Supplemental Material: Similarity-Distance-Magnitude Activations

Allen Schmaltz *allen@re.express*
Reexpress AI

A Appendix: Continued

A.4 Comparison to Bayesian Last-layer Networks

In the main text, we compare to representative Frequentist and empirically motivated estimators. In this section, we further compare 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 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[1], 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 & 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 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. 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 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).

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

Results. The results for the sentiment datasets appear in Table 5, and the results for the factcheck datasets appear in Table 6. 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 3 in the main text, 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 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 co-variate 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.

These results corroborate the patterns observed in the main text. In our controlled examination of representative Bayesian, Frequentist, and empirically motivated estimators, the SDM_{HR} estimator (i.e., the SDM activation function combined with the selection criteria of Eq. 9) is the only estimator that remains well-calibrated in high-probability regions across the in-distribution, co-variate shifted, and out-of-distribution data.

Table 5: 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.	$\frac{n}{\lvert\mathcal{D}_{\text{te}}\rvert}$
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 6: 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.

			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

References

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.07289.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx7BIiPzys.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.