



How Reliable Is LLM Confidence in Hindi and Marathi? A Study of Calibration and Answer Stability

Sneha Sanjay Deshmukh¹ and Dr. Mohammad Abuzar Ansari²

¹Master's Student, Data Science; ²Head of Department

Department of Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous), Sion (West),
Mumbai, India

sneha.deshmukh0310@gmail.com¹ abuzara@sies.edu.in²

Abstract

Language models are increasingly used to answer exam-style questions in Indian languages, and their confidence scores are often treated as a signal for when an answer can be trusted. We checked how far that trust is justified for two small open-weight models, Qwen2.5-1.5B-Instruct and Qwen2.5-3B-Instruct, using 300 native-language multiple-choice questions each in Hindi and Marathi from the MILU benchmark. Confidence was read directly from the probability the model assigns to each answer letter. The 3B model scored 45.0% in Hindi and 32.7% in Marathi while reporting an average confidence above 91%, which gives expected calibration errors (ECE) of 0.483 and 0.592. The Marathi–Hindi difference in ECE had a 95% bootstrap interval of 0.034 to 0.185. Reordering the answer options, a change that leaves the question unchanged, altered the chosen answer for about 38–40% of questions with the 3B model and 52–57% with the 1.5B model, even though overall accuracy and calibration hardly moved. A single temperature parameter fitted on half of the questions brought held-out ECE of the 3B model below 0.1 in both languages, including on reordered options, but it did not improve the model's ability to separate right answers from wrong ones (AUROC about 0.6). This is a small-scale study of one model family and one prompt format, and we state its limits openly. The main message is practical: an overall accuracy number can look steady while individual answers and confidence scores are not.

Keywords: *calibration; large language models; Hindi; Marathi; prompt sensitivity; selective prediction; temperature scaling*

1. Introduction

Anyone who has used a chatbot knows it will answer almost anything in a confident tone. In an exam-preparation tool, a tutoring app or a help desk, that

tone matters, because users and developers often decide whether to rely on an answer based on how sure the system sounds. For models that give a probability alongside the answer, the natural question is whether that probability means what it says: when a model reports 90% confidence, is it right about nine times out of ten?

This question has been studied extensively for English, but much less for Indian languages. Hindi and Marathi are spoken by hundreds of millions of people, yet evaluation resources for them are still thin, and a model's behaviour in English does not automatically carry over. Marathi in particular has far less online text than Hindi, so we might expect weaker accuracy and, more importantly, less trustworthy confidence.

There is a second concern. Benchmarks usually run each question once, in one fixed form. If a harmless change such as reordering the answer options changes the model's answer, then a single run tells us less than it appears to. We therefore ask three questions:

RQ1. How well calibrated are small open LLMs on native Hindi and Marathi exam questions?

RQ2. Do meaning-preserving changes to the prompt, specifically reordering the options, change the answer and the confidence for individual questions?

RQ3. Can a simple post-hoc correction repair the confidence, and does the repair still hold when the options are reordered?

We answer these with a deliberately small and fully reproducible setup that runs on a free cloud GPU and uses only public data and open models. Our contribution is an empirical one: a measurement of confidence quality in two Indian languages, evidence that per-question answers are less stable than aggregate scores suggest, and a check that temperature scaling transfers across option orderings. We do not propose a new model, and we make no claim about language models in general.

How to Cite this Article:

Deshmukh, S. S. (2026). How Reliable Is LLM Confidence in Hindi and Marathi? A Study of Calibration and Answer Stability. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(10), 1-9. <https://doi.org/10.55041/ijcope.v2i10.135>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i10.135>



2. Related Work

Calibration. Guo et al. [1] showed that modern neural networks tend to be overconfident and that a single temperature parameter fitted on held-out data often repairs this, which is the method we use. Expected calibration error [2] is the standard summary of the gap between confidence and accuracy. Work on language models has found that, in the right format, larger models can estimate whether their own answers are correct reasonably well [3], although this depends on the model, the task and the way confidence is extracted.

