Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

High Confidence, Low Accuracy: Probing Ownership Bias in LLMs on Banjarese and Dayak Ngaju

Muhammad Ihsan [a,*], *Didi Andriawan* [b]

[a] *Sistem Informasi, Politeknik Murakata, Hulu Sungai Tengah, 71315, Indonesia*
[b] *Sistem Informasi, Universitas Sapta Mandiri, Balangan, 71618, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:

Diterima Redaksi: 09 Juli 2026
Revisi Akhir: 20 Agustus 2026
Diterbitkan *Online*: 16 September 2026

KATA KUNCI

Confidence calibration,
Ownership bias,
Banjarese Dayak Ngaju language,
Large language models,
IndoMMLU

KORESPONDENSI

E-mail:
muhammad.ihsan@politeknikmurakata.ac.id*

A B S T R A C T

AI assistants often give wrong answers with high confidence, especially when answering in regional languages. One cause is ownership bias: a model trusts its own answer more when that answer appears in the assistant role. This effect is well documented in English, but it is unclear whether it also appears in low-resource languages such as Banjarese and Dayak Ngaju. This study measures the extent of ownership bias and the calibration quality of four open-weight 8-bit LLMs when tested on school exam questions in Banjarese and Dayak Ngaju, drawn from the IndoMMLU benchmark. Four models (Qwen3-8B, Llama-3.1-8B, SahabatAI-Llama-8B, and SahabatAI-Gemma2-9B) are evaluated under two conditions: Own, in which the model's answer is placed in the assistant role, and Reframed, in which the same answer is attributed to the user. We report accuracy, ownership bias, Expected Calibration Error (ECE), and Brier Score. Llama-3.1-8B shows the highest ownership bias (49.93 pp on Banjarese and 53.83 pp on Dayak Ngaju), while SahabatAI-Gemma2-9B shows almost no bias (0.91–1.33 pp) but assigns 100% confidence to every answer. The same pattern appears in both languages, indicating that ownership bias is driven by the model's architecture and training, not by the language being tested. AI assistants that are overconfident in English remain overconfident in Banjarese and Dayak Ngaju as well. Practitioners should treat verbalized confidence as an unreliable signal and use model-level calibration such as temperature scaling or logprob-based methods, rather than relying on prompt-level reframing alone.

1. INTRODUCTION

Imagine an AI tutoring assistant used in schools across South Kalimantan. A student asks for help with a reading question in their local language, either Banjarese or Dayak Ngaju. The model answers with great confidence, but the answer is wrong. The bigger problem is not the error itself, but the high confidence behind it: the student trusts the model and accepts the wrong answer. This situation is not imaginary; it is a direct result of the calibration failure that this paper measures.

Instruction-tuned large language models (LLMs) can now solve many reasoning and knowledge tasks. But one important quality is often overlooked: calibration, which refers to how well a model's stated confidence matches its actual accuracy. A well-calibrated model that says "I'm 80% confident" should be right about 80% of the time. Poor calibration is not just a small academic problem. In education, a tutoring system that sounds certain while wrong can mislead students. In healthcare, a confident but incorrect diagnosis can delay proper treatment. In public services, overconfident answers can push citizens toward bad decisions. The practical risk is clear: users trust confident machines, and low-resource-language users often have fewer tools to verify answers.

Recent work by Sanz-Guerrero et al. [1] identified a clear and repeatable calibration problem called ownership bias. When a model's own answer appears in the assistant role of the chat template, the model gives it higher confidence, up to 26 percentage points more, than when the same answer is shown as a user's answer. The likely cause is RLHF training, which teaches the model to "stand behind" its own answers. This creates a hidden bias that we cannot see directly but can measure by asking the model for its confidence. So far, however, all evidence for this effect comes from English and other high-resource European languages.

https://doi.org/10.25077/TEKNOSI.v12i2.2026.386-392

This leaves an open question: does ownership bias still happen, and how strong is it, when models are tested on languages they have rarely seen during training? Banjarese (Bahasa Banjar), the main language of South Kalimantan with about 4.1 million speakers, and Dayak Ngaju, another Austronesian language spoken by about 1.5 million people in Central Kalimantan, are exactly this kind of language. Both are strongly regional and largely absent from the large text corpora used to train LLMs. If ownership bias is a general feature of instruction-tuned models, and not just a side effect of English skill, it should also appear in these two languages, and it may be even stronger because the models know less about these topics. The IndoMMLU benchmark [2] provides standard school exam questions in both Banjarese and Dayak Ngaju, making them a good test case for this cross-language study.

This paper makes three practical contributions. First, we show that ownership bias and calibration problems known in English also appear in Banjarese and Dayak Ngaju, so users of low-resource languages face the same reliability risks. Second, we show that ownership bias is architecture-specific: its size depends mainly on the base model's training pipeline, not on whether the model is general-purpose or adapted to Indonesian. This helps practitioners choose models based on calibration behavior rather than just accuracy. Third, we show that reframing can lower Expected Calibration Error (ECE) in several ways. Hence, a drop in ECE alone is not enough to determine whether reframing has truly fixed the calibration.

