Beyond the Final Layer: Intermediate Representations Improve Multilingual Calibration

Anonymous authors
Paper under double-blind review

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), yet remains underexplored in multilingual contexts. In this work, we present the first systematic evaluation of multilingual calibration on human-translated benchmarks. 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 conducted a layer-wise analysis and uncovered a consistent pattern: **intermediate layers often yield better-calibrated outputs than final layers**, especially for low-resource languages. Motivated by this finding, we introduce a suite of novel calibration methods that leverage these intermediate representations, including ensemble strategies and contrastive decoding. Our methods substantially improve ECE, Brier Score, and AUROC, outperforming the final-layer baseline by wide margins. These findings challenge the conventional reliance on final-layer decoding and suggest 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-language settings (Tian et al., 2023; Li et al., 2024; Zhang et al., 2024b), or relied on machine-translated datasets (Xue et al., 2024). 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. Therefore, in this paper, we systematically investigate multilingual calibration by addressing the following research questions: **RQ1:** Do existing multilingual models exhibit different calibration performance in different languages? **RQ2:** What are the reasons of certain languages show worse calibration in transformer-based models? **RQ3:** Can we develop methods to achieve more robust and consistent confidence estimation across languages?

We first empirically analyze popular LLMs (Llama, Qwen, Mistral, Babel) calibration status using human-translated datasets MMMLU and MKQA, covering both multiple choice and short-form QA in Section 3. We demonstrate that Low-Resource Languages are with lower

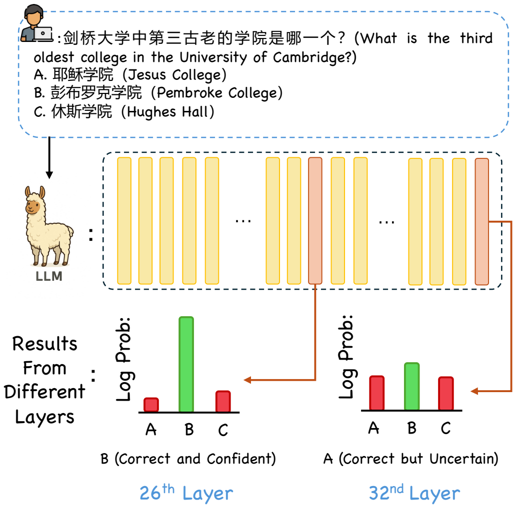

Figure 1: An LLM's layer-wise outputs for a question in Chinese. An intermediate layer (26th) correctly identifies the answer (B), while the final layer (32nd) becomes confidently wrong (A). This motivates our study of layer-wise calibration.

accuracy and lower calibration. Meanwhile, we point out that Latin languages show better calibration and accuracy compared with non-Latin languages.

Inspired by recent insights into layer-wise multilingual representations, we examine the calibration status for different layers to explore the reason behind last layer uncalibration. Recent study suggests that intermediate layers in LLMs encode cross-lingual semantic knowledge in a language-agnostic manner, whereas upper layers are typically language-specific (Bandarkar et al., 2024; Wendler et al., 2024). Leveraging this observation, in Section 4, we show that *different layers within multilingual models exhibit varying calibration quality across languages*. For low-resource languages, LLMs show better calibration results in intermediate layers, and dramatically turn bad in last layer.

Our finding inspired us to use intermediate layer representations to enhance calibration in multilingual LLMs, aiming to mitigate calibration disparities between high-resource and low-resource languages. In Section 5, we propose a series of novel calibration methods that leverage the intermediate layers to boost final calibration results. Our results demonstrate significant improvements in calibration performance, particularly for low-resource languages. This study provides valuable insights and methodological contributions towards achieving reliable multilingual calibration, paving the way for more equitable and trustworthy deployment of LLMs globally. 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 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., 2024). 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 Experiment Setup

Datasets and Models Previous work has mainly used machine-translated question-answering pairs (Xue et al., 2024), which may introduce potential biases. We therefore use human-translated datasets with both multiple-choice and short-form question answering: (1) MMMLU (Hendrycks et al., 2020) and (2) MKQA (Longpre et al., 2021). For our experiments, we evaluate a suite of recent large language models: Llama3-8B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Qwen2-7B (Yang et al., 2024a), and Babel (Zhao et al., 2025).

Confidence Elicitation Methods and Metrics For the MMMLU dataset, which consists of multiple-choice questions, we use the log probability of the chosen answer as the model’s confidence. For the MKQA dataset, which contains short-form answers, we explore three different confidence elicitation methods: (1) the log probability of the generated sequence (log prob), (2) the probability of the model generating a ”true” token after being presented with the question and its answer (ptrue), and (3) verbalized confidence where the model explicitly states its confidence level. To evaluate calibration and accuracy, we use four primary metrics: Area Under the Receiver Operating Characteristic Curve (AUROC), Expected Calibration Error (ECE), the Brier Score, and overall Accuracy.

Language	AUROC	ECE	BRIER	Accuracy
Arabic	61.00	33.06	24.37	38.20
Bengali	58.44	24.93	23.39	35.20
German	65.36	25.81	24.92	44.40
English	80.36	4.61	17.63	61.20
Spanish	71.65	18.21	21.89	52.00
French	71.39	13.87	22.75	51.30
Hindi	62.07	28.31	24.28	39.90
Indonesian	66.25	19.67	23.76	45.00
Italian	71.57	21.19	22.74	51.80
Japanese	61.73	28.36	27.27	43.00
Korean	62.59	30.86	25.06	42.50
Portuguese	71.37	10.51	21.76	50.40
Swahili	61.10	23.84	21.45	32.20
Yoruba	58.00	8.18	19.43	27.40
Chinese	50.63	41.94	19.56	23.10
Avg. Low-Resource	61.14	23.00	22.78	36.32
Avg. High-Resource	67.41	21.71	22.62	46.63
Avg. Latin-Script	71.14	16.27	22.21	50.87
Avg. Non-Latin-Script	59.44	27.44	23.10	35.19
Average (All Languages)	64.90	22.22	22.68	42.51

Table 1: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in LLaMA3, evaluated on the MMMLU dataset.

3.2 Results

Our evaluation, summarized in Table 1 for the LLaMA3 model on the MMMLU dataset, reveals notable performance disparities across various languages. We observe consistent patterns for Mistral 7B (Table 4), Qwen 2 7B (Table 6), and Babel (Table 5), which are provided in the Appendix.

LLM Calibration is Lacking in Low-Resource Languages As shown in Table 1, there is a clear trend of poorer calibration for low-resource languages. The average ECE for low-resource languages is 23.00%, which is substantially higher than the 4.61% ECE for English, indicating that the model's confidence scores in these languages are less aligned with the actual likelihood of correctness. Similarly, the average Brier score for low-resource languages is 22.78, again higher than that for high-resource languages. For instance, languages such as Arabic, Hindi, and Korean exhibit high ECE values of 33.06%, 28.31%, and 30.86%, respectively, underscoring this calibration challenge.