Selective prediction. If confidence is informative, a system can abstain on its least certain inputs and improve accuracy on the rest [4]. Recent work on abstention for black-box language models elicits verbal confidence and combines it with answer-or-abstain payoffs so that the model declines when it is likely to be wrong [5]. We instead read confidence from token probabilities and look at how stable it is.

Sensitivity to prompt details. Several studies show that multiple-choice results depend on the order of the options [6, 7]. Brittlebench [8] reports that meaning-preserving perturbations can lower benchmark accuracy by as much as 12%, and that even a single perturbation changed the relative ranking of models in 63% of cases. A large comparison of prompt-robustness methods [9] likewise finds that models are highly sensitive to small formatting changes, and that ranking the probabilities of the answer options is more robust than free generation, which is the approach we use. Most of this work measures accuracy. We look instead at what happens to the confidence and to the abstention decision, in languages other than English.

Indian-language evaluation. MILU [10] is a benchmark of roughly 80,000 multiple-choice questions in eleven Indian languages plus English, drawn largely from regional and state examinations. Its authors evaluated many models, mainly with a log-likelihood scoring method, and found that models do worse on lower-resource languages and on culturally specific subjects. Other groups have recently released Hindi versions of popular English benchmarks [11]. To the best of our knowledge, we did not find an earlier study of confidence calibration and option-order stability on Hindi and Marathi in the literature we reviewed, but our search was not exhaustive and we do not claim novelty beyond that.

3. Data and Models

Dataset. We used the Hindi and Marathi test splits of MILU [10] (CC-BY-4.0), downloaded from Hugging Face. Each question has four options and one correct answer. The data include a flag marking machine-translated items, and we kept only questions that were originally written in the target language. From these we drew a random sample of 300 questions per language with a fixed seed (42). The Hindi and Marathi samples are different questions, not translations of each other.

Models. We evaluated two instruction-tuned models from the Qwen2.5 family [12], with 1.5 and 3 billion parameters, both run in 16-bit precision on a free Google Colab T4 GPU. We chose one family in two sizes so that the answer-letter tokens are handled identically and any difference reflects model size rather than tokenisation. No model was trained or fine-tuned.

4. Method

Prompt and confidence. Each question was presented with its options labelled A to D and a short instruction asking the model to reply with only the letter. We took the model's next-token scores for the four letters, applied a softmax over just those four, and used the largest probability as the confidence and the corresponding letter as the answer. This needs one forward pass per question and gives a probability that is directly comparable across questions.

Perturbation. For every question we created a second version in which the four options were randomly permuted (one fixed permutation per question, seed 0), with the correct label updated accordingly. We call an answer stable if the model selects the same option text in both versions.

Metrics. We report accuracy, mean confidence, ECE with ten equal-width bins, and AUROC, which measures whether correct answers receive higher confidence than incorrect ones (0.5 means no information). Uncertainty is shown with 95% bootstrap intervals (1,000 resamples). For the effect of the shuffle we used a paired bootstrap on the change in each metric. We also looked at selective prediction by ranking questions by confidence and plotting accuracy against the fraction of questions answered.

Recalibration. We applied temperature scaling [1]: the four letter probabilities are divided by a temperature T before the softmax, with T chosen to minimise the negative log-likelihood on a calibration set. To avoid testing on the data used for fitting, we repeated 30 random 50/50 splits of the 300 questions (150 for fitting T on the original ordering, 150 for testing) and report the average held-out ECE. The same T was then applied to the shuffled version of the test questions, which tests whether the correction depends on one particular option layout. Because the splits reuse the same 300 questions, their spread should not be read as an independent confidence interval. Recalibration was run for the 3B model only.



