuffled datasets in Appendix A.3. Additional implementation details are included in Appendix A.4. Appendix A.5 provides an approach for analyzing the effective sample size for both the prediction-conditional and class-conditional estimates. Appendix A.6 discusses local updatability of the SDM activation function, highlighting the connection to instance-based metric learners. Appendix A.7 proposes an approach for constructing an ensemble over all J models created during training. Appendix A.8 provides additional details and comparisons for the experiments with variational Bayesian last-layer estimators (Harrison et al., 2024). We close with Appendix A.9, which briefly describes additional experiments appearing in the code repository.

A.1 Prompts

Sentiment. For the sentiment datasets, we prompt the LMs for a binary classification (`Yes` or `No`) as follows:

```
Here is a movie review. <review>
DOCUMENT </review> Is the
sentiment of the movie review
positive? Answer Yes if the
sentiment is positive. Answer
No if the sentiment is negative.
Start your response with Yes or
No.
```

We replace `DOCUMENT` with the corresponding text for each instance.

Factcheck. Similarly, for the factcheck datasets, we prompt the LMs for a binary classification (`Yes` or `No`) as follows:

```
Here is a statement that may
contain errors. <statement>
DOCUMENT </statement> Is the
statement true? Answer Yes if
the statement is true. Answer No
if the statement is false. Start
your response with Yes or No.
```

As above, we replace `DOCUMENT` with the corresponding text for each instance.

A.2 Additional Rows for Table 1

Additional rows for Table 1 appear in Table 3 for the sentiment datasets and Table 4 for the factcheck datasets.

A.3 Far OOD Shuffled Datasets

As discussed in Section 5 in the main text, Table 5 shows results for the SENTIMENTSHUFFLED and SENTIMENTOODSHUFFLED datasets, and Table 6 shows results for the FACTCHECKSHUFFLED datasets.

In the case of SENTIMENTSHUFFLED and SENTIMENTOODSHUFFLED, the semantics of the original labels are maintained. This requires the models and estimators to attempt a sentiment classification over the bag-of-words input, or reject the classification. This represents the setting where an LM is given far out-of-distribution input, and additionally provides a control on test-set contamination of the underlying LMs, which due to the shuffling, are relatively unlikely to have seen all of the long, contiguous n-gram sequences from these documents in training or fine-tuning.

In the case of FACTCHECKSHUFFLED, since the task prompt seeks to classify errors, we set the ground-truth labels of the shuffled counterparts to $y = 0$.

A.4 Additional Implementation Details

Replication code is available at the following URL: https://github.com/ReexpressAI/sdm_activations

We mean center the input to g, the 1-D CNN of the SDM activation layer and the otherwise identical CNN adaptors of the baseline comparison estimators, via the mean and standard deviation over $\mathcal{D}_{\text{tr}}$. In all experiments with adaptor layers, $M = 1000$ and we use a mini-batch size of 50. We use the Adam optimizer (Kingma and Ba, 2015) (without weight decay) with a learning rate of 1×10^{-5} for training.

A.4.1 Implementation of the SDM Activation Function

The SDM activation function can be calculated using existing numerically stable softmax implemenations via the following relation:

$$\begin{aligned}\text{SDM}(\boldsymbol{z}')_i &= \frac{(2+q)^{d \cdot z'_i}}{\sum_{c=1}^{C} (2+q)^{d \cdot z'_c}} \\ &= \frac{e^{\log_e(2+q) \cdot d \cdot z'_i}}{\sum_{c=1}^{C} e^{\log_e(2+q) \cdot d \cdot z'_c}}, 1 \leq i \leq C\end{aligned} \tag{10}$$

The change of base for the loss can be calculated by multiplying the $\log_e$ probabilities from standard optimized LogSoftmax operations by $\frac{1}{\log_e(2+q)}$.

A.4.2 Implementation of the Empirical CDF Function

In the present experiments, the empirical CDF functions are implemented such that the distance quantiles are exclusionary at the boundaries. When $d_{\text{nearest}} = 0$, the $1 - \text{eCDF}_{\text{ca}}(d_{\text{nearest}})$ quantile is 1, and when d_{nearest} is greater than the maximum observed distance (across $\mathcal{D}_{\text{ca}}$ for $\boldsymbol{x} \in \mathcal{D}_{\text{te}}$ and $\boldsymbol{x} \in \mathcal{D}_{\text{ca}}$, and across $\mathcal{D}_{\text{tr}}$ for $\boldsymbol{x} \in \mathcal{D}_{\text{tr}}$, the latter case only occurring during training), the $1 - \text{eCDF}_{\text{ca}}(d_{\text{nearest}})$ quantile is 0.

A.4.3 Training Dynamics

As indicated by our loss notation (Eq. 6) and described in § 3.1.2 and § 3.1.3, instance-wise q and d are not parameters learned directly via gradient descent. By design, we are not back-propagating through a bi- and/or cross-encoded search graph. The KNN approximations of Schmaltz (2021, §3.7.1) are trained with an iterative loss-masking approach to limit overfitting to outliers; with the SDM activation, instance-wise q and d serve as the regularizers during learning.

A.5 Analyzing the Effective Sample Size

In the context of the SDM estimator, to parameterize the prior belief that data points with a looser connection to $\mathcal{D}_{\text{tr}}$ reflect smaller effective sample sizes, while also explicitly accounting for the count of observed points in $\mathcal{D}_{\text{ca}}$, the effective sample size for each test instance can be estimated with the following conservative assumption:

Assumption A.1. The effective sample size is increasing in q', class-wise over $\mathcal{D}_{\text{ca}}$.

For each $\boldsymbol{x} \in \mathcal{D}_{\text{te}}$, using q', we calculate $\hat{\mathbf{n}}$, the vector of effective sample sizes across classes, relative to $\mathcal{D}_{\text{ca}}$, as:

$$\begin{aligned}\hat{\mathbf{n}} = &[|\mathcal{D}_{\text{ca}}|^{y_1} \cdot \text{eCDF}_{\text{ca}}^{y_1}(q'), \ldots, \\ &|\mathcal{D}_{\text{ca}}|^{y_C} \cdot \text{eCDF}_{\text{ca}}^{y_C}(q')]\end{aligned} \tag{11}$$

where $|\mathcal{D}_{\text{ca}}|^{y_c}$ is the count of calibration set points with true label $y = c$.

The estimate of the effective sample size for each label can then be used to estimate the Beta-distributed error term of split-conformal coverage (Vovk, 2012), providing a sample-size-based error estimate for the class-conditional estimate, assuming exchangeability.

For the prediction-conditional estimates, assuming independent and identically distributed (i.i.d.) data, these sample size estimates can be used to construct a band around the empirical CDFs over d_{nearest} (Eq. 5) using the sharp constant (Massart, 1990) of the distribution-free DKW inequality (Dvoretzky et al., 1956), with $\hat{n}_{\text{min}}$ taken as the minimum among the estimated sample sizes across classes for the test instance:

$$\epsilon = \sqrt{\frac{1}{2 \cdot \hat{n}_{\text{min}}} \log_e \left(\frac{2}{1-\alpha} \right)}, \tag{12}$$

$$\hat{n}_{\text{min}} = \min [\hat{n}_1, \ldots, \hat{n}_C] \tag{13}$$

If $\hat{n}_{\text{min}} = 0$, our convention is to set $\epsilon = 1$. We then construct the conservative lower and upper counterparts to the distance quantile of Eq. 5:

$$d_{\text{lower}} = \max(d - \epsilon, 0) \tag{14}$$

$$d_{\text{upper}} = \min(d + \epsilon, 1) \tag{15}$$

Eq. 6 can then be calculated by substituting d_{lower} and d_{upper}, in turn, for d, resulting in a band around the prediction-conditional estimate, $\text{SDM}(\boldsymbol{z}')$:

$$\begin{aligned}\text{SDM}(\boldsymbol{z}')_i^{\text{lower}} = \\ \frac{(2+q)^{d_{\text{lower}} \cdot z'_i}}{\sum_{c=1}^{C} (2+q)^{d_{\text{lower}} \cdot z'_c}}, 1 \leq i \leq C\end{aligned} \tag{16}$$

$$\text{SDM}(\boldsymbol{z}')_i^{\text{upper}} = \frac{(2+q)^{d_{\text{upper}} \cdot z'_i}}{\sum_{c=1}^{C} (2+q)^{d_{\text{upper}} \cdot z'_c}}, 1 \leq i \leq C \tag{17}$$

