Beyond the Final Layer: Using Intermediate Representations to Improve Multilingual Calibration

Anonymous Author(s)
Affiliation
Address
email

Abstract

Confidence calibration, the alignment between a model's predicted confidence and its empirical correctness, is crucial for the trustworthiness of Large Language Models (LLMs). Previous studies on multilingual calibration mainly use machine-translated data and are limited to a small number of languages. In this work, we present the first systematic evaluation of multilingual calibration on 3 high-quality datasets over 100 languages with 7 model families. Our analysis reveals that LLMs exhibit significant disparities across languages, particularly underperforming in low-resource and non-Latin-script settings. To understand the source of this miscalibration, we conduct a layer-wise analysis and uncovered a consistent pattern: intermediate layers often yield better-calibrated outputs than final layers, especially for low-resource languages. Inspired by this observation, we propose leveraging intermediate representations to enhance multilingual calibration. Our methods significantly improve Expected Calibration Error (ECE), Brier Score, and AUROC, outperforming final-layer baselines by large margins. Importantly, our approach is orthogonal to existing calibration methods, and combining them leads to further improvements. This work challenges the conventional reliance on final-layer decoding and opens a new direction for achieving robust and equitable multilingual calibration.

1 Introduction

Calibration in machine learning refers to the alignment between a model's confidence in its predictions and the actual probability of those predictions being correct [Guo et al., 2017, Tian et al., 2023, Geng et al., 2024]. For example, a perfectly calibrated model that assigns an 80% confidence to a prediction should indeed be correct approximately 80% of the time. Accurate calibration is crucial in practical applications of large language models (LLMs), particularly in high-stakes scenarios such as medical diagnosis, legal advice, or critical decision-making processes [Zhang et al., 2024a,b, Yang et al., 2024b]. Properly calibrated models can provide more reliable and interpretable confidence scores, increasing their trustworthiness and clearly indicating the reliability of generated responses.

However, existing research on calibration has primarily focused on English [Tian et al., 2023, Li et al., 2024, Zhang et al., 2024b]. Recent study on multilingual calibration relies on machine-translated datasets [Xue et al., 2024], which may introduce potential biases [Vanmassenhove et al., 2021, Choenni et al., 2024]. We argue that model calibration in more realistic multilingual scenarios, and the effectiveness of calibration methods in such environments, remain largely underexplored. This gap is especially concerning for low-resource languages, where limited training data often results in poorer calibration, increasing the risk of misleading or harmful outputs in critical applications.

We first empirically analyze the calibration of popular LLMs across 7 model families using 3 *human-translated* datasets: MMMLU [Hendrycks et al., 2020], Belebele [Bandarkar et al., 2024a], and

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

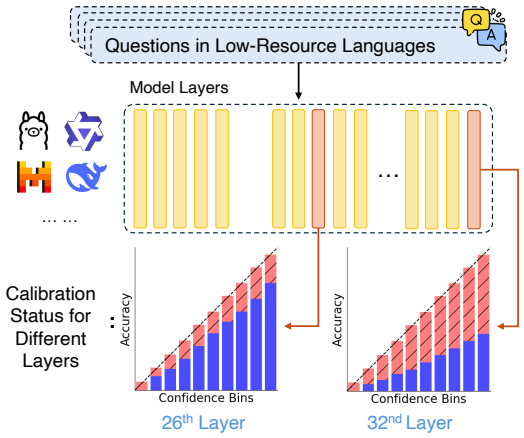

Figure 1: Different layers show various levels of calibration in multilingual LLMs for questions in low-resource languages. Intermediate layers usually exhibit better calibration, while last layers tend to be overconfident and poorly calibrated.

MKQA [Longpre et al., 2021], covering both multiple-choice and short-form QA tasks. Our study spans over 100 languages and provides further evidence that low-resource languages tend to exhibit lower accuracy and worse calibration. We also observe that Latin-script languages generally show better calibration and accuracy compared to non-Latin-script languages.

We further explore the underlying reasons of the consistently poor calibration observed at the final output layer. We draw inspiration from recent findings which suggest that intermediate layers in LLMs encode language-agnostic knowledge, while upper layers are more language-specific [Bandarkar et al., 2024b, Wendler et al., 2024]. Building on this insight, we analyze calibration across different layers in Section 4 and reveal that different layers exhibit varying calibration quality. Specifically for low-resource languages, models are better calibrated in intermediate layers before a significant drop-off in the final layer.

Our findings motivate us to use logits from intermediate layers as confidence scores to improve calibration in multilingual LLMs. In Section 5, we propose a series of novel calibration methods that leverage intermediate layers to enhance final calibration. Our results show *consistent improvements in calibration without affecting accuracy*, especially for **low-resource languages**. Furthermore, we demonstrate that our approach is orthogonal to traditional calibration methods, and combining them leads to even better performance. Our study offers valuable insights and methodological advances toward reliable multilingual calibration, supporting more equitable and trustworthy deployment of LLMs worldwide.

Our contributions are listed as follows:

- We provide a comprehensive empirical analysis of calibration in multilingual LLMs on human-translated datasets, revealing significant disparities between high-resource and low-resource languages.
- We are the first to investigate layer-wise calibration, showing that intermediate layers often exhibit better calibration for low-resource languages compared to the final layer.
- We propose novel calibration methods that leverage intermediate layer representations, demonstrating their effectiveness in improving calibration and reducing performance gaps across languages.

2 Related Work

Multilingual Calibration Recent work has highlighted that modern LLMs, despite their strong performance, often generate overconfident predictions [Xiong et al., 2024, Zhang et al., 2024a]. Calibration techniques are thus in need to mitigate the overconfidence issue Geng et al. [2023], but

Language	LLaMA3				Cohere			
	AUROC	ECE	BRIER	Accuracy	AUROC	ECE	BRIER	Accuracy
Arabic	61.00	33.06	24.37	38.20	71.49	28.41	33.79	45.20
Bengali	58.44	24.93	23.39	35.20	60.01	29.01	31.48	31.30
German	65.36	25.81	24.92	44.40	69.70	26.54	33.51	53.00
English	80.36	4.61	17.63	61.20	74.65	20.66	25.30	57.40
Spanish	71.65	18.21	21.89	52.00	71.12	28.17	31.86	51.10
French	71.39	13.87	22.75	51.30	70.69	23.80	32.72	53.40
Hindi	62.07	28.31	24.28	39.90	70.08	30.21	34.98	42.30
Indonesian	66.25	19.67	23.76	45.00	70.85	27.88	31.54	51.20
Italian	71.57	21.19	22.74	51.80	71.76	26.65	30.33	52.70
Japanese	61.73	28.36	27.27	43.00	69.92	16.30	26.26	46.70
Korean	62.59	30.86	25.06	42.50	72.06	32.07	37.09	45.00
Portuguese	71.37	10.51	21.76	50.40	70.71	27.33	31.42	53.50
Swahili	61.10	23.84	21.45	32.20	58.23	32.01	36.72	31.30
Yoruba	58.00	8.18	19.43	27.40	60.73	30.11	28.56	26.40
Chinese	50.63	41.94	19.56	23.10	67.35	17.12	28.75	52.20
Avg. Low-Resource	61.14	23.00	22.78	36.32	65.23	29.60	32.84	37.95
Avg. High-Resource	67.41	21.71	22.62	46.63	70.88	24.29	30.80	51.67
Avg. Non-Latin-Script	59.44	27.44	23.10	35.19	66.23	26.90	32.20	40.05
Avg. Latin-Script	71.14	16.27	22.21	50.87	71.35	25.86	30.95	53.19
Average (All Languages)	64.90	22.22	22.68	42.51	68.62	26.42	31.62	46.18

Table 1: Multilingual performance of **LLaMA3** (left) and **Aya** (right) on the MMMLU dataset. Metrics include AUROC, ECE, Brier Score, and Accuracy. All numbers are in percentages.

it is underexplored in multilingual setting. Seminal work by Ahuja et al. [2022] first established that massively multilingual models like mBERT and XLM-R are poorly calibrated, especially for low-resource and typologically distant languages. Subsequent research has confirmed that this problem persists and may even be amplified in modern generative models. For instance, Yang et al. [2023] specifically evaluated multilingual question-answering LLMs and found substantial calibration gaps between high-resource and low-resource languages. Expanding this line of research, Xue et al. [2024] conducted a comprehensive study across various models, covering both language-agnostic and language-specific tasks. However, all datasets in their study were translated by machine, which can potentially import bias. These studies collectively establish a critical performance bottleneck: even when models achieve reasonable accuracy, their reliability is undermined by poor multilingual calibration. However, they primarily focus on documenting this phenomenon at the final output layer. The architectural origins of this cross-lingual calibration deficit remain underexplored, motivating our work to investigate calibration dynamics within the internal layers of the model.

Layer-wise Representations A growing body of research investigates the functional specialization of layers within multilingual transformers. It is widely observed that intermediate layers encode cross-lingual semantic knowledge in a largely language-agnostic manner, forming a shared representational space [Bandarkar et al., 2024b]. In contrast, the final layers tend to be more language-specific, adapting these general representations to handle surface-level features like syntax and word order for the target language. Recent studies on predominantly English-trained LLMs, such as LLaMA, suggest a more specific mechanism: these models often process multilingual text by mapping it to an internal English-based representation in the middle layers, before translating it back to the target language in the final layers [Wendler et al., 2024, Kojima et al., 2024, Alabi et al., 2024]. This "latent English" hypothesis explains the empirical success of prompting strategies that explicitly ask the model to "think in English" before generating a response in another language, as this aligns with the model's internal processing pathway [Shi et al., 2022, Zhang et al., 2024c]. Our work builds on these insights by exploring the implications of this layer-wise specialization for model calibration.