2. LITERATURE REVIEW

2.1. Calibration in Large Language Model

Model calibration has been studied in the context of neural networks since Guo et al. [3] demonstrated that modern deep networks are systematically overconfident. In the LLM era, prior work has shown that large language models can verbalize uncertainty when explicitly prompted. That token-level log-probabilities have been explored as automated confidence proxies. Liu et al. [4] advanced verbalized confidence through uncertainty-aware instruction tuning (UaIT), aligning LLM self-expression with probabilistic uncertainty estimates derived from multi-sampling. Instruction tuning has been repeatedly found to worsen calibration, partly because RLHF-style training encourages models to appear confident and helpful. Beyond post hoc scaling, training-time calibration approaches such as MaC-Cal [5] have introduced stochastic masking and adaptive sparsity to regularize the classifier geometry. However, these techniques remain primarily validated in computer-vision settings. Zhang et al. [6] further demonstrated that the gradient decay rate in Softmax optimization directly governs model confidence: smaller decay rates amplify overconfidence by continuing to reward high-probability predictions, whereas larger decay rates smooth the training sequence and yield better calibrated outputs. In parallel, Fernando et al. [7] showed that class imbalance simultaneously degrades both classification accuracy and confidence calibration: their Dynamically Weighted Balanced (DWB) loss couples inverse-frequency class weights with a probability-dependent regularizer, yielding models that are not only fairer to minority classes but also better calibrated across all classes. On the post-hoc front, Balanya et al. [8] advanced Temperature Scaling by making the temperature factor input-dependent: their Entropy-based Temperature Scaling (HTS) assigns higher temperatures to predictions with greater predictive entropy, thereby selectively deflating the confidence of uncertain predictions while preserving well-separated predictions.

At the same time, uncertainty quantification methods for LLMs have expanded rapidly. Lin et al. [9] demonstrated that black-box LLMs can still be queried for reliable uncertainty estimates by generating multiple responses and measuring their semantic dispersion; this multi-sampling paradigm offers a practical alternative when only API-level access is available. Extending this line of investigation, Huang et al. [10] conducted a large-scale empirical comparison of twelve uncertainty estimation methods across nine LLMs and six tasks, confirming that sample-based techniques consistently outperform single-inference confidence metrics; their findings further highlight that model-specific idiosyncrasies and RLHF prompt templates can significantly distort uncertainty estimates, underscoring the fragility of relying on any single confidence signal. Shorinwa et al. [11] surveyed the growing UQ literature and highlighted that hallucinations are often expressed with high confidence, making calibration a safety-relevant property. Vashurin et al. [12] introduced LM-Polygraph, a unified benchmark that compares UQ methods across text-generation tasks, finding that sample-diversity-based methods fail on multiple-choice tasks such as MMLU because constrained output spaces offer insufficient variation, a limitation directly relevant to our forced-choice format. Bentegeac et al. [13] showed that token-level probabilities detect overconfidence in medical question answering better than verbalized confidence, and Qazi et al. [14] found a Dunning-Kruger-like pattern in multilingual fact-checking where smaller models are more confident when they are less accurate. Beyond the LLM domain, Bauer et al.[15] evaluated calibration methods on CNNs under data drift in manufacturing quality monitoring, underscoring that calibration reliability under domain shift is a cross-cutting challenge across machine learning applications. In the surveillance video domain, Mohanarangan and Palanisamy [16] demonstrated that temperature scaling combined with margin-based regularization effectively mitigates overconfidence in CNN-based anomaly detection, illustrating that overconfidence mitigation techniques are applicable across modalities. These findings collectively motivate our focus on verbalized confidence in low-resource settings, where calibration is understudied but safety-critical.

2.2. Ownership Bias

The ownership bias was formally characterized by Sanz-Guerrero et al. [1], who found that the chat template creates an asymmetry: answers placed in the assistant role receive significantly higher confidence scores than the same answers placed in the user role. This effect was demonstrated across six open-weight LLMs and three benchmarks (MMLU, TriviaQA, ARC). The authors proposed a simple mitigation: framing the model's answer as a user message during confidence elicitation, which they termed the reframing strategy. Complementing this line of evidence, Thelwall [17] showed that LLM evaluative judgments are highly sensitive to input presentation: providing an article abstract yielded more reliable quality scores than presenting the full text. Complex system prompts with explicit evaluation criteria outperformed terse instructions, underscoring that the way

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.386-392

information is framed for an LLM profoundly shapes its expressed confidence. In a related domain, Liu et al. [18] systematically compared AI content detectors and human reviewers in identifying AI-generated medical writing, finding that experienced reviewers and the most accurate detectors (Originality.ai, ZeroGPT) could reliably distinguish paraphrased AI texts from human-authored ones, although both occasionally misclassified human writing as AI-generated, highlighting the broader challenge of reliably assessing LLM-generated content.

