Beyond the Final Layer: Intermediate Representations for Better Multilingual Calibration in Large Language Models

Anonymous ACL submission

Abstract

Confidence calibration, the alignment of a model's predicted confidence with its actual accuracy, is critical for the reliable deployment of Large Language Models (LLMs), yet remains under-explored in multilingual contexts. We conduct the first large-scale systematic study of multilingual calibration across eight model families and over 100 languages, finding that non-English languages are worse-calibrated than English (e.g., LLaMA3's average Expected Calibration Error is 23.1% on non-English against 4.6% on English). To diagnose this, we investigate the model's internal representations and find that the final layer, biased by English-centric training, provides a poor signal for multilingual confidence, whereas late-intermediate layers offer a better-calibrated signal. Building on this, we introduce a suite of training-free methods, including Language-Aware Confidence Ensemble (LACE), which adaptively selects an ensemble of layers for each language. Both the multilingual calibration gap and the late-intermediate peak persist beyond MCQA, in open-ended QA, chain-of-thought reasoning, and alternative MCQA scoring protocols. Our study highlights the costs of English-centric alignment and offers a path toward more globally equitable LLMs.

1 Introduction

Calibration in machine learning denotes the alignment between a model's predicted confidence and the empirical probability that its predictions are correct (Guo et al., 2017; Geng et al., 2024; Zhang et al., 2025a,b).[1] A model is perfectly calibrated if predictions assigned 80% confidence are correct approximately 80% of the time. Calibration is particularly critical for large language models (LLMs) in high-stakes applications such as medical diagnosis, legal advice, and decision support, where miscalibration can amplify harm (Zhang et al., 2024b; Yang et al., 2024b, 2025b). Well-calibrated LLMs make their reliability explicit and interpretable, improving downstream trust and safety.

Despite its importance, most calibration work on LLMs has focused on English (Xue et al., 2024). This reflects what Ruder et al. (2022) termed *Square One Bias*: research often advances along a single axis (e.g., English alignment or multilingual coverage), while their intersection (multilingual calibration) remains under-explored. Prior studies also have practical limitations: they cover only a few languages, often rely on machine-translated data, and evaluate a limited set of models (Xue et al., 2024; Yang et al., 2023).

To address this gap, we conduct a large-scale systematic study of multilingual calibration. Our analysis spans eight major model families and over 100 languages, using human-curated benchmarks (MMMLU and Belebele) with $\sim 10^5$ instances. Non-English languages consistently exhibit lower accuracy and worse calibration: LLaMA3's average Expected Calibration Error (ECE) is 23.1% on non-English against 4.6% on English, a five-fold gap. The miscalibration pattern itself depends on training priorities. Prior work has largely framed LLM miscalibration as overconfidence (Zhang et al., 2024d; Chhikara, 2025); we observe two regimes. English-aligned models like LLaMA3 maintain calibration only for English; multilingual-first models like Aya are systematically over-confident across all languages, including English. Current alignment strategies thus fail to generalise to non-English contexts.

To identify the architectural source of this miscalibration, we examine the common practice of deriving confidence scores only from the final layer. Following work showing intermediate layers encode more language-agnostic representations (Ban-

[1] In this paper, we distinguish between calibration as a *property* and as a *process*. We use the term *calibration* to refer to the property of a model's confidence being well-aligned with its accuracy. In contrast, the methods used to achieve this alignment are referred to as *calibration methods*.

darkar et al., 2024b; Wendler et al., 2024), we hypothesise that the final layer, biased by English-dominated training, provides a sub-optimal signal for multilingual confidence. Our layer-wise analysis reveals a dichotomy: English calibration improves *monotonically* with model depth and peaks at the final layer, whereas for nearly all non-English languages, *late-intermediate layers* provide better-calibrated confidence estimates than the final layer.

This insight motivates our methodological contribution: training-free calibration methods that leverage these intermediate representations. We compare three confidence elicitation strategies: the *best layer* method selects the single most calibrated intermediate layer; the *good-layers ensemble* aggregates signals from multiple layers; and *LACE* (Language-Aware Confidence Ensemble) adaptively tailors layer selection to each target language. The three methods yield consistent multilingual calibration improvements across all models, and combining *LACE* with post-hoc methods such as Temperature Scaling (Guo et al., 2017) yields further gains. Finally, we verify that both the multilingual gap and the late-intermediate peak hold in free-form generation (open-ended QA, chain-of-thought) and under alternative MCQA scoring protocols (§6).

Our contributions are fourfold: (1) An empirical analysis of multilingual LLM calibration across eight model families and over 100 languages, revealing systematic disparities between English and other languages. (2) The first layer-wise investigation of multilingual calibration, showing that intermediate layers offer a more reliable calibration signal than the final, English-biased layer. (3) Training-free calibration methods that leverage intermediate representations and close part of the cross-lingual calibration gap. (4) Evidence that both findings hold beyond MCQA, in open-ended QA, chain-of-thought reasoning, and alternative MCQA scoring protocols.

2 Related Work

Multilingual Calibration Recent work has highlighted that modern LLMs, despite their strong performance, frequently struggle with calibration in their predictions (Xiong et al., 2024; Zhang et al., 2024c). Parallel studies document language-specific biases in LLMs (Zhang et al., 2024a; Qin et al., 2025; Zhou and Lu, 2026) and a broader linguistic hierarchy in the global AI ecosystem (Occhini et al., 2026). Yet calibration in multilingual settings remains underexplored. Ahuja et al. (2022) first established that multilingual models like mBERT and XLM-R are poorly calibrated, especially for low-resource languages like Swahili. Xue et al. (2024) conducted a confidence estimation study across various models, covering both language-agnostic and language-specific tasks, but datasets in their study included only 5 languages and were machine-translated which can potentially import bias (Vanmassenhove et al., 2021; Choenni et al., 2024). We extend prior research in two key dimensions. First, we provide more comprehensive evaluation using human-curated datasets across 100+ languages and eight LLM families. Second, while existing studies focus exclusively on final-layer outputs, we investigate the internal architectural origins of calibration through a novel layer-wise analysis.

Layer-wise Representations Recent work studies how layers specialize in multilingual transformers. Intermediate layers often encode cross-lingual semantics in a largely language-agnostic shared space (Bandarkar et al., 2024b). By contrast, later layers become more language-specific, adapting these representations to surface features such as syntax and word order. For predominantly English-trained LLMs (e.g., LLaMA), studies further suggest that multilingual inputs are mapped in middle layers to shared internal representations that often resemble English-like structures (Wendler et al., 2024; Kojima et al., 2024; Alabi et al., 2024). This view helps explain why prompting the model to "think in English" can improve cross-lingual generation: it matches the model's internal pathway before decoding back into the target language (Shi et al., 2023; Zhang et al., 2024e). Building on these findings, we examine how layer-wise specialization, especially the language-neutrality of intermediate layers, affects calibration across languages.

3 Benchmarking Multilingual Calibration

3.1 Experimental Setup

Task and Confidence Elicitation. We focus on multilingual Multiple-Choice Question Answering (MCQA) for two reasons. First, MCQA provides *unambiguous correctness labels*: in open-ended generation, correctness is approximated via metrics such as ROUGE, BLEU, or LLM-as-a-judge, which often disagree and can flip conclusions depending on the metric (Ielanskyi et al., 2025). Sec-

Language	LLaMA3				Aya			
	ECE↓	BRIER↓	AUROC↑	Accuracy	ECE↓	BRIER↓	AUROC↑	Accuracy
Arabic	33.06	24.37	61.00	38.20	28.41	33.79	71.49	45.20
Bengali	24.93	23.39	58.44	35.20	29.01	31.48	60.01	31.30
German	25.81	24.92	65.36	44.40	26.54	33.51	69.70	53.00
Spanish	18.21	21.89	71.65	52.00	28.17	31.86	71.12	51.10
French	13.87	22.75	71.39	51.30	23.80	32.72	70.69	53.40
Hindi	28.31	24.28	62.07	39.90	30.21	34.98	70.08	42.30
Indonesian	19.67	23.76	66.25	45.00	27.88	31.54	70.85	51.20
Italian	21.19	22.74	71.57	51.80	26.65	30.33	71.76	52.70
Japanese	28.36	27.27	61.73	43.00	16.30	26.26	69.92	46.70
Korean	30.86	25.06	62.59	42.50	32.07	37.09	72.06	45.00
Portuguese	10.51	21.76	71.37	50.40	27.33	31.42	70.71	53.50
Swahili	23.84	21.45	61.10	32.20	32.01	36.72	58.23	31.30
Yoruba	8.18	19.43	58.00	27.40	30.11	28.56	60.73	26.40
Chinese	41.94	19.56	50.63	23.10	17.12	28.75	67.35	52.20
English	**4.61**	**17.63**	**80.36**	**61.20**	**20.66**	**25.30**	**74.65**	**57.40**
Avg. Non-English	23.12	22.95	64.06	41.47	26.77	31.97	68.26	45.49
Avg. Low-Resource	23.00	22.78	61.14	36.32	29.60	32.84	65.23	37.95
Avg. High-Resource	21.71	22.62	67.41	46.63	24.29	30.80	70.88	51.67
Avg. Non-Latin-Script	27.44	23.10	59.44	35.19	26.90	32.20	66.23	40.05
Avg. Latin-Script	16.27	22.21	71.14	50.87	25.86	30.95	71.35	53.19
Average (All)	22.22	22.68	64.90	42.51	26.42	31.62	68.62	46.18

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

ond, MCQA admits a *clear definition of confidence* as the probability assigned to the selected option, whereas open-ended confidence estimation is method-dependent and unstable. Concretely, for each instance x with C choices we take the model's prediction $\hat{y} = \arg\max_i p_i$ at the final layer and elicit confidence as $\mathrm{Conf}(x) = \max_{i \in \{1,\ldots,C\}} p_i$. We later validate that our findings generalise beyond MCQA on generative tasks (§6) and across alternative MCQA scoring protocols (Appendix F.2).

Datasets and Models. We use two high-quality, human-translated benchmarks: (1) **MMMLU** (Hendrycks et al., 2021) (15 languages) and (2) **Belebele** (Bandarkar et al., 2024a) (122 languages). Together they cover a much larger range of languages and instances ($\sim 10^5$) than prior multilingual calibration studies. All experiments use eight-shot prompting in the respective language. We evaluate eight model families spanning 7–14B parameters: **LLaMA3**, **Qwen2.5**, **Qwen3**, **Mistral**, **Aya**, **DeepSeek**, **Phi**, and **Olmo3**. Full model identifiers and additional setup details are in Appendix A.

Metrics We evaluate calibration using expected calibration error (ECE; Guo et al., 2017) and the Brier score (Brier, 1950). To measure model's ability to discriminate between correct and incorrect predictions, we also report AUROC (Fawcett, 2006). We report ECE with number of bins $K = 10$; for bin size robustness, see Appendix F.1.

Model	ECE↓		Δ	AUROC↑		Δ
	EN	Non-EN		EN	Non-EN	
LLaMA3	4.61	23.12	-18.51	80.36	64.06	+16.30
Aya	20.66	26.77	-6.11	74.65	68.26	+6.39
Mistral	23.92	34.24	-10.32	73.75	63.67	+10.08
Qwen2.5	15.77	21.22	-5.45	78.23	70.18	+8.05
Phi	20.48	27.05	-6.57	71.13	57.41	+13.72
DeepSeek	9.10	26.21	-17.11	66.21	57.14	+9.07
Qwen3	12.09	13.65	-1.56	82.26	73.41	+8.85
Olmo3	23.70	27.24	-3.54	64.70	56.37	+8.33

Table 2: Per-model MMMLU calibration. Δ = EN − Non-EN for both metrics. Per-language results are in Table 1 and Tables 8–11.

3.2 Results

Table 1 reports per-language MMMLU results for **LLaMA3** and **Aya**; Table 2 extends the comparison to the English-vs.-Non-English gap across all eight model families.

Not only are LLMs more accurate but also more calibrated in English. Table 1 shows that non-English languages underperform English in both accuracy and calibration: LLaMA3's average ECE is 23.1% on non-English against 4.6% on English. The same pattern appears in Brier Score and AUROC and holds across all eight model families: Table 2 reports a negative ECE Δ and a positive AUROC Δ for every model.