3 Benchmarking Multilingual Calibration on Human-Translated Datasets

3.1 Experimental Setup

Datasets We use datasets that cover both multiple-choice and short-form question-answering formats across diverse languages, including (1) MMMLU [Hendrycks et al., 2020] (15 languages,

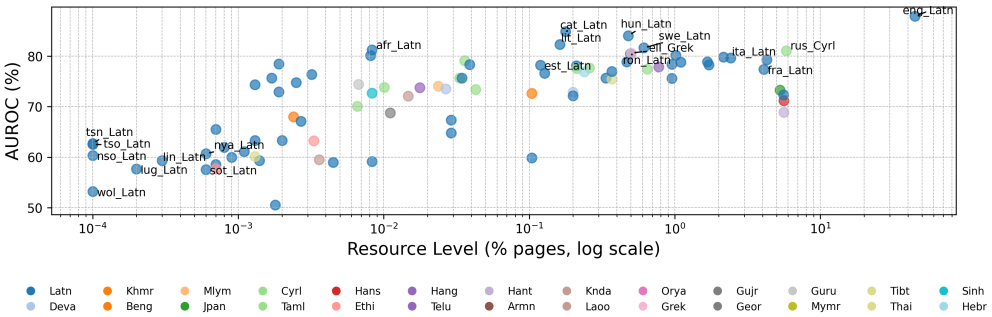

Figure 2: Relationship between language resource level and AUROC (%) for the LLaMA3 model on the Belebele benchmark. Each point represents a language, and languages sharing the same color use the same writing system.

multiple-choice), (2) Belebele [Bandarkar et al., 2024a] (122 languages, multiple-choice), (3) MKQA [Longpre et al., 2021] (26 langauges, short-form). All three datasets consist of high-quality human-translated items. All experiments were conducted using a eight-shot prompting setup in its respecitve language.

Models Our experiments involve evaluating several recent large language models: **LLaMA3** [Grattafiori et al., 2024] (`Llama-3.1-8B-Instruct`), **Qwen2.5** [Yang et al., 2024a] (`Qwen2.5-7B-Instruct`), **Mistral** [Jiang et al., 2023] (`Mistral-7B-Instruct-v0.3`), **Babel** [Zhao et al., 2025] (`Babel-9B-Chat`), **Aya** [Dang et al., 2024](`aya-expanse-8b`), **DeepSeek** [DeepSeek-AI, 2025] (`DeepSeek-R1-Distill-Qwen-7B`), and **Phi** [Abdin et al., 2024] (`phi-4`).

Confidence Elicitation Methods and Metrics For multiple-choice datasets such as MMMLU and Belebele, we adopt the standard confidence estimation approach, which uses the log-probability of the selected answer choice. For short-form datasets (MKQA), we experiment with three confidence elicitation approaches, following Xue et al. [2024]'s setup: (1) log probability of the generated answer sequence (***Prob***), (2) the probability of generating a "true" token given the question-answer pair (***True***), and (3) verbalized confidence (***Verb***), where the model explicitly articulates its confidence level. For all models, we restrict the answer format to short-form outputs by setting maximum response length to 48 during inference. We use PREM (Positive-Recall Exact Match) to evaluate accuracy, this is a relaxed evaluation metric that considers an answer correct if the predicted answer contains the reference or vice versa.

Metrics To evaluate calibration and accuracy, we use four primary metrics: Accuracy, ECE (expected calibration error; Guo et al., 2017), AUROC (area under the receiver operating characteristic curve; Fawcett, 2006), and the Brier Score (Brier, 1950).

To quantify resource availability across languages, we utilize the Common Crawl dataset (`CC-MAIN-2025-30`; Common Crawl Foundation, 2025), calculating resource levels as the percentage of web pages available per language from the crawl, as will be shown in Figure 2.

3.2 Results

Our results, presented in Table 1 for the LLaMA3 and Aya models on MMMLU, and visually summarized in Figure 2 for Belebele, reveal variations in accuracy and calibration across languages and resource categories. Additional results on MKQA are included in Appendix C.3. Similar patterns are observed for MMMLU across other models, including Mistral (Table 3), Babel (Table 4), Qwen2.5 (Table 5), Phi (Table 6), and Deepseek (Table 7). Complete Belebele results for LLaMA3 are also provided in Appendix 8.

Low-resource languages exhibit lower accuracy. As shown in Table 1, low-resource languages consistently underperform in terms of accuracy. The average accuracy in LLaMA3 across low-resource languages is just 36.32%, compared to 46.63% for high-resource languages and 61.20% for

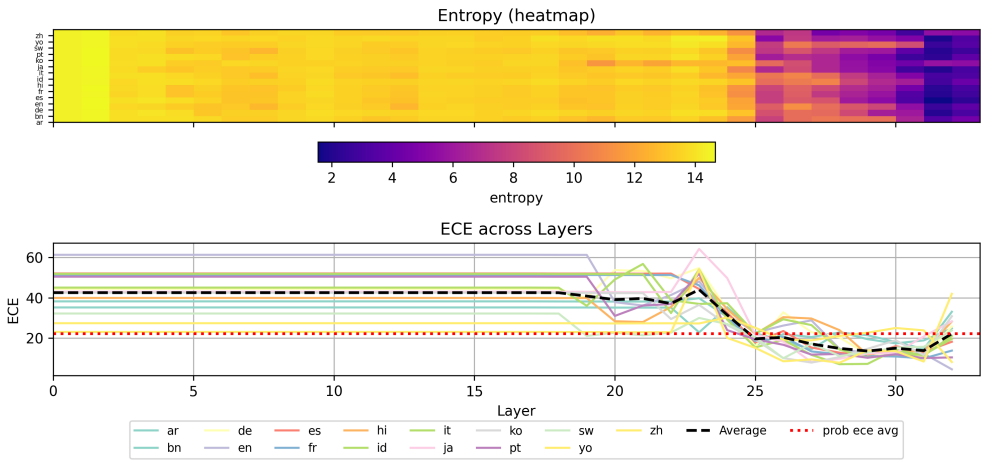

Figure 3: ECE vs. entropy across layers on the MMMLU subset for LLaMA3. In the multilingual setting, many languages achieve their lowest (best) ECE in intermediate layers (e.g., 22-26), after which calibration quality degrades towards the final layer. This contrasts with the English-only setting, where calibration improves monotonically.

English. Languages such as Swahili (32.20%), Yoruba (27.40%) demonstrate particularly low scores, highlighting substantial performance gaps in multilingual understanding and reasoning.

Low-resource languages suffer from worse calibration. In addition to reduced accuracy, calibration is also worse in low-resource languages. The average AUROC for these languages is 61.14, substantially lower than the 80.36 observed for English, suggesting that model confidence is far less reliable. Similarly, the average observed ECE for low-resource languages in Aya (29.60) exceeds that of the high-resource counterparts (24.29). This disparity is further illustrated in Figure 2, which shows a clear correlation between resource level and calibration performance, with low-resource languages consistently exhibiting higher calibration error.

Latin-script languages show better calibration and accuracy compared with non-Latin-script languages. Our results also highlight a performance gap between languages based on their script. In LLaMA3 Latin-script languages achieve an average accuracy of 50.87% and an average ECE of 16.27%. In contrast, non-Latin-script languages have a lower average accuracy of 35.19% and a much higher average ECE of 27.44%, indicating poorer calibration. This disparity is consistent across all metrics, with Latin-script languages showing a higher average AUROC (71.14% vs. 59.44%) and a slightly lower (better) Brier score (22.21% vs. 23.10%).

4 Mid-Layers Reveal Better Calibration

To understand the source of the poor calibration observed in the multilingual setting, we investigate how calibration evolves throughout the model's depth. Inspired by recent insights in layer-wise multilingual representations [Bandarkar et al., 2024b, Wendler et al., 2024]. We hypothesize that the final layers, which may over-specialize in high-resource languages like English, could be detrimental to the calibration of other languages.

4.1 Methodology for Layer-Wise Early Decoding

To investigate how calibration evolves across the depth of the model, we adopt a layer-wise probing technique inspired by the early exiting paradigm [Elbayad et al., 2020]. Instead of applying the modeling head only to the final hidden state, we attach it to each intermediate transformer layer. This allows us to extract logits and compute prediction confidence from every layer, providing a granular view of the model's decision-making process.

Formally, let $\mathbf{h}_\ell \in \mathbb{R}^d$ denote the hidden representation at layer ℓ, where $\ell = 1, \ldots, L$, and d is the dimensionality of the hidden state. We apply the original language modeling head, with weight matrix $W \in \mathbb{R}^{V \times d}$, to compute the logits at each layer:

$$\mathbf{z}_\ell = W\mathbf{h}_\ell$$

where $\mathbf{z}_\ell \in \mathbb{R}^V$ are the unnormalized token logits over the vocabulary of size V. These logits are then converted into probabilities using the softmax function, from which we derive the predicted token and its confidence at each layer:

$$\mathbf{p}_\ell = \mathrm{softmax}(\mathbf{z}_\ell), \quad \hat{y}_\ell = \arg\max_v [\mathbf{p}_\ell]_v$$

To quantify the model's uncertainty at each stage, we also compute the entropy of the probability distribution for each layer:

$$\mathcal{H}_\ell = -\sum_{v=1}^{V} [\mathbf{p}_\ell]_v \log_2 [\mathbf{p}_\ell]_v$$

4.2 Multilingual Language Models Calibrate Earlier

Calibration improves as expected in English-only settings. We first establish a baseline by conducting a layer-wise analysis in an English-only setting. As shown in Figure 4 in the Appendix for LLaMA3, we observe a clear and expected trend: calibration improves monotonically with layer depth. The ECE is high in the early layers and steadily decreases, reaching its minimum at the final layer. This aligns with the conventional understanding that representations become progressively more refined and task-specific, leading to greater confidence and better calibration as data propagates through the network.