5. Results

5.1 Accuracy and confidence

Table 1 shows the baseline results. Random guessing gives 25% accuracy. The 3B model is clearly above chance in both languages, while the 1.5B model is only a little above it (33.0% and 31.7%). In every case the average confidence is far higher than accuracy. The 3B model is the most extreme: it reports about 91–93% confidence while being right 33–45% of the time. Its confidence exceeded 0.9 on 80% of Hindi and 74% of Marathi questions, yet those answers were correct only 46.7% and 34.2% of the time. The reliability diagrams in Figure 1 show the same thing: the points sit well below the diagonal, and in Marathi accuracy stays around 0.3 even in the most confident bins.

Table 1. Baseline results on 300 native questions per language (original option order).

Model	Language	Accuracy	Mean conf.	ECE	AUROC	Same answer after shuffle
Qwen2.5-1.5B	Hindi	0.330	0.730	0.400	0.599	0.483
Qwen2.5-1.5B	Marathi	0.317	0.711	0.394	0.603	0.433
Qwen2.5-3B	Hindi	0.450	0.932	0.483	0.631	0.620
Qwen2.5-3B	Marathi	0.327	0.914	0.592	0.616	0.597

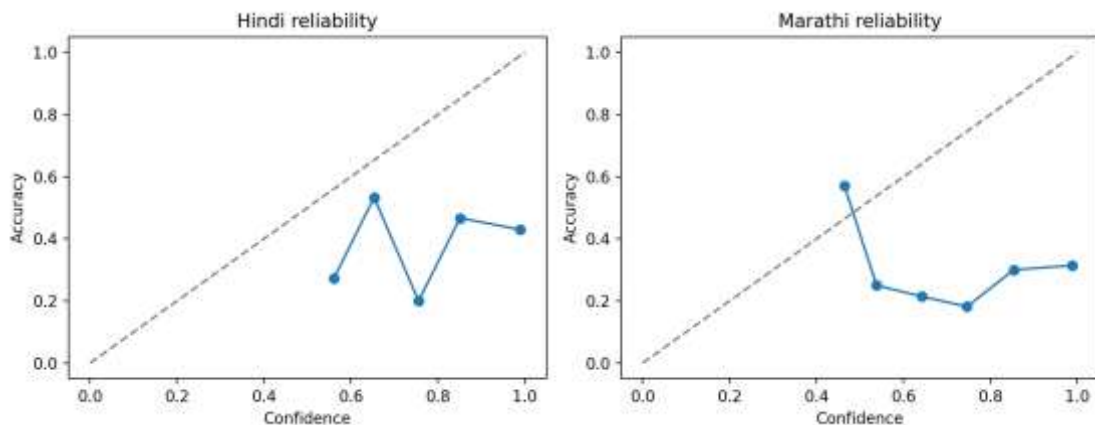


Figure 1. Reliability diagrams for the 3B model (bins with at least five questions). The dashed line is perfect calibration.

The larger model is more accurate and also more overconfident. Its ECE is higher than that of the 1.5B model, but this should not be read as the small model being better calibrated: it simply claims less certainty, and its 0.73 confidence against 0.33 accuracy is still a large gap. AUROC is close to 0.6 in all four settings, so neither model is good at telling which of its answers are likely to be correct. With only 300 questions, individual points in Figure 1 are noisy, and the overall pattern is more trustworthy than any single bin.

5.2 Hindi versus Marathi

For the 3B model, the 95% bootstrap interval for accuracy was 0.397–0.507 in Hindi and 0.270–0.377 in Marathi, and for ECE it was 0.432–0.538 and 0.544–0.648. The difference (Marathi minus Hindi) had an interval of -0.200 to -0.050 for accuracy and 0.034 to 0.185 for ECE, so Marathi was both less accurate and worse calibrated. The ECE intervals for the two languages nearly touch, so we treat that difference as suggestive rather than firm. The 1.5B model shows no language gap, but it is close to the chance floor in both languages, which would hide any difference. Since the two samples contain different questions, part of the gap may reflect question difficulty and not only the language.