Calibration correlates with language resource availability. The calibration gap also tracks re-

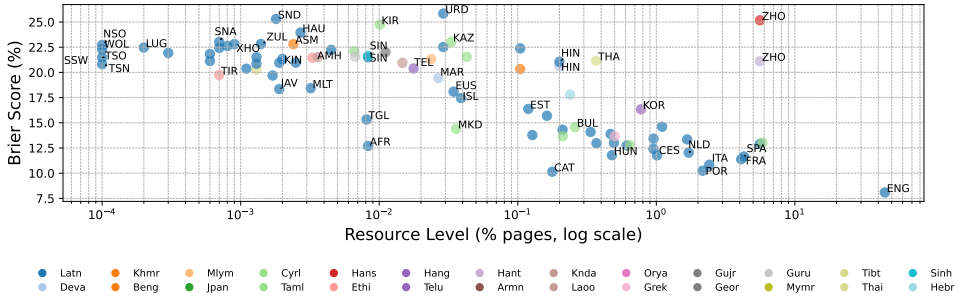
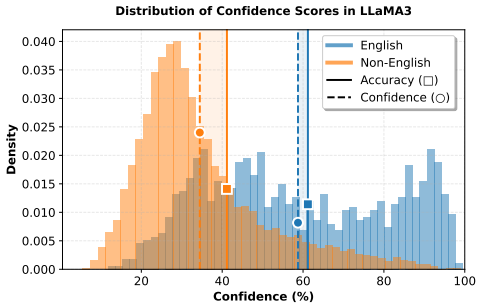
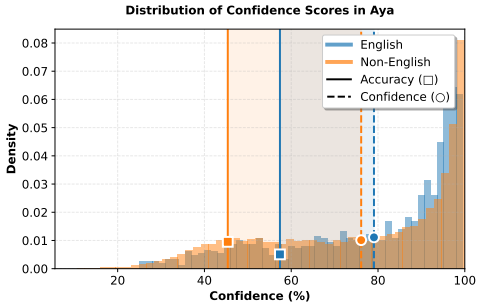


Figure 1: Brier score vs. resource level for LLaMA3 on Belebele. Each point is one language; colour indicates writing system. Correlations: Spearman $\rho = -0.59$, $p < 10^{-8}$; Kendall $\tau = -0.43$, $p < 10^{-8}$; Pearson $r = -0.39$, $p < 0.001$.

source availability: in Table 1, Hindi and Swahili are worse calibrated than the high-resource group. Plotting resource level[2] against calibration on Belebele (Figure 1; per-language tables in Appendix A.4) yields strong negative correlations with Brier (Spearman $\rho{=}-0.59$, $p{<}10^{-9}$) and positive correlations with AUROC ($\rho{=}0.66$, $p{<}10^{-12}$), suggesting that calibration is influenced by the representation of a language in the pre-training corpus.



(a) LLaMA3 confidence distribution



(b) Aya confidence distribution

Figure 2: Density of model confidence scores for English and Non-English samples. Dashed lines mark mean confidence; solid lines mark accuracy.

[2]LLaMA3 does not release its exact training data; we use Common Crawl (`CC-MAIN-2025-30`; Common Crawl Foundation, 2025) per-language page percentages as a proxy.

Miscalibration patterns reflect training and alignment priorities. Confidence distributions reveal two regimes. For LLaMA3 (Figure 2a), the English setting is well-calibrated (ECE = 4.6) while the non-English setting shows underconfidence. Aya's distributions in English and non-English are similar in shape (Figure 2b) but right-skewed (overconfident) in both. These patterns track training priorities. LLaMA3's alignment pipeline (supervised fine-tuning, preference ranking, safety) (Grattafiori et al., 2024) is primarily English-focused. Aya prioritises multilingual coverage (Üstün et al., 2024; Dang et al., 2024) with less attention to calibration. The two regimes echo the *Square One Bias* (Ruder et al., 2022): progress along either axis (English alignment or multilingual coverage) leaves the orthogonal dimension under-served. Confidence behaviour of other models is in Appendix A.3.

4 Mid-Layers Reveal Better Calibration

Inspired by recent insights in layer-wise multilingual representations (Bandarkar et al., 2024b; Wendler et al., 2024), we examine how confidence evolves throughout the model's depth to understand the source of the poor calibration observed in §3. Specifically, we ask: *is it possible to identify intermediate representations that elicit better calibrated confidence scores w.r.t. final-layer accuracy?*

4.1 Early-Decoded Confidence Estimation

We adopt a layer-wise probing technique inspired by the early exiting paradigm (Elbayad et al., 2020) to offer a new way of confidence estimation. Instead of applying the modelling 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

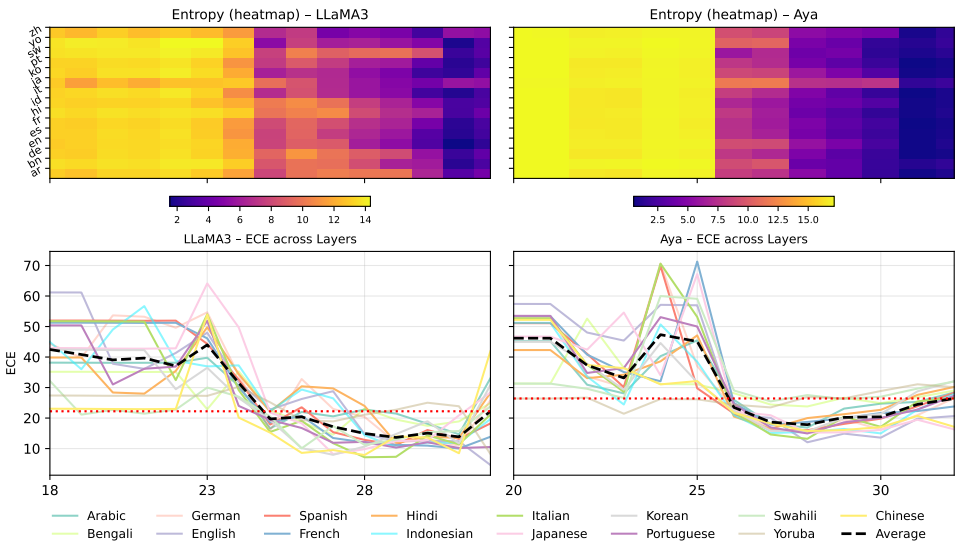


Figure 3: ECE vs. entropy across layers on the MMMLU subset for LLaMA3 and Aya. See Figure 7 for the English-only curve.

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:

$$\mathbf{p}_\ell = \mathrm{softmax}(\mathbf{z}_\ell), \quad \sum_{v=1}^{V} [\mathbf{p}_\ell]_v = 1.$$

With the $\mathbf{p}_\ell$, we *trace* the token ultimately predicted at the final layer, $\hat{y}_L = \arg\max_v [\mathbf{p}_L]_v$, back through the intermediate layers. At each layer ℓ, we then define the confidence score as the probability mass that this layer assigns to the final prediction $\mathrm{Conf}_\ell(x) = [\mathbf{p}_\ell]_{\hat{y}_L}$. Calibration is then evaluated by comparing these $\mathrm{Conf}_\ell(x)$ with the prediction accuracy determined at the final layer $\hat{y}_L$:

$$\mathrm{ECE}_\ell = \mathrm{ECE}\Big(\{(\mathrm{Conf}_\ell(x),\ \mathbf{1}\{\hat{y}_L = y\})\}\Big).$$

where $\mathbf{1}\{\hat{y}_L = y\}$ is the indicator function of whether the final-layer prediction is correct.

4.2 Layer-wise Multilingual Calibration Results

Multilingual settings reveal best calibration in late intermediate layers. The conventional understanding (Tenney et al., 2019) holds that intermediate representations become progressively more refined and task-specific, so calibration should improve with depth and peak at the final layer; we observe this pattern in the English-only setting (Appendix B.1). However, the multilingual picture differs: as Figure 3 shows, many languages achieve their best ECE in the late-intermediate layers (from layer 24 for LLaMA3, and from layer 26 for Aya, both 32-layer models), with calibration worsening again towards the final output layer. The turning point aligns with the entropy trend: as entropy drops sharply in these intermediate layers, ECE also decreases. Reliability diagrams are in Appendix B.3.

The mid-layer calibration peak is robust across model sizes and families. The same pattern holds across all eight model families we analyse layer-wise (Appendix B.2) and on LLaMA-3.1-70B (Appendix D): for most non-English languages, a late-intermediate layer yields lower ECE than the decoding layer. This questions the common practice of using final-layer probabilities as confidence, and motivates the layer-aware calibration strategies developed next.

5 Improving Multilingual Calibration

5.1 Multilingual Calibration Methods

Our observations in Section 4 motivate confidence estimators that exploit intermediate representations. We explore to extract confidence in three different

ways:

- **Final layer (baseline).** We follow the standard practice in prior work by using the model's final layer to derive probabilities.
- **Best layer.** We identify the *best layer* as the one that achieves the lowest average calibration error across languages. From this layer, we extract probability estimates following the procedure in Section 4.1. The best layer ℓ^* is selected using a separate validation set and defined as:

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

- **Good layers ensemble.** We consider the set of layers (*good layers*) whose multilingual calibration outperforms the final layer. To obtain confidence estimates, we average the predictive distributions across these layers $\mathcal{G}$, this reduces layer-specific noise while pooling calibration-aware signals:

$$\mathcal{G} = \{\ell : \mathrm{ECE}_\ell^{\mathrm{avg}} < \mathrm{ECE}_L^{\mathrm{avg}}\}, \ \mathbf{p}_{\mathrm{ensemble}} = \frac{1}{|\mathcal{G}|}\sum_{\ell \in \mathcal{G}} \mathbf{p}_\ell.$$

Combining with Classical Post-hoc Calibration. Since confidence elicitation and calibration are orthogonal approaches, we further test whether the proposed elicitation methods can be enhanced by standard post-hoc calibration techniques such as Temperature Scaling (Guo et al., 2017) and Isotonic Regression (Zadrozny and Elkan, 2002), which operate independently of how probabilities were obtained (Kadavath et al., 2022; Minderer et al., 2021). We adopt a two-stage pipeline:

$$\mathbf{p}_{\mathrm{final}} = \mathrm{Calibrate}(\mathbf{p}_{\mathrm{intermediate}}),$$

where $\mathbf{p}_{\mathrm{intermediate}}$ comes from the confidence scores discussed above. See Appendix C.1 for details.

Language-Aware Confidence Ensemble. The methods above optimise jointly across languages. *LACE* (Language-Aware Confidence Ensemble) instead tailors layer selection to each language. For each language k, we select the layers that are better calibrated than the final layer,

$$\mathcal{G}^{(k)} = \{\ell : \mathrm{ECE}_\ell^{(k)} < \mathrm{ECE}_L^{(k)}\},$$
$$\mathbf{p}_{\mathrm{ensemble}}^{(k)} = \frac{1}{|\mathcal{G}^{(k)}|}\sum_{\ell \in \mathcal{G}^{(k)}} \mathbf{p}_\ell^{(k)},$$

and apply a language-specific post-hoc calibrator:

$$\mathbf{p}_{\mathrm{final}}^{(k)} = \mathrm{Calibrate}^{(k)}(\mathbf{p}_{\mathrm{ensemble}}^{(k)}).$$

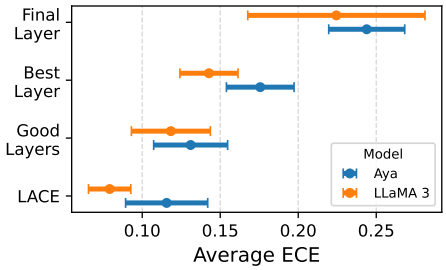


Figure 4: Forest plot of average ECE in MMMLU, with means and 95% CIs.

Per-language layer selection avoids negative transfer from layers that are miscalibrated for the target language, and ensembling reduces variance while preserving language-relevant signals. LACE is training-free and adds negligible latency: it reuses intermediate logits from the single forward pass already required for generation, with no retraining or weight modification.

Experiment Setup We use the data from Section 4 as a held-out validation set and evaluate on a separate MMMLU test split. Both splits are balanced across languages, with a total of 30K examples. For Belebele, we construct a comparable evaluation set with 24K examples overall with a similar validation/test split. We report macro-averaged results across languages.

5.2 LACE Results

Intermediate-layer confidence shows better calibration than final-layer confidence. Figure 4 shows that moving from Best Layer to Good-Layers Ensemble to LACE yields progressive improvements; full numbers for MMMLU and Belebele are in Table 3. Occasional drops in AUROC arise because discrimination and calibration are not necessarily correlated (Gao et al., 2022; Carriero et al., 2024) and our layer selection optimises ECE. By tailoring layer selection and calibration to each language, LACE achieves the best overall results, with ECE as low as 5.96% on LLaMA3.