Multilingual settings reveal a surprising calibration peak in middle layers. However, our analysis reveals a strikingly different pattern in the multilingual context. As illustrated in Figure 3, the best calibration performance for many languages **does not** occur at the final layer. Instead, we find that ECE often reaches its minimum in the late-intermediate layers (typically between layers 22 and 26 for a 32-layer model), after which calibration quality *worsens* as the signal proceeds to the final output layer.

Final-layer specialization may degrade multilingual calibration. This effect is particularly pronounced for low- and mid-resource languages, where the final layers exhibit a sharp degradation of calibration. It suggests that while intermediate layers may capture a well-calibrated, language-agnostic representation, the final layers might be overfitting to the patterns of dominant languages (i.e., English) or introducing noise during the final language-specific adaptation phase. This could harm calibration for less-represented languages, whose representations might be distorted by this final step.

The mid-layer calibration peak is a robust finding across models. This critical observation is not isolated to a single model or metric. We consistently find this pattern across multiple architectures and evaluation metrics, as detailed in the Appendix. For models like LLaMA3 (Figure 5), Cohere (Figure 6), Mistral (Figure 7), and more, calibration (measured by ECE, Brier score, and AUROC) improves through the deep layers, hits an optimal point in the middle, and then deteriorates. This core finding motivates the novel calibration methods proposed in the next section, which aim to leverage these better-calibrated intermediate representations.

5 Improving Multilingual Calibration

Our observations from the previous section suggests a promising direction: rather than relying solely on the final layer, we can develop calibration methods that explicitly leverage the strengths of intermediate representations. Below, we outline several such methods and their variations, each designed to enhance calibration in multilingual settings by taking advantage of these findings.

5.1 Traditional Post-hoc Calibrations

As classical calibration methods widely used in the literature, we incorporate Temperature Scaling [Guo et al., 2017], Isotonic Regression [Zadrozny and Elkan, 2002], Histogram Binning [Zadrozny and Elkan, 2001], and Platt Scaling [Platt, 2000] into our experiments. These traditional approaches are post-hoc calibration methods typically applied to directly adjusting the predicted probabilities from the model's final output layer. We include them to establish baseline performance levels.

5.2 Intermediate Representation Inspired Calibration Methods

Best Layer From our empirical analysis (Figure 3), we identify that the model achieves optimal calibration at certain intermediate layers. We define the "best" layer as the one that minimizes ECE on a held-out validation set. Formally, let ECE_ℓ denote the ECE computed from the output probabilities at layer ℓ. The best-performing layer ℓ^* is then selected as:

$$\ell^* = \arg\min_{\ell \in \{1,\ldots,L\}} \mathrm{ECE}_\ell$$

We then use the output probabilities from layer ℓ^* for downstream prediction and calibration-sensitive decision making. This approach is both simple and effective, requiring no additional parameters or training while leveraging empirical calibration dynamics.

Best+Last Ensemble To leverage complementary strengths of both intermediate and final layers, we propose a method that ensembles outputs from the best-calibrated layer ℓ^* and the final layer L. We explore two strategies:

(1) Probability Averaging: Compute the average of the softmax probabilities from both layers:

$$\mathbf{p}_{\text{ensemble}} = \frac{1}{2}\left(\mathrm{softmax}(W\mathbf{h}_{\ell^*}) + \mathrm{softmax}(W\mathbf{h}_L)\right)$$

(2) Hidden State Averaging: Compute the average of the hidden states before applying the output head and softmax:

$$\mathbf{p}_{\text{ensemble}} = \mathrm{softmax}\left(W \cdot \frac{1}{2}(\mathbf{h}_{\ell^*} + \mathbf{h}_L)\right)$$

This method allows the model to combine calibration-aware signals from intermediate layers with the semantic richness of the final layer, often resulting in improved overall calibration.

Good Layers Pooling Rather than selecting a single intermediate layer, we identify a set of layers that are better calibrated than the final layer and treat them collectively as "good" layers. Specifically, we define the set of good layers $\mathcal{G}$ as:

$$\mathcal{G} = \{\ell : \mathrm{ECE}_\ell < \mathrm{ECE}_L\}$$

We then explore two ensembling strategies, same as method 2:

(1) Probability Averaging:

$$\mathbf{p}_{\text{ensemble}} = \frac{\sum_{\ell \in \mathcal{G}} \mathrm{softmax}(W\mathbf{h}_\ell) + \mathrm{softmax}(W\mathbf{h}_L)}{|\mathcal{G}| + 1}$$

(2) Hidden State Averaging:

$$\mathbf{p}_{\text{ensemble}} = \mathrm{softmax}\left(W \cdot \frac{\sum_{\ell \in \mathcal{G}} \mathbf{h}_\ell + \mathbf{h}_L}{|\mathcal{G}| + 1}\right)$$

This approach integrates broader calibration-aware signals from multiple intermediate layers, potentially smoothing out noise from any individual layer and capturing more robust confidence estimates.

Contrastive Layer Decoding Inspired by contrastive decoding methods (e.g., Li et al. [2023]), we propose to enhance calibration by contrasting the final layer with the best-calibrated intermediate layer. The intuition is to use the calibrated intermediate signal to guide and correct the often overconfident final prediction.

Let $\mathbf{p}_{\ell^*}$ and $\mathbf{p}_L$ denote the softmax probability distributions from the best and final layers, respectively. We compute the contrastive log-probability vector as:

$$\mathbf{p}_{\text{contrast}} = \text{softmax}\left(\log \mathbf{p}_{\ell^*} - \alpha \cdot \log \mathbf{p}_L\right)$$

where α is a tunable contrastive strength parameter.

Hidden State Steering To improve calibration without modifying the model head, we steer the final hidden state toward the better-calibrated intermediate representation. Let $\mathbf{h}_L$ and $\mathbf{h}_{\ell^*}$ be the hidden states from the final and best layers, respectively. We compute a steering vector $\Delta_h = \mathbf{h}_{\ell^*} - \mathbf{h}_L$ and apply it with a tunable weight β:

$$\mathbf{p}_{\text{steered}} = \text{softmax}\left(W(\mathbf{h}_L + \beta \cdot \Delta_h)\right)$$

This method gently shifts the final representation in the direction of the calibrated intermediate signal, improving output confidence without disrupting task semantics.

5.3 Combining Intermediate Representation Methods with Traditional Calibrations

We predict that the intermediate representation-inspired calibration methods introduced above are complementary to traditional post-hoc calibration techniques in Section 5.1. Given that traditional methods operate on predicted probabilities independently of how those probabilities were derived, they can be straightforwardly applied as a second-stage calibration on top of our intermediate representation-based ensembles.

Specifically, we performed a two-step calibration procedure: First, we obtain calibrated predictions using our proposed intermediate representation methods (Good Layers Pooling as an example). This step exploits the inherent calibration benefits found in intermediate representations. Subsequently, we apply a traditional post-hoc calibration method to the probabilities obtained from the intermediate representation ensemble. Formally, given ensemble probabilities , we apply one of the classical calibration transformations:

$$\mathbf{p}_{\text{final}} = \text{Calibrate}(\mathbf{p}_{\text{ensemble}})$$

This combined approach leverages the strengths of both calibration strategies, potentially leading to further improvements in multilingual calibration performance. Additionally, it allows us to systematically investigate whether intermediate representation-inspired calibration provides benefits orthogonal to well-established post-hoc techniques.

5.4 Calibration Results

All proposed methods substantially outperform the final-layer baseline. As shown in Table 2, our evaluation on the MMMLU dataset with LLaMA3 and Mistral clearly demonstrates the effectiveness of leveraging intermediate representations for calibration. All the proposed intermediate-layer aggregation methods achieve substantial improvements compared to the final-layer baseline. Notably, the final-layer method exhibits high calibration error (ECE > 25% for both models), which is markedly reduced by employing our intermediate representation-inspired ensemble approaches.

Aggregating signals from multiple well-calibrated layers yields the most robust results. Among the evaluated ensemble approaches, **Good Layers Ensemble (Hidden Avg)** achieves superior overall performance on the LLaMA3 model, attaining the highest AUROC (75.55%) and a significantly improved ECE (10.03%), along with the lowest Brier Score (19.96%). Similarly, for the Mistral model, **Best+Last Ensemble (Hidden Avg)** provides substantial improvements, resulting in a notable reduction in calibration error (ECE of 7.32%) and the best Brier Score (20.48%). These results underscore the clear advantage of combining representations from multiple well-calibrated intermediate layers, significantly outperforming traditional post-hoc calibration methods.

Orthogonal combinations of ensemble and post-hoc methods further enhance calibration. Importantly, we observe that combining intermediate-layer ensembles orthogonally with traditional post-hoc calibration methods leads to even greater improvements. For example, the **Best Ensemble combined with Histogram Binning** achieves the lowest ECE (5.26% for LLaMA3 and 4.40% for Mistral) among all methods tested. Moreover, **Best Ensemble combined with Isotonic Regression**

Method	LLaMA3 ECE ↓	LLaMA3 Brier Score ↓	LLaMA3 AUROC ↑	Mistral ECE ↓	Mistral Brier Score ↓	Mistral AUROC ↑
Baseline						
FINAL LAYER (32)	26.42	31.62	68.62	25.55	28.15	69.77
Post-hoc Calibration						
TEMPERATURE SCALING	15.04	22.70	64.83	16.48	23.51	68.84
ISOTONIC REGRESSION	15.49	22.26	64.51	6.75	21.05	69.55
HISTOGRAM BINNING	12.24	22.29	63.66	5.59	21.11	68.70
PLATT SCALING	17.00	22.26	64.56	6.41	21.29	69.77
Ensembling Methods						
BEST LAYER (29)	20.22	28.75	68.11	21.75	27.45	72.33
BEST+LAST ENSEMBLE (PROB AVG)	12.26	20.32	72.76	21.00	25.88	**73.32**
GOOD LAYERS ENSEMBLE (PROB AVG)	12.33	**19.84**	74.68	25.13	31.19	72.99
BEST+LAST ENSEMBLE (HIDDEN AVG)	**9.95**	20.28	74.36	**7.32**	**20.48**	71.99
GOOD LAYERS ENSEMBLE (HIDDEN AVG)	**10.03**	**19.96**	**75.55**	13.28	22.01	72.43
CONTRASTIVE DECODING	14.97	22.55	72.76	22.53	28.44	71.18
HIDDEN STATE STEERING	17.11	24.05	73.90	25.79	32.38	72.92
Orthogonal: Ensembling + Post-hoc						
BEST ENSEMBLE + ISOTONIC	6.11	**18.97**	**75.44**	6.83	**19.82**	**72.91**
BEST ENSEMBLE + HISTOGRAM	**5.26**	19.09	75.11	**4.40**	20.93	69.23

Table 2: Calibration methods performance for LLaMA3 and Mistral. Languages with accuracy below 20% are excluded from this analysis to ensure that calibration metrics are meaningful and not confounded by extremely low prediction performance. For LLaMA3, Best Ensemble is Good Layers Pooling (Hidden Avg) and for Mistral it is Best+Last Ensemble (Hidden Avg).

yields the highest AUROC (75.44% for LLaMA3 and 72.91% for Mistral) and the best Brier Scores across both models (18.97% and 19.82% respectively). These findings indicate that intermediate-layer ensembles and post-hoc calibration methods capture complementary aspects of model confidence and uncertainty, enabling more robust and reliable calibration. Our results thus highlight the value of adopting a hybrid strategy particularly in multilingual scenarios.