5.3 Stability under option reordering

For the 3B model, reordering the options barely changed the aggregate numbers. In Hindi, accuracy moved from 0.450 to 0.447, ECE from 0.483 to 0.489 and AUROC from 0.631 to 0.657. In Marathi, accuracy moved from 0.327 to 0.357, ECE from 0.592 to 0.563 and AUROC from 0.616 to 0.572. The paired bootstrap intervals for the change in ECE and in AUROC all included zero (Hindi ECE -0.052 to 0.065, AUROC -0.045 to 0.094; Marathi ECE -0.082 to 0.020, AUROC -0.123 to 0.039), so we cannot claim that shuffling changed calibration on average.

Individual answers told a different story. The 3B model gave the same answer after the shuffle for 62.0% of Hindi and 59.7% of Marathi questions, meaning that roughly four in ten answers changed. For the 1.5B model the rates were 48.3% and 43.3%. Being very sure did not guarantee stability: among 3B questions where the original confidence exceeded 0.9, the answer survived the shuffle for only 67.5% (Hindi) and 62.6% (Marathi) of them. The 3B model was correct under both orderings for only 32.3% of Hindi and 24.7% of Marathi questions, compared with single-run accuracy of 45.0% and 32.7%. Mean absolute



change in confidence per question was 0.09 in Hindi and 0.10 in Marathi for the 3B model, and 0.18 and 0.16 for the 1.5B model.

For the 3B model this is not simple letter bias: the share of answers given to each letter stayed between 20% and 32%, close to the uniform 25%. The 1.5B model is different. It put 75–81% of its answers on A or B, and 41–48% on B alone, so letter preference is part of its instability.

5.4 Selective prediction

Confidence is a usable, if weak, abstention signal (Figure 2). When the 3B model answers only the questions it is most confident about, accuracy rises above the answer-everything baseline in both languages. At 20% coverage, accuracy was 0.65 in Hindi (baseline 0.45) and 0.57 in Marathi (baseline 0.33), and at 10% coverage it was 0.73 and 0.70. The very top of these curves rests on a handful of questions and is noisy. We also tried using agreement across the two orderings as an extra signal. Alone, it was weak (AUROC 0.59 in Hindi and 0.62 in Marathi). Combined with confidence it gave AUROC 0.653 and 0.645 against 0.631 and 0.616 for confidence alone. With 300 questions we cannot say this small gain is real, and the effect on the risk-coverage curves was mixed: the combined signal looked better in Hindi at 20% coverage (0.72 against 0.65) but worse in Marathi (0.50 against 0.57).

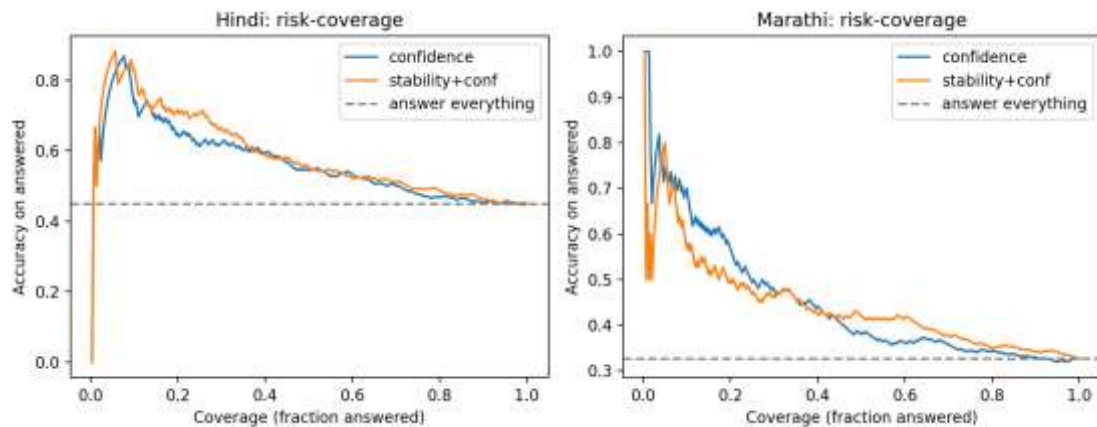


Figure 2. Risk-coverage curves for the 3B model: accuracy on the questions answered when only the most trusted fraction is answered. The dashed line is the answer-everything baseline.