2.3. Low-Resource Language Evaluation

The IndoMMLU benchmark [2] covers 63 tasks across multiple Indonesian regional languages, including Banjarese and Dayak Ngaju, providing the first standardized multiple-choice evaluation suite for these languages. Prior work has documented that LLMs generally perform worse on low-resource languages, but calibration behavior specifically has not been studied in this context. In the multilingual entity-linking domain, Ramamoorthy et al. [19] introduced MERLIN, a testbed for linking named entities in five low-resource languages (including Indonesian) to a knowledge base, demonstrating that even massively multilingual models struggle with ambiguity in non-English contexts and that supplementary visual or contextual signals are required to boost accuracy. More broadly, the design of LLM evaluation benchmarks has received increasing attention: Zhang et al. [20] introduced AcademicEval. This live benchmark continuously updates from arXiv to prevent data contamination. It employs both automatic metrics and LLM-as-a-Judge evaluation to assess long-context academic writing tasks across hierarchical abstraction levels. Beyond accuracy-centric evaluation, de Zarza et al. [21] demonstrated that multilingual LLM deployment introduces additional operational dimensions: their energy–performance analysis across thirteen typologically diverse languages (including low-resource varieties such as Basque and Luxembourgish) revealed that energy consumption varies up to threefold across models and exhibits only weak correlation with task performance (rs = 0.109), while language choice itself acts as a measurable deployment parameter with Romance languages achieving lower energy consumption than English on average; although their work focuses on inference-time energy rather than calibration, it underscores that comprehensive multilingual evaluation must account for language-dependent operational trade-offs beyond pure accuracy, a perspective that complements our calibration-focused assessment.

At the broader landscape level, Qin et al. [22] provided a comprehensive survey of multilingual large language models (MLLMs), introducing a unified taxonomy that categorizes alignment strategies into parameter-tuning alignment (PTA), which involves fine-tuning model parameters through pretraining, supervised fine-tuning, RLHF, and downstream fine-tuning, and parameter-frozen alignment (PFA), which achieves multilingual capability through prompting without parameter adjustment; their survey identifies low-resource languages as a critical emerging frontier and emphasizes that multilingual alignment remains a key challenge, contextualizing our evaluation of Indonesian-adapted models (SahabatAI, representing PTA) against general-purpose baselines within the broader MLLM research trajectory.

3. METHOD

3.1. Dataset

We evaluated two low-resource Austronesian language subsets from IndoMMLU [2]. The Banjarese subset is a collection of multiple-choice questions drawn from school examinations at SD (elementary school), SMP (junior high school), and SMA (senior high school) levels. Each question consists of a question stem, four or five answer choices (A–E), and a ground-truth answer label. After filtering for data quality (removing items with missing choices or ambiguous answer keys), the Banjarese evaluation set contains N = 143 questions. The Dayak Ngaju subset (N = 109) is another under-resourced Austronesian variety with the same structure, drawn from school examinations across SD–SMP–SMA levels. We evaluate on both subsets to test whether ownership-bias and calibration patterns generalize across low-resource Indonesian regional languages.

Table 1. Distribution of IndoMMLU questions used in this study

Subject	Level Label	Count
Bahasa Banjar	SD-SMP-SMA	143
Bahasa Dayak Ngaju	SD-SMP-SMA	109

3.2. Model

We evaluate four open-weight instruction-tuned language models across two families, all run locally via mlx-lm on an Apple Silicon Mac (Metal/GPU inference). We deliberately include both general-purpose models and models with Indonesian continual pre-training (CPT) to test whether language-specific adaptation moderates ownership bias.

Table 2. Models used in the experiment

Model Name	Provider	Description
Qwen3-8B	Alibaba	Model with featuring hybrid thinking/non-thinking modes; we use non-thinking mode (enable thinking=False) for consistent single-token responses
Llama-3.1-8B	Meta	An instruction-tuned model via RLHF with broad multilingual exposure
SahabatAI-Llama-8B	GoTo	Llama-3-8B was continually pre-trained on Indonesian and Southeast Asian corpora via the SEA-LION and SahabatAI programs. Designed to improve factual and linguistic coverage of Indonesian regional languages
SahabatAI-Gemma2-9B	Goto	Gemma-2-9B with the same Indonesian CPT treatment, representing a different base architecture under the same language-adaptation regime

All four models are evaluated using their instruction-tuned chat variants with greedy decoding (temperature = 0) to ensure reproducibility. All models are quantized to 8-bit precision for a fair computational comparison.

3.3. Experimental Design

For each question q in the dataset, we conduct a two-stage, two-condition inference procedure:

Stage 1: Answer Elicitation. The model is prompted with the question-and-answer choices in a single user turn and instructed

to answer with a single letter (A–E). The raw text response is recorded.

```
[User]: Ini adalah soal {subject} untuk {level}.
Pilihlah salah satu jawaban yang dianggap benar!
{question}
A. {choice_a}
B. {choice_b}
...
Jawab hanya dengan huruf pilihan (A, B, C, D, atau E).
```

Stage 2a: Own Condition. The model's Stage 1 response is appended to the conversation history as an assistant message. A new user turns on and then asks the model to rate its confidence:

```
[User]: [question + choices]
[Asst]: [model's answer from Stage 1]
[User]: Seberapa yakin Anda dengan jawaban tersebut?
Berikan angka dari 0 hingga 100.
Jawab hanya dengan satu angka bulat.
```

Stage 2b: Reframed Condition. A fresh, single-turn prompt presents the identical model answer as a student's response, stripping all chat-template ownership signals [1]:

```
[User]: [question + choices]
Seorang siswa memberikan jawaban: {model_answer}.
Seberapa yakin Anda bahwa jawaban {model_answer}
tersebut benar? Berikan angka dari 0 hingga 100.
Jawab hanya dengan satu angka bulat.
```

3.4. Calibration Metrics

To make the metrics concrete, consider a teacher who answers a set of multiple-choice questions and is asked each time, "How sure are you?" If the teacher says "80% sure" on ten questions, we expect about eight of those answers to be correct. Calibration is simply the match between claimed confidence and actual accuracy. Ownership bias is a different distortion: the same teacher trusts an answer more when it is written in his or her own notebook than when the identical answer is written in a student's notebook. Our experiment measures both effects for LLMs.