Classical post-hoc methods offer improvements but face challenges in multilingual settings. Temperature scaling and isotonic regression noticeably reduce calibration error on Aya, which is systematically overconfident (avg. confidence 76.9% on MMMLU, 82.5% on Belebele; Figure 2). However, they offer marginal benefit or even degradation for LLaMA3 (avg. confidence 36.0% / 58.3%): a globally fit temperature gives

	MMMLU						Belebele					
	LLaMA3 (Acc. = 43.2%)			**Aya** (Acc. = 48.8%)			**LLaMA3** (Acc. = 68.6%)			**Aya** (Acc. = 67.8%)		
Method	**ECE**	**Brier**	**AUROC**	**ECE**	**Brier**	**AUROC**	**ECE**	**Brier**	**AUROC**	**ECE**	**Brier**	**AUROC**
FINAL LAYER BASELINE†	22.44	23.03	64.05	24.39	30.45	68.31	17.68	20.38	69.01	15.73	19.08	**72.92**
▷ *(+Temperature Scaling)*	23.35	22.59	64.05	15.40	23.76	68.31	17.63	20.39	69.01	10.66	17.19	72.92
▷ *(+Isotonic Regression)*	20.23	22.55	63.47	9.15	22.02	68.07	11.09	19.07	68.85	8.33	16.44	72.78
Intermediate Layer Calibration (Global Selection)												
BEST LAYER†	14.28	22.78	71.44	17.57	27.08	66.68	13.67	18.28	71.33	15.40	20.06	66.79
▷ *(+Temperature Scaling)*	13.71	20.60	71.44	9.34	22.52	66.68	13.12	17.52	71.34	14.99	19.16	66.79
▷ *(+Isotonic Regression)*	13.12	20.80	71.28	10.66	22.46	66.26	12.40	17.39	71.48	14.02	19.00	66.80
GOOD LAYERS ENSEMBLE†	11.84	20.23	**73.91**	13.10	23.78	**68.62**	10.78	15.59	**76.26**	11.47	18.00	70.24
▷ *(+Temperature Scaling)*	11.30	20.01	73.91	10.23	22.21	68.62	10.49	15.62	76.26	11.81	17.57	70.24
▷ *(+Isotonic Regression)*	9.60	19.90	73.49	7.71	21.82	68.25	10.16	15.54	76.07	10.42	17.48	70.31
Intermediate Layer Calibration (LACE)												
LANGUAGE-AWARE ENSEMBLE‡	5.96	20.51	72.94	11.42	22.70	68.38	7.05	**14.35**	75.61	10.22	17.77	69.79
▷ *(+Temperature Scaling)*	4.34	**19.73**	73.40	4.88	21.87	68.49	6.05	14.47	74.98	5.46	**16.36**	70.45
▷ *(+Isotonic Regression)*	**3.09**	20.51	69.13	**3.45**	**21.46**	67.10	**5.79**	14.53	73.70	**4.80**	16.73	68.40

Table 3: Calibration results for **LLaMA3** and **Aya** on MMMLU (left) and the Belebele subset (right). † denote methods that do not assume access to language identity. ‡ denote methods with given language identity. Indented italic rows correspond to post-hoc calibration adjustments.

Method	**ECE**	**Brier**	**AUROC**
ORIG. PROB.	23.51	22.10	65.19
(+Temp. Scal.)	22.56	21.97	65.19
(+Isotonic)	19.01	21.34	65.07
LACE	6.14	15.76	79.42
(+Temp. Scal.)	4.93	15.60	79.92
(+Isotonic)	2.81	15.38	78.79

Table 4: Calibration for LLaMA-3.1 70B on MMMLU.

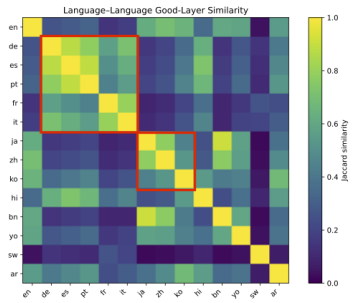


Figure 5: Jaccard Similarity between different language's layer-sets in LLaMA-3.1 70B.

limited per-language gains under distributional heterogeneity (cf. Simpson, 1951).

LACE is complementary to post-hoc approaches and delivers the best calibration performance. Importantly, our method *reshapes*, rather than *rescales*, the calibration signal. Therefore, our method does not compete with post-hoc calibrators but provides orthogonal improvements. As shown in Table 3, all methods combined with temperature scaling or isotonic regression yields further incremental gains, and notably LACE is further boosted to consistently deliver the lowest calibration error across benchmarks: ECE to 3.09% on LLaMA3 and 3.45% on Aya.

5.3 Analysis

We confirm that LACE scales to larger models, and analyse robustness.

LACE scales effectively to larger models. Consistent with our findings at the 7–14B scales, our layer-based methods continue to outperform final-layer confidence on LLaMA-3.1-70B (Table 4; full results in Appendix D). LACE provides the largest overall improvement at this scale, reducing ECE from 23.51% (baseline) to as low as 2.81% when combined with Isotonic Regression.

Layer-set overlap and robustness across languages and datasets. To assess whether LACE's layer sets generalise beyond specific samples or datasets, we examine their consistency across languages and benchmarks. Similar languages share overlapping layer subsets: Figure 5 shows clusters of high overlap between Western European languages and CJK (Chinese, Japanese, Korean) languages (red boxes). Layer sets derived from Belebele also retain strong calibration when transferred to MMMLU. The overlapping structure makes LACE robust to language-ID noise: in a controlled corruption experiment that randomly replaces the language tag of test examples, LACE degrades only minimally across all metrics. Full discussion is in

Dataset	Group	Perp.	Margin	P(True)	Verb.
MKQA	English	**13.5**	**10.6**	**19.2**	**26.9**
	Avg. Non-Eng.	33.9	34.8	31.7	41.1
	Δ	-20.4	-24.2	-12.5	-14.2
MGSM	English	**18.9**	**15.9**	11.0	**12.8**
	Avg. Non-Eng.	42.9	41.1	**12.6**	23.8
	Δ	-24.0	-25.2	+1.6	-11.0

Table 5: Generative-task calibration on **LLaMA3**: ECE (%, ↓) for four confidence elicitation methods on MKQA and MGSM. Δ = EN − Non-EN; negative values mean English is better-calibrated. Per-language numbers are in Appendix E.

Method	Lang	Acc.	Last	Best ($\ell^\star$)	ΔECE
Margin	fr	41.0	42.5	5.1 (ℓ=27)	−37.4
	ru	53.0	30.3	8.2 (ℓ=30)	−22.1
	th	41.0	42.7	9.0 (ℓ=27)	−33.7
Perplexity	fr	41.0	43.2	23.1 (ℓ=31)	−20.1
	ru	53.0	29.5	10.2 (ℓ=31)	−19.3
	th	41.0	42.9	33.2 (ℓ=31)	−9.6

Table 6: Layer-wise calibration on MGSM with **LLaMA3** (32 layers). ECE in percent. Last = final-layer ECE; Best = minimum ECE over all layers (best-layer index in parentheses); ΔECE = Best − Last.

Appendix C.

6 Generalisation Beyond MCQA

Our main analysis focuses on MCQA because it offers unambiguous correctness labels and a well-defined notion of confidence (§3). To probe whether our two core findings (the multilingual calibration gap and the late-intermediate sweet spot) generalise to free-form generation, we extend the analysis to two generative benchmarks.

Setup. We use **MKQA** (Longpre et al., 2021) (open-ended multilingual QA) and **MGSM** (Shi et al., 2023) (multilingual chain-of-thought math) on **LLaMA3**. Correctness is scored against gold answers: for MKQA we apply Unicode-aware normalization and require gold-answer containment; for MGSM we extract the final numeric token and require an exact match. Following Lewis-Lim et al. (2025), we evaluate four standard generative-confidence elicitation methods: *Perplexity* (geometric mean of next-token probabilities), *Margin* (top-1 vs. top-2 softmax gap, averaged across generated tokens), *P(True)* (Kadavath et al., 2022) (re-prompted true/false probability), and *Verbalised* (Tian et al., 2023) (expected confidence on a 1–9 scale, mapped to $[0, 1]$). Further details are in Appendix E.1.

The multilingual calibration gap persists in generation. Table 5 reports group-averaged ECE for each confidence method on the two datasets. English is the best-calibrated language for most cells. Averaging across the four methods, English ECE is 17.6% on MKQA and 14.7% on MGSM, while non-English ECE is 35.4% and 30.1%, roughly 2× worse on both datasets. The gap is widest on the lowest-resource languages (Japanese on MKQA, avg. ECE 47.8%; Bengali on MGSM, avg. ECE 46.7%), mirroring the resource correlation we report on MCQA (§3).

Late-intermediate layers remain better-calibrated. We apply the early-decoded confidence-estimation procedure of §4.1 to the two probability-space methods that admit a per-layer formulation (margin and perplexity; P(True) and Verbalised require re-prompting and therefore have no natural layer analogue). Table 6 reports the result for French, Russian, and Thai: for every (language × method) pair, a late-intermediate layer is substantially better calibrated than the final layer (ΔECE ranges from −9.6 to −37.4 percentage points). The best layers concentrate in ℓ=27–31 on a 32-layer model, the same region we identified for MCQA. Per-layer ECE curves are in Figure 16 (Appendix E).

Both findings also hold under two alternative MCQA scoring protocols (full-option-likelihood and text-generation; Appendix F.2).

7 Conclusion

This work provides a systematic evaluation of multilingual calibration in LLMs across eight model families and over 100 languages, uncovering disparities between English and other languages in both accuracy and reliability. Our layer-wise analysis identifies a mid-layer calibration peak: for non-English languages, late-intermediate layers are better-calibrated than the final layer. Leveraging this insight, we introduce LACE, a training-free, layer-aware ensemble that complements post-hoc techniques such as Temperature Scaling. We further show that the multilingual gap and the mid-layer effect generalise beyond MCQA to generative settings. We hope these results encourage further work on multilingual calibration to support more equitable language technologies.

Limitations

Our primary analysis is on MCQA. §6 extends the findings to MKQA and MGSM, but a larger study across more models, languages, and generative confidence-aggregation schemes remains future work. LACE is post-hoc and requires a small per-language validation set (Appendix C.5). A natural next step is to address the gap at training time, e.g. via per-language ECE penalties in RLHF/DPO, or via mechanistic interventions (Templeton et al., 2024) that localise the final-layer English bias. LACE also selects layers at the language level; a per-instance scoring mechanism could further improve robustness for code-mixed inputs. Finally, our evaluation is statistical (ECE, Brier, AUROC); we do not measure whether improved calibration translates into better human decision-making under deployment, which would require a targeted user study.

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. We only use LLMs for grammar checking during the paper writing.

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, and 8 others. 2024. Phi-4 technical report.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba Alabi, Marius Mosbach, Matan Eyal, Dietrich Klakow, and Mor Geva. 2024. The hidden space of transformer language adapters. pages 6588–6607, Bangkok, Thailand.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024a. 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.

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

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

Alex Carriero, Kim Luijken, Anne de Hond, Karel GM Moons, Ben van Calster, and Maarten van Smeden. 2024. The harms of class imbalance corrections for machine learning based prediction models: a simulation study.

Prateek Chhikara. 2025. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models.

Rochelle Choenni, Sara Rajaee, Christof Monz, and Ekaterina Shutova. 2024. On the evaluation practices in multilingual nlp: Can machine translation offer an alternative to human translations? *Preprint*, arXiv:2406.14267.

Common Crawl Foundation. 2025. Common Crawl. `https://commoncrawl.org`. 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, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. Depth-adaptive transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.

Riqiang Gao, Thomas Li, Yucheng Tang, Zhoubing Xu, Michael Kammer, Sanja L. Antic, Kim Sandler, Fabien Moldonado, Thomas A. Lasko, and Bennett Landman. 2022. A comparative study of confidence calibration in deep learning: From computer vision to medical imaging.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. pages 6577–6595, Mexico City, Mexico.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. 2025. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

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. 2023. Mistral 7b.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. pages 6919–6971, Mexico City, Mexico.

Samuel Lewis-Lim, Xingwei Tan, Zhixue Zhao, and Nikolaos Aletras. 2025. Analysing chain of thought dynamics: Active guidance or unfaithful post-hoc rationalisation? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29838–29853, Suzhou, China. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15682–15694.

Giulia Occhini, Kumiko Tanaka-Ishii, Anna Barford, Refael Tikochinski, Songbo Hu, Roi Reichart, Yijie Zhou, Hannah Claus, Ulla Petti, Ivan Vulić, Ramit Debnath, and Anna Korhonen. 2026. Artificial intelligence is creating a new global linguistic hierarchy. *Preprint*, arXiv:2602.12018.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1).

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. pages 2340–2354, Dublin, Ireland.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

E. H. Simpson. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jermyn, Catherine Olsson, Tristan Hume, Jared Kaplan, Tom Henighan, Sam McCandlish, and Chris Olah. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/. Anthropic Research.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. pages 2203–2213, Online.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. pages 15366–15394, Bangkok, Thailand.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. 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.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, and 43 others. 2024a. Qwen2 technical report.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Sen Yang, Nigel Collier, Dong Yu, and Deqing Yang. 2024b. Logu: Long-form generation with uncertainty expressions.