5.5 Temperature scaling

Fitted temperatures were large (mean T of 12.4 for Hindi and 18.0 for Marathi), reflecting how sharp the raw probabilities were. After scaling, held-out ECE dropped from about 0.5 to below 0.1 in both languages (Table 2). The temperature was fitted on the original ordering only, yet it reduced ECE equally well on the shuffled version, so the correction does not depend on a particular option layout. Scaling changes the level of confidence but not the ranking of questions, so AUROC is unaffected; the model is not better at spotting its mistakes, it just no longer claims certainty it does not have. With 150 test questions per split, differences between values such as 0.060 and 0.095 are within noise, and we do not compare the two languages after scaling.

Table 2. Held-out ECE of the 3B model before and after temperature scaling (mean of 30 random splits).

Language	Mean T	Option order	ECE before	ECE after
Hindi	12.37	Original	0.482	0.080
Hindi	12.37	Shuffled	0.493	0.083
Marathi	18.02	Original	0.601	0.095
Marathi	18.02	Shuffled	0.573	0.060

6. Interactive Demonstration

To make the main finding easy to see, we built a small web demonstration with Gradio that runs on the same free GPU session. The user selects a language and a question, and the app shows the model's probabilities for the original option order and for a shuffled order side by side, with a statement of whether the answer changed. A temperature slider applies the recalibration. It is a teaching and inspection tool for the experiments in this paper, not a production system. Figure 3 shows the interface.



Figure 3. The demonstration app on start-up. The controls on the left set the language, question number, shuffle seed and temperature; the question and the original and shuffled option orders appear on the right.

Figures 4 and 5 show a Marathi question asking which material the segments of a commutator are made of. In the original order the model puts 100.0% of its probability on iron segments, which is wrong. After the options are shuffled it puts 99.9% on copper segments, which is correct. The model is essentially certain of two different answers to the same question. With the recalibration temperature set to 18, the same two answers appear with 37.1% and 36.6% confidence, a far more honest level. Figures 6 and 7 show the same behaviour on a Hindi question: 99.8% on the correct answer in the original order, 100.0% on a wrong answer after the shuffle, and 42.4% and 44.6% after recalibration with $T = 12.4$.