A corresponding $\text{SDM}_{\text{HR}}^{\text{lower}}$ estimator then follows:

$$q'_{\text{lower}} = \min\left(q, (2+q)^{\text{SDM}(\boldsymbol{z}')_{\hat{y}}^{\text{lower}}}\right) \tag{18}$$

$$\text{SDM}_{\text{HR}}^{\text{lower}} := \begin{cases} \hat{y} & \text{if } q'_{\text{lower}} \geq q'_{\text{min}} \wedge \text{SDM}(\boldsymbol{z}')_{\hat{y}}^{\text{lower}} \geq \psi_{\hat{y}} \\ \perp & \text{otherwise} \end{cases} \tag{19}$$

In the experiments, we focus on the SDM_{HR} estimator to simplify the presentation, and since a well-calibrated SDM_{HR} estimator implies a well-calibrated $\text{SDM}_{\text{HR}}^{\text{lower}}$ estimator. For applications, we recommend taking into account the effective sample size at test-time via d_{lower}, $\text{SDM}(\boldsymbol{z}')^{\text{lower}}$, q'_{lower}, and $\text{SDM}_{\text{HR}}^{\text{lower}}$. The publicly available code (Appendix A.4) calculates these additional values.

A.6 Updatability

The SDM activation function inherits the *updatability* property of instance-based metric learners. Instances with labels $y \in \mathcal{Y} = \{1, \ldots, C\}$ can be dynamically added to $\mathcal{D}_{\text{tr}}$ after training. This can change the SIMILARITY and DISTANCE values, and by extension the uncertainty estimates, while the MAGNITUDE, and by extension the $\arg\max$ prediction, remains unchanged provided the weights of the CNN adaptor are held fixed. We hypothesize this can be a useful tradeoff between fast moving weights and slow moving weights in continual learning settings.

Relatedly, instances with labels $y \notin \mathcal{Y}$ can also be added to $\mathcal{D}_{\text{tr}}$ after training. Given Eq. 4, test instances matching to such instances will have reduced q values, ceteris paribus, since the model can never predict such labels. This is potentially a useful, lightweight alternative to adding an explicit "out-of-distribution class" to $\mathcal{Y}$, a comparison to which we leave to future work.

A.7 Estimator Ensembles

At the cost of additional compute, we can control for uncertainty across shuffles of $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$ and parameter initializations by constructing an ensemble across all J models saved during training. The publicly available code (Appendix A.4) provides an option for constructing an ensembled $\text{SDM}_{\text{HR}}^{\text{lower}}$ estimator by only admitting a point if $\hat{y}$ is identical across all J models *and* the point falls into the $\text{SDM}_{\text{HR}}^{\text{lower}}$ region for all J models.

A.8 Comparison to Bayesian Last-layer Networks

In this section, we provide additional details for the comparisons to variational Bayesian last-layer neural networks (VBLL), a computationally efficient Bayesian approach (Harrison et al., 2024). The basic setup is similar to the Frequentist and empirically motivated approaches using CNN-based adaptors examined in the main text in that it involves training a small final-layer adaptor over the frozen parameters of the language model. However, in this case, the adaptor network is a multi-layer perceptron (MLP) combined with the VBLL estimator. We consider both the discriminative and generative VBLL estimators.

Models. We follow the parameter choices and architecture of Harrison et al. (2024), and its associated code tutorial[8], with applicable changes to match the experimental settings of the other models and estimators. Specifically, the VBLL models consist of an input linear layer, 2 core linear layers, and a final output linear layer. The hidden dimension is set at 795, which yields a similar number of parameters as the SDM activation layers (approximately 6 million parameters for the `Phi-3.5-mini-instruct` adaptors and approximately 8 million parameters for the `Mixtral 8x7B Instruct v0.1` adaptors). The input to the VBLL models is the same mean-centered embeddings used with the SDM activation layers and the baseline CNN adaptor layers of the main text. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 1×10^{-4}, which following the code tutorial, is not applied to the final layer. We use a learning rate of 1×10^{-5} for training, matching that used with the SDM activation layers. Following the code tutorial, we use gradient clipping with a max norm of 1, and we use the exponential linear unit (ELU) as the activation function (Clevert et al., 2016). We train for 200 epochs, choosing the epoch with the lowest held-out validation loss over $\mathcal{D}_{\text{ca}}$ as the chosen weights. We repeat this process for $J = 10$ shuffles

[8] Available at https://github.com/VectorInstitute/vbll

of the data (i.e., splits of $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$), choosing the model with the globally lowest overall held-out validation loss over $\mathcal{D}_{\text{ca}}$ as the final model.

For the discriminative models over `Phi-3.5-mini-instruct` and `Mixtral 8x7B Instruct v0.1`, we use the labels PHI3.5+DISCVBLLMLP and MIXTRAL8X7B+DISCVBLLMLP, respectively. Those are the models appearing in Table 1 in the main text. For the generative models over `Phi-3.5-mini-instruct` and `Mixtral 8x7B Instruct v0.1`, we use the labels PHI3.5+GENVBLLMLP and MIXTRAL8X7B+GENVBLLMLP, respectively. For all of the aforementioned models, we use a KL regularization weight set at $\frac{1}{|\mathcal{D}_{\text{tr}}|}$, as in the original paper. We also consider analogous models that increase the KL regularization weight by a multiplicative factor of 50. Increasing the KL regularization weight is suggested in the code tutorial as the "simplest and most effective method to control the scale of uncertainty" of VBLL models. For these models with the larger KL regularization weight, we use the labels PHI3.5+DISCVBLLMLP$_{\text{rw50}}$, MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$, PHI3.5+GENVBLLMLP$_{\text{rw50}}$, and MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$, respectively.

We use the label VBLL for the estimator that thresholds the output of the variational Bayesian last-layer neural network at α for the predicted class, analogous to the softmax and SDM$_\alpha$ estimators. The NO-REJECT estimator provides a reference point without any selection criteria applied (i.e., the standard marginal and class- and prediction-conditional accuracies over the given dataset).

Results. The results for the sentiment datasets appear in Table 7, and the results for the factcheck datasets appear in Table 8. For reference, in all cases we also provide the results over the held-out calibration set, $\mathcal{D}_{\text{ca}}$, which is the held-out split used to determine the final model weights, as noted above.

Comparing to Table 2, the accuracies over $\mathcal{D}_{\text{ca}}$ for the NO-REJECT estimators are similar for the VBLL models and the models using SDM activation functions. The MLPs of the VBLL models and the 1-D CNNs of the SDM activation layers have a similar number of parameters and are trained with the same maximum number of epochs and the same number of iterated shuffles of $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{ca}}$. The accuracies without selection for the SENTIMENT calibration set are already at least α, and those for the FACTCHECK calibration set are below α, but at least $\alpha - 0.05$. As such, differences in calibration effectiveness of the respective methods are not directly attributable to substantively different baseline accuracies for the in-distribution tasks.

We find that the VBLL estimators are well-calibrated in high-probability regions over in-distribution data, but generally fare poorly over covariate shifts and out-of-distribution data. We find no clear advantages nor disadvantages for the discriminative vs. generative variants, and modifying the KL regularization weight has a minimal impact, at least at this scale. In Harrison et al. (2024), VBLL estimators over LMs for sentiment classification are only compared to an MLP baseline on in-distribution data; our evaluation setting is significantly more challenging and closer to real-world conditions encountered with LM applications.

A.9 Additional Experiments

The publicly available code repository (Appendix A.4) includes additional experiments to further examine the behavior of SDM activations and estimators. In the interest of length since these experiments are secondary to the main experiments, we defer a full discussion to the code repository.

First, we demonstrate that the behavior of the SDM estimator is not restricted to binary classification. This is shown by examining the training dynamics, calibration results, and interpretability-by-exemplar behavior on the standard AGNews (4-class classification) dataset from Zhang et al. (2015).