Ruihan Yang, Caiqi Zhang, Zhisong Zhang, Xinting Huang, Dong Yu, Nigel Collier, and Deqing Yang. 2025b. Uncle: Uncertainty expressions in long-form generation. *Preprint*, arXiv:2505.16922.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2023. On the calibration of multilingual question answering llms. *Preprint*, arXiv:2311.08669.

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

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024a. Do we need language-specific fact-checking models? the case of Chinese. pages 1899–1914, Miami, Florida, USA.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024b. LUQ: Long-text uncertainty quantification for LLMs. pages 5244–5262, Miami, Florida, USA.

Caiqi Zhang, Chang Shu, Ehsan Shareghi, and Nigel Collier. 2025a. All roads lead to rome: Graph-based confidence estimation for large language model reasoning. *Preprint*, arXiv:2509.12908.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. 2024c. Atomic calibration of llms in long-form generations. *Preprint*, arXiv:2410.13246.

Caiqi Zhang, Xiaochen Zhu, Chengzu Li, Nigel Collier, and Andreas Vlachos. 2025b. Reinforcement learning for better verbalized confidence in long-form generation. *Preprint*, arXiv:2505.23912.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024d. Calibrating the confidence of large language models by eliciting fidelity. pages 2959–2979, Miami, Florida, USA.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024e. PLUG: Leveraging pivot language in cross-lingual instruction tuning. pages 7025–7046, Bangkok, Thailand.

Ej Zhou and Weiming Lu. 2026. Bias beyond english: Evaluating social bias and debiasing methods in a low-resource setting. In *Natural Language Processing and Chinese Computing*, pages 214–227, Singapore. Springer Nature Singapore.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

A Benchmarking Multilingual Calibration

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

A.1 Experimental Details

A.1.1 Task Motivation and Confidence Elicitation

In this study, we focus on MCQA for two key reasons: first, MCQA provides unambiguous correctness labels for calibration. In open-ended generation, correctness is typically approximated using metrics such as ROUGE, BLEU, or LLM-as-a-judge, which often disagree and can even lead to flipped conclusions depending on the choice of metric (Ielanskyi et al., 2025). Second, MCQA enables a clear definition of confidence as the probability assigned to the selected option. In contrast, confidence estimation for open-ended generation remains unstable and method-dependent (e.g., verbalized confidence, perplexity-based methods, or sampling-based consistency). By using MCQA, we can isolate the effect of layer-wise signals on calibration without additional noise from generation-based confidence heuristics.

Correctness Calculation and Confidence Elicitation We take the output of the final layer as the model answer when evaluating correctness. We do not compute calibration using each layer's own prediction, because our goal is to provide a reliability signal for the deployed model in a standard inference setting, where the user only observes the final-layer answer. Therefore, the confidence score should approximate *the probability that this final-layer answer is correct*. We follow the standard MCQA practice and use the output probability of the selected answer choice: $\mathrm{Conf}(x) = \max_{i \in \{1,\ldots,C\}} p_i, \quad \text{where } \sum_{i=1}^{C} p_i = 1$, where C is the number of choices.

A.1.2 Models

Our experiments evaluate eight model families, plus a larger LLaMA-3.1-70B variant for scale analysis:

- **LLaMA3** (Grattafiori et al., 2024) (if **not** specified: `Llama-3.1-8BInstruct`)
- **Qwen2.5** (Yang et al., 2024a) (`Qwen2.5-7B-Instruct`)
- **Qwen3** (Yang et al., 2025a) (`Qwen3-8B`)
- **Mistral** (Jiang et al., 2023) (`Mistral-7B-Instruct-v0.3`)
- **Aya** (Dang et al., 2024) (`aya-expanse-8b`)
- **DeepSeek** (DeepSeek-AI, 2025) (`DeepSeek-R1-DistillQwen-7B`)
- **Phi** (Abdin et al., 2024) (`phi-4`)
- **Olmo3** (Olmo et al., 2026) (`Olmo-3-7B`)
- **LLaMA-3.1 70B** (Grattafiori et al., 2024) (`Llama-3.1-70BInstruct`)

A.1.3 Language Group Definitions

We randomly sampled 1,000 examples per language for MMMLU. We group languages in the MMMLU dataset according to resource availability and script type, as summarised in Table 7.

Group	Languages
Low-Resource	Arabic, Bengali, Swahili, Yoruba, Hindi, Indonesian
High-Resource	German, French, English, Spanish, Chinese, Italian, Japanese, Korean, Portuguese
Latin-Script	German, English, Spanish, French, Indonesian, Italian, Portuguese, Swahili, Yoruba
Non-Latin-Script	Arabic, Bengali, Hindi, Japanese, Korean, Chinese

Table 7: MMMLU language group definitions by resource availability and script type.

Language Codes The FLORES-200 language codes (3-letter form) for all evaluated languages are provided in the Belebele dataset documentation (Bandarkar et al., 2024a).

A.2 MMMLU Results

Results for each model are reported in the following tables: Mistral (Table 8), Qwen2.5 (Table 9), Phi (Table 10), and DeepSeek (Table 11).

A.3 Model Confidence Behaviours

Figure 6 illustrates the distribution of confidence scores and accuracies across English and non-English settings for Qwen2.5, DeepSeek, Mistral, and Phi. Solid vertical lines indicate mean accuracies, while dashed vertical lines indicate mean confidences. The divergence between confidence and accuracy highlights calibration behaviour: underconfidence when the dashed line falls left of the solid line, and overconfidence when it falls to the right.

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 8: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in **Mistral**, 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 9: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in **Qwen 2.5**, evaluated on the MMMLU dataset.

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

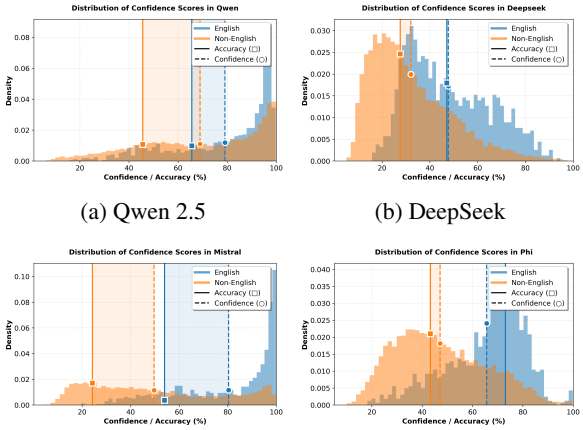
Language	AUROC	ECE	BRIER	Accuracy
Arabic	55.33	32.74	21.54	26.40
Bengali	58.50	40.80	14.41	13.70
German	60.28	18.50	23.91	39.80
English	66.21	9.10	22.92	47.10
Spanish	62.24	12.47	23.51	40.80
French	62.93	10.84	23.12	41.40
Hindi	56.08	30.42	20.62	26.40
Indonesian	61.00	31.61	21.11	27.30
Italian	63.14	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 11: Performance comparison across languages for AUROC, ECE, BRIER score, and Accuracy in **DeepSeek**, evaluated on the MMMLU dataset.

Table 12 reports detailed calibration and confidence statistics for LLaMA3. In addition to standard accuracy, we provide the model's average confidence, the confidence gap (accuracy minus confidence), the proportion of correct predictions made with low confidence ("Underconf"), and the mean confidence levels assigned to correct vs. incorrect predictions. We further include the difference between these two distributions ("Corr–Inc Gap"), which captures how well the model separates correct from incorrect responses. English shows a relatively small confidence gap (2.5%), with strong separation between correct and incorrect predictions (23.8% Corr–Inc Gap). In contrast, most non-English languages show lower accuracy, larger underconfidence rates, and much smaller separation between correct and incorrect predictions (average Corr–Inc Gap of 6.3%).

A.4 Belebele Results

Belebele (Bandarkar et al., 2024a) is a multiple-choice 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. Per-language calibration metrics on Belebele are reported in Table 13 for LLaMA3 and Table 14 for Aya.



(a) Qwen 2.5 (b) DeepSeek

(c) Mistral (d) Phi

Figure 6: Distribution of confidence scores (predicted probabilities) versus accuracies in English (blue) and non-English (orange) for four models: (a) Qwen 2.5, (b) DeepSeek, (c) Mistral, and (d) Phi.

Language	Acc. (%)	Avg Conf (%)	Conf Gap (%)	Underconf (%)	Corr. Conf (%)	Inc. Conf (%)	Corr–Inc Gap (%)
Arabic	38.2	29.5	8.8	89.0	31.7	28.0	3.7
Chinese	23.1	23.2	-0.1	99.6	22.2	23.6	-1.4
Korean	42.5	31.2	11.3	92.0	33.5	29.5	4.0
Japanese	43.0	24.0	19.0	92.3	26.8	21.8	5.0
Swahili	32.2	31.5	0.7	88.8	34.2	30.3	3.9
Italian	51.8	41.8	10.0	58.3	47.4	35.8	11.6
Bengali	35.2	34.6	0.6	88.4	36.1	33.8	2.3
Spanish	52.0	46.1	5.9	50.0	52.6	39.0	13.6
Portuguese	50.4	47.1	3.3	47.0	53.6	40.6	13.0
Indonesian	45.0	37.0	8.0	75.3	41.1	33.7	7.4
Hindi	39.9	31.7	8.2	88.2	34.1	30.1	4.0
German	44.4	33.8	10.6	80.4	37.1	31.2	5.9
Yoruba	27.4	29.4	-2.0	93.8	31.5	28.7	2.8
French	51.3	41.5	9.8	59.5	47.4	35.3	12.1
English	61.2	58.8	2.5	**25.7**	68.0	44.2	**23.8**
Non-English	41.2	34.5	6.7	**78.8**	37.8	31.5	**6.3**

Table 12: LLaMA3 Calibration and underconfidence analysis across languages. Metrics include accuracy, average confidence, confidence gap (accuracy minus confidence), proportion of underconfident correct predictions (confidence < 0.5 when correct), average confidence for correct vs. incorrect predictions, and their difference (Corr–Inc Gap).

B Layer-Wise Calibration Analysis

B.1 English Calibration Improves as Layer Deepens

Figure 7 shows the layer-wise Expected Calibration Error (ECE) for English in LLaMA3, illustrating how calibration improves progressively in deeper layers.

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

We visualize calibration performance across layers by plotting metrics against entropy on the MMMLU dataset for LLaMA3 (Figure 8a), Aya (Figure 8b), Mistral (Figure 8c), Phi (Figure 9a), DeepSeek (Figure 9b), and Qwen2.5 (Figure 9c).

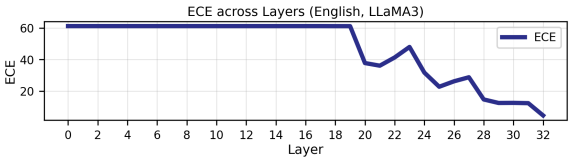


Figure 7: Layer-wise predicted confidence ECE for English in LLaMA3.

B.3 Reliability Diagrams

Per-language calibration trends reveal that late-intermediate layer improves calibration for most non-English languages, with a slight trade-off for English. To further explore these dynamics, Figure 10 presents per-language reliability diagrams for nine languages, comparing the selected intermediate layer (Layer 29) against the final layer of LLaMA3. Nearly all non-English languages benefit greatly from moving away from the final layer: their reliability curves align more closely with the diagonal, and ECE decreases substantially. For example, German and Hindi show reductions in ECE of more than 13% and 16%, respectively. In contrast, calibration for English appears to degrade slightly at intermediate layers. Since the final layer already exhibits strong calibration for English, earlier layers offer no additional benefit. This highlights a potential bias introduced during pretraining, where the model overfits to English patterns or introduces noise during the final stages of adaptation. While the degree of improvement varies across languages, the trend remains consistent across diverse linguistic families and scripts. This suggests that the effect is not language-specific but rather a systematic property of multilingual calibration. Additional examples are provided in Appendix B.3.