Low-resource languages show lower accuracy. A direct correlation between the resource level of a language and the model's accuracy is also evident. The average accuracy for low-resource languages is a mere 36.32%, starkly contrasting with the 61.20% accuracy achieved in English and the 46.63% average for high-resource languages. Languages like Swahili, Yoruba, and Chinese show particularly low accuracy scores of 32.20%, 27.40%, and 23.10%, respectively. This suggests that the model's reasoning and knowledge retrieval capabilities are significantly weaker in these languages.

Latin languages show better calibration and accuracy compared with non-Latin languages. Our results also highlight a performance gap between languages based on their script. 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 significantly 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%). This suggests that the predominantly Latin-character-based pre-training of many foundational models may disadvantage languages with different writing systems.

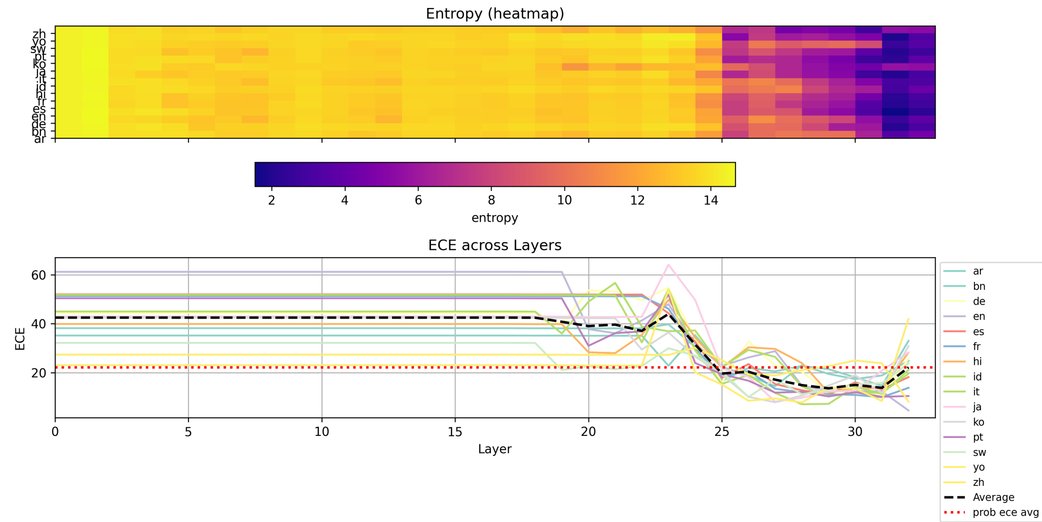

Figure 2: ECE vs. entropy across layers on the MMMLU subset for LLaMA3. In the multilingual setting, many languages achieve their lowest (best) ECE in intermediate layers (e.g., 22-26), after which calibration quality degrades towards the final layer. This contrasts with the English-only setting, where calibration improves monotonically (see Figure 3).

4 Mid-Layers Reveal Better Calibration

To understand the source of the poor calibration observed in the final layer, especially for low-resource languages, we investigate how calibration evolves throughout the model's depth. 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:

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

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 3 for Llama 3, we observe a clear and expected trend: calibration improves monotonically with layer depth. 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 2, **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 phenomenon is particularly pronounced for low- and mid-resource languages. 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 4), Cohere (Figure 5), Mistral (Figure 6), and others, 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 Low-Resource Calibration

Building on our observations from the previous section, we find that calibration performance often peaks at intermediate layers, particularly for low-resource languages. This 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 Layer-wise Calibration Methods

Method 1: Best Layer

From our empirical analysis (Figure 2), 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:

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

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.

Method 2: 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:

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

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

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

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.

Method 3: 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:

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

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

(1) Probability Averaging:

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

(2) Hidden State Averaging:

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

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.

Method 4: 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.

Method 5: 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.

Method	ECE ↓	Brier Score ↓	AUROC ↑
Best Layer (29)	13.51	21.92	73.01
Best+Last Ensemble (Prob Avg)	12.26	20.32	72.76
Good Layers Ensemble (Prob Avg)	12.33	**19.84**	74.68
Best+Last Ensemble (Hidden Avg)	**9.95**	20.28	74.36
Good Layers Ensemble (Hidden Avg)	10.03	19.96	**75.55**
Contrastive Decoding	14.97	22.55	72.76
Hidden State Steering	17.11	24.05	73.90
Calibration Head (Trained)	27.96	39.64	54.83
Final Layer (32)	22.28	22.79	64.56

Table 2: Calibration performance of proposed methods on MMMLU using LLaMA3. Lower is better for ECE and Brier; higher is better for AUROC. Best values in bold.

Method 6: Calibration Head Training

We propose training a lightweight MLP that operates directly on the best intermediate representation to predict a small set of target classes. Given the hidden state $\mathbf{h}_{\ell^*}$ from the best layer, we define a learnable projection head $W_{\text{cal}} \in \mathbb{R}^{C \times d}$, where C is the number of task-specific classes (e.g., $C = 4$ for MMMLU). The calibrated prediction is computed as:

$$\mathbf{p}_{\text{cal}} = \text{softmax}(W_{\text{cal}}\mathbf{h}_{\ell^*})$$

This calibration head is trained using a supervised loss (cross-entropy) on held-out data.

5.2 Calibration Results

Our proposed methods substantially outperform the final-layer baseline. As shown in Table 2, our evaluation on the MMMLU dataset with LLaMA3 confirms the effectiveness of our approach. This demonstrates a consistent advantage in moving beyond final-layer outputs for calibration.

Aggregating signals from multiple well-calibrated layers yields the most robust results. Among our methods, the **Good Layers Ensemble (Hidden Avg)** emerges as the top performer in overall metrics. It achieves the best AUROC (75.55) and Brier Score (19.96), supporting our hypothesis that combining the representations from multiple high-quality intermediate layers leads to more stable and reliable predictions.

A simpler ensemble of the best and final layers also offers strong performance. The **Best+Last Ensemble (Hidden Avg)** also proves highly competitive, securing the lowest ECE of just 9.95. This result is particularly compelling as it suggests that even a simple, two-layer combination can dramatically improve calibration without introducing significant complexity, making it a practical and effective solution.

Our findings confirm the value of leveraging intermediate representations. Ultimately, the results validate our central thesis: using intermediate representations—whether through direct selection, ensembling, or other decoding strategies—is a powerful technique for enhancing multilingual calibration. By empirically identifying and utilizing the better-calibrated parts of the model, we can mitigate the issues observed at the final layer.

5.3 Intermediate Representations Also Improve Accuracy

We find that better calibration can also lead to improved task accuracy. Beyond improving calibration, we investigated whether these intermediate representations could enhance task performance itself. To test this, we replaced the final-layer hidden state with the states derived from our top-performing methods (Best Layer, Best+Last Ensemble, and Good Layers Ensemble) and used them for final prediction without any re-training.