The SDM estimator is intended to be robust to the optimization process that determines the parameters of the adaptor of the SDM activation. We examine this with an ablation using the hidden states of the underlying model as the representations used for matching, instead of those of a learned CNN. The result is the SDM estimator remains well-calibrated at α, but with the additional cost of L^2 matching with a higher-dimensional $\boldsymbol{h}'$, and reduced statistical efficiency, as reflected in fewer points in the HIGH-RELIABILITY region, on most of the examined datasets, ceteris paribus. In this way, we demonstrate that Alg. 1 is robust to representations held constant during the optimization of Eq. 7.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
SENTIMENT	PHI3.5	NO-REJECT	0.98	0.50	0.85	0.50	0.86	0.57	0.98	0.43	0.91	1.
SENTIMENT	PHI3.5	SOFTMAX	0.98	0.50	0.86	0.48	0.88	0.56	0.98	0.42	0.93	0.98
SENTIMENT	PHI3.5	TEMPSCALING	0.99	0.49	0.91	0.41	0.93	0.52	0.99	0.38	0.95	0.90
SENTIMENT	PHI3.5	APS	0.99	0.49	0.92	0.40	0.94	0.51	0.99	0.37	0.96	0.89
SENTIMENT	PHI3.5	RAPS	0.99	0.48	0.91	0.41	0.93	0.51	0.99	0.38	0.95	0.90
SENTIMENT	PHI3.5+ADAPTOR	NO-REJECT	0.97	0.50	0.95	0.50	0.96	0.51	0.97	0.49	0.96	1.
SENTIMENT	PHI3.5+ADAPTOR	SOFTMAX	0.99	0.42	1.00	0.42	1.00	0.42	0.99	0.42	0.99	0.84
SENTIMENT	PHI3.5+ADAPTOR	TEMPSCALING	0.99	0.42	1.00	0.41	1.00	0.42	0.99	0.41	0.99	0.83
SENTIMENT	PHI3.5+ADAPTOR	APS	0.98	0.45	0.98	0.45	0.98	0.45	0.98	0.45	0.98	0.90
SENTIMENT	PHI3.5+ADAPTOR	RAPS	0.98	0.45	0.98	0.44	0.98	0.45	0.98	0.44	0.98	0.89
SENTIMENT	PHI3.5+SDM	NO-REJECT	0.96	0.50	0.96	0.50	0.96	0.50	0.96	0.50	0.96	1.
SENTIMENT	PHI3.5+SDM	SOFTMAX	0.97	0.48	0.97	0.48	0.97	0.48	0.97	0.48	0.97	0.96
SENTIMENT	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	0.99	0.30	0.99	0.24	1.00	0.30	0.99	0.24	0.99	0.54
SENTIMENT	PHI3.5+SDM	SDM_{α}	0.99	0.43	0.99	0.38	0.99	0.43	0.99	0.38	0.99	0.81
SENTIMENT	PHI3.5+SDM	SDM_{HR}	1.00	0.37	0.99	0.30	0.99	0.38	1.00	0.30	0.99	0.68
SENTIMENT	MIXTRAL8X7B	NO-REJECT	0.98	0.50	0.88	0.50	0.89	0.55	0.98	0.45	0.93	1.
SENTIMENT	MIXTRAL8X7B	SOFTMAX	0.98	0.50	0.88	0.50	0.89	0.55	0.98	0.45	0.93	1.00
SENTIMENT	MIXTRAL8X7B	TEMPSCALING	0.99	0.50	0.90	0.48	0.91	0.54	0.98	0.44	0.94	0.98
SENTIMENT	MIXTRAL8X7B	APS	0.98	0.49	0.91	0.47	0.92	0.52	0.98	0.44	0.95	0.96
SENTIMENT	MIXTRAL8X7B	RAPS	0.99	0.49	0.92	0.47	0.93	0.52	0.98	0.44	0.95	0.96
SENTIMENT	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.97	0.50	0.96	0.50	0.96	0.51	0.97	0.49	0.97	1.
SENTIMENT	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.99	0.45	0.99	0.43	0.99	0.45	0.99	0.43	0.99	0.87
SENTIMENT	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.99	0.43	0.99	0.41	0.99	0.43	0.99	0.41	0.99	0.84
SENTIMENT	MIXTRAL8X7B+ADAPTOR	APS	0.99	0.46	0.98	0.45	0.98	0.46	0.99	0.44	0.99	0.91
SENTIMENT	MIXTRAL8X7B+ADAPTOR	RAPS	0.99	0.46	0.98	0.45	0.98	0.47	0.98	0.45	0.98	0.92
SENTIMENT	MIXTRAL8X7B+SDM	NO-REJECT	0.96	0.50	0.95	0.50	0.95	0.51	0.96	0.49	0.96	1.
SENTIMENT	MIXTRAL8X7B+SDM	SOFTMAX	0.97	0.49	0.96	0.49	0.96	0.50	0.97	0.49	0.97	0.98
SENTIMENT	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	0.99	0.43	0.99	0.33	0.99	0.43	0.99	0.33	0.99	0.77
SENTIMENT	MIXTRAL8X7B+SDM	SDM_{α}	0.98	0.48	0.98	0.43	0.98	0.47	0.98	0.43	0.98	0.90
SENTIMENT	MIXTRAL8X7B+SDM	SDM_{HR}	0.99	0.41	0.98	0.33	0.99	0.41	0.98	0.33	0.99	0.74
SENTIMENTOOD	PHI3.5	NO-REJECT	1.00	0.50	0.53	0.50	0.68	0.73	0.99	0.27	0.76	1.
SENTIMENTOOD	PHI3.5	SOFTMAX	1.00	0.50	0.54	0.46	0.70	0.71	0.99	0.25	0.78	0.96
SENTIMENTOOD	PHI3.5	TEMPSCALING	1.00	0.49	0.58	0.30	0.80	0.62	0.99	0.17	0.84	0.79
SENTIMENTOOD	PHI3.5	APS	1.00	0.49	0.59	0.28	0.81	0.60	0.99	0.17	0.85	0.77
SENTIMENTOOD	PHI3.5	RAPS	1.00	0.49	0.59	0.28	0.81	0.60	0.99	0.17	0.85	0.77
SENTIMENTOOD	PHI3.5+ADAPTOR	NO-REJECT	0.47	0.50	0.70	0.50	0.61	0.38	0.57	0.62	0.59	1.
SENTIMENTOOD	PHI3.5+ADAPTOR	SOFTMAX	0.57	0.03	0.96	0.07	0.84	0.02	0.85	0.07	0.85	0.09
SENTIMENTOOD	PHI3.5+ADAPTOR	TEMPSCALING	0.60	0.02	0.97	0.05	0.86	0.01	0.87	0.06	0.87	0.07
SENTIMENTOOD	PHI3.5+ADAPTOR	APS	0.46	0.14	0.83	0.18	0.67	0.09	0.68	0.22	0.68	0.32
SENTIMENTOOD	PHI3.5+ADAPTOR	RAPS	0.48	0.13	0.82	0.18	0.66	0.10	0.68	0.22	0.68	0.32
SENTIMENTOOD	PHI3.5+SDM	NO-REJECT	0.92	0.50	0.84	0.50	0.85	0.54	0.91	0.46	0.88	1.
SENTIMENTOOD	PHI3.5+SDM	SOFTMAX	0.96	0.42	0.87	0.45	0.87	0.46	0.96	0.41	0.91	0.87
SENTIMENTOOD	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01
SENTIMENTOOD	PHI3.5+SDM	SDM_{α}	1.	0.01	0.98	0.01	0.98	0.01	1.	0.01	0.99	0.02
SENTIMENTOOD	PHI3.5+SDM	SDM_{HR}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	0.01
SENTIMENTOOD	MIXTRAL8X7B	NO-REJECT	1.00	0.50	0.35	0.50	0.61	0.82	1.00	0.18	0.67	1.
SENTIMENTOOD	MIXTRAL8X7B	SOFTMAX	1.00	0.50	0.35	0.49	0.61	0.82	1.00	0.17	0.68	0.99
SENTIMENTOOD	MIXTRAL8X7B	TEMPSCALING	1.00	0.49	0.37	0.41	0.66	0.75	0.99	0.15	0.71	0.90
SENTIMENTOOD	MIXTRAL8X7B	APS	1.00	0.45	0.44	0.32	0.71	0.63	0.99	0.14	0.77	0.77
SENTIMENTOOD	MIXTRAL8X7B	RAPS	1.00	0.45	0.44	0.32	0.72	0.63	0.99	0.14	0.77	0.77
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.88	0.50	0.51	0.50	0.64	0.69	0.82	0.31	0.70	1.
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.98	0.02	0.83	0.07	0.66	0.04	0.99	0.06	0.87	0.10
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.98	0.01	0.90	0.05	0.67	0.02	1.00	0.05	0.91	0.06
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	APS	0.94	0.14	0.63	0.18	0.67	0.20	0.93	0.12	0.77	0.32
SENTIMENTOOD	MIXTRAL8X7B+ADAPTOR	RAPS	0.94	0.14	0.63	0.18	0.67	0.20	0.93	0.12	0.76	0.32
SENTIMENTOOD	MIXTRAL8X7B+SDM	NO-REJECT	0.71	0.50	0.83	0.50	0.81	0.44	0.74	0.56	0.77	1.
SENTIMENTOOD	MIXTRAL8X7B+SDM	SOFTMAX	0.74	0.43	0.86	0.47	0.83	0.39	0.78	0.52	0.80	0.91
SENTIMENTOOD	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	1.	<0.01	0.98	0.02	0.78	<0.01	1.	0.02	0.98	0.02
SENTIMENTOOD	MIXTRAL8X7B+SDM	SDM_{α}	0.98	0.05	0.96	0.04	0.97	0.05	0.98	0.04	0.97	0.08
SENTIMENTOOD	MIXTRAL8X7B+SDM	SDM_{HR}	0.9487	0.01	0.96	0.01	0.9487	0.01	0.96	0.01	0.95	0.02

