Estimating Uncertainty from Reasoning: A Large-Scale Study of Multi- and Crosslingual MCQA Performance in LLMs

Andrea Alfarano[*] [†] **Andrea Bacciu**[*1] **Saab Mansour**[1] **Amin Mantrach**[1] **Marcello Federico**[1]
[1]Amazon
{andbac, saabm, mantrach, marcfede}@amazon.com

Abstract

Uncertainty estimation (UE) enables LLM-powered systems to recognize when to abstain, yet existing research has predominantly focused on English. We present the first large-scale evaluation of UE methods across 22 languages, spanning high-, mid-, and low-resource settings. Using two human-curated Q&A datasets, we compare open and closed box UE methods (nine in total) across different model sizes and architectures while eliciting long-form reasoning, avoiding LLM-as-a-judge and embedding-based scoring, which can introduce evaluation noise. We report three main actionable findings. First, we find that prompting models to reason in English while keeping questions in low-resource languages substantially improves UE performance, suggesting that comprehension of low-resource languages is largely intact, and that the reliability bottleneck lies in generation rather than understanding. Second, prompting models to reason in English closes the UE performance gap between low and high-resource languages, demonstrating that generation language matters more than the question language. Third, the choice of UE method should depend on model scale: at smaller scales, open-box probability-based methods outperform alternatives; at larger scales, closed-box self-verbalized uncertainty becomes superior. Finally, we provide an analysis of threshold selection for selective prediction, offering guidance on calibrating abstention in multilingual settings.

1 Introduction

Large Language Models (LLMs) have changed how people access and interact with information, supporting tasks from everyday planning to complex question-answering (Bommasani et al., 2022). Recent research has primarily focused on improving task performance, for example through chain-of-thought prompting (Wei et al., 2022) and instruction tuning (Ouyang et al., 2022), enabling models to solve increasingly complex problems. However, high accuracy alone is insufficient: downstream systems must recognize when an LLM's answer is not grounded in its knowledge, enabling them to abstain, defer to humans, or fall back to safer behavior. This has motivated a parallel line of work on uncertainty estimation (UE), which seeks to determine when models know the answer, enabling LLM-powered systems to recognize and communicate their lack of knowledge (Kuhn et al., 2023). Existing research on LLM uncertainty has predominantly focused on English, leaving limited evidence on whether UE methods maintain their efficacy in other languages, especially in low-resource settings (Kuhn et al. (2023); Kossen et al. (2024); Santilli et al. (2025); Cecere et al. (2025), inter alia). The only dedicated multilingual UE study we are aware of (Xue et al., 2025) evaluates just three methods on five languages, relies on machine-translated data without human post-editing, and uses a multiple-choice, short-answer format rather than open-ended generation. Their dataset yields a median answer length of just one word, meaning this setup primarily evaluates how language affects question comprehension while offering limited evidence about uncertainty during longer, more linguistically rich generation. Establishing multilingual trustworthiness through UE is also methodologically challenging because standard metrics such as AUROC rely on ground truth labels. When ground truth is approximated via LLM-as-judge, BERTScore, or n-gram overlap, these proxies introduce noise that can misrank uncertainty methods (Santilli et al., 2025). In multilingual settings, this problem is compounded: neural judges may behave inconsistently across languages, introducing language-dependent bias into UE comparisons. As a result,

[*]Equal contribution; authors listed in alphabetical order.
[†]Work done during an internship at Amazon.

existing English-centric evaluations do not yet provide decisive evidence that UE methods transfer reliably to low-resource languages and long-form generation. To address these challenges, we evaluate nine UE methods on two human-curated QA datasets across 22 languages, covering high-, mid-, and low-resource settings and spanning diverse language families and resource levels. We elicit long-form, language-rich reasoning (on average 150 words) while preserving objective correctness from QA labels, thereby avoiding LLM-as-judge and embedding-based proxies.