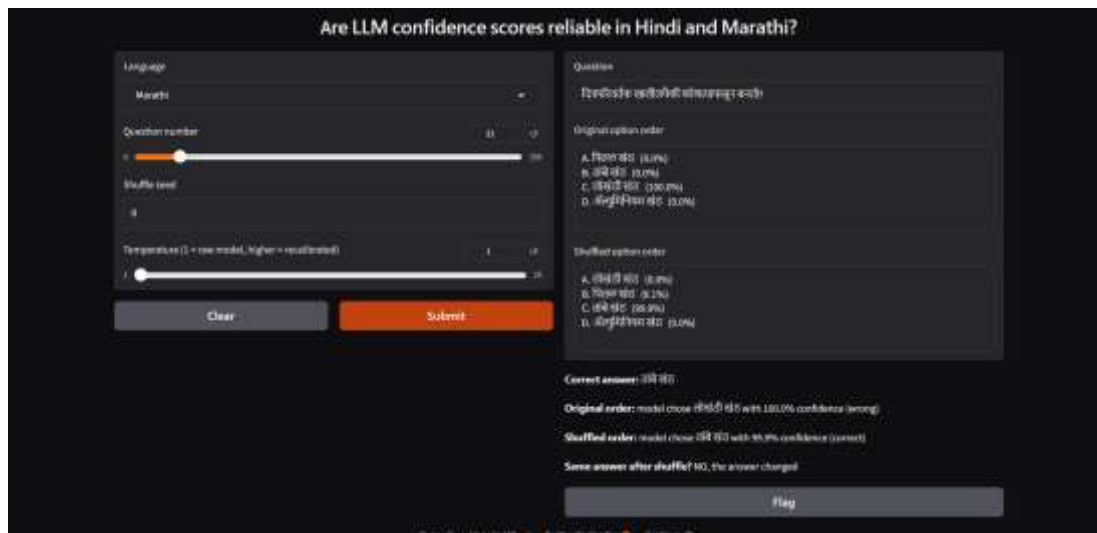


Figure 4. Demonstration screen for a Marathi question (question 31 of the Marathi sample, shuffle seed 0) with the raw model (temperature 1). The model gives 100.0% to a wrong answer in the original order and 99.9% to the correct answer after the shuffle.

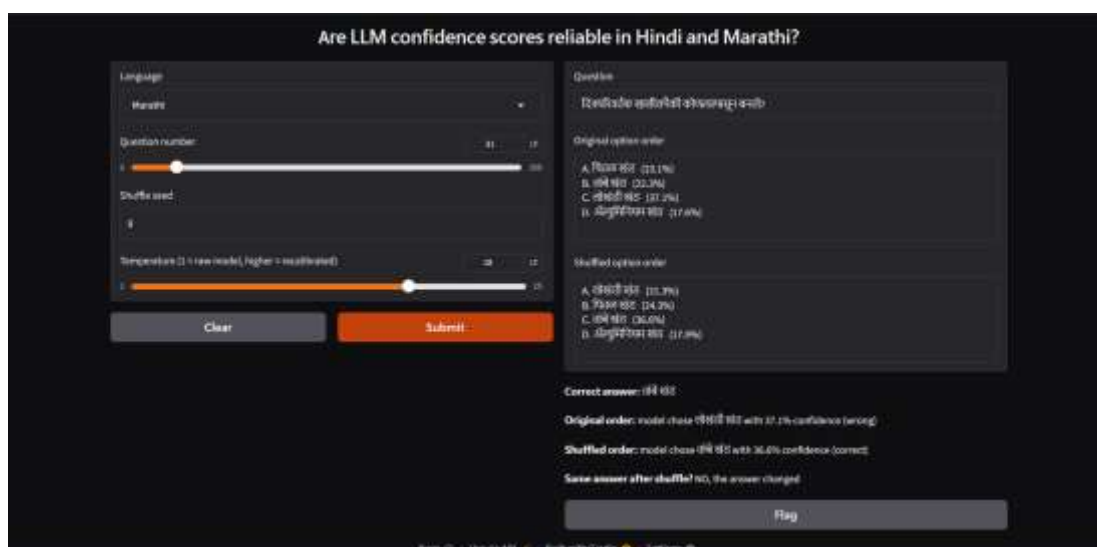


Figure 5. The same Marathi question after recalibration (temperature 18). The two answers now receive 37.1% and 36.6% confidence.