Table 3: Comparison of estimators for the sentiment datasets, with $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
FACTCHECK	PHI3.5	NO-REJECT	0.94	0.51	0.71	0.49	0.78	0.62	0.92	0.38	0.83	1.
FACTCHECK	PHI3.5	SOFTMAX	0.94	0.51	0.73	0.46	0.79	0.60	0.92	0.36	0.84	0.97
FACTCHECK	PHI3.5	TEMPSCALING	0.97	0.38	0.79	0.37	0.83	0.45	0.96	0.31	0.88	0.76
FACTCHECK	PHI3.5	APS	0.98	0.22	0.82	0.27	0.82	0.27	0.98	0.23	0.89	0.50
FACTCHECK	PHI3.5	RAPS	0.98	0.20	0.84	0.28	0.81	0.24	0.98	0.24	0.90	0.47
FACTCHECK	PHI3.5+ADAPTOR	NO-REJECT	0.33	0.51	0.94	0.49	0.85	0.20	0.57	0.80	0.62	1.
FACTCHECK	PHI3.5+ADAPTOR	SOFTMAX	0.40	0.08	0.99	0.33	0.89	0.04	0.87	0.37	0.87	0.41
FACTCHECK	PHI3.5+ADAPTOR	TEMPSCALING	0.38	0.07	0.99	0.29	0.86	0.03	0.88	0.33	0.88	0.36
FACTCHECK	PHI3.5+ADAPTOR	APS	0.26	0.14	0.99	0.38	0.90	0.04	0.78	0.48	0.79	0.52
FACTCHECK	PHI3.5+ADAPTOR	RAPS	0.36	0.18	0.98	0.35	0.89	0.07	0.74	0.46	0.76	0.53
FACTCHECK	PHI3.5+SDM	NO-REJECT	0.70	0.51	0.88	0.49	0.86	0.42	0.73	0.58	0.79	1.
FACTCHECK	PHI3.5+SDM	SOFTMAX	0.75	0.27	0.94	0.39	0.89	0.22	0. 85	0.43	0.86	0.65
FACTCHECK	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	1.	0.03	R	0.	1.	0.03	1.	0.03
FACTCHECK	PHI3.5+SDM	SDM_{α}	1.	0.01	0.97	0.14	0.75	0.02	1.	0.14	0.97	0.16
FACTCHECK	PHI3.5+SDM	SDM_{HR}	R	0.	1.	0.12	R	0.	1.	0.12	1.	0.12
FACTCHECK	MIXTRAL8X7B	NO-REJECT	0.98	0.51	0.48	0.49	0.66	0.76	0.95	0.24	0.73	1.
FACTCHECK	MIXTRAL8X7B	SOFTMAX	0.98	0.51	0.48	0.49	0.66	0.76	0.95	0.24	0.73	1.
FACTCHECK	MIXTRAL8X7B	TEMPSCALING	0.99	0.50	0.46	0.43	0.68	0.73	0.98	0.20	0.75	0.93
FACTCHECK	MIXTRAL8X7B	APS	1.	0.18	0.80	0.16	0.84	0.21	1.	0.13	0.90	0.34
FACTCHECK	MIXTRAL8X7B	RAPS	1.	0.14	0.66	0.20	0.67	0.21	1.	0.13	0.80	0.35
FACTCHECK	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.56	0.51	0.87	0.49	0.82	0.36	0.65	0.64	0.71	1.
FACTCHECK	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.68	0.11	0.97	0.31	0.90	0.09	0.89	0.34	0.89	0.42
FACTCHECK	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.70	0.09	0.97	0.29	0.89	0.07	0.91	0.31	0.91	0.39
FACTCHECK	MIXTRAL8X7B+ADAPTOR	APS	0.62	0.22	0.96	0.37	0.89	0.16	0.80	0.44	0.83	0.59
FACTCHECK	MIXTRAL8X7B+ADAPTOR	RAPS	0.65	0.22	0.96	0.35	0.92	0.16	0.81	0.41	0.84	0.57
FACTCHECK	MIXTRAL8X7B+SDM	NO-REJECT	0.63	0.51	0.90	0.49	0.87	0.38	0.70	0.62	0.76	1.
FACTCHECK	MIXTRAL8X7B+SDM	SOFTMAX	0.67	0.34	0.96	0.40	0.93	0.24	0.78	0.50	0.83	0.74
FACTCHECK	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	0.80	0.04	0.	0.01	1.	0.03	0.80	0.04
FACTCHECK	MIXTRAL8X7B+SDM	SDM_{α}	0.88	0.10	0.95	0.18	0.91	0.09	0.93	0.18	0.93	0.27
FACTCHECK	MIXTRAL8X7B+SDM	SDM_{HR}	R	0.	R	0.	R	0.	R	0.	R	0.
FACTCHECK	MIXTRAL8X7B+SDM	$\text{SDM}_{\alpha}, \alpha = 0.94$	0.85	0.11	0.95	0.18	0.92	0.10	0.91	0.19	0.91	0.29
FACTCHECK	MIXTRAL8X7B+SDM	$\text{SDM}_{\text{HR}}, \alpha = 0.94$	1.	0.03	0.95	0.16	0.80	0.04	1.	0.15	0.96	0.19