Let n denote the total number of valid responses, c_i The normalized confidence (0–1) elicited from the model for the question i , and $a_i \in \{0, 1\}$ the correctness indicator.

Expected Calibration Error (ECE):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\overline{a}(B_m) - \overline{c}(B_m)| \tag{1}$$

where M = 10 equal-width bins, Bm is the set of samples in bin m, and a(Bm), c(Bm) are the mean accuracy and mean confidence in that bin, respectively

Maximum Calibration Error (MCE):

$$\mathrm{MCE} = \max_{m \in 1..M} |\overline{a}(B_m) - \overline{c}(B_m)| \tag{2}$$

Brier Score (ECE):

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} (c_i - a_i)^2 \tag{3}$$

Ownership Bias (OB): The mean signed difference in verbalized confidence (in percentage points) between the Own and Reframed conditions:

$$\mathrm{OB} = \frac{1}{n} \sum_{i=1}^{n} \left(c_i^{\mathrm{own}} - c_i^{\mathrm{reframed}}\right) \tag{4}$$

4. RESULTS AND DISCUSSIONS

4.1. Answer Accuracy

Before we can interpret a model's confidence, we need to know how much it knows; Tables 3 and 4 report answer accuracy for all four models on the Banjarese and Dayak Ngaju subsets, respectively.

Table 3: Accuracy on Banjarese IndoMMLU by model family

Model Name	Family	Accuracy
Qwen3-8B	General	35.7%
Llama-3.1-8B	General	47.2%
SahabatAI-Llama-8B	Indonesian CPT	46.9%
SahabatAI-Gemma2-9B	Indonesian CPT	53.1%

Table 4: Accuracy on Dayak Ngaju IndoMMLU by model family

Model Name	Family	Accuracy
Qwen3-8B	General	35.8%
Llama-3.1-8B	General	29.0%
SahabatAI-Llama-8B	Indonesian CPT	31.2%
SahabatAI-Gemma2-9B	Indonesian CPT	38.5%

On Banjarese, all four models clear the 25% random-chance baseline, yet only SahabatAI-Gemma2-9B convincingly passes 50%, reaching 53.1%. This suggests that Indonesian continual pre-training on the Gemma-2-9B base confers a genuine factual advantage on regional school content. Llama-3.1-8B and SahabatAI-Llama-8B settle near 47%, while the newer-generation Qwen3-8B trails at 35.7%. The lesson is a useful caution: gains in English-centric reasoning do not automatically carry over to low-resource regional knowledge.

Dayak Ngaju proves harder for every model, which fits its smaller speaker base and thinner textual footprint. SahabatAI-Gemma2-9B again leads at 38.5%, and the family ranking is preserved, but three of the four models lose between 14.6 and 18.2 percentage points relative to Banjarese: 14.6 pp for SahabatAI- Gemma2-9B, 15.7 pp for SahabatAI-Llama-8B, and a steep 18.2 pp for Llama-

https://doi.org/10.25077/TEKNOSI.v12i2.2026.386-392

3.1-8B, whose accuracy falls from 47.2% to 29.0%. The one exception is Qwen3-8B, which stays essentially flat across the two languages (35.7% versus 35.8%). Accuracy is therefore sensitive to language in its magnitude but stable in its ordering, and this stability of ordering is a thread that runs through every result that follows.

4.2. Ownership Bias

The central question of this paper is how much extra confidence a model grants an answer simply because that answer appears in its own assistant role. Tables 5 and 6 report ownership-bias statistics for Banjarese and Dayak Ngaju, respectively.

Table 5: Ownership bias statistics for Banjarese

Model	μ Conf Own	μ Conf Ref	OB (pp)	*p*	Cohen *d*
Qwen3	93.7%	79.4%	14.3	<0.0001	1.628
Llama-3.1	79.5%	29.6%	49.93	<0.0001	1.815
SahabatAI-Llama	88%	68.5%	19.48	<0.0001	0.874
SahabatAI-Gemma2	100%	99.1%	0.91	0.0002	0.316

Table 6: Ownership bias statistics for Dayak Ngaju.