Our results reveal three key patterns: (1) UE method performance varies largely across language families, consistent with effects of resource availability and linguistic distance; (2) generation language has a larger impact on UE than question language, a factor largely overlooked by prior work; and (3) while the closed-box Self Verbalized method achieves the best overall performance, sampling-based closed-box methods degrade substantially on low-resource languages, whereas open-box probability-based methods maintain more consistent performance across all language resource levels.

We make the following contributions: (i) we propose a label-grounded framework for evaluating UE on open-ended multilingual generation without model-based correctness metrics; (ii) a large-scale study of UE across 9 models and 22 languages; and (iii) practical guidance for deploying uncertainty-aware systems in multilingual settings. Specifically, we address: (RQ1) which UE methods are robust across models and languages, (RQ2) how model scale affects UE efficacy, (RQ3) how the language of reasoning shapes UE quality, (RQ4) UE robustness under cross-lingual answer settings, and (RQ5) threshold selection strategies for selective prediction.

2 Related Work and Background

Uncertainty estimation (UE) is critical for deploying large language models (LLMs) in high-stakes environments, enabling systems to flag unreliable predictions. Existing literature is commonly categorized by the degree of model access available: *open-box methods*, which utilize internal model representations, and *closed-box methods*, which rely solely on textual outputs.

2.1 Open-Box Methods

Open-box approaches derive uncertainty directly from model internals, such as hidden states or token probabilities.

Inference method approaches require no additional training and aggregate token-level logits into sequence-level scores. Standard metrics include length-normalized log-probabilities, the geometric mean of token probabilities, and minimum token probability (Fomicheva et al., 2020; Vazhentsev et al., 2023). Inference-only methods remain the standard for open-weights models due to their computational efficiency and lack of supervision requirements.

2.2 Closed-Box Methods

Closed-box techniques are essential when access to logits or gradients is restricted (e.g., commercial APIs). These methods rely exclusively on the generated text.

Self verbalized uncertainty This approach prompts the LLM to explicitly state its confidence as a score or linguistic expression (Kadavath et al., 2022; Tian et al., 2023). While accessible, self-verbalization is frequently miscalibrated and often struggles to discriminate correctness effectively compared to logit-based baselines (Xiong et al., 2024; Groot and Valdenegro Toro, 2024).

Consistency and Sampling Strategies Consistency-based methods rely on the intuition that correct reasoning is stable across stochastic samples. *Semantic Entropy* and its variants (Kuhn et al., 2023; Farquhar et al., 2024; Cecere et al., 2025) cluster multiple generations by meaning and compute entropy over these clusters. Other works extend this method using graph theory, modeling generations as nodes in a similarity graph to derive uncertainty from structural properties (Lin et al., 2024; Vashurin et al., 2025).

2.3 Evaluation and Multilingual Challenges

UE performance is typically evaluated via *discrimination* AUROC but reliable evaluation remains a challenge.

Ground Truth Matching Santilli et al. (2025) demonstrate that automated ground truth matching like ROUGE or BERTScore correlate strongly with superficial features (e.g., length) rather than factual accuracy, inflating reported UE performance.

To address this, we avoid automated ground truth matching and rely on exact matching of multiple-choice answers against human-curated labels. Prior work has also used MCQA exact matching to evaluate model reliability. For example, Madhusudhan et al. (2025) study whether models select an explicit "I Don't Know" option as a binary abstention behavior. Our framework differs in three ways: we evaluate continuous-valued uncertainty signals (rather than binary abstention), extract them from long-form reasoning traces (rather than the answer choice itself), and operate across 22 languages (rather than English only). We estimate uncertainty from reasoning generated *before* the answer, following the inference paradigm of modern reasoning LLMs where the reasoning trace is part of the computation that produces the prediction. We do not consider explanations elicited after the answer, as such post-hoc rationalizations are known to be unfaithful to the underlying computation (Atanasova et al., 2023).