Language	True Acc. (%)	Best Layer (%)	Best+Last (%)	Good Layers (%)
Arabic	38.2	38.9	40.4	40.9
Bengali	35.3	34.6	35.5	37.4
German	44.6	47.7	49.1	51.0
English	60.8	60.3	61.1	61.3
Spanish	52.2	52.9	53.1	53.4
French	51.5	52.6	53.2	52.7
Hindi	39.0	39.6	41.1	41.6
Indonesian	45.1	46.2	46.5	46.7
Italian	51.9	54.8	54.4	55.0
Japanese	44.0	49.2	50.4	50.8
Korean	42.4	45.4	46.3	47.1
Portuguese	50.3	51.3	51.1	51.3
Swahili	32.3	37.9	37.6	37.6
Yoruba	27.0	29.4	29.8	29.9
Chinese	23.1	47.8	48.2	49.9
Average	**42.51**	**45.91**	**46.52**	**47.11**

Table 3: True accuracy vs. predicted accuracy across languages and calibration strategies on MMMLU (LLaMA3). Predictions are based the top-1 probabilities from each method.

The ensembling methods provide consistent accuracy gains across languages. The results, presented in Table 3, are striking. These alternative representations lead to consistent accuracy improvements across nearly every language. The GOOD LAYERS ENSEMBLE is again a standout, boosting the average accuracy to 47.11%—a 4.6% absolute improvement over the final-layer baseline (42.51%). This demonstrates that the benefits of our methods are not confined to calibration alone.

Improved accuracy likely stems from more robust and less noisy representations. This finding is particularly noteworthy because the hidden states were optimized purely for calibration, not accuracy. We hypothesize this dual benefit arises because: (1) intermediate representations retain richer multilingual signals before final-layer overspecialization, (2) ensembling averages out layer-specific noise, leading to more stable predictions, and (3) better-calibrated representations are inherently more discriminative, which directly aids task performance. This suggests that pursuing better calibration can be a pathway to more accurate and reliable multilingual models overall.

6 Conclusion

We present the first systematic evaluation of multilingual calibration on human-translated benchmarks, confirming that large language models are poorly 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 and contrastive decoding. Our experiments demonstrate that these approaches not only substantially improve calibration metrics such as ECE and Brier score but also yield significant gains in task accuracy across languages. This research challenges the conventional wisdom of relying solely on the final layer for multilingual generation and suggests a new direction for building more robust and equitable models by harnessing the well-calibrated knowledge within the network's intermediate layers.

References

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 (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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, 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*, 2024.

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

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 (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 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 (eds.), *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*, pp. 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 (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 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/.

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 (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 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/.

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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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.

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 (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 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 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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 Full Results

A.1 LLMs Are Not Calibrated in Low-Resource Languages

- **Dataset 1: MMMLU (Hendrycks et al., 2020)**
  - Table 1: LLaMA3 calibration metrics across languages
  - Table 4: Mistral calibration metrics across languages
  - Table 5: Babel calibration metrics across languages
  - Table 6: Qwen calibration metrics across languages

Language	AUROC	ECE	BRIER	Accuracy
Bengali	64.56	49.70	11.72	0.10
German	70.84	24.14	29.32	43.00
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	61.99	38.29	19.39	16.58
Avg. High-Resource	64.58	30.85	24.58	31.58
Avg. Latin-Script	71.35	24.91	28.33	43.78
Avg. Non-Latin-Script	56.92	41.25	17.66	10.40
Average (All Languages)	**63.61**	**33.74**	**22.56**	**25.76**

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

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
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	72.55	15.74	19.10	45.26
Avg. Latin-Script	76.78	6.86	19.42	58.23
Avg. Non-Latin-Script	61.33	22.67	20.57	29.21
Average (All Languages)	**67.99**	**15.77**	**20.08**	**41.81**

Table 5: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy using Babel 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
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	77.53	16.81	22.54	53.54
Avg. Latin-Script	76.07	18.92	23.92	58.38
Avg. Non-Latin-Script	65.77	22.95	27.12	36.08
Average (All Languages)	**70.13**	**21.27**	**25.79**	**45.67**

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

Dataset	Conf.	Avg ARC.	Avg ECE	Avg BRR	en ARC.	en ECE	en BRR	fr ARC.	fr ECE	fr BRR	ja ARC.	ja ECE	ja BRR	th ARC.	th ECE	th BRR	zh ARC.	zh ECE	zh BRR
SciQ	***Accuracy***	30.07	30.07	30.07	60.13	60.13	60.13	41.14	41.14	41.14	14.56	14.56	14.56	11.87	11.87	11.87	22.63	22.63	22.63
	Prob	73.01	24.83	25.46	71.73	6.90	20.74	74.05	15.72	23.80	75.19	31.60	25.79	74.89	38.52	29.19	69.17	31.43	27.77
	True	71.67	43.62	39.20	69.66	21.29	27.00	67.76	40.54	39.31	76.37	50.43	42.73	68.80	57.73	45.45	75.76	48.11	41.49
	Verb	62.68	31.51	40.44	67.02	21.60	29.51	60.63	25.03	32.94	67.37	38.54	51.89	65.23	40.27	50.31	53.17	32.12	37.56
common	***Accuracy***	35.28	35.28	35.28	75.35	75.35	75.35	46.68	46.68	46.68	15.12	15.12	15.12	17.31	17.31	17.31	21.94	21.94	21.94
	Prob	70.60	28.07	25.71	79.69	18.23	16.35	67.51	16.17	25.01	64.84	35.05	26.87	75.32	37.74	27.19	65.66	33.17	33.11
	True	64.81	33.07	33.33	63.49	4.98	17.94	64.89	27.44	30.79	70.99	50.48	41.30	56.31	38.69	38.74	68.35	43.77	37.87
	Verb	62.97	31.37	38.51	61.45	26.11	19.58	57.71	24.00	35.80	68.59	37.11	52.08	71.99	37.16	44.10	55.09	32.48	40.99
triviaqa	***Accuracy***	31.02	31.02	31.02	66.18	66.18	66.18	48.94	48.94	48.94	15.61	15.61	15.61	10.65	10.65	10.65	13.74	13.74	13.74
	Prob	82.73	24.08	21.27	80.48	10.82	17.27	77.91	15.64	21.85	87.45	23.70	18.09	88.57	34.23	21.14	79.22	36.02	28.01
	True	74.69	42.27	36.74	74.60	26.05	21.52	70.23	32.64	32.35	72.92	50.91	42.88	74.93	53.82	43.90	80.78	47.92	43.06
	Verb	71.16	33.87	41.05	78.94	21.18	23.47	70.05	30.50	32.78	69.39	34.05	50.25	73.09	40.08	55.85	64.31	43.56	42.92

Table 7: Experimental results of AUROC (ARC.), ECE and Brier on various datasets. meta-llama/**Llama-3.1-8B-Instruct** Accracy is RPEM.