Table 4: Comparison of estimators for the factcheck datasets. Unless specified otherwise, $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
SENTIMENTSHUFFLED	PHI3.5	NO-REJECT	1.00	0.50	0.18	0.50	0.55	0.91	0.98	0.09	0.59	1.
SENTIMENTSHUFFLED	PHI3.5	SOFTMAX	1.00	0.50	0.16	0.45	0.57	0.88	0.99	0.07	0.60	0.95
SENTIMENTSHUFFLED	PHI3.5	TEMPSCALING	1.	0.44	0.12	0.18	0.74	0.60	1.	0.02	0.75	0.62
SENTIMENTSHUFFLED	PHI3.5	APS	1.00	0.42	0.14	0.15	0.76	0.55	0.97	0.02	0.77	0.57
SENTIMENTSHUFFLED	PHI3.5	RAPS	1.	0.41	0.13	0.16	0.75	0.55	1.	0.02	0.76	0.57
SENTIMENTSHUFFLED	PHI3.5+ADAPTOR	NO-REJECT	0.83	0.50	0.81	0.50	0.81	0.51	0.82	0.49	0.82	1.
SENTIMENTSHUFFLED	PHI3.5+ADAPTOR	SOFTMAX	0.98	0.13	0.97	0.12	0.98	0.13	0.97	0.12	0.97	0.25
SENTIMENTSHUFFLED	PHI3.5+ADAPTOR	TEMPSCALING	0.98	0.11	0.97	0.10	0.97	0.11	0.97	0.10	0.97	0.21
SENTIMENTSHUFFLED	PHI3.5+ADAPTOR	APS	0.93	0.23	0.90	0.24	0.90	0.24	0.93	0.23	0.91	0.48
SENTIMENTSHUFFLED	PHI3.5+ADAPTOR	RAPS	0.92	0.24	0.91	0.24	0.91	0.24	0.92	0.24	0.92	0.48
SENTIMENTSHUFFLED	PHI3.5+SDM	NO-REJECT	0.99	0.50	0.32	0.50	0.59	0.83	0.97	0.17	0.66	1.
SENTIMENTSHUFFLED	PHI3.5+SDM	SOFTMAX	0.99	0.49	0.29	0.36	0.65	0.74	0.98	0.11	0.69	0.85
SENTIMENTSHUFFLED	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	R	0.	R	0.	R	0.	R	0.
SENTIMENTSHUFFLED	PHI3.5+SDM	SDM_{α}	1.	0.01	1.	<0.01	1.	0.01	1.	<0.01	1.	0.01
SENTIMENTSHUFFLED	PHI3.5+SDM	SDM_{HR}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01
SENTIMENTSHUFFLED	MIXTRAL8X7B	NO-REJECT	0.99	0.50	0.35	0.50	0.60	0.82	0.97	0.18	0.67	1.
SENTIMENTSHUFFLED	MIXTRAL8X7B	SOFTMAX	0.99	0.50	0.34	0.48	0.61	0.81	0.97	0.17	0.67	0.98
SENTIMENTSHUFFLED	MIXTRAL8X7B	TEMPSCALING	0.99	0.48	0.37	0.38	0.67	0.72	0.98	0.14	0.72	0.86
SENTIMENTSHUFFLED	MIXTRAL8X7B	APS	0.99	0.45	0.44	0.27	0.75	0.60	0.97	0.12	0.79	0.72
SENTIMENTSHUFFLED	MIXTRAL8X7B	RAPS	0.99	0.44	0.43	0.29	0.73	0.60	0.98	0.13	0.77	0.73
SENTIMENTSHUFFLED	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.91	0.50	0.72	0.50	0.76	0.60	0.89	0.40	0.81	1.
SENTIMENTSHUFFLED	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.99	0.28	0.83	0.14	0.92	0.31	0.98	0.12	0.94	0.43
SENTIMENTSHUFFLED	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.99	0.26	0.82	0.11	0.93	0.28	0.98	0.09	0.94	0.37
SENTIMENTSHUFFLED	MIXTRAL8X7B+ADAPTOR	APS	0.97	0.34	0.78	0.24	0.86	0.38	0.95	0.19	0.89	0.58
SENTIMENTSHUFFLED	MIXTRAL8X7B+ADAPTOR	RAPS	0.96	0.35	0.79	0.23	0.87	0.38	0.93	0.20	0.89	0.58
SENTIMENTSHUFFLED	MIXTRAL8X7B+SDM	NO-REJECT	0.79	0.50	0.82	0.50	0.82	0.48	0.79	0.52	0.80	1.
SENTIMENTSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX	0.80	0.48	0.83	0.49	0.82	0.47	0.80	0.50	0.81	0.97
SENTIMENTSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	R	0.	R	0.	R	0.	R	0.
SENTIMENTSHUFFLED	MIXTRAL8X7B+SDM	SDM_{α}	1.	<0.01	R	0.	1.	<0.01	R	0.	1.	<0.01
SENTIMENTSHUFFLED	MIXTRAL8X7B+SDM	SDM_{HR}	R	0.	R	0.	R	0.	R	0.	R	0.
SENTIMENTOODSHUFFLED	PHI3.5	NO-REJECT	1.00	0.50	0.35	0.50	0.60	0.82	0.99	0.18	0.67	1.
SENTIMENTOODSHUFFLED	PHI3.5	SOFTMAX	1.00	0.50	0.34	0.47	0.62	0.81	0.99	0.16	0.68	0.97
SENTIMENTOODSHUFFLED	PHI3.5	TEMPSCALING	1.00	0.49	0.34	0.29	0.72	0.68	0.99	0.10	0.75	0.78
SENTIMENTOODSHUFFLED	PHI3.5	APS	1.00	0.48	0.36	0.27	0.74	0.65	0.99	0.10	0.77	0.75
SENTIMENTOODSHUFFLED	PHI3.5	RAPS	1.00	0.49	0.36	0.27	0.74	0.66	0.99	0.10	0.77	0.76
SENTIMENTOODSHUFFLED	PHI3.5+ADAPTOR	NO-REJECT	0.64	0.50	0.64	0.50	0.64	0.50	0.64	0.50	0.64	1.
SENTIMENTOODSHUFFLED	PHI3.5+ADAPTOR	SOFTMAX	0.78	0.01	0.93	0.03	0.81	0.01	0.92	0.03	0.89	0.04
SENTIMENTOODSHUFFLED	PHI3.5+ADAPTOR	TEMPSCALING	0.79	0.01	0.94	0.03	0.83	0.01	0.93	0.03	0.90	0.03
SENTIMENTOODSHUFFLED	PHI3.5+ADAPTOR	APS	0.68	0.11	0.73	0.14	0.67	0.12	0.74	0.14	0.71	0.26
SENTIMENTOODSHUFFLED	PHI3.5+ADAPTOR	RAPS	0.70	0.12	0.72	0.14	0.68	0.12	0.74	0.13	0.71	0.25
SENTIMENTOODSHUFFLED	PHI3.5+SDM	NO-REJECT	0.97	0.50	0.63	0.50	0.72	0.67	0.95	0.33	0.80	1.
SENTIMENTOODSHUFFLED	PHI3.5+SDM	SOFTMAX	0.98	0.47	0.64	0.43	0.75	0.62	0.97	0.28	0.82	0.89
SENTIMENTOODSHUFFLED	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	R	0.	R	0.	R	0.	R	0.
SENTIMENTOODSHUFFLED	PHI3.5+SDM	SDM_{α}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	0.01
SENTIMENTOODSHUFFLED	PHI3.5+SDM	SDM_{HR}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01
SENTIMENTOODSHUFFLED	MIXTRAL8X7B	NO-REJECT	1.00	0.50	0.12	0.50	0.53	0.94	1.00	0.06	0.56	1.
SENTIMENTOODSHUFFLED	MIXTRAL8X7B	SOFTMAX	1.00	0.50	0.12	0.49	0.54	0.93	1.00	0.06	0.56	0.99
SENTIMENTOODSHUFFLED	MIXTRAL8X7B	TEMPSCALING	1.	0.49	0.14	0.34	0.63	0.78	1.	0.05	0.65	0.83
SENTIMENTOODSHUFFLED	MIXTRAL8X7B	APS	1.	0.36	0.24	0.18	0.72	0.51	1.	0.04	0.74	0.55
SENTIMENTOODSHUFFLED	MIXTRAL8X7B	RAPS	1.	0.36	0.23	0.19	0.72	0.51	1.	0.04	0.74	0.55
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.97	0.50	0.25	0.50	0.56	0.86	0.89	0.14	0.61	1.
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+ADAPTOR	SOFTMAX	1.	0.04	0.23	0.03	0.62	0.07	1.	0.01	0.66	0.07
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	1.	0.02	0.27	0.02	0.58	0.03	1.	0.01	0.64	0.04
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+ADAPTOR	APS	0.99	0.18	0.21	0.15	0.60	0.29	0.94	0.03	0.64	0.32
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+ADAPTOR	RAPS	0.99	0.18	0.21	0.14	0.61	0.29	0.93	0.03	0.64	0.32
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+SDM	NO-REJECT	0.75	0.50	0.71	0.50	0.72	0.52	0.74	0.48	0.73	1.
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX	0.79	0.43	0.73	0.45	0.73	0.46	0.78	0.42	0.76	0.89
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+SDM	SDM_{α}	1.	<0.01	0.83	<0.01	0.88	0.01	1.	<0.01	0.93	0.01
SENTIMENTOODSHUFFLED	MIXTRAL8X7B+SDM	SDM_{HR}	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01	1.	<0.01