Model	μ Conf Own	μ Conf Ref	OB (pp)	*p*	Cohen *d*
Qwen3	91.2%	83.0%	8.26	<0.0001	1.596
Llama-3.1	77.7%	23.8%	53.83	<0.0001	2.076
SahabatAI-Llama	77.3%	59.9%	17.43	<0.0001	0.600
SahabatAI-Gemma2	99.9%	98.6%	1.33	0.0001	0.396

On Banjarese, the first lesson is that this ownership bias is architecture-specific rather than a simple function of model family. Llama-3.1-8B opens an extraordinary49.93 pp gap (Cohen's d = 1.815), nearly twice the roughly 26 pp reported for English MMLU [1], whereas Qwen3-8B, also a general-purpose model, shows a far more modest 14.30 pp, comparable to the Indonesian adapted SahabatAI-Llama-8B at 19.48 pp. SahabatAI-Gemma2-9B looks almost unbiased at 0.91 pp. Still, the number is deceptive: both its Own and its Reframed confidences are pinned near 100%. Hence, the ownership signal is not absent but hidden beneath a confidence-saturation pathology. Figure 1 makes these dynamics visible by contrasting each model's per-question confidence distributions across both languages, where the saturation of SahabatAI-Gemma2-9B and the large Own-to-Reframed shift of Llama-3.1-8B stand out briefly.

Does this pattern survive a change of language? Dayak Ngaju answers clearly: it does. The ranking of models by ownership bias is identical (Llama-3.1-8B far above SahabatAI-Llama-8B, then Qwen3-8B, then SahabatAI-Gemma2-9B), and for the most biased model, the effect even intensifies, with Llama-3.1-8B reaching 53.83 pp (d = 2.076). SahabatAI-Gemma2-9B again collapses to a near-zero 1.33 pp under saturation, SahabatAI-Llama-8B holds at 17.43 pp, and Qwen3-8B settles at 8.26 pp, with every effect statistically significant (all $p < 0.001$). That the same ordering reappears in a second, even lower-resource language is the central piece of evidence that ownership bias belongs to a model's architecture and training regime, not to any single language. Figure 2 states this directly, placing the two languages side by side for every model.

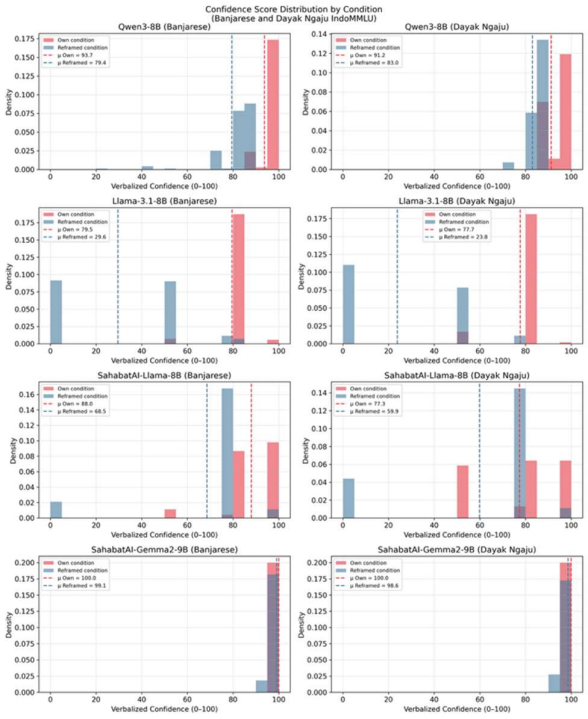

Figure 1: Per-question verbalized confidence distributions for each model under the Own and Reframed conditions, shown for Banjarese (left column) and Dayak Ngaju (right column).

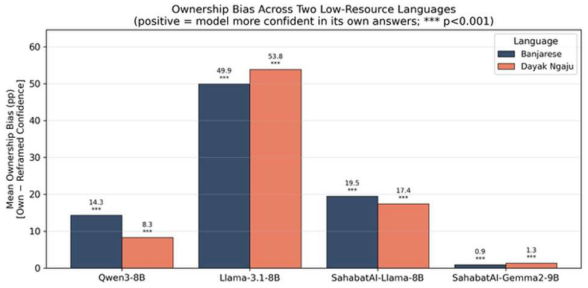

Figure 2: Mean ownership bias (Own minus Reframed confidence, in percentage points) for each model on Banjarese and Dayak Ngaju.

4.3. Calibration Metrics

A large ownership gap tells us that confidence is malleable, but not whether it is accurate. We therefore ask how closely each model's stated confidence tracks its real accuracy, and whether reframing the answer as a user message repairs the mismatch; Tables 7 and 8 report ECE, MCE, and Brier Score under both conditions.

Table 7: Calibration metrics for Banjarese

Model	ECE Own	ECE Ref	ΔECE	MCE Own	MCE Ref	BS Own	BS Ref
Qwen3	0.5808	0.4378	+0.1430	0.6147	0.7500	0.5658	**0.4162**
Llama-3.1	0.3232	0.2042	+0.1190	0.3338	0.3846	0.3533	**0.3077**
SahabatAI-Llama	0.4115	0.2867	+0.1248	0.7500	0.3333	0.4040	**0.3252**
SahabatAI-Gemma2	**0.4685**	**0.4594**	**+0.0091**	**0.4685**	**0.4594**	**0.4685**	**0.4583**