- **Dataset 2: SciQ, Common, TriviaQA (Xue et al., 2024)**
  - Table 7: LLaMA3 PREM results in SciQ, Common, TriviaQA
  - Table 8: Mistral PREM results in SciQ, Common, TriviaQA
  - Table 9: Qwen PREM results in SciQ, Common, TriviaQA
  - Table 10: Babel PREM results in SciQ, Common, TriviaQA
- **Dataset 3: MKQA (Longpre et al., 2021)**
  - Table 11: MKQA results with ECE metrics with three models: LLaMA3, Mistral and Qwen

Dataset	Conf.	Avg			en			fr			ja			th			zh		
		ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR
SciQ	*Accuracy*	27.69	27.69	27.69	57.28	57.28	57.28	37.66	37.66	37.66	16.46	16.46	16.46	2.69	2.69	2.69	24.37	24.37	24.37
	Prob	73.10	35.03	29.86	76.15	26.51	24.86	69.93	29.62	31.42	77.09	37.33	29.48	66.07	47.93	33.31	76.26	33.76	30.22
	True	64.43	39.93	43.30	64.48	35.62	36.42	66.84	30.81	45.55	64.16	43.74	49.68	64.40	53.13	33.32	62.29	36.35	51.52
	Verb	64.41	40.54	43.77	62.27	39.66	36.07	67.27	30.05	44.74	64.15	43.67	53.36	63.18	45.55	34.86	65.19	43.77	49.80
common	*Accuracy*	48.74	48.74	48.74	74.13	74.13	74.13	49.83	49.83	49.83	40.38	40.38	40.38	30.68	30.68	30.68	48.69	48.69	48.69
	Prob	59.29	33.63	37.29	61.50	10.87	19.24	61.86	24.56	29.10	58.50	41.55	45.35	55.82	51.90	54.55	58.77	39.26	38.21
	True	56.61	27.87	40.24	55.74	23.29	23.39	56.33	26.54	41.31	55.47	30.69	48.42	58.62	30.96	44.59	56.91	27.86	43.51
	Verb	53.32	29.74	43.28	50.73	27.62	24.55	56.25	24.08	42.67	51.92	37.58	51.22	53.94	30.09	53.41	53.77	29.31	44.55
triviaqa	*Accuracy*	27.79	27.79	27.79	68.37	68.37	68.37	45.69	45.69	45.69	10.16	10.16	10.16	4.23	4.23	4.23	10.49	10.49	10.49
	Prob	81.37	30.68	23.39	74.67	15.56	20.08	73.76	28.22	26.45	86.30	33.92	25.14	84.16	40.71	22.12	87.95	34.99	23.14
	True	68.13	36.15	40.07	70.48	16.58	21.69	70.48	29.22	33.59	66.61	43.65	49.56	65.03	45.52	36.95	68.07	45.80	58.58
	Verb	66.82	43.66	45.39	72.04	30.88	23.79	69.83	43.53	36.83	62.96	44.59	57.21	59.76	49.28	50.41	69.49	50.04	58.71

Table 8: Experimental results of AUROC (ARC.), ECE and brier on various datasets. Inference & Confidence done on mistralai/**Mistral-7B-Instruct-v0.3**. Accracy is RPEM.

Dataset	Conf.	Avg			en			fr			ja			th			zh		
		ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR	ARC.	ECE	BRR
SciQ	*Accuracy*	39.69	39.69	39.69	62.82	62.82	62.82	43.83	43.83	43.83	27.69	27.69	27.69	23.58	23.58	23.58	40.51	40.51	40.51
	Prob	63.63	29.15	30.86	61.67	18.50	23.92	70.23	29.33	32.11	60.72	32.87	30.71	67.09	37.38	34.77	58.45	27.66	32.77
	True	46.81	34.37	50.24	51.69	28.70	34.67	53.32	35.19	50.81	46.89	39.36	61.41	35.35	37.08	53.83	46.82	31.50	50.50
	Verb	64.32	32.89	37.63	60.29	26.21	29.01	64.49	38.15	37.96	69.09	36.87	38.71	64.96	35.29	42.41	62.76	27.93	40.08
common	*Accuracy*	58.09	58.09	58.09	80.86	80.86	80.86	61.89	61.89	61.89	43.88	43.88	43.88	50.17	50.17	50.17	53.67	53.67	53.67
	Prob	63.51	28.33	29.88	75.66	23.28	14.28	59.12	28.71	29.88	60.09	27.75	34.86	65.51	38.67	38.15	57.18	23.26	32.21
	True	56.10	25.28	38.50	60.31	13.82	17.68	56.47	24.17	35.17	55.90	35.97	51.30	51.31	26.47	45.34	56.50	25.99	42.99
	Verb	62.79	18.70	28.21	71.27	14.53	13.67	58.61	22.99	27.97	62.69	21.22	34.98	60.13	19.02	32.52	61.26	15.75	31.90
triviaqa	*Accuracy*	25.56	25.56	25.56	45.93	45.93	45.93	30.08	30.08	30.08	15.45	15.45	15.45	14.31	14.31	14.31	22.03	22.03	22.03
	Prob	81.01	34.89	29.00	85.51	37.26	25.41	78.59	31.37	29.56	81.18	31.98	26.73	85.49	34.90	26.93	74.30	38.92	36.38
	True	42.71	40.39	61.27	58.44	35.99	49.94	48.62	38.55	65.48	34.81	47.95	68.56	35.44	35.48	57.25	36.24	44.00	65.11
	Verb	81.39	33.78	25.76	81.72	25.58	20.63	82.87	31.85	23.97	82.31	38.54	25.25	81.26	38.79	29.94	78.79	34.15	29.03

Table 9: Experimental results of AUROC (ARC.), ECE and Brier on various datasets. Inference & Confidence done on Qwen/**Qwen2.5-7B-Instruct**. Accracy is RPEM.

A.2 Layer-Wise Calibration Analysis

A.2.1 English Calibration improves as layer deepens

As shown in Figure 3, calibration in English steadily improves as the model progresses through deeper layers, with lower ECE observed alongside increasing entropy.

A.2.2 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 4), Cohere (Figure 5), Mistral (Figure 6), Phi (Figure 8), Deepseek-Qwen-Distilled (Figure 8) but not in Qwen3 (Figure 9). These findings support our hypothesis that calibration benefits most from late-intermediate layers rather than the final decoder output.