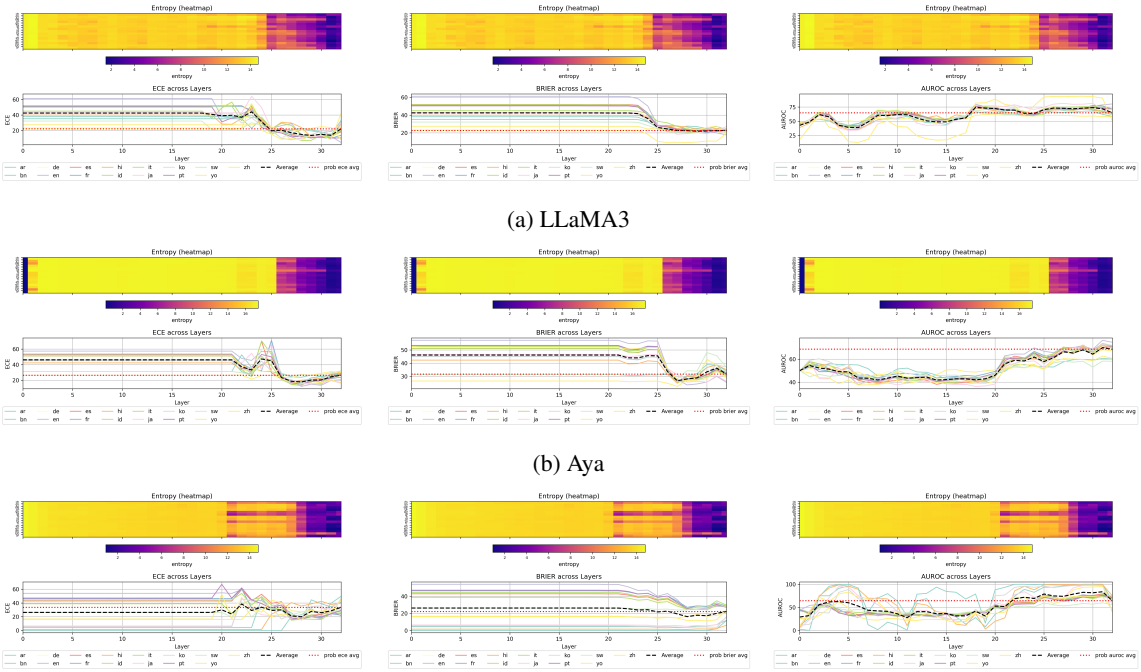
Figures 10 and 11 present reliability diagrams for LLaMA3 and Aya, respectively, illustrating calibration behaviour across languages and comparing intermediate versus final layers.

C LACE — Supporting Details & Robustness

This section provides implementation details for the post-hoc calibration baselines used alongside LACE, followed by a series of robustness and generalization analyses.

C.1 Post-hoc Calibration Baselines

We include two widely used post-hoc calibration methods as baselines: **Temperature Scaling** (Guo



(a) LLaMA3

(b) Aya

(c) Mistral

Figure 8: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for LLaMA3, Aya, and Mistral.

et al., 2017) and **Isotonic Regression** (Zadrozny and Elkan, 2002). Both are trained on a held-out validation set (15k examples from MMMLU and 12k examples from Belebele, non-overlapping with the evaluation sets). The fitted calibrators are then applied to test-set predictions, and metrics (ECE, AUROC, Brier score, Accuracy) are computed using the calibrated probabilities.

Temperature Scaling Temperature scaling applies a single scalar parameter $T > 0$ to rescale logits before computing probabilities. The parameter is optimized by minimizing the **negative log-likelihood (NLL)** on the validation set. In practice, we perform a coarse-to-fine grid search: first over a wide range ($T \in [0.05, 5.0]$ with 60 candidates), then locally refining around the best value with a denser grid. This procedure provides stable estimates across languages and avoids degenerate minima. The resulting optimal temperature is then used to rescale logits of all models prior to evaluation.

Isotonic Regression Isotonic regression learns a non-parametric, monotone mapping from predicted probabilities to calibrated probabilities in $[0, 1]$. We use the `scikit-learn` implementation with out-of-bounds clipping and monotonicity constraints. The model is fitted on the validation set and then applied to test-set predictions.

C.2 Cross-Lingual Layer-Sharing and Transfer

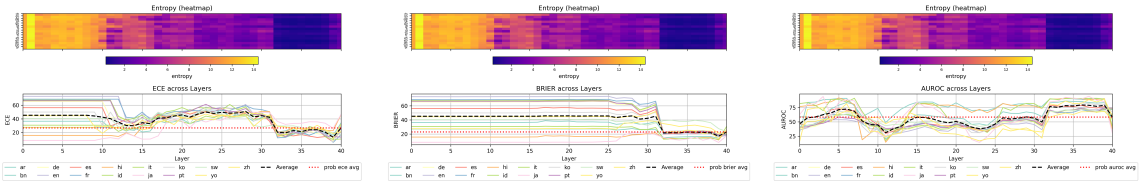
To better understand whether calibration improvements transfer across languages, we analyze the extent to which languages share similar "good layer" sets under our intermediate-layer calibration framework. If two languages select highly overlapping sets of good layers, we would expect stronger cross-lingual transfer: optimizing for one language could benefit the other. Conversely, disjoint sets would suggest that calibration signals are language-specific.

Measuring cross-lingual similarity. For each language ℓ, let G_ℓ denote its selected set of good layers. We compute a language–language similarity matrix using the Jaccard index:

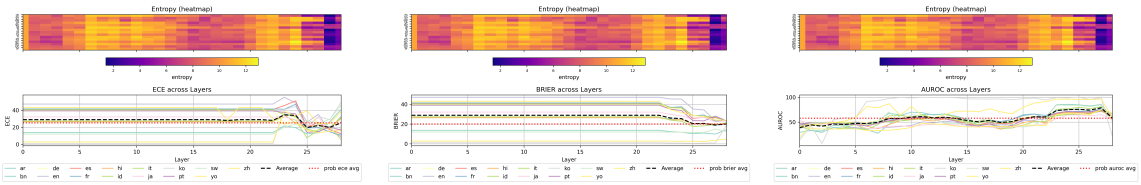
$$\text{Jaccard}(\ell_i, \ell_j) = \frac{|G_{\ell_i} \cap G_{\ell_j}|}{|G_{\ell_i} \cup G_{\ell_j}|}.$$

A score of 1.0 indicates identical layer sets, while 0.0 means no overlap. Higher similarity reflects stronger potential cross-lingual transfer.

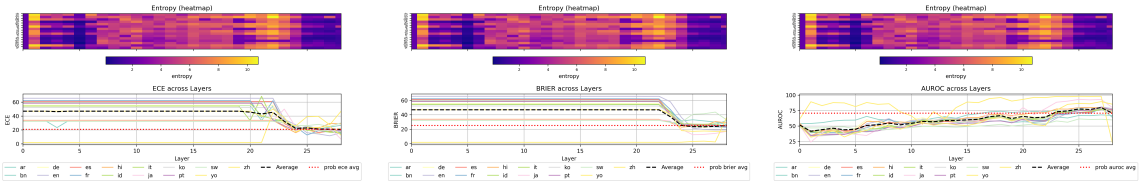
Findings. Figure 12 shows the cross-lingual similarity matrices for LLaMA-3 8B and 70B. While both models show some alignment among typologically related languages, the 70B model exhibits



(a) Phi



(b) Deepseek



(c) Qwen 2.5

Figure 9: Calibration metrics (ECE, Brier score, AUROC) vs. entropy across layers on the MMMLU dataset for Phi, Deepseek, and Qwen 2.5.

far clearer and more coherent clustering, reflecting stronger structural organization in its intermediate representations. In particular:

- **Western European languages** (e.g., German, Spanish, Portuguese, French, Italian) share overlapping good-layer subsets.
- **CJK languages** (Japanese, Chinese, Korean) also form a consistent cluster.
- Languages across distant families (e.g., English vs. CJK; Romance vs. Indic or African languages) show much lower overlap.

Example good-layer sets within Romance languages in the 70B model illustrate this structured overlap:

- `de`: [72, 75, 76, 77, 78, 79, 80]
- `es`: [72, 73, 75, 76, 77, 78, 79, 80]
- `pt`: [71, 72, 73, 75, 76, 77, 78, 79, 80]
- `fr`: [77, 78, 79, 80]
- `it`: [72, 77, 78, 79, 80]

Overall, these results confirm that **similar languages tend to share partially overlapping calibration-relevant layer patterns**, while more distant languages exhibit far less similarity.

C.3 Cross-Dataset Consistency of Good-Layer Selection

We assess whether "good layers" for calibration are consistent across different tasks by performing a cross-dataset layer-overlap analysis. For each language ℓ common to both MMMLU and Belebele, we extract the good-layer sets G^{MMMLU} and G^{Belebele} and compute their Jaccard similarity:

$$\text{Jaccard}(m) = \frac{1}{|\mathcal{L}_m|} \sum_{\ell \in \mathcal{L}_m} \frac{\left| G_{m,\ell}^{\text{MMMLU}} \cap G_{m,\ell}^{\text{Belebele}} \right|}{\left| G_{m,\ell}^{\text{MMMLU}} \cup G_{m,\ell}^{\text{Belebele}} \right|}.$$

This evaluation was conducted across the 12 languages common to both datasets. The results show moderate cross-dataset consistency, with an average Jaccard similarity of **0.5227** for LLaMA-3 and **0.4704** for Aya. This indicates that despite the different task formulations, the benchmarks identify partially overlapping sets of calibration-relevant layers. This suggests that certain structural properties of intermediate representations support robust calibration across tasks, though task-specific effects remain.

Transfer evaluation on MMMLU using Belebele-derived layers. To further test generalization, we applied the good-layer sets identified on Belebele

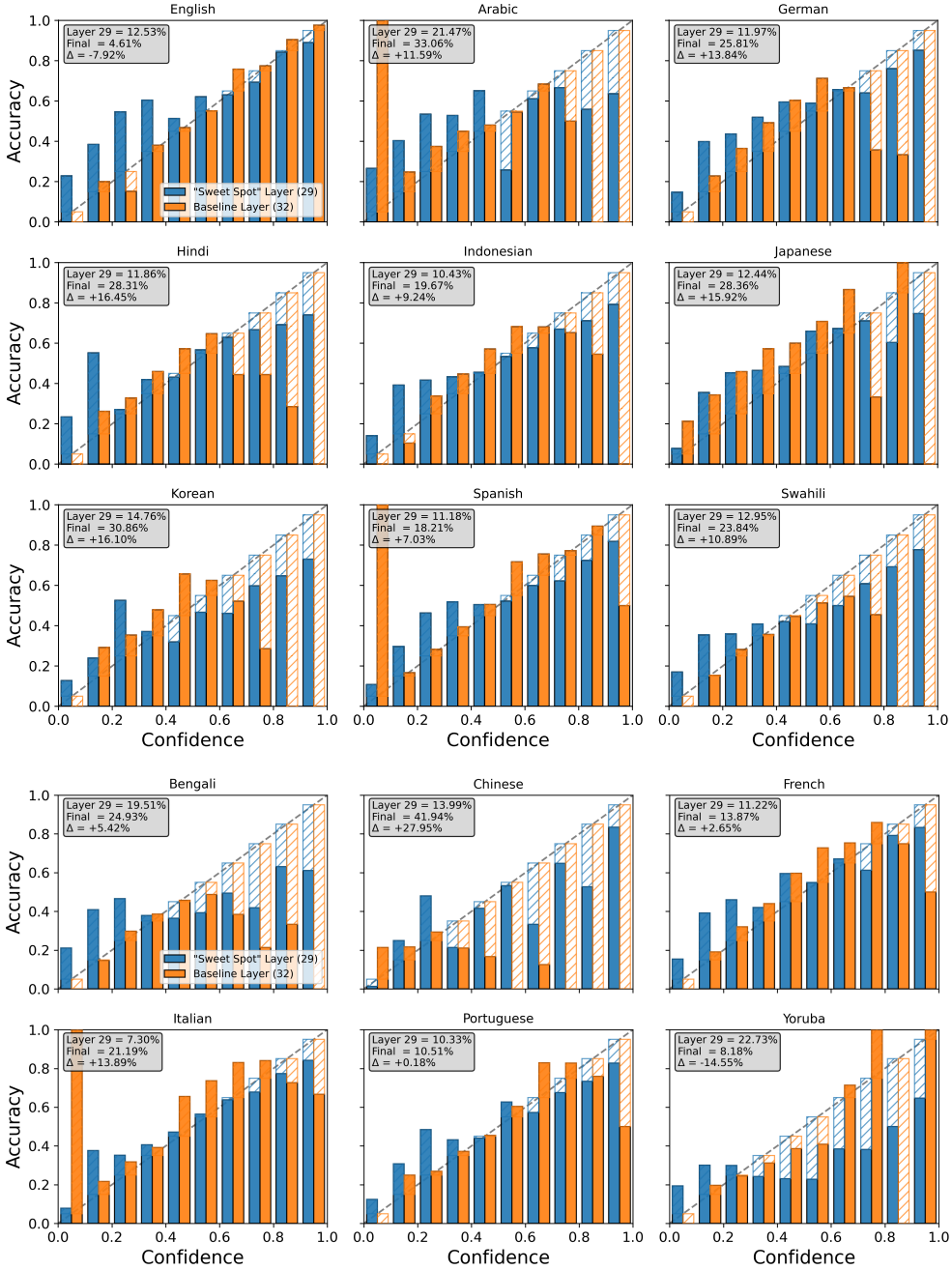


Figure 10: **Per-language calibration reliability diagrams for LLaMA3.** Each panel shows a reliability histogram with evenly spaced confidence bins. **Blue** bars correspond to the chosen intermediate layer (Layer 29), and **orange** bars correspond to the original final layer. The dashed diagonal is the perfectly calibrated line ($y{=}\,x$). Hatched overlays indicate the absolute calibration gap within each bin. The inset reports ECE (%) for both layers and the change $\Delta\text{ECE} = \text{ECE}_{\text{Final}} - \text{ECE}_{29}$ (positive values denote improved calibration at Layer 29).