6 Conclusion

We present the first systematic evaluation of multilingual calibration on human-translated benchmarks, confirming that large language models are multilingually poor-calibrated, particularly for low-resource and non-Latin-script languages. Our key finding is that calibration quality does not monotonically improve with model depth; instead, for many languages, it peaks at intermediate layers before degrading at the final output. Motivated by this discovery, we propose a suite of novel methods that leverage these more reliable intermediate representations including layer ensembling, which can be orthogonally deployed with traditional post-hoc calibration methods. Our experiments demonstrate that these approaches substantially improve performance in calibration metrics, significantly reducing ECE and improving AUROC across diverse multilingual settings. Crucially, combining intermediate-layer ensembling with traditional post-hoc calibration methods yields complementary gains, delivering the most robust and reliable calibration outcomes. These results not only advance our understanding of calibration behavior in multilingual contexts but also offer practical guidance for deploying large language models reliably in linguistically diverse scenarios.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. On the calibration of massively multilingual language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.290. URL https://aclanthology.org/2022.emnlp-main.290/.

Jesujoba Alabi, Marius Mosbach, Matan Eyal, Dietrich Klakow, and Mor Geva. The hidden space of transformer language adapters. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6607, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.356. URL https://aclanthology.org/2024.acl-long.356/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 749–775. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.44. URL http://dx.doi.org/10.18653/v1/2024.acl-long.44.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models. *arXiv preprint arXiv:2410.01335*, 2024b.

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

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations?, 2024.

Common Crawl Foundation. Common Crawl. https://commoncrawl.org, 2025. Accessed: 2025-07-26.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive transformer, 2020. URL https://arxiv.org/abs/1910.10073.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.10.010. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X. ROC Analysis in Pattern Recognition.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of language model confidence estimation and calibration. *ArXiv preprint*, abs/2311.08298, 2023. URL https://arxiv.org/abs/2311.08298.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*

Long Papers), pages 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL https://aclanthology.org/2024.naacl-long.366/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.384. URL https://aclanthology.org/2024.naacl-long.384/.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. Can large language models achieve calibration with in-context learning? In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization, 2023. URL https://arxiv.org/abs/2210.15097.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406, 2021. doi: 10.1162/tacl_a_00433. URL https://aclanthology.org/2021.tacl-1.82/.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of*

the Association for Computational Linguistics: Main Volume, pages 2203–2213, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.188. URL https://aclanthology.org/2021.eacl-main.188/.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.820. URL https://aclanthology.org/2024.acl-long.820/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, and Kam-Fai Wong. Mlingconf: A comprehensive study of multilingual confidence estimation on large language models. *arXiv preprint arXiv:2410.12478*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. Logu: Long-form generation with uncertainty expressions, 2024b. URL https://arxiv.org/abs/2410.14309.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. On the calibration of multilingual question answering llms, 2023.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 08 2002. doi: 10.1145/775047.775151.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. LUQ: Long-text uncertainty quantification for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.299. URL https://aclanthology.org/2024.emnlp-main.299/.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. Atomic calibration of llms in long-form generations, 2024b.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.379. URL https://aclanthology.org/2024.acl-long.379/.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and Wenxuan Zhang. Babel: Open multilingual large language models serving over 90

A Limitations

We conduct experiments on mid-scale models (7B–8B parameters), leaving larger model sizes out of the current picture; larger models may exhibit different internal dynamics. Further, our focus is on standard multiple-choice and short-answer QA tasks as with such tasks model correctness is well-defined and easy to measure. The observed benefits of using intermediate layers may not directly extend to open-ended generative tasks such as dialogue, summarization, or long-form QA: we leave those tasks for future research. Finally, our proposed methods are post-hoc interventions that correct poor calibration, rather than fundamental solutions that integrate multilingual calibration objectives into the model's training process to address the issue at its root. This constitutes another very compelling direction for future research.

B Ethics Statement

Our research adheres to strict ethical guidelines. We verified the licenses of all software and datasets used in this study to ensure full compliance with their terms. No privacy concerns have been identified. We have conducted a thorough assessment of the project and do not anticipate any further risks.

C Additional Results on Multilingual Calibration Evaluation

In this section, we present the detailed multilingual evaluation results for the models and benchmarks discussed in the main text.

C.1 MMMLU Results

Language	AUROC	ECE	BRIER	Accuracy
Arabic	64.91	41.18	11.87	4.50
Bengali	64.56	49.70	11.72	0.10
German	70.84	24.14	29.32	43.00
English	73.75	23.92	27.95	54.00
Spanish	71.33	21.64	26.79	42.90
French	71.25	22.20	28.36	46.40
Hindi	75.08	39.77	6.23	1.60
Indonesian	69.48	26.98	29.69	38.80
Italian	74.08	25.24	28.25	44.50
Japanese	56.09	44.15	15.48	6.50
Korean	39.78	46.62	16.25	5.50
Portuguese	71.11	29.25	27.59	47.10
Swahili	56.02	30.81	27.34	26.30
Yoruba	44.79	44.18	21.99	16.10
Chinese	62.12	33.55	24.58	16.70
Avg. Low-Resource	62.47	38.77	18.14	14.57
Avg. High-Resource	65.59	30.08	24.95	34.07
Avg. Latin-Script	71.69	24.77	28.28	45.24
Avg. Non-Latin-Script	57.92	41.24	16.93	9.66
Average (All Languages)	64.35	33.56	22.23	26.27

Table 3: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in Mistral, evaluated on the MMMLU dataset.

- **Mistral Results** are provided in Table 3.
- **Babel Results** are provided in Table 4.

Language	AUROC	ECE	BRIER	Accuracy
Arabic	72.52	5.12	21.32	51.70
Bengali	69.42	14.08	19.35	31.00
German	75.66	8.22	19.85	57.00
English	77.38	7.09	18.61	65.50
Spanish	78.22	6.65	18.94	59.10
French	74.35	7.23	20.04	59.60
Hindi	64.91	16.07	22.01	37.20
Indonesian	79.00	5.22	18.64	56.80
Italian	77.86	4.74	18.92	59.50
Japanese	67.60	37.98	15.96	19.20
Korean	60.43	35.34	20.31	26.10
Portuguese	75.60	9.09	20.11	57.40
Swahili	66.53	6.04	21.65	38.80
Yoruba	18.59	50.08	25.27	5.50
Chinese	70.67	16.63	18.67	24.20
Avg. Low-Resource	61.83	16.10	21.37	36.83
Avg. High-Resource	73.09	14.77	19.05	47.51
Avg. Latin-Script	76.87	6.89	19.30	59.27
Avg. Non-Latin-Script	61.33	22.67	20.57	29.21
Average (All Languages)	**68.58**	**15.31**	**19.98**	**43.24**

Table 4: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in Babel, evaluated on the MMMLU dataset.

Language	AUROC	ECE	BRIER	Accuracy
Arabic	67.15	14.30	26.67	54.90
Bengali	64.10	26.68	31.98	33.20
German	76.94	21.59	25.08	55.60
English	78.23	15.77	19.25	65.60
Spanish	76.95	19.26	23.98	61.10
French	75.65	16.92	22.88	62.20
Hindi	72.01	28.73	28.86	33.90
Indonesian	75.69	15.83	23.53	54.30
Italian	75.32	21.07	24.46	58.70
Japanese	80.03	6.71	17.10	33.10
Korean	74.15	17.60	25.75	52.20
Portuguese	75.85	18.86	23.61	58.40
Swahili	59.93	30.12	33.09	32.30
Yoruba	23.49	46.99	36.11	2.00
Chinese	85.31	12.47	17.42	47.00
Avg. Low-Resource	60.40	27.11	30.04	35.10
Avg. High-Resource	77.60	16.69	22.17	54.88
Avg. Latin-Script	76.38	18.47	23.26	59.41
Avg. Non-Latin-Script	65.77	22.95	27.12	36.08
Average (All Languages)	**70.72**	**20.86**	**25.32**	**46.97**

Table 5: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in **Qwen**, evaluated on the MMMLU dataset.

- **Qwen2.5 Results** are provided in Table 5.
- **Phi Results** are provided in Table 6.
- **DeepSeek Results** are provided in Table 7.

C.2 Belebele Results

Complete multilingual results for LLaMA3 on the Belebele dataset are shown in Table 8.