Dataset	Conf.	Avg ARC.	Avg ECE	Avg BRR	en ARC.	en ECE	en BRR	fr ARC.	fr ECE	fr BRR	ja ARC.	ja ECE	ja BRR	th ARC.	th ECE	th BRR	zh ARC.	zh ECE	zh BRR
SciQ	*Accuracy*	34.34	34.34	34.34	60.60	60.60	60.60	41.93	41.93	41.93	19.78	19.78	19.78	23.10	23.10	23.10	26.27	26.27	26.27
	Prob	71.00	24.46	28.14	69.29	8.49	21.57	75.70	20.72	25.98	66.87	27.20	27.44	82.70	35.47	34.16	60.46	30.41	31.56
	True	48.90	29.60	28.92	46.68	25.16	34.85	53.23	23.91	26.96	40.75	38.12	28.09	53.70	31.29	28.73	50.16	29.54	25.97
	Verb	57.45	35.25	49.28	55.82	24.05	32.54	61.23	36.89	44.27	58.70	38.54	56.52	52.43	38.98	58.02	59.08	37.81	55.07
common	*Accuracy*	43.74	43.74	43.74	78.32	78.32	78.32	53.23	53.23	53.23	18.01	18.01	18.01	41.35	41.35	41.35	27.80	27.80	27.80
	Prob	66.05	24.73	28.86	70.78	8.03	15.58	63.90	13.46	25.48	62.44	33.78	30.35	72.36	34.75	36.08	60.79	33.62	36.81
	True	47.04	32.03	35.74	52.95	34.79	41.75	50.73	26.60	30.31	40.51	36.08	38.85	51.27	25.14	30.30	39.75	37.55	37.50
	Verb	63.28	33.04	39.97	61.41	14.12	18.31	56.04	28.56	37.97	69.16	41.79	48.83	58.37	41.27	47.72	71.41	39.45	47.01
triviaqa	*Accuracy*	21.42	21.42	21.42	43.01	43.01	43.01	28.78	28.78	28.78	9.35	9.35	9.35	11.79	11.79	11.79	14.15	14.15	14.15
	Prob	81.01	29.73	25.07	85.11	21.57	22.05	80.54	20.36	23.01	82.68	33.41	21.93	85.93	33.59	26.97	70.77	39.74	31.40
	True	48.46	33.93	28.24	47.27	27.49	30.80	51.86	29.71	27.00	45.91	38.45	23.00	48.45	38.79	36.06	48.81	35.22	24.34
	Verb	63.74	40.79	48.32	71.67	32.54	33.68	66.35	35.29	43.16	61.58	45.51	52.28	56.29	45.51	57.85	62.80	45.12	54.64

Table 10: Experimental results of AUROC (ARC.), ECE and Brier on various datasets. Inference & Confidence done on Tower-Babel/**Babel-9B-Chat**. Accracy is RPEM.

Language	LLaMA3 Prob ECE	LLaMA3 True ECE	LLaMA3 Verb ECE	LLaMA3 Acc.	Mistral Prob ECE	Mistral True ECE	Mistral Verb ECE	Mistral Acc.	Qwen Prob ECE	Qwen True ECE	Qwen Verb ECE	Qwen 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 11: Expected Calibration Error (ECE) and Accuracy results across LLaMA3, Mistral, and Qwen on the MKQA dataset.

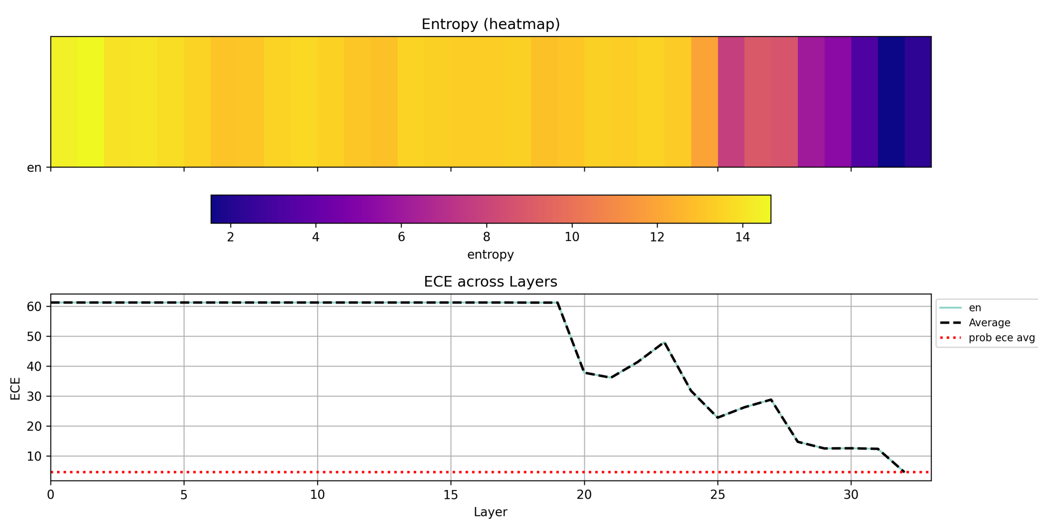

Figure 3: ECE vs. Entropy across layers in LLaMA3 on the MMMLU English subset.