Multilingual Uncertainty Estimation While UE is well-studied in English, multilingual evaluation remains underexplored. To the best of our knowledge, Xue et al. (2025) present the only dedicated multilingual UE study, introducing *MlingConf* to benchmark three uncertainty methods across five languages. However, their evaluation relies on machine-translated data without human post-editing and employs a short-answer format with a median response length of just one word. This setup primarily assesses how language affects question comprehension, offering limited insight into uncertainty during extended generation. Our benchmark addresses this gap by eliciting long-form reasoning in the target language while preserving objective correctness labels.

3 Methodology

Our goal is to evaluate LLM UE methods across multiple languages, including low-resource ones, while supporting long text generation and using model-free metrics to maintain an unbiased setup.

To achieve this, we leverage human-curated Multiple-Choice Question Answer (MCQA) datasets where the set of possible labels is a fixed set of options (such as A, B, C, and D). MCQA labels can be interpreted without relying on string-matching, LLM-as-a-judge, or embedding similarity techniques, thus providing ground truth without any approximation or language bias. MCQA labels are too short to derive uncertainty estimates from the model; for this reason, we apply UE methods to the LLM's reasoning text, similarly to the approach of Podolak and Verma (2025).

Specifically, we prompt the model to generate a reasoning explanation before producing its answer (e.g., Chain of Thoughts (Wei et al., 2022)). We then apply uncertainty estimation methods to this reasoning text (on average 150 words), while using exact matching against MCQA labels to determine answer correctness. This allows us to retain the reliability of MCQA labels as ground truth while extending UE evaluation to long text generation. We deliberately choose MCQA over open-ended QA datasets for two reasons: (i) open-ended answers are extremely short factoids (median 1–2 words), so they do not provide longer text for UE than MCQA labels do, and (ii) their evaluation does not ensure a language- and correctness-unbiased setup. A detailed justification is provided in Appendix B. Finally, we evaluate the performance of each UE method using the Area Under the Receiver Operating Characteristic (AUROC) curve, which is the standard metric in the UE literature for measuring how well uncertainty scores correlate with answer correctness. Statistical significance is assessed with DeLong tests on the ROC curves (paired when methods are compared on identical samples, unpaired across model scales), applying Benjamini–Hochberg FDR correction across languages; per-language results are reported in Appendix L. The complete process is illustrated in Figure 1.

Datasets: We use two human-curated MCQA datasets, namely Global-MMLU (Singh et al., 2025) (with 4 possible choices) and MMLU-ProX (Xuan et al., 2025) (with 10 possible choices). We focus on the 22 languages that are common to both datasets (the complete list of languages is available in Appendix C). These two datasets are parallel, meaning that every question-answer pair is available in every language, making our multilingual evaluation comparable across languages. These two datasets cover a broad set of categories: Global-MMLU contains 6 categories, while MMLU-ProX contains 14 categories (see Appendix E). To ensure balanced representation, we sample 100 questions per category from each dataset, yielding 600 questions from Global-MMLU and 1,400 from MMLU-ProX, for a total of 2,000 unique questions. Since both datasets are available in all 22 languages, this results in 44,000 language-specific instances. We