Language	AUROC	ECE	BRIER	Accuracy
Arabic	52.66	30.21	25.35	36.50
Bengali	52.62	34.13	24.73	27.20
German	63.47	22.86	22.86	65.60
English	71.13	20.48	17.92	73.10
Spanish	61.29	27.15	25.32	56.40
French	71.57	17.07	20.21	68.90
Hindi	37.74	46.43	26.16	15.70
Indonesian	42.89	32.36	30.63	30.70
Italian	72.25	10.51	19.13	67.50
Japanese	30.62	46.69	17.59	8.30
Korean	66.95	29.00	24.50	50.00
Portuguese	73.79	13.24	18.77	66.60
Swahili	64.42	16.18	23.61	40.50
Yoruba	53.76	20.83	21.01	27.60
Chinese	59.73	31.98	26.17	44.60
Avg. Low-Resource	50.68	30.02	25.25	29.70
Avg. High-Resource	63.42	24.33	21.39	55.67
Avg. Latin-Script	65.20	20.52	22.12	61.26
Avg. Non-Latin-Script	52.31	31.93	23.64	31.30
Average (All Languages)	**58.33**	**26.61**	**22.93**	**45.28**

Table 6: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in Phi, evaluated on the MMMLU dataset.

Language	AUROC	ECE	BRIER	Accuracy
Arabic	55.33	32.74	21.54	26.40
Bengali	58.50	40.80	14.41	13.70
German	60.28	18.50	23.91	39.80
English	66.21	9.10	22.92	47.10
Spanish	62.24	12.47	23.51	40.80
French	62.93	10.84	23.12	41.40
Hindi	56.08	30.42	20.62	26.40
Indonesian	61.00	31.61	21.11	27.30
Italian	63.14	5.65	22.85	40.40
Japanese	55.56	18.05	23.14	32.10
Korean	21.56	49.09	18.66	1.10
Portuguese	62.37	16.78	23.26	39.10
Swahili	51.67	45.76	12.45	12.00
Yoruba	60.35	38.16	4.94	2.80
Chinese	69.00	16.13	23.97	43.10
Avg. Low-Resource	57.16	36.58	15.84	18.10
Avg. High-Resource	58.14	17.40	22.82	36.10
Avg. Latin-Script	62.60	14.99	22.95	39.41
Avg. Non-Latin-Script	53.51	33.89	17.47	19.70
Average (All Languages)	**57.75**	**25.07**	**20.03**	**28.90**

Table 7: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in DeepSeek, evaluated on the MMMLU dataset.

C.3 MKQA Results

For short-form datasets (MKQA and MlingConf), we experiment with three confidence elicitation approaches, following Xue et al. [2024]'s setup: (1) log probability of the generated answer sequence (***Prob***), (2) the probability of generating a "true" token given the question-answer pair (***True***), and (3) verbalized confidence (***Verb***), where the model explicitly articulates its confidence level. For all models, we restrict the answer format to short-form outputs by setting maximum response length to 48 during inference. We use PREM (Positive-Recall Exact Match) to evaluate accuracy, this is

Set 1					Set 2					Set 3				
Lang	**Acc**	**AUR.**	**ECE**	**Brier**	**Lang**	**Acc**	**AUR.**	**ECE**	**Brier**	**Lang**	**Acc**	**AUR.**	**ECE**	**Brier**
acm	58.2	78.4	8.7	19.2	arz	69.8	77.8	11.9	18.2	ceb	65.8	75.6	12.5	19.7
fin	80.8	75.6	10.4	14.1	hin	67.0	72.9	18.1	20.6	ita	86.0	79.6	10.0	10.8
khm	4.0	5.5	50.3	51.4	lvs	74.8	79.7	10.7	16.2	npi	59.8	74.8	5.8	20.0
pol	80.0	78.9	8.8	13.3	slv	81.0	76.6	9.2	13.8	swe	79.2	81.6	15.9	12.8
tso	34.0	62.5	9.9	21.5	xho	37.0	60.0	12.9	22.8	afr	80.8	81.2	10.9	12.7
asm	45.8	68.0	11.2	22.8	ces	82.2	80.2	10.4	11.8	fra	86.2	77.4	14.4	11.4
hin	57.0	72.1	4.2	21.1	jav	68.0	78.4	10.8	18.4	kin	34.8	63.3	12.7	21.3
mal	60.5	74.0	12.5	21.3	npi	32.5	59.9	11.6	21.4	por	86.2	79.8	16.0	10.2
sna	35.8	58.6	25.0	23.0	swh	67.0	75.0	8.2	19.1	tur	78.2	78.8	8.1	14.6
yor	31.2	61.1	12.0	20.4	als	73.5	78.0	7.4	16.2	azj	66.8	71.8	17.5	21.4
ckb	46.0	71.8	17.0	22.2	fuv	28.0	51.9	16.4	20.4	hrv	79.8	78.1	11.5	14.3
jpn	66.5	73.3	28.7	28.0	kir	63.0	73.8	27.8	24.7	mar	67.5	73.5	8.8	19.4
nso	37.8	60.3	7.5	22.7	snd	17.5	50.5	34.7	25.3	tam	65.5	73.4	13.4	21.5
ukr	84.2	77.4	13.3	12.8	zho	76.5	71.2	24.0	25.2	amh	34.8	63.2	18.4	21.4
bam	31.2	60.6	16.2	20.8	dan	79.8	78.8	9.1	13.9	gaz	31.8	53.1	20.3	21.7
hun	82.5	84.0	11.2	11.8	kac	30.2	61.8	9.9	20.5	kor	77.8	77.9	13.1	16.3
mkd	77.8	79.1	10.7	14.4	nya	32.0	60.7	12.8	21.1	ron	80.0	80.4	9.9	13.0
som	35.2	59.0	12.2	22.2	tel	59.5	73.7	10.4	20.4	urd	59.5	67.3	29.3	25.8
zho	81.2	68.9	23.1	21.1	apc	65.0	78.3	9.2	18.4	ben	65.5	72.6	9.7	20.3
deu	86.8	72.3	23.4	12.9	grn	39.8	65.5	10.2	22.4	hye	0.2	1.5	54.0	42.1
kan	58.5	72.1	6.4	20.9	lao	32.5	59.5	14.1	21.5	mlt	69.8	76.4	10.7	18.4
ory	55.8	70.0	14.4	24.7	rus	81.2	81.0	12.7	13.0	sot	32.8	57.5	17.1	21.8
tgk	63.8	70.1	11.9	22.1	urd	41.2	64.8	12.7	22.5	zsm	82.5	82.9	11.3	11.8
arb	79.5	74.2	10.3	15.5	ben	35.2	59.9	25.0	22.4	ell	80.5	80.6	10.6	13.7
guj	58.0	68.8	10.2	22.0	ibo	40.2	62.0	8.8	22.6	kat	1.5	6.0	55.0	51.6
lin	34.2	59.3	11.5	21.9	mri	35.5	63.3	6.2	21.5	pan	58.0	74.4	21.1	21.6
shn	16.8	47.9	32.9	16.9	spa	84.0	79.3	8.6	11.7	tgl	75.2	80.1	7.5	15.3
uzn	69.0	77.0	12.8	19.2	zul	36.5	59.3	8.2	22.8	arb	29.8	56.4	15.9	20.8
bod	29.0	60.2	12.1	20.3	eng	87.8	87.9	9.9	8.1	hat	55.8	72.9	6.1	20.9
ilo	54.0	69.8	17.9	22.6	kaz	63.5	75.6	20.9	23.0	lit	73.8	82.3	10.6	15.7
mya	0.8	7.2	58.0	55.7	pbt	47.5	67.7	8.6	22.6	sin	32.2	59.2	17.1	21.5
srp	83.2	77.5	15.1	13.7	tha	71.8	75.5	16.4	21.2	vie	83.5	78.4	10.0	12.4
ars	62.0	78.5	11.0	18.8	bul	80.8	77.7	12.7	14.6	est	71.8	78.2	4.2	16.4
hau	45.2	67.1	14.2	23.9	ind	81.8	75.6	8.4	13.4	kea	48.8	73.0	8.2	21.1
lug	35.5	57.6	11.2	22.5	nld	83.0	78.2	8.1	12.0	pes	79.2	77.8	8.0	14.4
sin	58.8	72.7	10.8	21.6	ssw	31.8	61.5	10.9	20.8	tir	28.0	57.7	12.4	19.7
war	62.2	74.7	12.7	21.0	ary	58.5	72.0	9.6	21.2	cat	86.2	84.8	9.5	10.2
eus	69.5	75.7	15.1	18.1	heb	77.2	76.9	22.3	17.8	isl	67.0	78.3	6.2	17.4
khk	48.0	70.2	10.7	22.9	luo	31.8	55.7	8.3	21.5	nob	79.2	77.8	9.6	14.7
plt	44.2	65.6	12.1	23.1	slk	82.0	76.9	9.0	13.0	sun	65.5	74.3	14.9	20.9
tsn	31.8	62.7	6.4	20.8	wol	33.0	53.2	20.7	22.3	Avg.	57.6	68.9	14.3	19.8

Table 8: Per-language performance on the belebele test set for the LLaMA3 model, reporting AUROC, ECE, and Brier score. Each row is color-coded by language category, based on resource availability (high, medium, low) and script type (Latin vs. non-Latin). The categories are shaded with soft pastel colors: high-resource Latin (light blue), high-resource non-Latin (light pink), medium-resource Latin (light green), medium-resource non-Latin (lavender), low-resource Latin (cream), and low-resource non-Latin (tan).

a relaxed evaluation metric that considers an answer correct if the predicted answer contains the reference or vice versa.

Results for LLaMA3, Mistral, and Qwen2.5 on the MKQA dataset are shown in Table 9.

D Additional Results on Layer-Wise Calibration Analysis

D.1 English Calibration improves as layer deepens

As shown in Figure 4, calibration in English steadily improves as the model progresses through deeper layers, with lower ECE observed alongside increasing entropy.