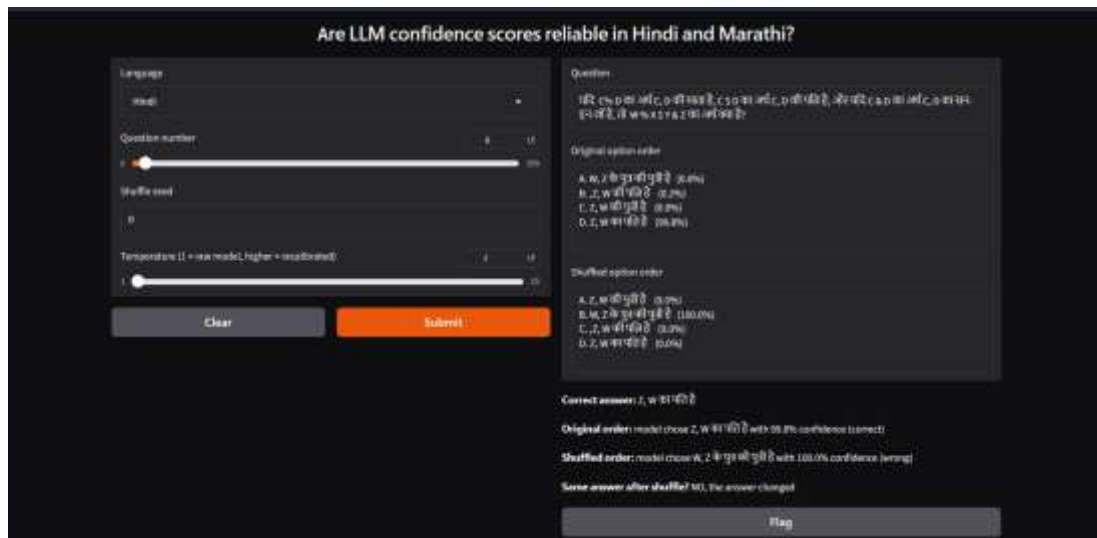


Figure 6. Demonstration screen for a Hindi question (question 6 of the Hindi sample, shuffle seed 0) with the raw model (temperature 1). The model gives 99.8% to the correct answer in the original order and 100.0% to a wrong answer after the shuffle.

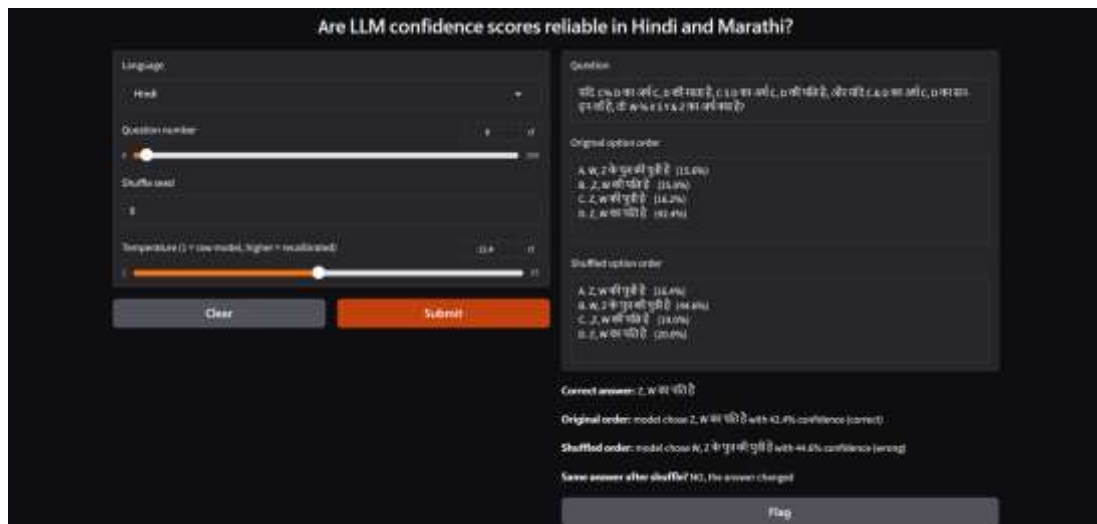


Figure 7. The same Hindi question after recalibration (temperature 12.4). The two answers now receive 42.4% and 44.6% confidence.

7. Discussion

Three points stand out. First, a model can look steady on aggregate while being unsteady question by question. Accuracy and ECE barely moved when we reordered the options, yet four in ten answers changed. Anyone who tunes an abstention threshold or reports a benchmark score from a single prompt layout is relying on an average that hides this. Reporting how often answers survive harmless changes, alongside accuracy, is cheap and informative.

Second, raw confidence from these models should not be shown to users as a probability. Even with the larger model, answers given with more than 90% confidence were correct only 46.7% of the time in Hindi and 34.2% in Marathi. Temperature scaling is an easy fix for the level, and it held up when the options were reordered. But because it does not improve ranking, it cannot make a weak model a good judge of its own answers.