Table 8: Calibration metrics for Dayak Ngaju

Model	ECE Own	ECE Ref	ΔECE	MCE Own	MCE Ref	BS Own	BS Ref
Qwen3	0.5546	0.4757	+0.0789	0.5796	0.7500	0.5658	0.4162
Llama-3.1	0.4869	0.2103	+0.2766	1.0000	0.3846	0.3533	0.3077
SahabatAI-Llama	0.4610	0.3417	+0.1193	0.6571	0.3333	0.4040	0.3252
SahabatAI-Gemma2	0.6142	0.6009	+0.0133	0.4685	0.4594	0.4685	0.4583

On Banjarese, reframing lowers ECE for three of the four models, yet the improvement means something different in every case. Llama-3.1-8B improves on paper (0.3232 to 0.2042) only by overshooting into underconfidence, dropping to 29.6% confidence against 47.2% accuracy. Qwen3-8B reduces its overconfidence (from 0.5808 to 0.4378) but remains badly miscalibrated, with 79.4% confidence against 35.7% accuracy. Only SahabatAI-Llama-8B lands somewhere useful, with reframed confidence (68.5%) approaching its accuracy (46.9%); notably, its MCE drops sharply from 0.7500 to 0.3333, the largest single MCE improvement in the Banjarese set.

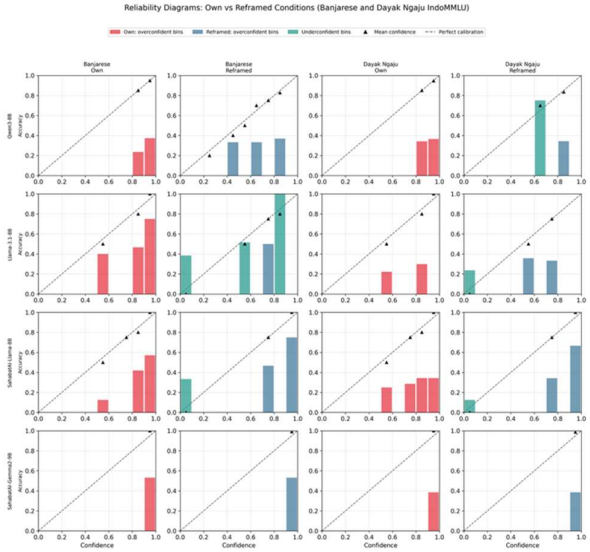

Figure 3: Reliability diagrams for each model and condition. Rows are models; columns are Banjarese Own, Banjarese Reframed, Dayak Ngaju Own, and Dayak Ngaju Reframed.

The MCE picture for Qwen3-8B moves in the opposite direction, from 0.6147 to 0.7500, because reframing redistributes confidence into a new worst-case bin even as ECE overall decreases, a reminder that ECE and MCE can diverge when the confidence distribution shifts non-uniformly across bins. SahabatAI-Gemma2-9B, pinned at 100% confidence on every question, posts the worst Brier Score (0.4685) despite the best accuracy, and reframing barely moves it to 99.1% because the saturation is structural rather than a prompt artifact. Figure 3 shows these patterns bin by bin for both languages: the Own-condition bars sit well below the perfect-calibration diagonal, and reframing narrows the gap without always closing it.

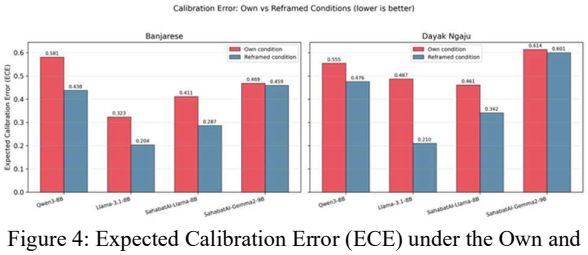

Figure 4: Expected Calibration Error (ECE) under the Own and Reframed conditions for each model, shown for Banjarese (left) and Dayak Ngaju (right).

The Dayak Ngaju results reinforce rather than complicate this picture. Reframing again lowers ECE for all four models, and the same four behaviors reappear. Llama-3.1-8B shows the most dramatic shift (ΔECE = +0.2766) as confidence plunges from 77.7% to 23.8%, once more overshooting into underconfidence against 29.0% accuracy. Its MCE Own reaches the theoretical maximum of 1.0000, driven by a single question answered with 100% confidence that turned out to be incorrect, an extreme but isolated outlier that MCE, unlike ECE, does not dilute. Qwen3-8B and SahabatAI-Llama-8B improve moderately (+0.0789 and +0.1193), while SahabatAI-Gemma2-9B barely moves (+0.0133) under persistent saturation. SahabatAI-Gemma2-9B records the highest ECE of any model in any condition across both languages (0.6142 Own, 0.6009 Reframed), a consequence of assigning near-100% confidence to questions it answers correctly only 38.5% of the time. Absolute ECE is higher than on Banjarese for every model, a direct reflection of weaker factual grounding in the lower-resource language. Taken together, the calibration evidence shows that reframing is a useful diagnostic that reduces ECE across both languages and all models, but its numerical gains conceal four distinct failure modes, namely underconfidence (Llama-3.1-8B), residual overconfidence (Qwen3-8B), near-calibration (SahabatAI-Llama-8B), and saturation (SahabatAI-Gemma2-9B), that persist regardless of language. Figure 4 contrasts Own and Reframed ECE across both languages, and Figure 5 consolidates accuracy, expressed confidence, and ownership bias for all four models in a single view.

https://doi.org/10.25077/TEKNOSI.v12i2.2026.386-392

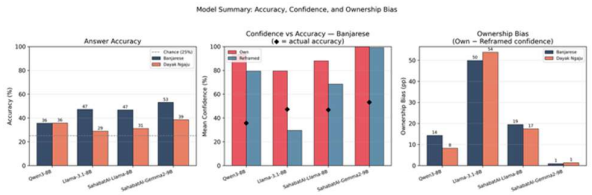

Figure 5: Model summary across all four models: answer accuracy on both languages (left), mean confidence in the Own and Reframed conditions against actual accuracy on Banjarese (center), and ownership bias on both languages (right).