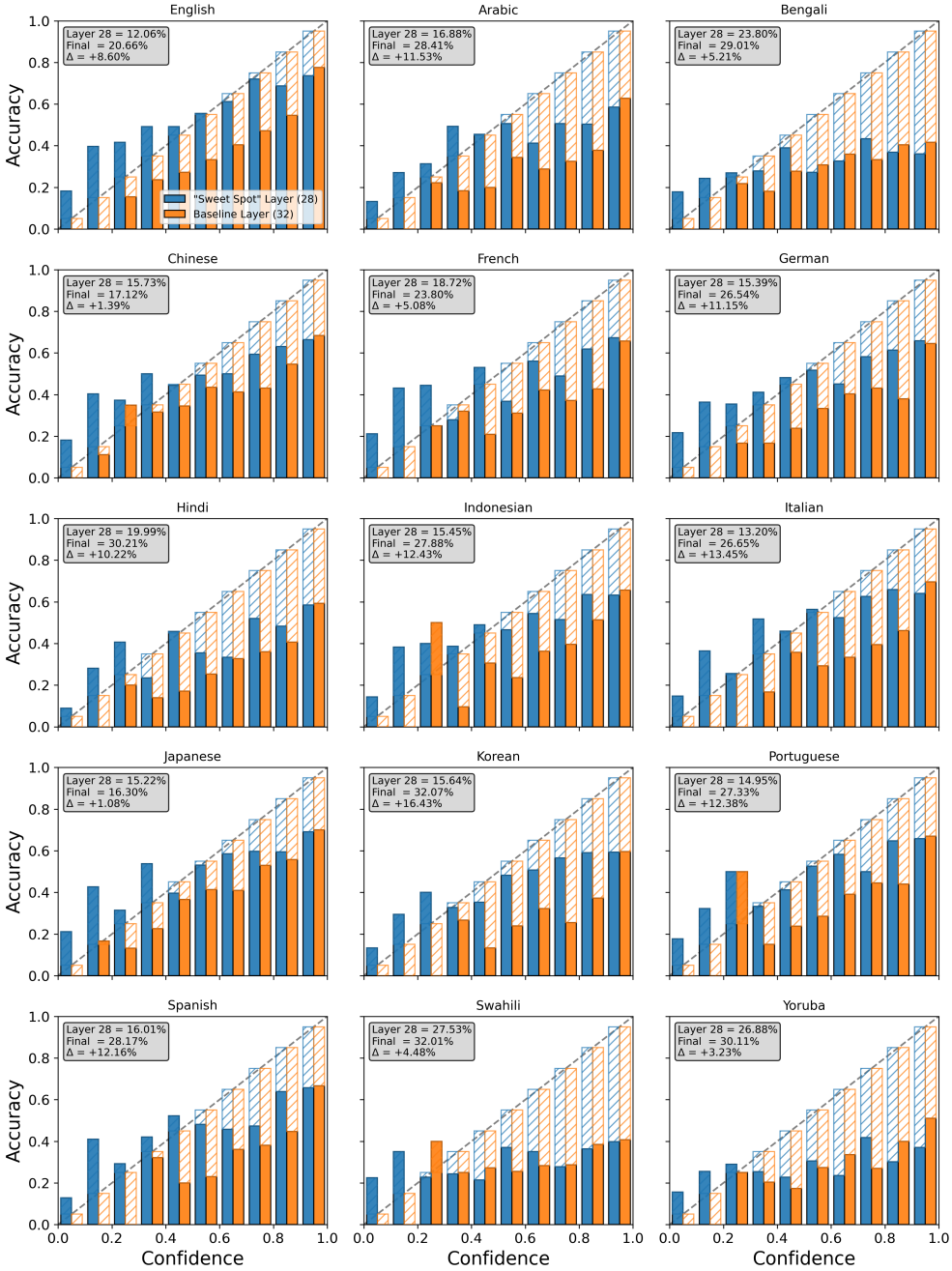
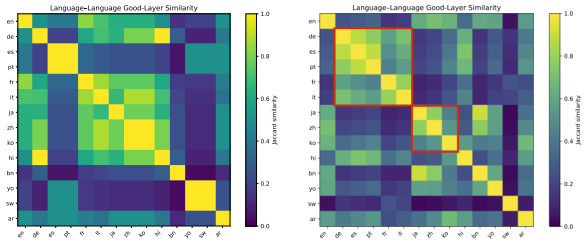


Figure 11: Reliability diagrams for Aya across all evaluated languages. Each plot compares the "sweet spot" intermediate layer with the final layer. Blue bars represent accuracy across confidence bins for the intermediate layer, while orange bars represent the final layer. The diagonal dashed line indicates perfect calibration. Reported values denote the ECE for Layer 28 and the final layer, along with the relative improvement (Δ), highlighting how calibration changes across layers in different languages.



(a) LLaMA-3 8B (b) LLaMA-3 70B

Figure 12: Cross-lingual Jaccard similarity between languages' good-layer sets. Clear correlation between Western European and CJK language groups indicate stronger organization of calibration-relevant information in deeper layers.

directly to MMMLU without any post-hoc calibration. The results, shown in Table 15, confirm that Belebele-derived layers retain meaningful calibration performance on MMMLU, demonstrating clear cross-task transferability.

C.4 Robustness to Language-ID Noise

To quantify how sensitive LACE is to noisy language labels, we run a controlled corruption experiment on MMMLU for two base models (Aya and LLaMA-3).

We start from our setup in the main paper where language-specific layer sets are selected on the validation split using gold language labels. Then at evaluation time, we simulate language-ID mistakes by randomly corrupting the language tag for a proportion $p \in \{0.0, 0.1, 0.2\}$ of test examples: for each selected example, we replace its true language label with a *different* language sampled from the remaining languages. LACE uses these (possibly incorrect) language tags to choose which language-specific layer set to average, while the underlying model logits remain unchanged.

Therefore, $p = 0$ corresponds to perfect language identification, whereas $p = 0.1$ and $p = 0.2$ approximate moderate error rates in an external language-ID system. We compute ECE, AUROC, and Brier score for each corruption level. The results are shown in Table 16.

The impact of language-ID noise is mild. For Aya, corrupting 10% of tags increases ECE by only 0.18 (about 1.5% relative) and Brier by 0.14, while AUROC decreases by 0.23. Even at 20% corruption, ECE rises by only 0.65 compared to the clean setting, and Brier changes by 0.23. For LLaMA-3, the effects are even smaller: ECE changes by $+0.01$ at 10% and $+0.33$ at 20%, with AUROC decreasing by less than 0.60 and Brier increasing by at most 0.24. These variations are negligible relative to the large cross-lingual calibration gaps reported in the main paper.

Intuitively, LACE remains robust because it averages over multiple "good" layers per language rather than relying on a single, highly language-specific layer: misrouting an example to another language's layer set still yields a reasonable ensemble, so moderate language-ID noise only weakly perturbs the final predictions.

C.5 Sensitivity to Validation Set Size

A potential limitation of LACE is that it requires a labeled validation set for each language in order to identify the corresponding Good Layers. This requirement may appear burdensome for low-resource languages, which are often the settings where improved calibration is most needed. However, in our main experiments, we use only 1,000 labeled validation examples per language, which we believe is already a practically attainable amount of data for many languages.

To further examine the data requirement of LACE, we conduct a sensitivity analysis by reducing the validation set size from the full 1,000 examples to 500, 200, and 100 examples. For each reduced setting, we recompute the selected Good Layers and compare them against those obtained from the full 1,000-example validation set. As a measure of agreement, we report the Jaccard similarity between the two selected layer sets.

Table 17 shows that the selected Good Layers remain highly consistent when using only 500 validation examples, with an average Jaccard similarity of 0.88 relative to the 1,000-example baseline. When the validation size is further reduced to 200 examples, the agreement remains moderate (0.65), while at 100 examples the consistency drops to 0.51. These results suggest that LACE is robust to substantial reductions in validation data, and that in practice as few as 500 labeled examples are sufficient to recover nearly the same layer selection as the full validation set.

Overall, this analysis indicates that the validation data requirement of LACE is lower than it may initially seem. While performance naturally degrades as the validation set becomes very small, the method remains highly stable at 500 examples, substantially lowering the barrier to applying LACE in lower-resource settings.

D Experiments on the 70B Model

To answer whether the calibration patterns observed in 7B–8B models persist in much larger models, we conducted a layer-wise calibration study on the **LLaMA-3.1 70B Instruct** model, using the same multilingual MMMLU setup.

Experimental Setup. We evaluated the 70B model under identical conditions as the 8B model: 15 languages and 15,000 test examples. To manage computational demands, the model was loaded using 4-bit quantization, which reduces memory usage while preserving performance. All experiments were performed on NVIDIA H100 80GB GPUs. LLaMA-3.1 70B has **80 transformer layers** (compared to 32 in the smaller models).

Layer-wise Calibration in a Large Model. The results, shown in Figures 13, 14, and 15, clearly demonstrate that the phenomenon observed in smaller models is also present in the 70B model: Calibration improves substantially in the late-intermediate layers, well before the final layer.

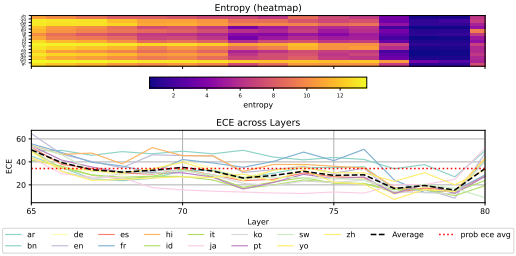


Figure 13: Expected Calibration Error (ECE) across the 80 layers of LLaMA-3.1 70B. A clear calibration minimum is visible in the intermediate layers across all language groups.

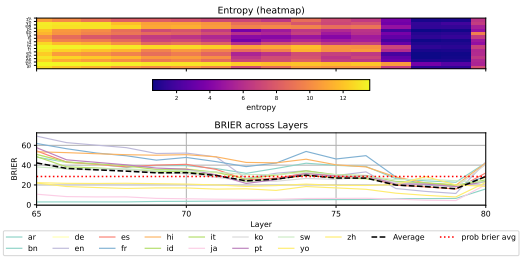


Figure 14: Brier score across the 80 layers of LLaMA-3.1 70B. The best-calibrated layers occur before the final layer, mirroring the pattern seen in ECE.

Quantifying the Improvement. To evaluate whether our calibration methods continue to be effective at larger scales, we also applied BEST LAYER, GOOD LAYER, and LACE to the 70B model. The results are reported in Table 18. Following the same format as Table 3 in the main text, this table shows that the intermediate-layer signal remains strong in the 70B model: using probabilities from the best-performing intermediate layer yields substantially better calibration than relying on the final layer. Moreover, LACE continues to provide the largest overall improvement, achieving the best calibration among all methods. These results confirm that our proposed layer-based calibration strategies, including LACE, scale effectively to much larger models.

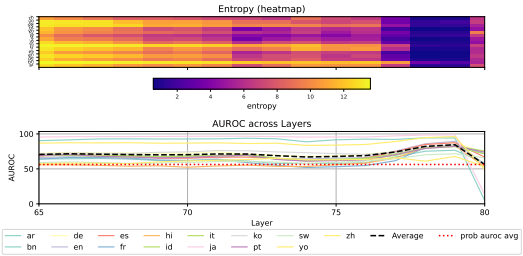


Figure 15: AUROC across the 80 layers of LLaMA-3.1 70B. The improvement in discriminative performance at intermediate layers is consistent with the trends in smaller models.

E Generalization Beyond MCQA: Full Details

This appendix supplements the generative-task analysis presented in §6.

E.1 Experimental Setup

Datasets. We use two human-translated multilingual generative benchmarks. **MKQA** (Longpre et al., 2021) is an open-ended multilingual question-answering benchmark. We sample 500 examples per language for seven languages (en, de, es, fr, ru, ja, th), spanning Latin, Cyrillic, Japanese, and Thai scripts. **MGSM** (Shi et al., 2023) is a multilingual chain-of-thought math benchmark derived from GSM8K. We use 250 examples per language for eight languages (en, de, es, fr, ru, ja, th, bn).