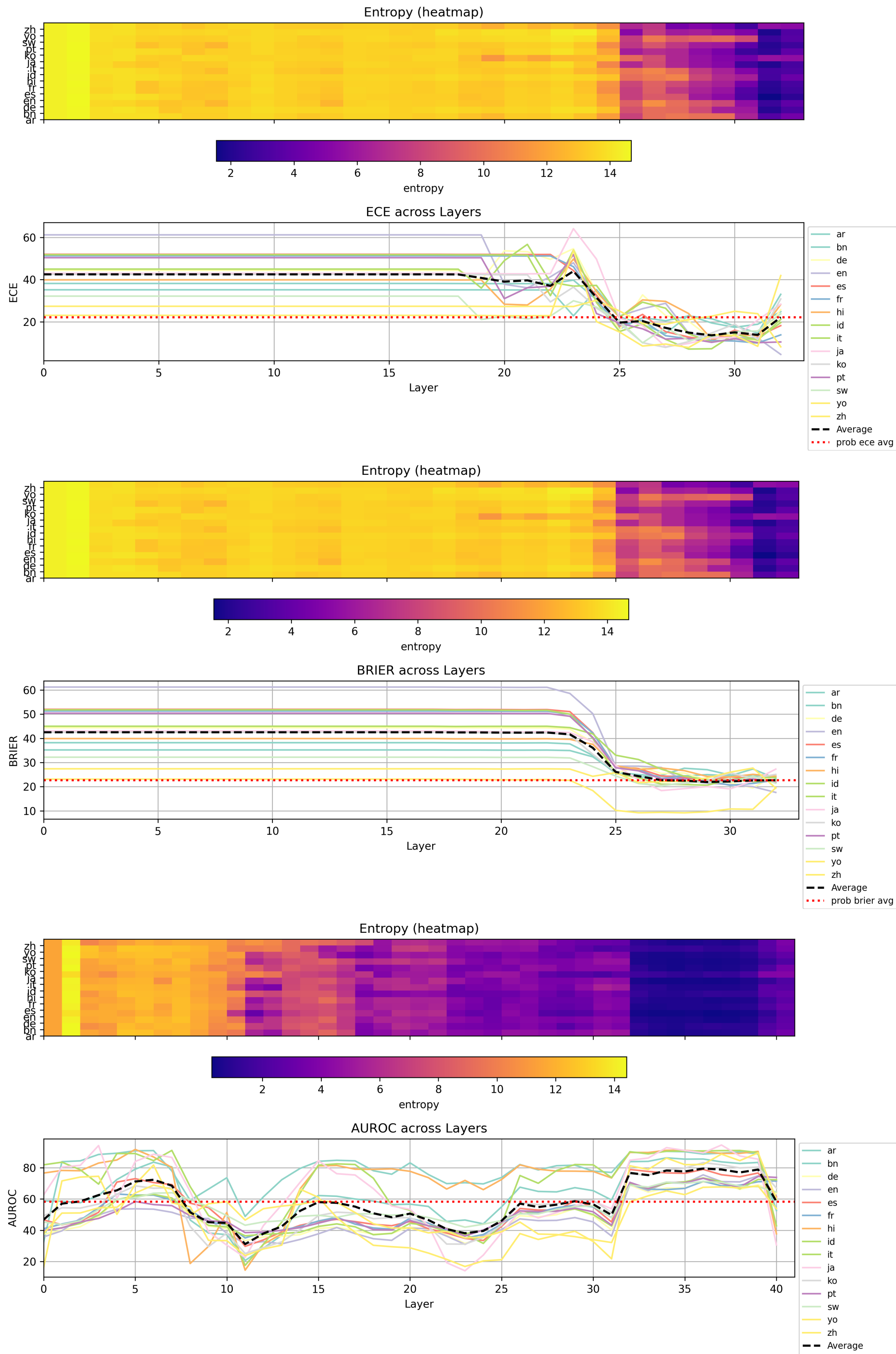

Figure 4: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU subset for LLaMA3.

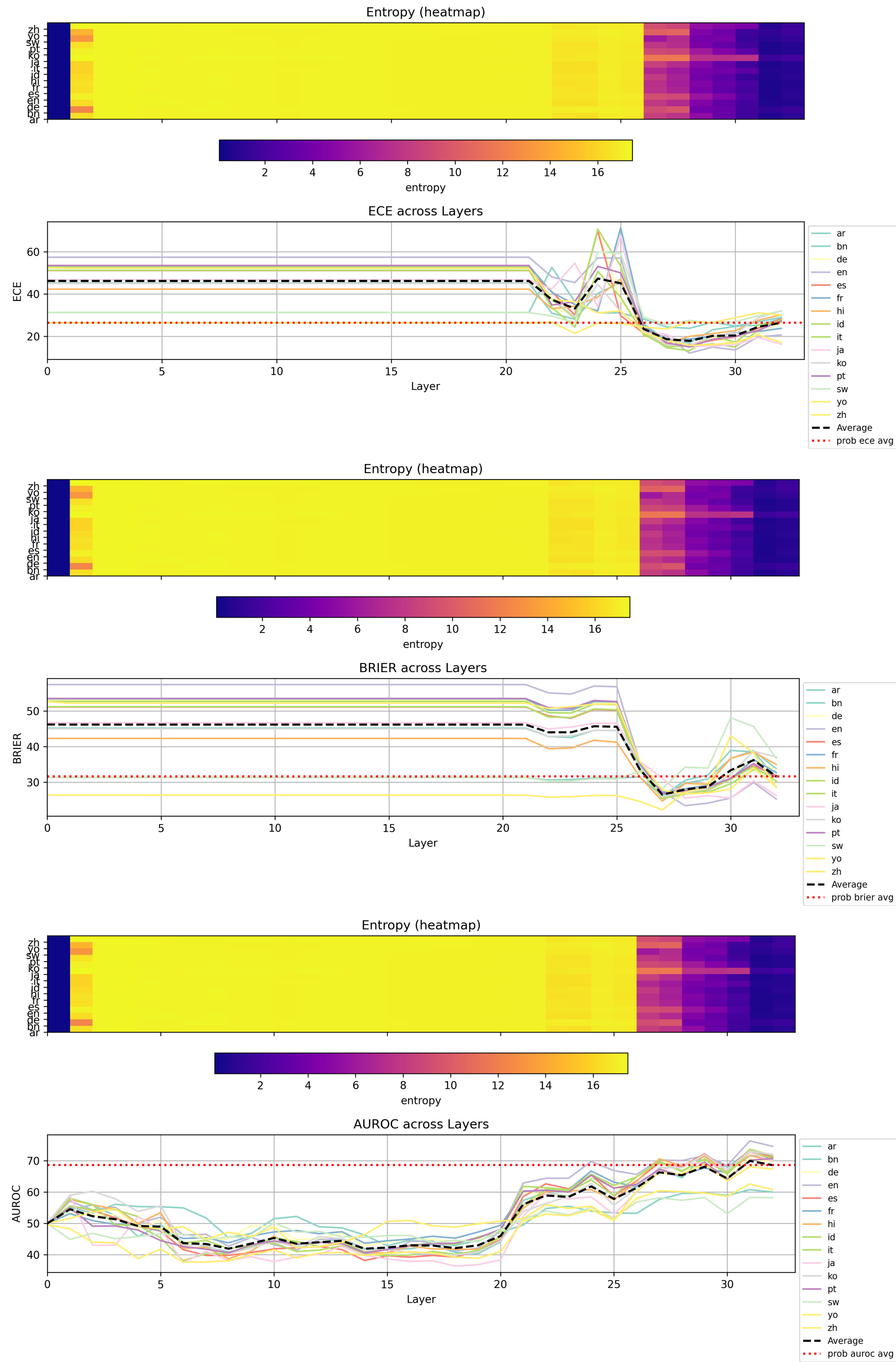

Figure 5: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Cohere.