4.4. Discussions

Four findings emerge from this experiment, each with distinct implications. First, Ownership bias is architecture-specific, not uniformly tied to the model family. Bias ranges from 0.91 pp (SahabatAI-Gemma2-9B) to 49.93 pp (Llama-3.1-8B). Crucially, Llama-3.1-8B, a general-purpose model, shows the most extreme bias in the experiment (d = 1.815). In contrast, Qwen3-8B, another general-purpose model, shows only moderate bias (14.30 pp, d = 1.628), comparable to SahabatAI-Llama-8B (19.48 pp). This demolishes a simple general-vs-CPT dichotomy: the Llama post-training regime instills a stronger ownership prior than either the Qwen3 or the SahabatAI-Llama training pipelines, regardless of language coverage. Indonesian CPT on the Llama base reduces bias (19.48 vs. 49.93 pp, a 61% reduction), confirming that language-adaptive pre-training has a moderating effect. Still, the base model's training pipeline remains the primary determinant.

Second, Reframing improves ECE for three models but through different mechanisms. ΔECE is positive for Qwen3-8B (+0.143), Llama-3.1-8B (+0.119), and SahabatAI-Llama-8B (+0.125). However, the nature of improvement differs: Llama-3.1-8B achieves ECE gain by plunging into underconfidence (Ref=29.6% vs Acc=47.2%); Qwen3-8B reduces overconfidence but remains miscalibrated (Ref=79.4% vs Acc=35.7%); only SahabatAI-Llama-8B achieves a genuinely useful reframing (Ref=68.5% ≈ Acc=46.9%). ECE is a scalar that does not distinguish between the direction of miscalibration. The reframing intervention [1] is most effective when it brings expressed confidence close to actual accuracy. This condition holds only when domain knowledge is sufficient to produce a grounded uncertainty estimate without the ownership scaffold.

Third, Confidence saturation is a distinct failure mode, orthogonal to ownership bias. SahabatAI-Gemma2-9B achieves the highest accuracy (53.1%) but outputs 100% confidence on every question under both conditions, yielding the worst Brier Score (0.4685). Saturation is not an ownership artifact (it persists after reframing, at 99.1%) but instead reflects the model's verbalized confidence architecture under the Indonesian CPT chat template. A model can be both the most accurate and most comprehensively miscalibrated in the set. Verbalized confidence protocols must be independently validated per model-architecture before use in downstream applications.

Fourth, Reframing is a diagnostic instrument, not a calibration solution [1]. Reframing improves ECE numerically for 3/4 models, but improvement does not imply calibration. It either overshoots (creating underconfidence), undershoots (reducing but not eliminating overconfidence), or produces calibrated output only under favorable conditions (sufficient domain knowledge). Logprob-based confidence, temperature scaling [1], and Platt calibration operate at the model level and are likely more reliable as calibration interventions for low-resource language settings.

Limitations. This study elicits verbalized confidence, which depends on the model's linguistic ability to express uncertainty. Token-logprob methods may yield different bias profiles. The evaluation covers four models across two families, two languages (Banjarese and Dayak Ngaju), and one benchmark; generalization to other IndoMMLU regional languages (Sundanese, Minangkabau, Makassar) remains an open question.

5. CONCLUSION

This paper presented a systematic study of ownership bias and confidence calibration for four open-weight 8-bit LLMs across two low-resource Indonesian regional languages, Banjarese (N=143) and Dayak Ngaju (N=109), evaluated on IndoMMLU [2] school-examination questions. Across both languages, the evidence converges on a single message: calibration behavior is governed by a model's architecture and post-training regime, not by its language coverage. Ownership bias spanned a negligible 0.91 pp to an extreme 49.93 pp on Banjarese, and the same model ordering reappeared on Dayak Ngaju, where Llama-3.1-8B reached 53.83 pp, confirming that the effect is architectural rather than language-specific. Reframing lowered ECE for most models yet rarely produced genuine calibration, instead exposing four distinct outcomes: underconfidence, residual overconfidence, near-calibration, and saturation. The case of SahabatAI-Gemma2-9 B, the most accurate model yet the least calibrated due to its confidence saturation, makes clear that high accuracy and trustworthy confidence are separate properties.

Together, these findings establish that calibration pathology in low-resource language settings is neither uniform nor reducible to a single mechanism. Indonesia has over 700 regional languages; each may expose a different vulnerability in an LLM's calibration architecture. Practitioners must treat verbalized confidence as unreliable by default and pursue model-level calibration, logprob-based methods, temperature scaling [3], or language-targeted CPT, rather than relying on prompt-level heuristics.