Table 5: Comparison of estimators for the shuffled sentiment datasets, with $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
FACTCHECKSHUFFLED	PHI3.5	NO-REJECT	0.91	1.	-	0.	1.	0.91	0.	0.09	0.91	1.
FACTCHECKSHUFFLED	PHI3.5	SOFTMAX	0.92	0.99	-	0.	1.	0.91	0.	0.08	0.92	0.99
FACTCHECKSHUFFLED	PHI3.5	TEMPSCALING	0.93	0.87	-	0.	1.	0.81	0.	0.06	0.93	0.87
FACTCHECKSHUFFLED	PHI3.5	APS	0.93	0.45	-	0.	1.	0.42	0.	0.03	0.93	0.45
FACTCHECKSHUFFLED	PHI3.5	RAPS	0.95	0.52	-	0.	1.	0.50	0.	0.02	0.95	0.52
FACTCHECKSHUFFLED	PHI3.5+ADAPTOR	NO-REJECT	0.34	1.	-	0.	1.	0.34	0.	0.66	0.34	1.
FACTCHECKSHUFFLED	PHI3.5+ADAPTOR	SOFTMAX	0.20	0.24	-	0.	1.	0.05	0.	0.19	0.20	0.24
FACTCHECKSHUFFLED	PHI3.5+ADAPTOR	TEMPSCALING	0.13	0.19	-	0.	1.	0.02	0.	0.17	0.13	0.19
FACTCHECKSHUFFLED	PHI3.5+ADAPTOR	APS	0.24	0.38	-	0.	1.	0.09	0.	0.29	0.24	0.38
FACTCHECKSHUFFLED	PHI3.5+ADAPTOR	RAPS	0.27	0.39	-	0.	1.	0.11	0.	0.29	0.27	0.39
FACTCHECKSHUFFLED	PHI3.5+SDM	NO-REJECT	0.66	1.	-	0.	1.	0.66	0.	0.34	0.66	1.
FACTCHECKSHUFFLED	PHI3.5+SDM	SOFTMAX	0.69	0.64	-	0.	1.	0.44	0.	0.20	0.69	0.64
FACTCHECKSHUFFLED	PHI3.5+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	-	0.	R	0.	R	0.	R	0.
FACTCHECKSHUFFLED	PHI3.5+SDM	SDM_{α}	R	0.	-	0.	R	0.	R	0.	R	0.
FACTCHECKSHUFFLED	PHI3.5+SDM	SDM_{HR}	R	0.	-	0.	R	0.	R	0.	R	0.
FACTCHECKSHUFFLED	MIXTRAL8X7B	NO-REJECT	0.98	1.	-	0.	1.	0.98	0.	0.02	0.98	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B	SOFTMAX	0.98	1.	-	0.	1.	0.98	0.	0.02	0.98	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B	TEMPSCALING	0.98	0.98	-	0.	1.	0.96	0.	0.02	0.98	0.98
FACTCHECKSHUFFLED	MIXTRAL8X7B	APS	0.98	0.18	-	0.	1.	0.18	0.	<0.01	0.98	0.18
FACTCHECKSHUFFLED	MIXTRAL8X7B	RAPS	0.98	0.23	-	0.	1.	0.23	0.	<0.01	0.98	0.23
FACTCHECKSHUFFLED	MIXTRAL8X7B+ADAPTOR	NO-REJECT	0.79	1.	-	0.	1.	0.79	0.	0.21	0.79	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+ADAPTOR	SOFTMAX	0.69	0.13	-	0.	1.	0.09	0.	0.04	0.69	0.13
FACTCHECKSHUFFLED	MIXTRAL8X7B+ADAPTOR	TEMPSCALING	0.55	0.09	-	0.	1.	0.05	0.	0.04	0.55	0.09
FACTCHECKSHUFFLED	MIXTRAL8X7B+ADAPTOR	APS	0.77	0.40	-	0.	1.	0.31	0.	0.09	0.77	0.40
FACTCHECKSHUFFLED	MIXTRAL8X7B+ADAPTOR	RAPS	0.79	0.39	-	0.	1.	0.31	0.	0.08	0.79	0.39
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	NO-REJECT	0.76	1.	-	0.	1.	0.76	0.	0.24	0.76	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX	0.79	0.65	-	0.	1.	0.51	0.	0.13	0.79	0.65
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	SOFTMAX$(d \cdot \boldsymbol{z}')$	R	0.	-	0.	R	0.	R	0.	R	0.
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	SDM_{α}	1.	0.01	-	0.	1.	0.01	R	0.	1.	0.01
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	SDM_{HR}	R	0.	-	0.	R	0.	R	0.	R	0.
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	$\text{SDM}_{\alpha}, \alpha = 0.94$	1.	0.01	-	0.	1.	0.01	R	0.	1.	0.01
FACTCHECKSHUFFLED	MIXTRAL8X7B+SDM	$\text{SDM}_{\text{HR}}, \alpha = 0.94$	1.	0.01	-	0.	1.	0.01	R	0.	1.	0.01