Language	LLaMA3				Mistral				Qwen2.5			
	Prob	**True**	**Verb**	**Acc.**	**Prob**	**True**	**Verb**	**Acc.**	**Prob**	**True**	**Verb**	**Acc.**
Arabic	26.16	57.02	42.06	7.62	49.90	48.32	47.07	1.35	49.23	48.50	46.79	2.61
Danish	15.26	38.63	30.41	34.54	38.18	38.00	43.83	29.06	40.11	55.92	41.82	14.08
German	13.90	34.77	27.66	37.84	35.79	37.28	37.49	31.61	42.05	53.42	40.57	15.98
English	11.86	20.73	27.79	43.01	40.18	35.07	36.90	37.07	43.41	47.68	43.55	16.68
Spanish	11.88	32.76	24.06	35.99	36.74	39.74	39.81	28.51	44.81	51.55	44.15	14.38
Finnish	17.77	36.13	29.78	31.03	37.07	30.89	36.04	22.44	36.90	55.08	36.71	15.33
French	13.48	31.04	28.16	37.04	31.92	36.58	43.95	31.61	46.27	51.68	42.75	13.23
Hebrew	33.97	49.33	50.16	8.67	50.39	48.98	48.28	0.95	40.19	50.54	43.72	3.06
Hungarian	17.10	42.23	40.36	30.33	36.75	38.44	38.52	23.15	39.59	53.47	38.78	11.82
Italian	17.53	32.80	31.28	35.19	35.79	34.18	45.41	31.51	46.39	52.68	44.27	12.93
Japanese	36.25	50.18	46.27	8.27	41.12	48.42	52.17	3.01	51.18	56.16	46.22	3.51
Khmer	52.01	69.72	51.77	0.35	58.62	49.92	48.42	0.05	59.30	65.01	47.29	0.40
Korean	29.12	51.92	48.52	7.17	47.48	48.74	41.59	1.85	51.90	50.20	46.95	2.45
Malay	14.62	28.65	31.80	36.29	34.96	36.53	39.47	28.01	36.30	50.96	41.72	19.44
Dutch	14.47	25.64	39.04	36.19	33.66	29.97	37.83	32.41	42.20	53.21	41.81	15.58
Norwegian	16.83	30.69	40.82	32.78	34.91	38.65	40.26	27.91	38.11	54.58	38.98	15.33
Polish	16.27	28.45	29.68	35.14	36.04	35.17	46.81	31.56	38.50	57.20	39.78	17.13
Portuguese	14.46	30.12	31.57	34.94	37.77	35.38	37.72	29.81	49.98	49.46	41.57	14.68
Russian	20.86	45.11	37.98	17.34	37.23	43.95	39.70	16.28	47.02	54.67	44.67	7.21
Swedish	14.93	30.79	39.36	31.83	37.09	33.42	38.03	29.01	38.14	51.20	44.04	13.98
Thai	45.94	63.49	49.56	4.91	41.97	47.64	53.47	1.55	54.93	46.70	47.39	2.45
Turkish	16.49	36.48	31.90	33.13	39.85	39.76	36.56	17.99	39.06	61.84	42.79	13.03
Vietnamese	15.01	29.73	33.08	35.34	39.53	37.53	48.19	17.69	42.81	51.95	42.84	12.17
Chinese (CN)	34.87	59.53	44.98	4.41	32.43	47.76	49.48	3.06	51.47	59.82	44.51	6.51
Chinese (HK)	36.24	57.55	43.36	5.96	43.87	49.37	43.69	2.20	49.18	56.02	45.95	4.46
Chinese (TW)	40.14	55.68	45.54	3.51	39.31	48.83	51.88	2.25	50.20	56.64	44.35	5.46
Average	**22.98**	**41.12**	**37.57**	**24.19**	**39.56**	**40.71**	**43.18**	**18.53**	**44.97**	**53.70**	**43.23**	**10.53**

Table 9: Per-language evaluation of model calibration and accuracy on the MKQA dataset across three models: LLaMA3, Mistral, and Qwen2.5. For each language, we report the ECE score of three uncertainty evaluation methods—*Prob*, *True*, and *Verb*—alongside accuracy.

D.2 Multilingual Calibration is Best at Late-Intermediate Layers

We visualize calibration performance across layers by plotting metrics against entropy on the MMMLU dataset. Across all models, we observe that ECE, Brier score, and AUROC improve (lower ECE/Brier, higher AUROC) at deeper layers before slightly degrading toward the final layers.

This trend is consistent in LLaMA3 (Figure 5), Cohere (Figure 6), Mistral (Figure 7), Phi (Figure 9), Deepseek (Figure 9). These findings support our hypothesis that multilingual calibration benefits most from late-intermediate layers rather than the final decoder output.

E Dataset Details

E.1 MMMLU Language Group Definitions

We group languages in the MMLU dataset according to resource availability and script as follows:

Low-Resource Languages Languages with relatively limited annotated data and pretrained model support: Arabic, Bengali, Swahili, Yoruba, Hindi, Indonesian

High-Resource Languages Languages with substantial resources and strong support in major multilingual models: German, French, English, Spanish, Chinese, Italian, Japanese, Korean, Portuguese

Latin-Script Languages Languages primarily written using the Latin script: German, English, Spanish, French, Indonesian, Italian, Portuguese

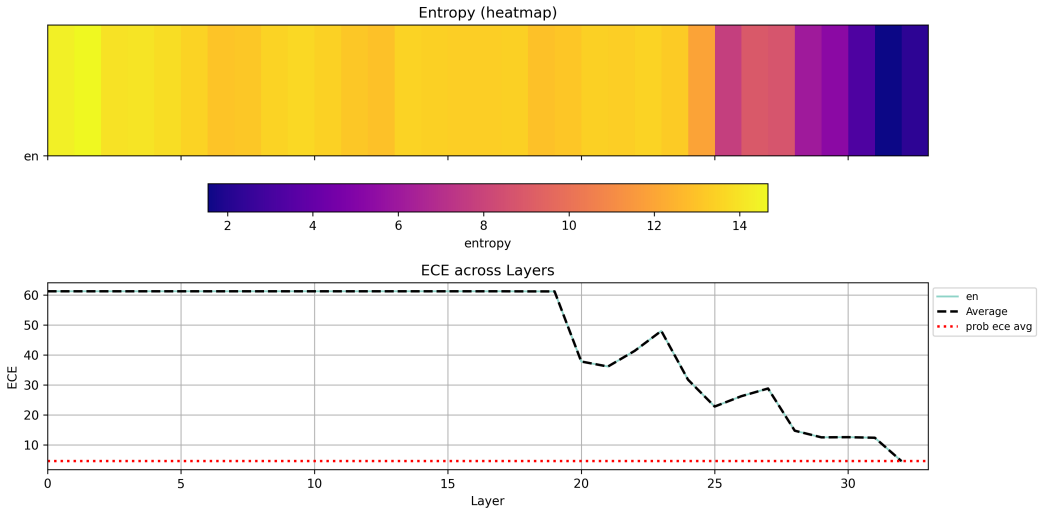

Figure 4: ECE vs. Entropy across layers in LLaMA3 on the MMMLU English subset.

Non-Latin-Script Languages Languages primarily written using non-Latin scripts (e.g., Arabic script, Devanagari, Hangul, Han characters): Arabic, Bengali, Hindi, Japanese, Korean, Swahili, Yoruba, Chinese

Note: Some languages fall into multiple categories. For example, Indonesian is both low-resource and Latin-script, while Chinese is high-resource and non-Latin-script.

E.2 Belebele

Belebele [Bandarkar et al., 2024a] is a multiple-choice machine reading comprehension (MRC) dataset covering 122 language variants, enabling robust evaluation of NLU across high-, medium-, and low-resource languages. The dataset is fully parallel, allowing for direct cross-linguistic comparison of model performance. In our experiments, we sample 400 examples per language and evaluate the LLaMA3 model using eight-shot inference, where eight in-context examples are provided for each test instance

F Baseline Detail

F.1 Temperature Scaling

We used temperature scaling as a baseline method on a held-out validation set sampled from the MMMLU dataset (1000 examples per language, non-overlapping with the main evaluation set).

We performed a grid search over temperatures in the range: Temperature_start = 0.5, Temperature_end = 1.5, At each temperature value, we evaluated model predictions with the ECE per language. The average ECE across languages was used to select the optimal temperature. Best temperature was found at 0.72. This value was then used to rescale the logits of all models before computing final evaluation metrics (AUROC, ECE, Brier score, Accuracy) on the test set.

F.2 Isotonic Regression

We applied **Isotonic Regression** as a post-hoc calibration method. The procedure was as follows:

1. An `IsotonicRegression` model from `scikit-learn` was fitted on the a held-out calibration set (1000 examples per language, same as the test set) using predicted probabilities and binary ground-truth labels.
2. The fitted model was used to recalibrate probabilities on the test set.