Third, thresholds and temperatures should be set per language. The fitted temperature for Marathi was noticeably larger than for Hindi, so a setting tuned on one language would misjudge the other.

8. Limitations

This study is limited in scope and its conclusions are narrow. We tested two sizes of one model family, one prompt format and one kind of perturbation. We used 300 questions per language, which gives wide intervals, and the Hindi and Marathi samples are different questions, so the language comparison is confounded with question difficulty. Confidence came only from answer-letter probabilities; we did not evaluate verbalised confidence or agreement across sampled answers. The 1.5B model is close to chance and shows a strong letter preference, which limits what it can tell us. Recalibration was run for the 3B model only. Two model sizes cannot establish a scaling trend. We did not study larger models, other families or other Indian languages, and we did not run a full literature search, so we cannot rule out closely related prior work. All results are specific to the setup described here.



9. Conclusion

On Hindi and Marathi exam questions, two small open language models reported far more confidence than their accuracy justified, and the larger one was both more accurate and more overconfident. Marathi was harder than Hindi for the 3B model, in both accuracy and calibration. Reordering the answer options left the averages almost unchanged but changed a large share of individual answers, including answers given with high confidence. A single fitted temperature repaired the calibration level and transferred to reordered options, but did not improve the model's ability to rank its own answers. We hope the numbers and the demonstration are useful as a starting point. The natural next steps are more models and languages, larger samples with translated question pairs, other confidence signals and other perturbations.

Acknowledgements

The first author ran all experiments on a free cloud GPU using the public MILU dataset. An AI assistant (Claude, Anthropic) was used for code, analysis and drafting support. The authors reviewed all content and are responsible for it.

References

- [1] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th International Conference on Machine Learning (ICML), 2017.
- [2] M. P. Naeini, G. F. Cooper and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in Proc. AAAI Conference on Artificial Intelligence, 2015.
- [3] S. Kadavath et al., "Language models (mostly) know what they know," arXiv preprint arXiv:2207.05221, 2022.
- [4] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [5] H. Zong, B. Li, Y. Long, S. Chang, J. Wu and G. K. Hadfield, "I-CALM: Incentivizing confidence-aware abstention for LLM selective answering," arXiv preprint arXiv:2604.03904, 2026.
- [6] P. Pezeshkpour and E. Hruschka, "Large language models sensitivity to the order of options in multiple-choice questions," arXiv preprint arXiv:2308.11483, 2023.
- [7] C. Zheng, H. Zhou, F. Meng, J. Zhou and M. Huang, "Large language models are not robust multiple choice selectors," in Proc. International Conference on Learning Representations (ICLR), 2024.
- [8] A. Romanou, M. Ibrahim, C. Ross, C. Shaib, K. Oktar, S. J. Bell, A. Ovalle, J. Dodge, A. Bosselut, K. Sinha and A. Williams, "Brittlebench: Quantifying LLM robustness via prompt sensitivity," arXiv preprint arXiv:2603.13285, 2026.
- [9] M. Seleznyov, M. Chaichuk, G. Ershov, A. Panchenko, E. Tutubalina and O. Somov, "When punctuation matters: A large-scale comparison of prompt robustness methods for LLMs," arXiv preprint arXiv:2508.11383, 2025.
- [10] S. Verma, M. S. U. R. Khan, V. Kumar, R. Murthy and J. Sen, "MILU: A multi-task Indic language understanding benchmark," in Proc. NAACL (Long Papers), 2025.
- [11] A. Kamath, K. Singla, R. Paul, R. Joshi, U. Vaidya, S. S. Chauhan and N. Wartikar, "Benchmarking Hindi LLMs: A new suite of datasets and a comparative analysis," in Proc. 1st Workshop on Benchmarks, Harmonization, Annotation, and Standardization for Human-Centric AI in Indian Languages (BHASHA), 2025.
- [12] Qwen Team, "Qwen2.5 technical report," arXiv preprint arXiv:2412.15115, 2024.