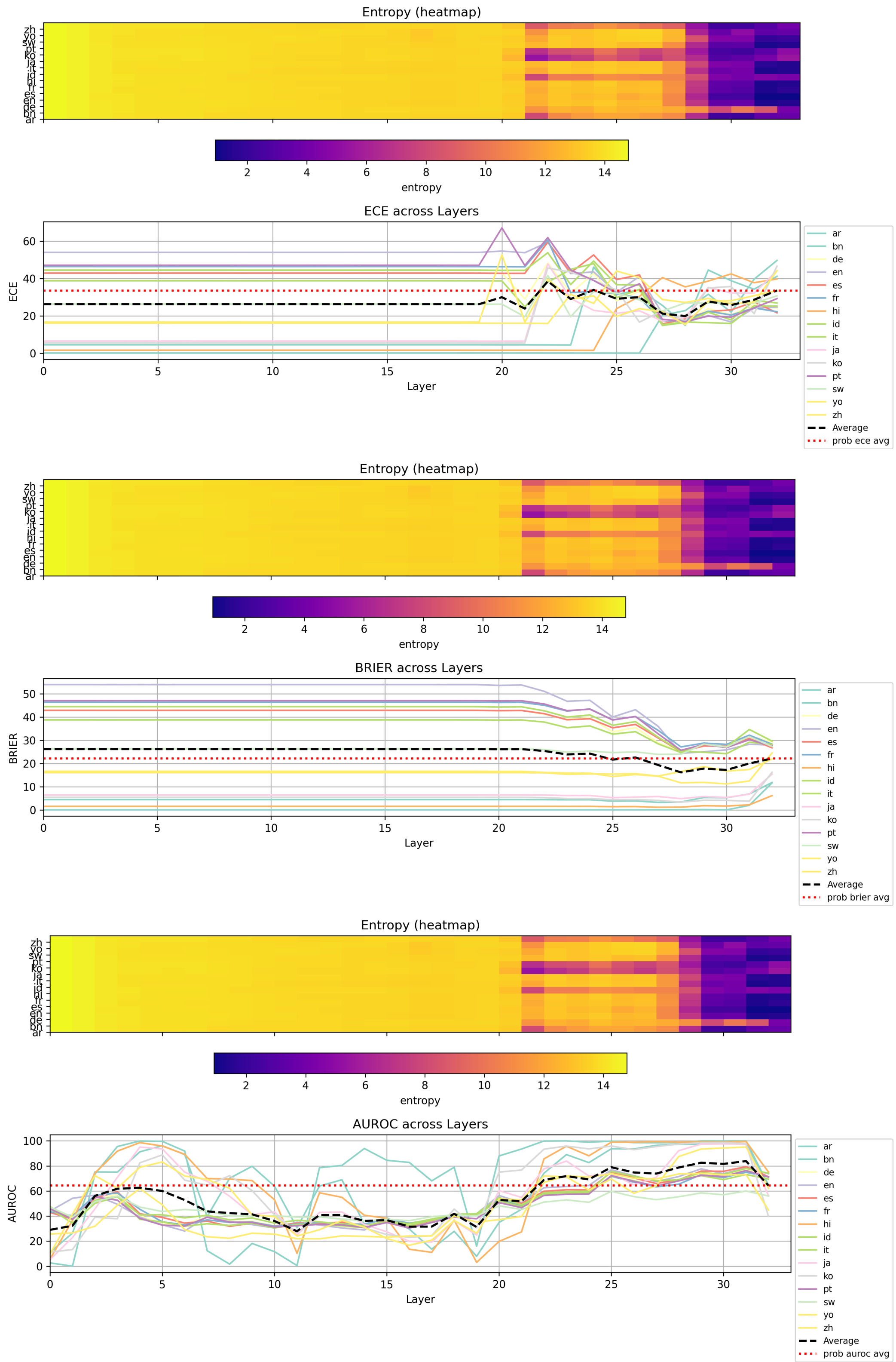

Figure 6: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Mistral.

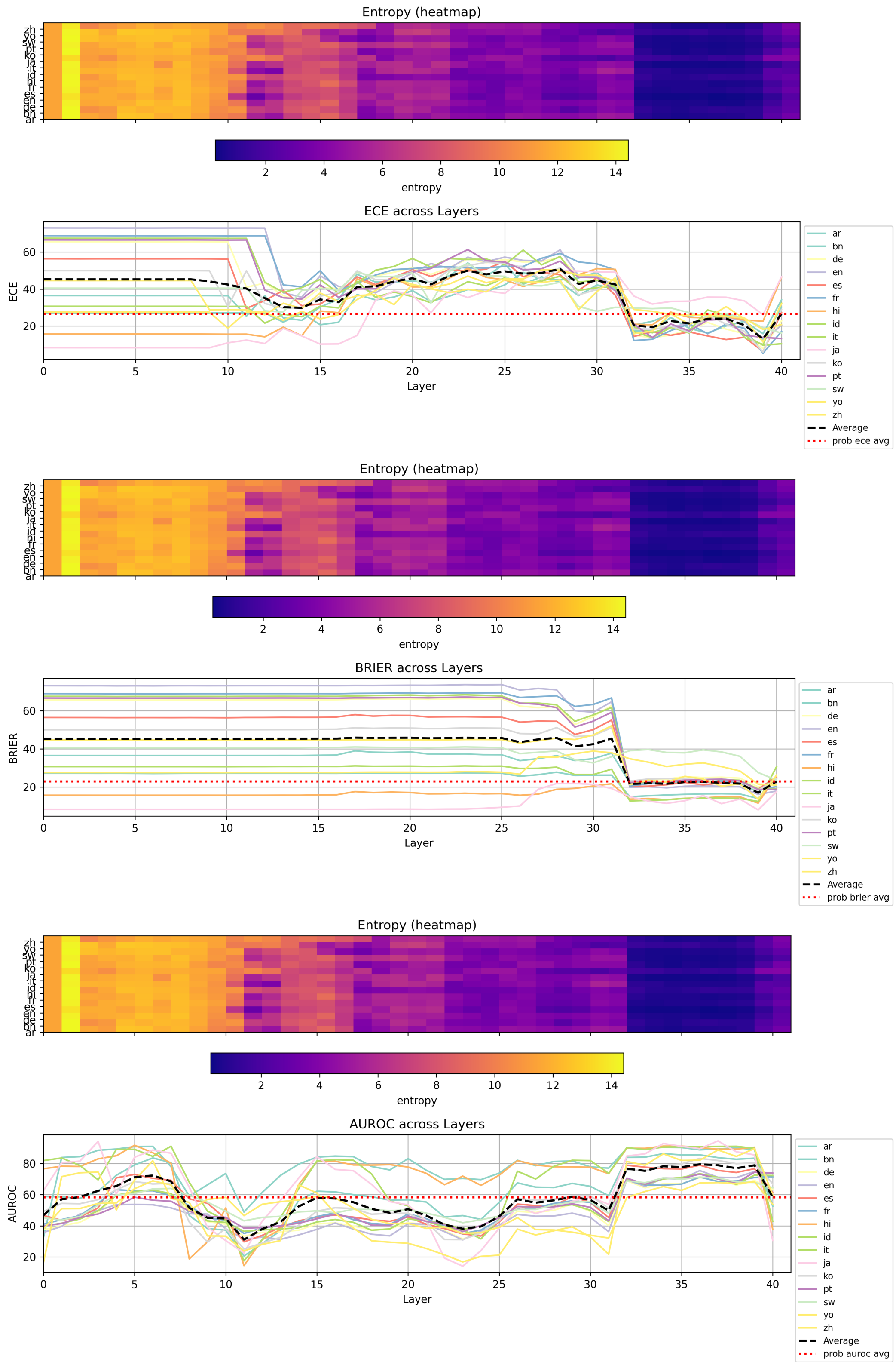

Figure 7: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Phi.