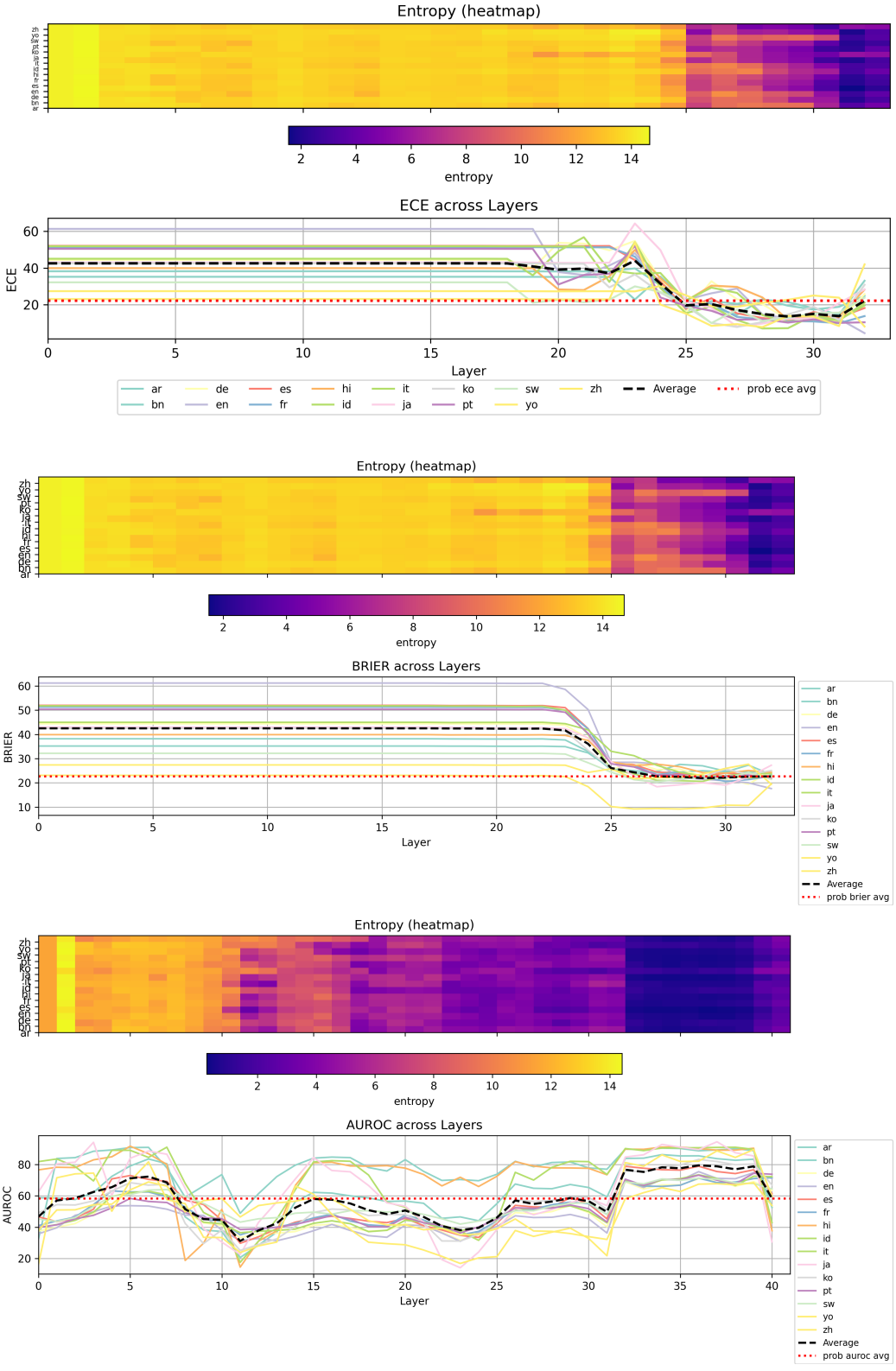

Figure 5: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU subset for LLaMA3.

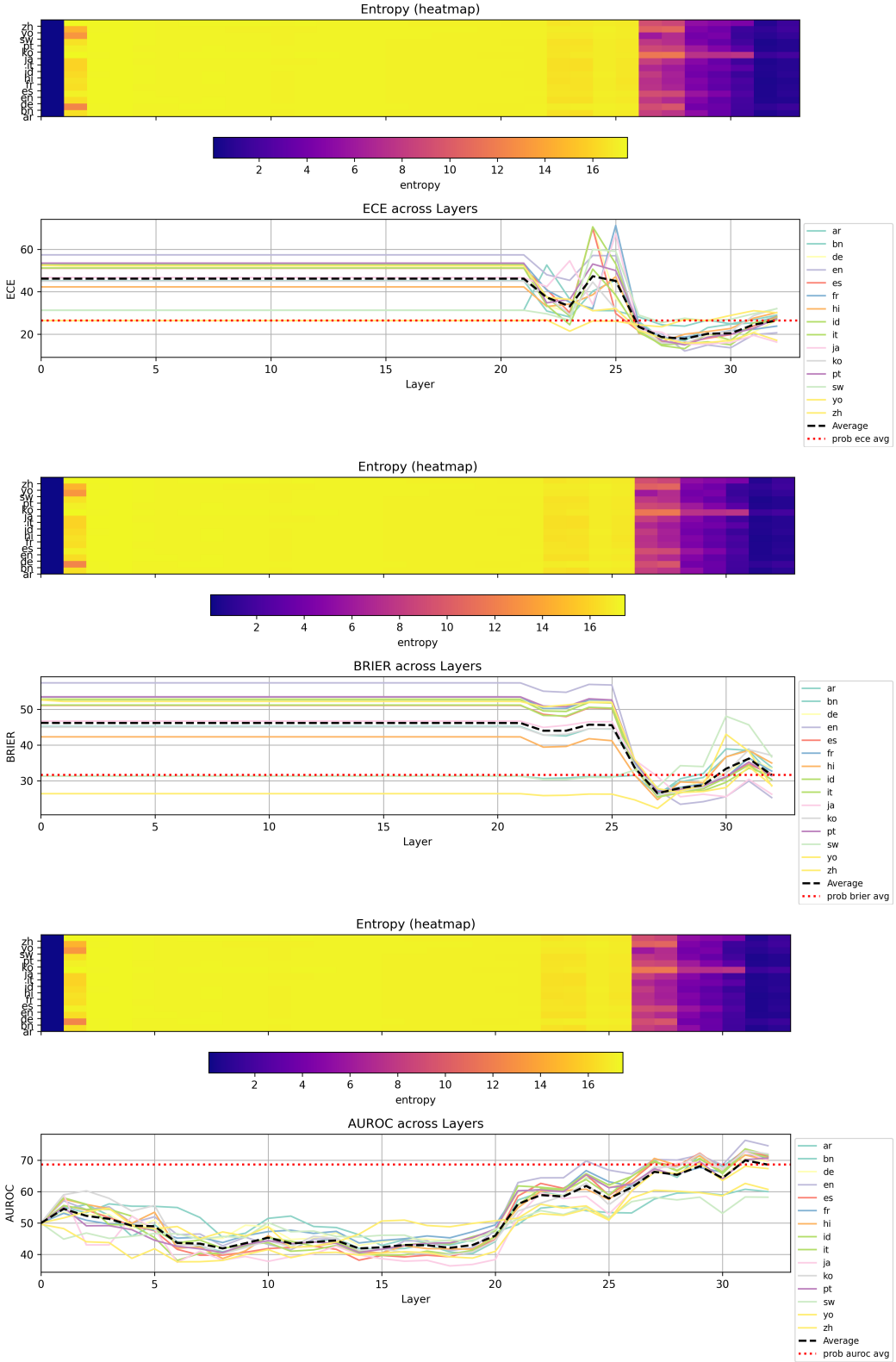

Figure 6: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Cohere.

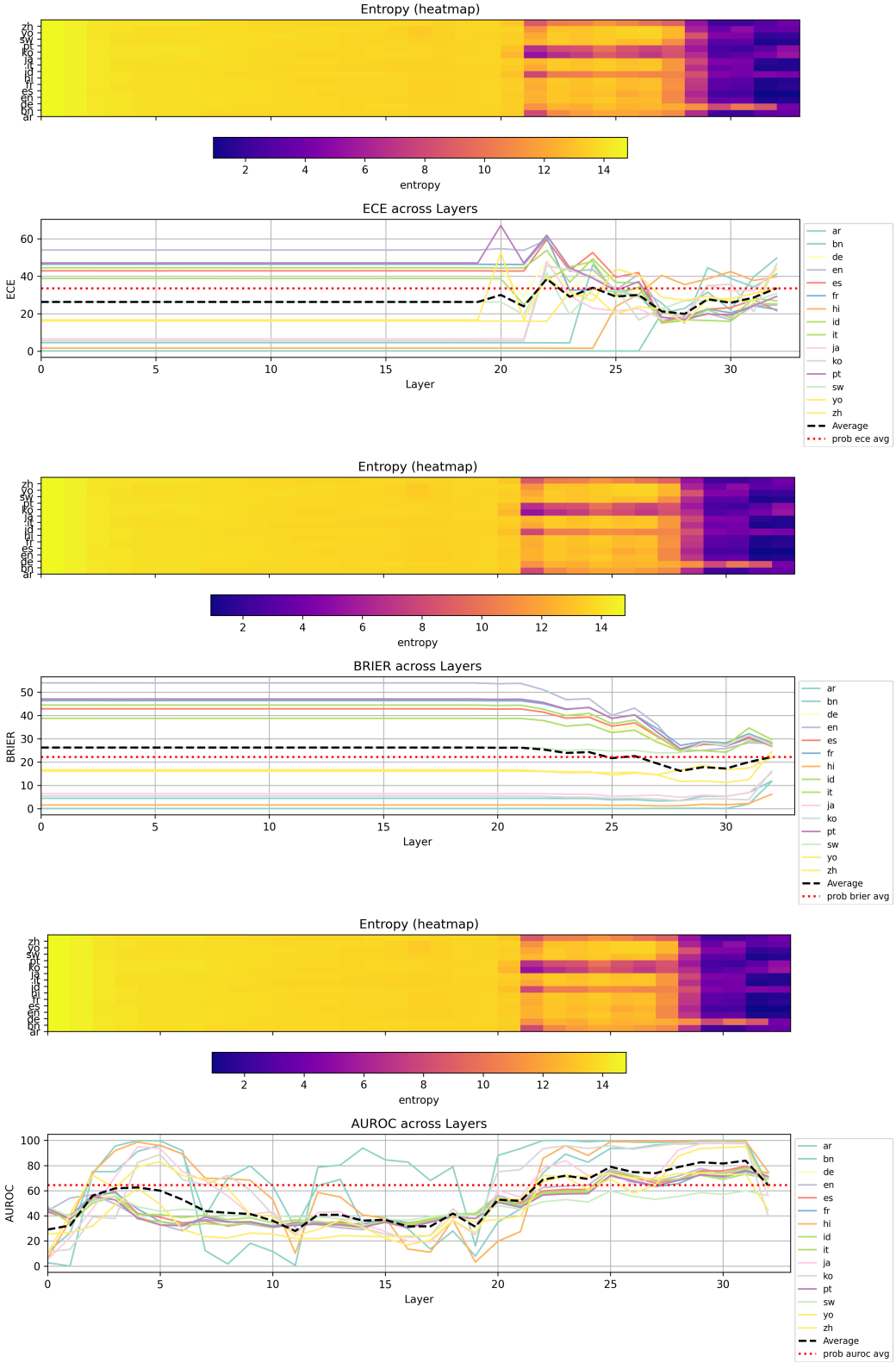

Figure 7: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Mistral.