Table 6: Comparison of estimators for the shuffled factcheck datasets. Unless specified otherwise, $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	Acc.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	Acc.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	Acc.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	Acc.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	Acc.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+DiscVBLLMLP	no-reject	0.97	0.51	0.96	0.49	0.96	0.51	0.96	0.49	0.96	1.
Sentiment	phi3.5+DiscVBLLMLP	no-reject	0.97	0.50	0.95	0.50	0.95	0.51	0.97	0.49	0.96	1.
SentimentOOD	phi3.5+DiscVBLLMLP	no-reject	0.61	0.50	0.67	0.50	0.65	0.47	0.63	0.53	0.64	1.
SentimentShuffled	phi3.5+DiscVBLLMLP	no-reject	0.87	0.50	0.75	0.50	0.78	0.56	0.85	0.44	0.81	1.
SentimentOODShuffled	phi3.5+DiscVBLLMLP	no-reject	0.77	0.50	0.56	0.50	0.64	0.61	0.71	0.39	0.67	1.
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+DiscVBLLMLP	VBLL	0.99	0.42	0.99	0.40	0.99	0.42	0.99	0.40	0.99	0.82
Sentiment	phi3.5+DiscVBLLMLP	VBLL	0.99	0.42	1.00	0.40	1.00	0.42	0.99	0.40	0.99	0.82
SentimentOOD	phi3.5+DiscVBLLMLP	VBLL	0.89	0.03	0.96	0.04	0.93	0.02	0.94	0.05	0.94	0.07
SentimentShuffled	phi3.5+DiscVBLLMLP	VBLL	0.99	0.13	0.95	0.05	0.98	0.13	0.97	0.05	0.98	0.18
SentimentOODShuffled	phi3.5+DiscVBLLMLP	VBLL	0.98	0.02	0.86	0.02	0.89	0.02	0.97	0.01	0.92	0.04
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.96	0.51	0.96	0.49	0.96	0.51	0.96	0.49	0.96	1.
Sentiment	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.97	0.50	0.95	0.50	0.95	0.51	0.97	0.49	0.96	1.
SentimentOOD	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.64	0.50	0.71	0.50	0.69	0.47	0.66	0.53	0.67	1.
SentimentShuffled	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.85	0.50	0.77	0.50	0.79	0.54	0.84	0.46	0.81	1.
SentimentOODShuffled	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.80	0.50	0.60	0.50	0.67	0.60	0.75	0.40	0.70	1.
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.43	0.99	0.41	0.99	0.42	0.99	0.41	0.99	0.84
Sentiment	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.42	1.00	0.41	1.00	0.42	0.99	0.41	0.99	0.83
SentimentOOD	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.95	0.02	0.96	0.04	0.92	0.02	0.98	0.04	0.96	0.07
SentimentShuffled	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.13	0.96	0.09	0.98	0.13	0.98	0.09	0.98	0.22
SentimentOODShuffled	phi3.5+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.02	0.86	0.02	0.90	0.02	0.97	0.02	0.93	0.04
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+GenVBLLMLP	no-reject	0.96	0.50	0.96	0.50	0.96	0.50	0.96	0.50	0.96	1.
Sentiment	phi3.5+GenVBLLMLP	no-reject	0.97	0.50	0.96	0.50	0.96	0.51	0.97	0.49	0.97	1.
SentimentOOD	phi3.5+GenVBLLMLP	no-reject	0.28	0.50	0.93	0.50	0.80	0.17	0.56	0.83	0.60	1.
SentimentShuffled	phi3.5+GenVBLLMLP	no-reject	0.94	0.50	0.60	0.50	0.70	0.67	0.91	0.33	0.77	1.
SentimentOODShuffled	phi3.5+GenVBLLMLP	no-reject	0.43	0.50	0.85	0.50	0.74	0.29	0.60	0.71	0.64	1.
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+GenVBLLMLP	VBLL	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.87
Sentiment	phi3.5+GenVBLLMLP	VBLL	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.87
SentimentOOD	phi3.5+GenVBLLMLP	VBLL	0.13	0.16	0.99	0.28	0.89	0.02	0.67	0.41	0.68	0.43
SentimentShuffled	phi3.5+GenVBLLMLP	VBLL	1.00	0.31	0.70	0.14	0.88	0.35	0.99	0.10	0.90	0.45
SentimentOODShuffled	phi3.5+GenVBLLMLP	VBLL	0.32	0.08	0.97	0.19	0.80	0.03	0.77	0.24	0.78	0.27
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.96	0.51	0.97	0.49	0.97	0.50	0.96	0.50	0.96	1.
Sentiment	phi3.5+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.96	0.50	0.96	0.50	0.96	0.50	0.96	0.50	0.96	1.
SentimentOOD	phi3.5+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.43	0.50	0.78	0.50	0.66	0.32	0.58	0.68	0.61	1.
SentimentShuffled	phi3.5+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.78	0.50	0.84	0.50	0.83	0.47	0.79	0.53	0.81	1.
SentimentOODShuffled	phi3.5+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.64	0.50	0.76	0.50	0.72	0.44	0.68	0.56	0.70	1.
Sentiment $\mathcal{D}_{\text{ca}}$	phi3.5+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.45	0.99	0.43	0.99	0.45	0.99	0.43	0.99	0.88
Sentiment	phi3.5+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.44	0.99	0.44	0.99	0.44	0.99	0.44	0.99	0.89
SentimentOOD	phi3.5+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.36	0.12	0.94	0.15	0.83	0.05	0.65	0.22	0.68	0.27
SentimentShuffled	phi3.5+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.91	0.21	0.96	0.26	0.95	0.20	0.93	0.27	0.94	0.47
SentimentOODShuffled	phi3.5+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.76	0.09	0.86	0.14	0.78	0.09	0.85	0.14	0.82	0.24
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+DiscVBLLMLP	no-reject	0.96	0.50	0.96	0.50	0.96	0.50	0.96	0.50	0.96	1.
Sentiment	Mixtral8x7B+DiscVBLLMLP	no-reject	0.97	0.50	0.96	0.50	0.96	0.50	0.97	0.50	0.97	1.
SentimentOOD	Mixtral8x7B+DiscVBLLMLP	no-reject	0.84	0.50	0.60	0.50	0.68	0.62	0.79	0.38	0.72	1.
SentimentShuffled	Mixtral8x7B+DiscVBLLMLP	no-reject	0.89	0.50	0.78	0.50	0.80	0.55	0.87	0.45	0.83	1.
SentimentOODShuffled	Mixtral8x7B+DiscVBLLMLP	no-reject	0.91	0.50	0.40	0.50	0.60	0.76	0.82	0.24	0.66	1.
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+DiscVBLLMLP	VBLL	0.99	0.44	0.99	0.44	0.99	0.44	0.99	0.44	0.99	0.88
Sentiment	Mixtral8x7B+DiscVBLLMLP	VBLL	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.87
SentimentOOD	Mixtral8x7B+DiscVBLLMLP	VBLL	0.94	0.01	0.97	0.11	0.79	0.02	0.99	0.10	0.96	0.12
SentimentShuffled	Mixtral8x7B+DiscVBLLMLP	VBLL	0.99	0.24	0.91	0.12	0.95	0.25	0.98	0.11	0.96	0.36
SentimentOODShuffled	Mixtral8x7B+DiscVBLLMLP	VBLL	1.	0.01	0.82	0.01	0.68	0.01	1.	0.01	0.87	0.02
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.96	0.50	0.97	0.50	0.97	0.50	0.96	0.50	0.96	1.
Sentiment	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.97	0.50	0.97	0.50	0.97	0.50	0.97	0.50	0.97	1.
SentimentOOD	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.89	0.50	0.59	0.50	0.69	0.65	0.84	0.35	0.74	1.
SentimentShuffled	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.86	0.50	0.81	0.50	0.81	0.53	0.85	0.47	0.83	1.
SentimentOODShuffled	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	no-reject	0.94	0.50	0.44	0.50	0.63	0.75	0.88	0.25	0.69	1.
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.87
Sentiment	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.44	0.99	0.43	0.99	0.44	0.99	0.43	0.99	0.87
SentimentOOD	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.03	0.93	0.07	0.87	0.04	0.99	0.06	0.95	0.10
SentimentShuffled	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.24	0.92	0.14	0.95	0.25	0.98	0.13	0.96	0.38
SentimentOODShuffled	Mixtral8x7B+DiscVBLLMLP$_{\text{rw50}}$	VBLL	1.	0.02	0.69	0.01	0.87	0.02	1.	0.01	0.90	0.03
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+GenVBLLMLP	no-reject	0.97	0.50	0.97	0.50	0.97	0.50	0.97	0.50	0.97	1.
Sentiment	Mixtral8x7B+GenVBLLMLP	no-reject	0.97	0.50	0.96	0.50	0.96	0.51	0.97	0.49	0.97	1.
SentimentOOD	Mixtral8x7B+GenVBLLMLP	no-reject	0.80	0.50	0.72	0.50	0.74	0.54	0.78	0.46	0.76	1.
SentimentShuffled	Mixtral8x7B+GenVBLLMLP	no-reject	0.84	0.50	0.82	0.50	0.82	0.51	0.83	0.49	0.83	1.
SentimentOODShuffled	Mixtral8x7B+GenVBLLMLP	no-reject	0.85	0.50	0.54	0.50	0.65	0.65	0.78	0.35	0.69	1.
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+GenVBLLMLP	VBLL	0.99	0.44	0.99	0.45	0.99	0.44	0.99	0.46	0.99	0.89
Sentiment	Mixtral8x7B+GenVBLLMLP	VBLL	0.99	0.44	0.99	0.44	0.99	0.44	0.99	0.45	0.99	0.88
SentimentOOD	Mixtral8x7B+GenVBLLMLP	VBLL	0.93	0.04	0.99	0.15	0.95	0.04	0.98	0.15	0.97	0.19
SentimentShuffled	Mixtral8x7B+GenVBLLMLP	VBLL	0.96	0.23	0.91	0.20	0.93	0.24	0.95	0.20	0.94	0.44
SentimentOODShuffled	Mixtral8x7B+GenVBLLMLP	VBLL	0.99	0.02	0.81	0.04	0.73	0.03	0.99	0.03	0.87	0.06
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.97	0.50	0.97	0.50	0.97	0.50	0.97	0.50	0.97	1.
Sentiment	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.97	0.50	0.96	0.50	0.96	0.50	0.97	0.50	0.97	1.
SentimentOOD	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.79	0.50	0.73	0.50	0.75	0.53	0.78	0.47	0.76	1.
SentimentShuffled	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.83	0.50	0.82	0.50	0.82	0.51	0.83	0.49	0.83	1.
SentimentOODShuffled	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	no-reject	0.83	0.50	0.56	0.50	0.65	0.64	0.77	0.36	0.69	1.
Sentiment $\mathcal{D}_{\text{ca}}$	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.44	0.99	0.46	0.99	0.44	0.98	0.46	0.99	0.90
Sentiment	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.99	0.44	0.99	0.45	0.99	0.44	0.99	0.45	0.99	0.89
SentimentOOD	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.91	0.04	0.99	0.16	0.95	0.04	0.98	0.16	0.97	0.20
SentimentShuffled	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.95	0.24	0.91	0.21	0.92	0.24	0.95	0.21	0.94	0.45
SentimentOODShuffled	Mixtral8x7B+GenVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.02	0.82	0.04	0.73	0.03	0.99	0.04	0.88	0.06

Table 7: Comparison of Bayesian last-layer estimators (Harrison et al., 2024) for the sentiment datasets, including the shuffled challenge sets, with $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.