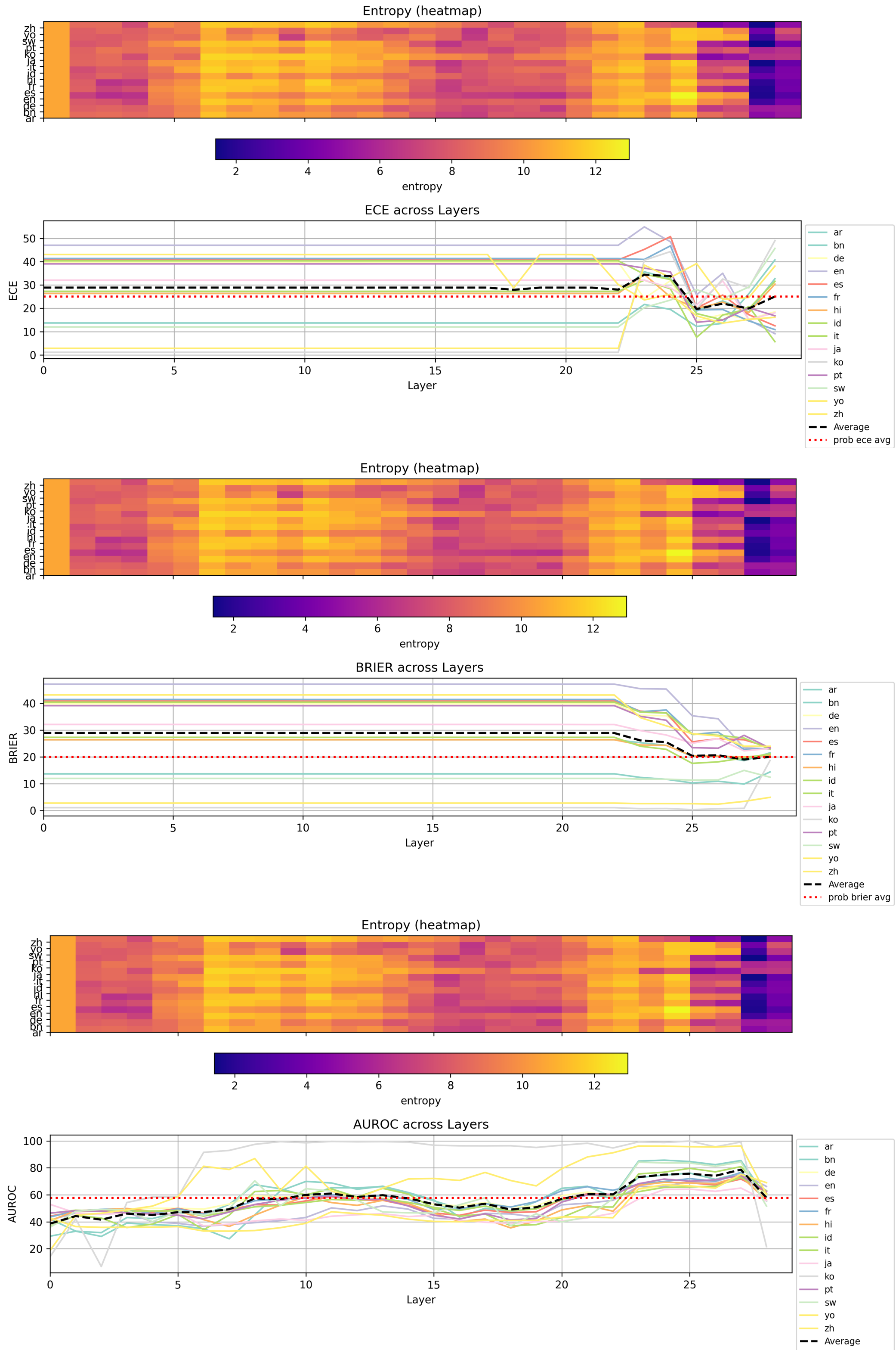

Figure 8: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Deepseek-qwen-distilled.

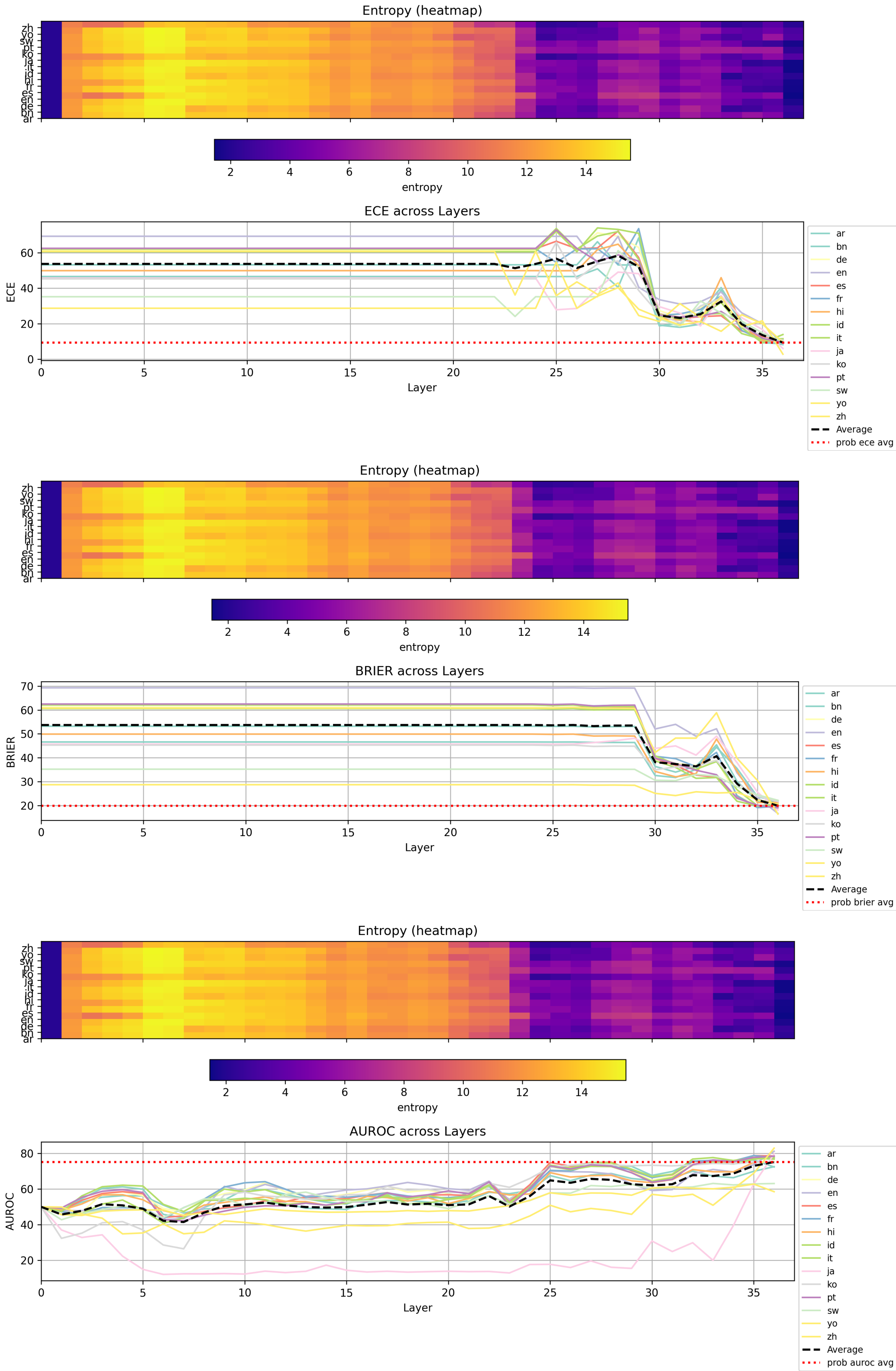

Figure 9: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Qwen3.