BIBLIOGRAPHY

[1] J. Sanz-Guerrero and others, "Ownership Bias in Instruction-Tuned Large Language Models," 2026.

[2] F. Koto and others, "IndoMMLU: A Massive Multilingual Language Understanding Benchmark for Indonesian," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *International Conference on Machine Learning*, 2017, pp. 1321–1330.

[4] Z. Liu and others, "UAIT: Uncertainty-Aware Instruction Tuning for Large Language Models," in

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[5] J. Ni, H. Zhao, Y. Yang, and D. Guo, "Deep neural network calibration by reducing classifier shift with stochastic masking," *Pattern Recognit.*, vol. 177, p. 113217, 2026.

[6] S. Zhang and L. Xie, "Advancing neural network calibration: The role of gradient decay in large-margin Softmax optimization," *Neural Networks*, vol. 178, p. 106457, 2024.

[7] K. R. M. Fernando and C. P. Tsokos, "Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[8] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Comput. Appl.*, vol. 36, pp. 8073–8095, 2024.

[9] Z. Lin, S. Trivedi, and J. Sun, "Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models," *Transactions on Machine Learning Research*, 2024.

[10] Y. Huang *et al.*, "Look Before You Leap: An Exploratory Study of Uncertainty Analysis for Large Language Models," *IEEE Transactions on Software Engineering*, 2024.

[11] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar, "A Survey on Uncertainty Quantification of Large Language Models: Taxonomy, Open Research Challenges, and Future Directions," *ACM Comput. Surv.*, vol. 58, no. 3, p. 63, 2025.

[12] R. Vashurin, E. Fadeeva, A. Vazhentsev, and others, "Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph," in *Transactions of the Association for Computational Linguistics*, 2025.

[13] R. Bentegeac, B. Le Guellec, G. Kuchcinski, and others, "Token Probabilities to Mitigate Large Language Models Overconfidence in Answering Medical Questions: Quantitative Study," *J. Med. Internet Res.*, 2025.

[14] I. A. Qazi, Z. Khan, A. Ghani, and others, "Large language models show Dunning-Kruger-like effects in multilingual fact-checking," *Sci. Rep.*, 2026.

[15] J. C. Bauer, S. Trattnig, F. Vieltorf, and R. Daub, "Handling data drift in deep learning-based quality monitoring: evaluating calibration methods using the example of friction stir welding," *J. Intell. Manuf.*, vol. 37, pp. 759–774, 2026.

[16] K. Mohanarangan and P. Palanisamy, "Confidence-Aware Cascaded Learning for Real-Time Weakly Supervised Anomaly Detection in Surveillance Video," *IEEE Access*, 2026.

[17] M. Thelwall, "Evaluating research quality with Large Language Models: An analysis of ChatGPT's effectiveness with different settings and inputs," *Journal of Data and Information Science*, vol. 10, no. 1, pp. 7–25, 2025.

[18] J. Q. J. Liu, K. T. K. Hui, and others, "The great detectives: humans versus AI detectors in catching large language model-generated medical writing," *International Journal for Educational Integrity*, vol. 20, no. 8, 2024.

[19] S. Ramamoorthy, V. Shah, S. Khanuja, Z. Sheikh, and others, "MERLIN: A Testbed for Multilingual Multimodal Entity Recognition and Linking," in *Transactions of the Association for Computational Linguistics*, 2026.

[20] H. Zhang, T. Feng, P. Han, and J. You, "AcademicEval: Live Long-Context LLM Benchmark," *Transactions on Machine Learning Research*, 2025.

[21] I. de Zarzà, M. Liz, J. de Curtò, and C. T. Calafate, "Energy-Aware Multilingual Evaluation of Large Language Models," *Energies (Basel).*, 2026.

[22] L. Qin, Q. Chen, Y. Zhou, Z. Chen, and others, "A survey of multilingual large language models," *Patterns*, 2025.

NOMENKLATUR

• π = Total number of valid responses

• c_i = Normalized confidence (0–1) elicited from the model for question i

• α_i = Correctness indicator (1 if correct, 0 if incorrect)

• M = Number of equal-width bins (M = 10)

• B_m = Set of samples in bin m

• $\overline{\boldsymbol{a}}(\boldsymbol{B_m})$= Mean accuracy in bin m

• $\overline{\boldsymbol{c}}(\boldsymbol{B_m})$ = Mean confidence in bin m

• c_i^{own} = Model confidence under the Own condition (answer in assistant role)

• $c_i^{reframed}$ = Model confidence under the Reframed condition (answer in user role)

• ΔECE = Difference in ECE between the Own and Reframed conditions

• p = p-value (statistical significance)

• d = Cohen's d (effect size)

• α = Significance threshold (α = 0.05)

• H_o = Null hypothesis (no ownership bias)

• ECE = Expected Calibration Error: weighted average calibration error across bins

• MCE = Maximum Calibration Error: worst-case calibration error across all bins

• BS = Brier Score: mean squared error between predicted confidence and correctness

• OB = Ownership Bias: mean signed confidence difference (in percentage points) between Own and Reframed Conditions

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.386-392