Correctness measure. For MKQA, we follow the standard open-domain QA convention: a generation is scored correct if any gold answer (or alias) is contained in the normalized prediction, or vice versa. Normalization uses Unicode-aware case folding, punctuation stripping, and whitespace

normalization (consistent with the reference implementation released by the dataset authors). For MGSM, we extract the final numeric token from the generated chain-of-thought and require an exact match to the gold answer up to a tolerance of 10^{-4}.

Confidence methods. We evaluate four standard generative-confidence elicitation methods, following the taxonomy of Lewis-Lim et al. (2025):

- **Perplexity**: $\exp\left(\frac{1}{T}\sum_{t=1}^{T}\log p(y_t|y_{<t}, x)\right)$, the geometric mean of next-token probabilities over the generated answer of length T.
- **Margin**: the per-token softmax margin between the top-1 and top-2 candidate tokens, averaged across generated tokens. The output lies in $[0, 1]$.
- **P(True)** (Kadavath et al., 2022): we re-prompt the model with "Question: ... Proposed answer: ... Is the proposed answer correct?" and read the softmax probability assigned to the "True" token at the next position.
- **Verbalised** (Tian et al., 2023): we append "How confident are you on a scale of 1–9?" and decode an expected confidence $\hat{c} = \mathbb{E}[s|s \in \{1, \ldots, 9\}]$, mapped to $[0, 1]$.

E.2 Per-Language Results

Tables 19 and 20 report per-language accuracy and ECE for each of the four confidence methods.

E.3 Discussion of the P(True) Outlier on MGSM

On MGSM, P(True) reports nearly identical ECE for every language ($\sim$10–17%), in contrast to the wide spread observed for perplexity, margin, and verbalised. This is consistent with the documented brittleness of self-reprompting probes: Kadavath et al. (2022) and Lin et al. (2022) both note that the "Is the proposed answer correct?" template is read by the model in the language of the template (here, English), independently of the language of the original question. The resulting probabilities are dominated by the English-language meta-question rather than by the multilingual content of the answer, so the protocol effectively short-circuits the language-conditional confidence signal. We retain the method to faithfully report the benchmark and note that the other three methods, which read confidence directly from the generated tokens, all preserve the multilingual gap.

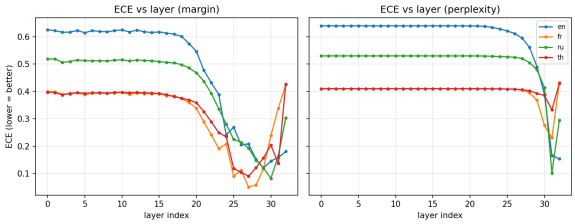


Figure 16: Layer-wise ECE on MGSM (LLaMA3). Left: margin confidence. Right: perplexity confidence. x-axis is layer index (0–32); y-axis is ECE.

E.4 Layer-wise Generative Calibration

Figure 16 shows the ECE-by-layer curves on MGSM that accompany the magnitudes in Table 6. The full numerical results, including AUROC and Brier scores at every layer, are released alongside the code (`layerwise_mgsm_llama3.csv`).

F Robustness of the Multilingual Calibration Gap

The English/Non-English calibration gap reported in §3 is the central empirical claim of the paper. This appendix collects the robustness checks that establish the gap is preserved under different choices of (i) the ECE bin count and (ii) the MCQA scoring protocol.

F.1 ECE Binning Sensitivity

In the main paper, following Guo et al. (2017), we report ECE using $K = 10$ uniformly spaced bins over $[0, 1]$, applying the identical binning scheme across all models, layers, languages, and calibration methods.

To evaluate robustness with respect to this hyperparameter, we recomputed ECE for MMMLU predictions for both LLaMA-3 and Aya using $K \in \{5, 10, 15, 20, 25\}$. Table 21 presents the complete results for both LLaMA-3 and Aya, while Figure 17 visualizes the per-layer trends for LLaMA-3.

Although absolute values shift somewhat under different bin counts, the *qualitative pattern remains unchanged*: non-English languages remain substantially less calibrated than English for every choice of K, and non-English calibration is highly stable once $K \geq 10$ (LLaMA-3 within 22.83–23.27%, Aya within 26.77–28.27%). Brier score and AUROC (both bin-free) exhibit the same cross-lingual trends, so our conclusions are robust to reasonable variations in the ECE computation scheme.

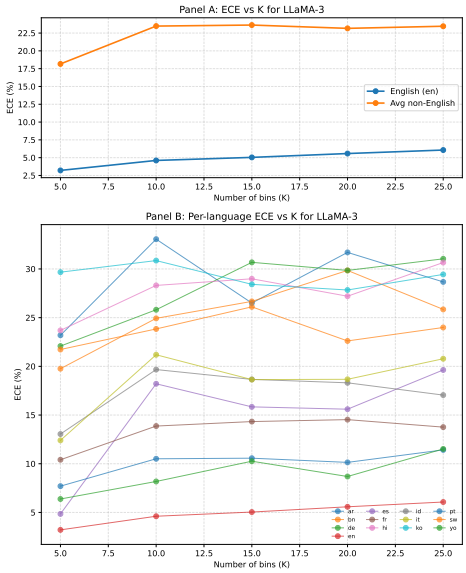


Figure 17: Per-layer ECE for LLaMA-3 on MMMLU across different numbers of bins K. The calibration gap between English and non-English languages persists across all binning configurations.

F.2 Alternative MCQA Scoring Protocols

The main paper elicits MCQA confidence as the softmax probability assigned to the option-letter token at the choice position (Appendix A). This is a standard convention in MCQA evaluation, but not the only one. To verify that the multilingual calibration gap holds under alternative scoring protocols, we evaluate LLaMA3 under three protocols:

- **P1 – Letter probability** (paper baseline): softmax over the four option-letter tokens (`A`, `B`, `C`, `D`) at the choice position; confidence is the max-class probability. This is the protocol used throughout the main paper, reported here at full MMMLU scale (1000 examples per language; numbers reproduced from Table 1).
- **P2 – Option likelihood**: each full option text is scored under the model (sum of token log-probabilities), then softmax is taken across the four options; the model's prediction is $\arg\max$, the confidence is the max softmax value. Evaluated on a balanced 14-language $\times$ 200-example MMMLU subset.
- **P3 – Text generation**: greedy-decode up to 8 new tokens, regex-extract the predicted letter, and use the softmax probability of that decoded letter at the first generated position as confidence. Evaluated on the same 14-language $\times$ 200-example subset.

Finding. As Table 22 shows, the EN $-$ Non-EN ordering documented in the main paper is preserved on *every* metric under all three scoring protocols. The averaged gap across protocols is $+13.78$ points of accuracy and $+16.13$ points of AUROC in favour of English, with ECE and Brier both reduced for English. Reading down the table, the P1 row reproduces the main-paper gap (full scale); P2 and P3 verify on a controlled subset that the gap persists even when confidence is derived from full-option likelihoods (P2) or from text-generation outputs (P3) rather than letter-token probabilities. The multilingual calibration gap therefore holds across these three scoring protocols.

Per-language values for P2 and P3. Table 23 reports the per-language breakdown for the two new scoring protocols on the 14-language subset. The qualitative resource-level effect from the main paper (high-resource languages are better calibrated than low-resource) persists: under P3, for instance, English (ECE 6.71%) and Italian (9.71%) sit at one end, while Swahili (18.39%) and Chinese (30.50%) sit at the other.

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	7.1	19.2	arz	69.8	77.8	12.2	18.2	ceb	65.8	75.6	11.1	19.7
fin	80.8	75.6	6.0	14.1	hin	67.0	72.9	14.0	20.6	ita	86.0	79.6	7.4	10.8
khm	4.0	5.5	66.2	51.4	lvs	74.8	79.7	11.3	16.2	npi	59.8	74.8	5.1	20.0
pol	80.0	78.9	6.3	13.3	slv	81.0	76.6	8.7	13.8	swe	79.2	81.6	7.3	12.8
tso	34.0	62.5	3.4	21.5	xho	37.0	60.0	5.0	22.8	afr	80.8	81.2	8.3	12.7
asm	45.8	68.0	6.1	22.8	ces	82.2	80.2	7.8	11.8	fra	86.2	77.4	9.9	11.4
hin	57.0	72.1	4.7	21.1	jav	68.0	78.4	11.1	18.4	kin	34.8	63.3	4.7	21.3
mal	60.5	74.0	12.3	21.3	npi	32.5	59.9	6.3	21.4	por	86.2	79.8	5.6	10.2
sna	35.8	58.6	6.2	23.0	swh	67.0	75.0	8.1	19.1	tur	78.2	78.8	8.5	14.6
yor	31.2	61.1	4.4	20.4	als	73.5	78.0	7.2	16.2	azj	66.8	71.8	14.3	21.4
ckb	46.0	71.8	8.3	22.2	fuv	28.0	51.9	5.2	20.4	hrv	79.8	78.1	8.2	14.3
jpn	66.5	73.3	29.9	28.0	kir	63.0	73.8	21.9	24.7	mar	67.5	73.5	9.2	19.4
nso	37.8	60.3	3.2	22.7	snd	17.5	50.5	30.0	25.3	tam	65.5	73.4	14.5	21.5
ukr	84.2	77.4	12.1	12.8	zho	76.5	71.2	30.6	25.2	amh	34.8	63.2	5.0	21.5
bam	31.2	60.6	4.8	20.8	dan	79.8	78.8	8.2	13.9	gaz	31.8	53.1	2.4	21.7
hun	82.5	84.0	9.7	11.8	kac	30.2	61.8	3.5	20.5	kor	77.8	77.9	14.5	16.3
mkd	77.8	79.1	7.8	14.4	nya	32.0	60.7	3.3	21.1	ron	80.0	80.4	6.3	13.0
som	35.2	59.0	3.9	22.2	tel	59.5	73.7	7.2	20.4	urd	59.5	67.3	19.7	25.8
zho	81.2	68.9	24.4	21.1	apc	65.0	78.3	9.7	18.4	ben	65.5	72.6	10.1	20.3
deu	86.8	72.3	13.4	12.9	grn	39.8	65.5	6.6	22.4	hye	0.2	1.5	63.8	42.1
kan	58.5	72.1	5.2	20.9	lao	32.5	59.5	2.1	21.5	mlt	69.8	76.4	9.9	18.4
ory	55.8	70.0	16.4	24.7	rus	81.2	81.0	10.4	13.0	sot	32.8	57.5	2.6	21.8
tgk	63.8	70.1	12.5	22.1	urd	41.2	64.8	4.3	22.5	zsm	82.5	82.9	9.5	11.8
arb	79.5	74.2	11.8	15.5	ben	35.2	59.9	4.9	22.4	ell	80.5	80.6	10.1	13.7
guj	58.0	68.8	8.2	22.0	ibo	40.2	62.0	3.7	22.6	kat	1.5	6.0	68.4	51.6
lin	34.2	59.3	3.5	21.9	mri	35.5	63.3	4.3	21.5	pan	58.0	74.4	11.4	21.6
shn	16.8	47.9	15.1	16.9	spa	84.0	79.3	5.8	11.7	tgl	75.2	80.1	7.8	15.3
uzn	69.0	77.0	13.4	19.2	zul	36.5	59.3	5.1	22.8	arb	29.8	56.4	3.6	20.8
bod	29.0	60.2	6.8	20.3	sun	65.5	74.3	13.5	20.9	hat	55.8	72.9	4.6	20.9
ilo	54.0	69.8	9.2	22.6	kaz	63.5	75.6	19.3	23.0	lit	73.8	82.3	11.0	15.7
mya	0.8	7.2	72.5	55.7	pbt	47.5	67.7	3.1	22.6	sin	32.2	59.2	3.8	21.5
srp	83.2	77.5	13.0	13.7	tha	71.8	75.5	19.8	21.2	vie	83.5	78.4	8.7	12.4
ars	62.0	78.5	9.6	18.8	bul	80.8	77.7	13.6	14.6	est	71.8	78.2	4.2	16.4
hau	45.2	67.1	11.7	23.9	ind	81.8	75.6	6.4	13.4	kea	48.8	73.0	8.0	21.1
lug	35.5	57.6	2.5	22.5	nld	83.0	78.2	5.9	12.0	pes	79.2	77.8	8.5	14.4
sin	58.8	72.7	10.3	21.6	ssw	31.8	61.5	3.0	20.8	tir	28.0	57.7	4.9	19.7
war	62.2	74.7	12.1	21.0	ary	58.5	72.0	7.7	21.2	cat	86.2	84.8	9.4	10.2
eus	69.5	75.7	7.0	18.1	heb	77.2	76.9	17.2	17.8	isl	67.0	78.3	5.2	17.4
khk	48.0	70.2	10.0	22.9	luo	31.8	55.7	5.7	21.5	nob	79.2	77.8	9.5	14.7
plt	44.2	65.6	5.8	23.1	slk	82.0	76.9	4.7	13.0	**eng**	**87.8**	**87.9**	**4.0**	**8.1**
tsn	31.8	62.7	4.6	20.8	wol	33.0	53.2	5.7	22.3	**Avg.**	**57.6**	**68.9**	**10.9**	**19.8**

Table 13: 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). English line and the Average line is **bolded**.