			Class-conditional $y=0$		Class-conditional $y=1$		Prediction-conditional $\hat{y}=0$		Prediction-conditional $\hat{y}=1$		Marginal $y \in \{0,1\}$	
Dataset	Model	Estimator	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$	ACC.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+DISCVBLLMLP	NO-REJECT	0.92	0.50	0.91	0.50	0.91	0.50	0.92	0.50	0.91	1.
FACTCHECK	PHI3.5+DISCVBLLMLP	NO-REJECT	0.38	0.51	0.92	0.49	0.84	0.23	0.59	0.77	0.64	1.
FACTCHECKSHUFFLED	PHI3.5+DISCVBLLMLP	NO-REJECT	0.33	1.	R	0.	1.	0.33	0.	0.67	0.33	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+DISCVBLLMLP	VBLL	0.98	0.32	0.99	0.28	0.99	0.32	0.98	0.29	0.98	0.61
FACTCHECK	PHI3.5+DISCVBLLMLP	VBLL	0.35	0.07	1.	0.31	1.	0.02	0.88	0.36	0.88	0.38
FACTCHECKSHUFFLED	PHI3.5+DISCVBLLMLP	VBLL	0.15	0.21	R	0.	1.	0.03	0.	0.18	0.15	0.21
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.91	0.50	0.92	0.50	0.92	0.49	0.91	0.51	0.91	1.
FACTCHECK	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.45	0.51	0.94	0.49	0.89	0.26	0.62	0.74	0.69	1.
FACTCHECKSHUFFLED	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.43	1.	R	0.	1.	0.43	0.	0.57	0.43	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.34	0.98	0.33	0.98	0.34	0.98	0.33	0.98	0.67
FACTCHECK	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.35	0.11	0.99	0.32	0.90	0.04	0.82	0.39	0.83	0.43
FACTCHECKSHUFFLED	PHI3.5+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.33	0.33	R	0.	1.	0.11	0.	0.22	0.33	0.33
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+GENVBLLMLP	NO-REJECT	0.90	0.50	0.93	0.50	0.93	0.49	0.91	0.51	0.92	1.
FACTCHECK	PHI3.5+GENVBLLMLP	NO-REJECT	0.41	0.51	0.93	0.49	0.87	0.24	0.60	0.76	0.67	1.
FACTCHECKSHUFFLED	PHI3.5+GENVBLLMLP	NO-REJECT	0.40	1.	R	0.	1.	0.40	0.	0.60	0.40	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+GENVBLLMLP	VBLL	0.98	0.33	0.99	0.32	0.99	0.33	0.98	0.33	0.98	0.65
FACTCHECK	PHI3.5+GENVBLLMLP	VBLL	0.31	0.07	0.99	0.30	0.83	0.02	0.87	0.34	0.87	0.37
FACTCHECKSHUFFLED	PHI3.5+GENVBLLMLP	VBLL	0.20	0.22	R	0.	1.	0.04	0.	0.18	0.20	0.22
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.91	0.50	0.93	0.50	0.93	0.49	0.91	0.51	0.92	1.
FACTCHECK	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.41	0.51	0.93	0.49	0.87	0.24	0.60	0.76	0.67	1.
FACTCHECKSHUFFLED	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.44	1.	R	0.	1.	0.44	0.	0.56	0.44	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.33	0.99	0.29	0.99	0.33	0.98	0.29	0.99	0.62
FACTCHECK	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.29	0.06	0.98	0.24	0.80	0.02	0.86	0.28	0.85	0.30
FACTCHECKSHUFFLED	PHI3.5+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.22	0.21	R	0.	1.	0.04	0.	0.16	0.22	0.21
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+DISCVBLLMLP	NO-REJECT	0.91	0.50	0.93	0.50	0.93	0.49	0.91	0.51	0.92	1.
FACTCHECK	MIXTRAL8X7B+DISCVBLLMLP	NO-REJECT	0.66	0.51	0.88	0.49	0.86	0.40	0.71	0.60	0.77	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+DISCVBLLMLP	NO-REJECT	0.84	1.	R	0.	1.	0.84	0.	0.16	0.84	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.98	0.32	0.99	0.31	0.99	0.31	0.98	0.31	0.99	0.62
FACTCHECK	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.85	0.08	0.97	0.27	0.89	0.08	0.95	0.27	0.94	0.35
FACTCHECKSHUFFLED	MIXTRAL8X7B+DISCVBLLMLP	VBLL	0.91	0.14	R	0.	1.	0.13	0.	0.01	0.91	0.14
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.90	0.50	0.94	0.50	0.94	0.48	0.90	0.52	0.92	1.
FACTCHECK	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.62	0.51	0.89	0.49	0.86	0.37	0.69	0.63	0.75	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.81	1.	R	0.	1.	0.81	0.	0.19	0.81	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.98	0.33	0.99	0.34	0.99	0.32	0.98	0.34	0.99	0.67
FACTCHECK	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.72	0.12	0.97	0.30	0.91	0.09	0.90	0.32	0.90	0.42
FACTCHECKSHUFFLED	MIXTRAL8X7B+DISCVBLLMLP$_{\text{rw50}}$	VBLL	0.82	0.16	R	0.	1.	0.13	0.	0.03	0.82	0.16
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+GENVBLLMLP	NO-REJECT	0.91	0.48	0.93	0.52	0.92	0.48	0.92	0.52	0.92	1.
FACTCHECK	MIXTRAL8X7B+GENVBLLMLP	NO-REJECT	0.63	0.51	0.88	0.49	0.85	0.38	0.70	0.62	0.76	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+GENVBLLMLP	NO-REJECT	0.73	1.	R	0.	1.	0.73	0.	0.27	0.73	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+GENVBLLMLP	VBLL	0.96	0.32	0.99	0.40	0.99	0.31	0.97	0.41	0.98	0.72
FACTCHECK	MIXTRAL8X7B+GENVBLLMLP	VBLL	0.58	0.13	0.99	0.33	0.95	0.08	0.85	0.38	0.87	0.47
FACTCHECKSHUFFLED	MIXTRAL8X7B+GENVBLLMLP	VBLL	0.54	0.11	R	0.	1.	0.06	0.	0.05	0.54	0.11
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.92	0.48	0.91	0.52	0.90	0.49	0.93	0.51	0.91	1.
FACTCHECK	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.67	0.51	0.89	0.49	0.87	0.40	0.72	0.60	0.78	1.
FACTCHECKSHUFFLED	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	NO-REJECT	0.69	1.	R	0.	1.	0.69	0.	0.31	0.69	1.
FACTCHECK $\mathcal{D}_{\text{ca}}$	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.97	0.38	0.98	0.40	0.98	0.37	0.97	0.41	0.97	0.78
FACTCHECK	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.67	0.19	0.95	0.35	0.89	0.14	0.84	0.39	0.85	0.53
FACTCHECKSHUFFLED	MIXTRAL8X7B+GENVBLLMLP$_{\text{rw50}}$	VBLL	0.67	0.15	R	0.	1.	0.10	0.	0.05	0.67	0.15

Table 8: Comparison of Bayesian last-layer estimators (Harrison et al., 2024) for the factcheck datasets, including the shuffled challenge sets, with $\alpha = 0.95$. R indicates all predictions were rejected, which is preferred over falling under the expected accuracy. $n = |\text{Admitted}|$, the count of non-rejected documents.