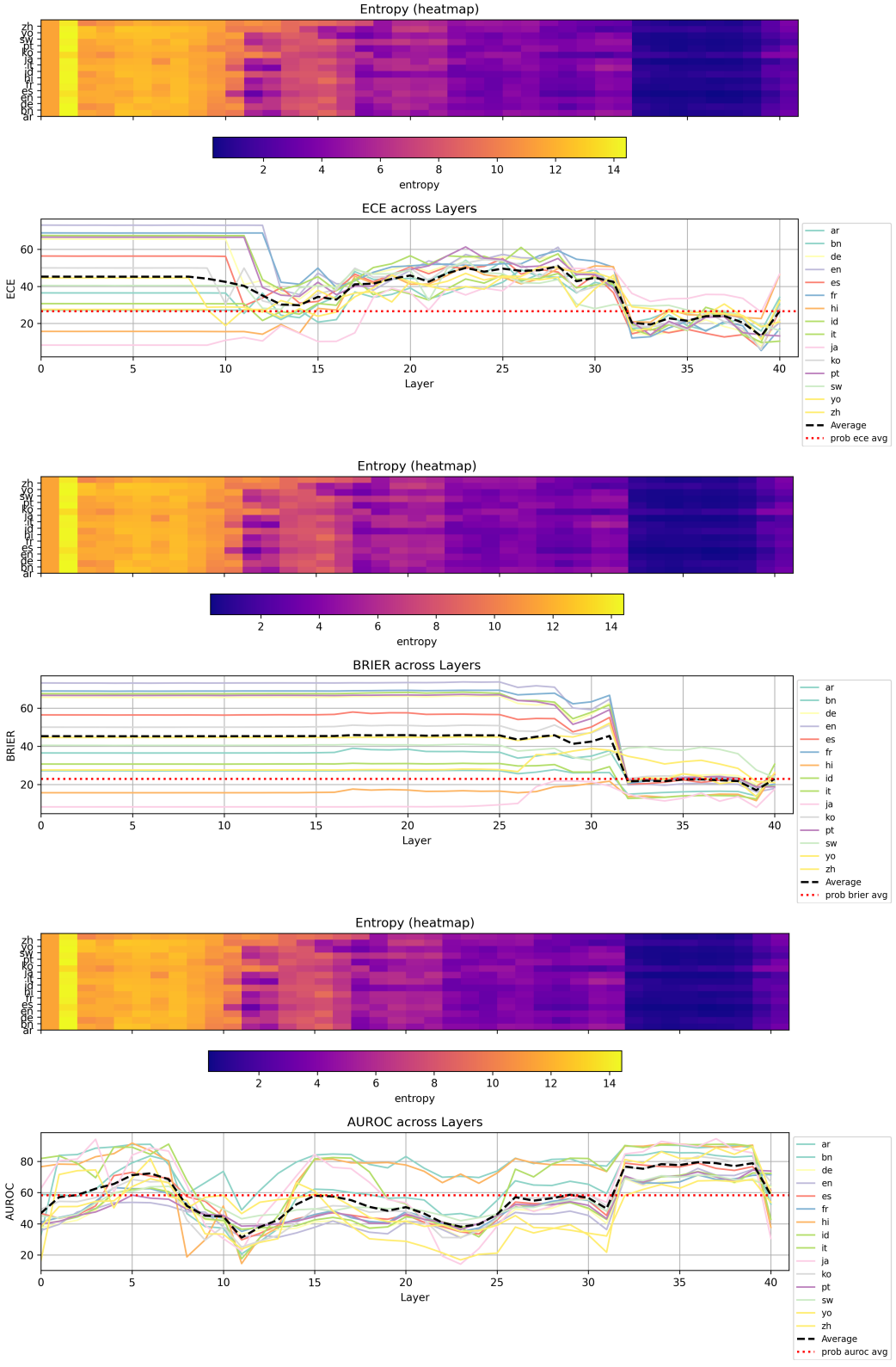

Figure 8: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Phi.

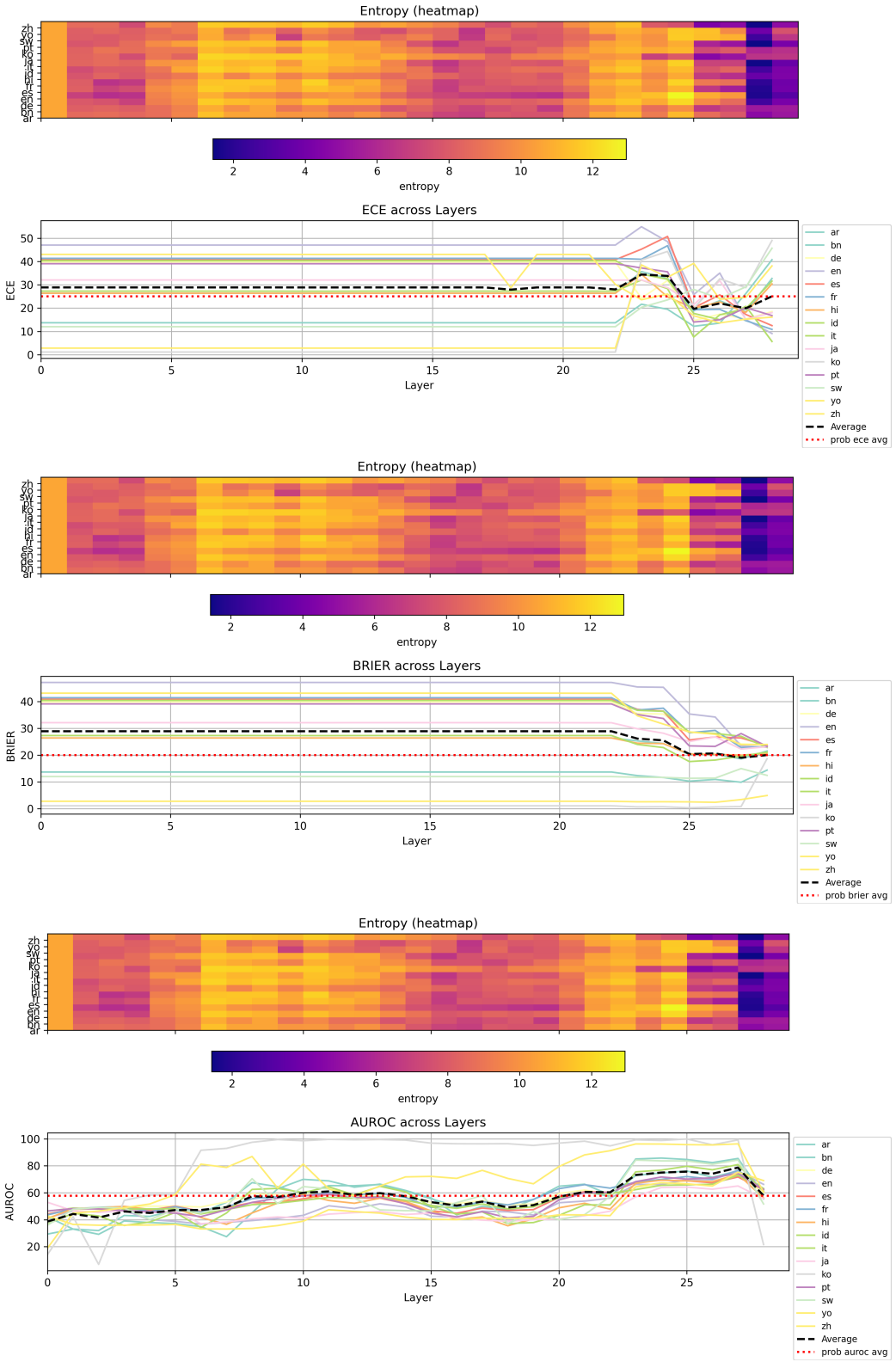

Figure 9: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Deepseek.

3. Calibration metrics (ECE, AUROC, and Brier score) were computed using the calibrated probabilities on the held-out test set.

F.3 Histogram Binning

We applied **Histogram Binning** as a non-parametric post-hoc calibration method. The procedure was as follows:

1. Predicted probabilities and binary ground-truth labels from the held-out calibration set (1000 examples per language) were used to construct the binning model.
2. The probability range $[0, 1]$ was divided into 10 equal-width bins. For each bin, we computed the empirical accuracy (mean label).
3. For inference, each predicted probability was mapped to the corresponding bin and replaced with the empirical accuracy of that bin. Empty bins were assigned a default value of 0.5.
4. The calibrated probabilities were evaluated on the test set using ECE, AUROC, and Brier score.

F.4 Platt Scaling

We applied **Platt Scaling** as a parametric post-hoc calibration method using logistic regression. The procedure was as follows:

1. A `LogisticRegression` model from `scikit-learn` was fitted on the held-out calibration set (1000 examples per language) using predicted probabilities and binary ground-truth labels.
2. The model was trained with the `"lbfgs"` solver and outputs calibrated probabilities via `predict_proba`.
3. The fitted model was used to recalibrate probabilities on the test set.
4. Calibration metrics (ECE, AUROC, and Brier score) were computed using the calibrated probabilities on the held-out test set.