Set 1					Set 2					Set 3				
Lang	**Acc**	**AUR.**	**ECE**	**Brier**	**Lang**	**Acc**	**AUR.**	**ECE**	**Brier**	**Lang**	**Acc**	**AUR.**	**ECE**	**Brier**
acm	64.7	77.4	25.6	26.8	arz	75.7	82.3	16.7	17.5	ceb	47.3	72.9	29.9	30.7
fin	63.0	64.7	17.1	25.2	hin	67.3	75.0	22.7	24.1	ita	80.3	80.1	12.7	14.4
lvs	55.3	68.6	24.1	28.4	npi	43.3	67.7	32.7	33.4	pol	79.7	80.0	14.5	16.1
slv	63.3	75.4	23.5	25.9	swe	74.3	69.2	18.5	20.8	tso	34.3	57.7	29.4	32.3
xho	34.0	59.0	25.2	29.4	afr	71.0	80.5	18.7	20.1	asm	38.3	59.8	25.7	30.2
ces	80.7	81.5	13.4	14.2	fra	86.7	82.0	8.5	10.3	hin	54.0	68.5	28.6	30.4
jav	56.7	75.5	25.2	26.6	kin	36.7	58.0	26.8	30.7	npi	37.3	61.7	28.9	31.7
por	83.0	79.0	12.2	12.9	sna	34.0	66.5	28.5	29.1	swh	37.0	66.9	32.2	32.1
tur	78.3	82.0	12.9	15.6	yor	29.7	53.6	29.6	31.8	als	49.0	69.7	26.4	29.6
azj	58.7	67.0	18.4	26.0	ckb	47.0	63.6	29.4	33.0	fuv	27.7	54.9	34.1	34.0
hrv	68.0	76.6	21.4	23.1	jpn	78.3	75.8	13.1	16.1	kir	44.7	63.3	32.8	34.6
mar	45.3	66.6	32.3	33.9	nso	34.7	58.0	27.8	31.4	snd	40.7	62.8	22.8	28.6
tam	28.3	54.9	32.6	33.4	ukr	81.3	83.5	13.0	14.0	zho	84.3	86.8	8.6	10.6
amh	26.0	48.2	16.9	23.8	bam	35.3	62.1	25.8	29.2	dan	72.0	77.2	14.1	18.4
gaz	32.3	51.2	26.1	30.8	hun	60.0	72.2	23.7	26.2	kac	33.0	63.2	23.9	27.2
kor	78.7	77.2	14.7	15.8	mkd	61.0	74.9	22.7	25.1	nya	33.7	54.6	25.7	31.0
ron	82.0	82.7	12.8	13.5	som	33.0	63.9	29.7	30.6	urd	50.3	68.9	16.1	25.1
zho	80.3	76.2	11.6	14.9	apc	67.7	78.7	20.9	21.6	ben	49.7	65.0	20.7	28.1
deu	83.3	77.7	12.2	13.6	grn	35.7	64.7	31.9	32.3	hye	36.7	57.6	21.9	28.5
mlt	41.3	65.6	29.1	31.7	rus	83.3	82.0	12.4	13.3	sot	33.7	55.2	27.1	31.6
tgk	38.7	68.5	23.7	27.3	urd	41.3	63.0	26.0	30.9	zsm	79.7	81.3	12.3	14.4
arb	80.0	78.9	14.8	15.3	ben	36.0	62.8	29.3	31.5	ell	81.7	84.9	10.1	12.7
wol	31.6	56.5	22.0	27.7	ibo	30.0	50.7	27.4	31.9	kat	46.7	64.8	26.8	30.1
lin	33.0	63.0	34.9	34.2	mri	35.3	60.0	26.7	30.2	spa	79.0	78.2	19.1	18.3
tgl	69.0	73.6	17.5	21.6	uzn	46.9	66.0	29.1	32.3	zul	31.1	58.1	25.6	28.7
arb	28.3	54.7	33.8	34.0	plt	35.2	58.9	21.5	27.2	hat	46.4	67.9	27.2	30.3
ilo	37.8	68.2	30.7	31.1	kaz	42.1	67.3	30.4	32.3	lit	63.5	74.2	23.5	25.8
pbt	36.0	60.6	31.6	33.3	sin	36.0	58.2	23.0	29.0	srp	66.1	71.9	17.8	23.3
tha	49.0	69.2	24.9	28.7	vie	82.7	80.8	10.7	12.8	ars	65.6	78.3	25.7	25.6
bul	67.6	77.4	21.8	23.4	est	51.8	67.8	25.1	29.4	hau	29.3	60.8	33.2	32.3
ind	80.6	81.0	13.8	14.9	kea	47.5	67.2	33.1	34.4	lug	31.1	55.8	23.6	27.9
nld	81.1	79.4	12.6	14.7	pes	80.1	78.7	12.4	15.0	tsn	33.7	60.9	23.3	28.0
ssw	36.2	55.7	21.1	29.0	tir	27.6	54.7	23.9	27.0	war	49.2	66.5	26.3	30.6
ary	64.3	77.2	24.2	25.4	cat	76.5	81.4	14.7	16.3	eus	47.2	64.8	22.6	29.1
heb	79.8	79.6	11.5	14.7	isl	51.5	70.9	27.0	29.6	khk	37.5	57.6	21.7	28.8
luo	29.1	63.9	26.3	26.5	nob	72.7	77.0	16.2	19.0	**eng**	**87.0**	**83.6**	**4.8**	**8.7**
slk	76.8	80.7	15.1	16.9	sun	47.5	74.6	30.9	30.9	**Avg.**	**53.0**	**68.0**	**22.8**	**25.2**

Table 14: Per-language performance on the belebele test set for the **Aya** model, reporting AUROC, ECE, and Brier score. Each row is color-coded, same as Table 13. Language entries with lower than 5% accuracy is excluded. English line and the Average line is **bolded**.

Model	Setting	ECE ↓	AUROC ↑	Brier ↓
LLaMA-3	baseline	22.2	66.1	23.3
	transferred-LACE	**15.6**	**70.6**	**21.8**
Aya	baseline	27.4	68.9	32.3
	transferred-LACE	**19.8**	**69.3**	**25.2**

Table 15: Calibration transfer results when applying **Belebele-derived good layers** directly to **MMMLU** (without post-hoc calibration).

Corruption	Model	ECE (%)	AUROC (%)	Brier (%)
0%	LLaMA-3	5.96	72.94	19.91
	Aya	11.42	68.61	23.16
10%	LLaMA-3	5.97	72.80	19.97
	Aya	11.60	68.38	23.31
20%	LLaMA-3	6.29	72.39	20.15
	Aya	12.07	68.25	23.39

Table 16: Robustness of LACE to language-ID corruption on MMMLU.

Comparison	Avg. Jaccard	Interpretation
1000 vs. 500	0.88	High consistency
1000 vs. 200	0.65	Moderate consistency
1000 vs. 100	0.51	Lower consistency

Table 17: Stability of Good Layer selection under reduced validation set sizes, measured by Jaccard similarity against the full 1,000-example validation set.

Method	ECE	Brier	AUROC
ORIGINAL PROBABILITY†	23.51	22.10	65.19
(+Temperature Scaling)	22.56	21.97	65.19
(+Isotonic Regression)	19.01	21.34	65.07
BEST LAYER†	16.31	16.14	84.57
(+Temperature Scaling)	13.81	14.94	84.57
(+Isotonic Regression)	7.62	13.77	84.40
GOOD LAYER†	13.90	16.43	79.30
(+Temperature Scaling)	15.00	16.42	79.30
(+Isotonic Regression)	13.44	16.25	78.98
LACE‡	6.14	15.76	79.42
(+Temperature Scaling)	4.93	15.60	79.92
(+Isotonic Regression)	2.81	15.38	78.79

Table 18: Calibration performance of different layer-based methods on the LLaMA-3.1 70B model using MMMLU. Intermediate-layer methods continue to yield substantial calibration improvements even at large model scale.

Lang.	Acc.	Perp.	Margin	P(True)	Verb.
English	52.0	**13.5**	**10.6**	**19.2**	**26.9**
German	43.0	23.4	23.7	26.3	31.2
Spanish	42.8	25.2	24.1	25.7	31.0
French	43.2	25.1	24.7	24.2	32.3
Russian	26.4	39.8	41.2	37.4	49.0
Japanese	18.4	45.3	47.7	42.4	55.7
Thai	24.0	44.5	47.6	34.1	47.2
Avg. Non-Eng.	33.0	33.9	34.8	31.7	41.1

Table 19: Per-language MKQA results (LLaMA3, 500 examples per language). ECE (%, ↓) for four confidence methods. Bold marks the column-minimum.

Lang.	Acc.	Perp.	Margin	P(True)	Verb.
English	65.6	**18.9**	**15.9**	11.0	**12.8**
German	51.6	32.5	30.3	10.9	21.7
Spanish	56.4	27.8	26.0	13.0	16.3
French	44.8	40.1	38.7	**9.5**	24.4
Russian	47.2	37.9	36.4	13.5	19.2
Japanese	44.0	40.0	37.9	17.1	25.6
Thai	36.4	48.4	47.6	12.2	28.8
Bengali	14.0	73.3	70.7	12.2	30.4
Avg. Non-Eng.	42.1	42.9	41.1	12.6	23.8

Table 20: Per-language MGSM results (LLaMA3, 250 examples per language). ECE (%, ↓) for four confidence methods. Bold marks the column-minimum.

Bins (K)	LLaMA-3 ECE (%)		Aya ECE (%)	
	English	Non-Eng.	English	Non-Eng.
5	3.21	17.87	19.90	25.93
10	4.61	23.12	20.66	26.77
15	5.04	23.27	20.34	27.45
20	5.58	22.83	21.72	28.26
25	6.07	23.13	21.34	28.27

Table 21: ECE binning sensitivity for LLaMA-3 and Aya on MMMLU. English consistently shows lower calibration error than non-English across all binning configurations.

Protocol	Group	Acc.	ECE↓	AUROC↑	Brier↓
P1: Letter prob.†	EN	**61.20**	**4.61**	**80.36**	**17.63**
	Non-EN avg.	41.47	23.12	64.06	22.95
P2: Option lik.	EN	**39.50**	**20.10**	**59.70**	**29.10**
	Non-EN avg.	31.40	27.00	54.10	30.90
P3: Text gen.	EN	**40.50**	**6.70**	**76.10**	**18.30**
	Non-EN avg.	27.00	12.20	49.60	19.00
Avg. Δ (EN − Non-EN)		+13.78	-10.30	+16.13	-2.61

Table 22: MMMLU calibration under three MCQA scoring protocols on LLaMA3. P1 is reported at full MMMLU scale (1000 examples per language; same numbers as Table 1). P2 and P3 are evaluated on a controlled 14-language × 200-example MMMLU subset. The bottom row averages the EN − Non-EN gap across protocols. †P1 values come from the main paper's full-scale evaluation.

Lang.	P2: Option lik. (200 ex.)			P3: Text gen. (200 ex.)		
	Acc.	ECE	AUROC	Acc.	ECE	AUROC
Arabic	26.0	30.25	55.52	25.5	10.14	50.45
Bengali	29.0	22.44	56.51	23.0	10.63	53.49
German	36.5	22.82	55.02	30.0	9.03	42.45
Spanish	32.5	26.13	61.45	38.5	3.79	68.75
French	33.0	24.68	60.64	29.5	6.14	65.83
Hindi	27.5	39.14	55.50	23.5	10.45	56.08
Indonesian	34.0	26.19	48.02	31.5	5.61	59.63
Italian	34.0	23.53	51.99	34.5	9.71	75.86
Japanese	28.0	26.59	57.89	26.0	25.13	24.62
Korean	31.5	25.67	52.62	17.5	16.56	28.54
Portuguese	37.5	22.71	54.18	38.0	10.76	76.32
Swahili	22.0	40.97	42.92	15.5	18.39	22.20
Yoruba	31.5	26.35	51.73	24.0	4.11	44.70
Chinese	37.0	20.89	53.89	21.0	30.50	25.92
English	39.5	20.08	59.68	40.5	6.71	76.08
Avg. Non-Eng.	31.4	27.03	54.13	27.0	12.21	49.63

Table 23: Per-language MMMLU results under P2 (option likelihood) and P3 (text generation) on the controlled 14-language $\times$ 200-example subset (LLaMA3). ECE and AUROC in percent. Per-language P1 (letter probability) values at full MMMLU scale are in Table 1.