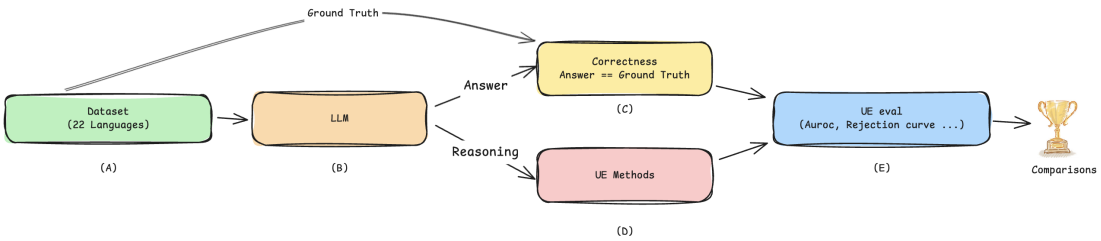


Figure 1: Evaluation pipeline. Unlike prior work that estimates uncertainty from predicted labels in English-only settings, our framework (1) extracts uncertainty from the model's *reasoning trace* and (2) evaluates across 22 languages. The pipeline proceeds as: multilingual prompts (A) → LLM generation (B) → correctness evaluation against ground truth (C) → uncertainty estimation on reasoning (D) → method comparison via AUROC (E).

further verify that the MCQA labels are uniformly distributed in both datasets, ruling out positional bias (see Appendix J).

Models: Our evaluation spans 9 models from both closed-source and open-source families. For closed-source models, we employ Claude Sonnet 4.5 (Anthropic, 2025) - the state-of-the-art model at the time of writing. For open-source models, we consider instruction-tuned models including five Gemma3 models (0.27B, 1B, 4B, 12B, 27B) (Team et al., 2025) and three Qwen3 models (4B, 30B-A3B, 235B-A22B) (Yang et al., 2025), encompassing both dense and Mixture of Experts (MoE) architectures. We report a scaling study on the Q&A accuracy in Figure 2. In our experimentation, we applied the quantization only to Qwen3 235B-A22B to 4-bit due to hardware constraints (see our Hardware Infrastructure in Appendix H). In Appendix G, we investigate whether quantization impacts multilingual UE performance. We selected Qwen3-30B for this analysis as it is the largest model that fits in memory at full precision. Results show no statistically significant effect when quantizing to 8-bit and 4-bit.

Uncertainty Estimation Methods: Following the taxonomy in Section 2, we evaluate nine methods, including both open-box methods (with access to model internal states) and closed-box methods. Open-box probability baselines include Mean Token Entropy (Token Entropy) (Fomicheva et al., 2020), Maximum Sequence Probability (Max Prob) (Vashurin et al., 2025), and Self Certainty (Kang et al., 2025). Closed-box methods include Self-verbalized uncertainty (Tian et al., 2023), multi-sample disagreement based on lexical overlap as Lexical Similarity (ROUGE-L) (Fomicheva et al., 2020; Kuhn et al., 2023), Semantic Entropy with Jaccard-based clustering (Kuhn et al., 2023), and graph-based dispersion (EigValLaplacian_Jaccard, DegMat_Jaccard, Eccentricity_NLI_Jaccard) (Lin et al., 2024). More detailed method definitions are available in Appendix A. All prompts used are listed in Appendix D. All UE methods are implemented through LM-Polygraph (Fadeeva et al., 2023; Vashurin et al., 2025).

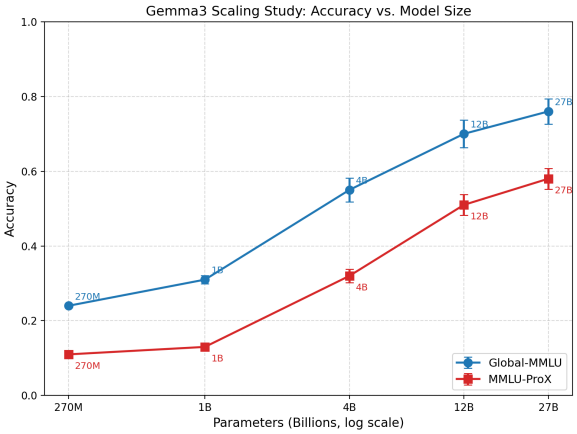


Figure 2: Parameter scaling study on MCQA accuracy on both datasets (mean ± 95% CI) over 22 languages using Gemma 3. Similar trends are observed in Qwen models.

4 Results And Discussion

RQ1: Which UE methods are robust across models and languages? To address this question, Figure 3 compares nine UE methods across 22 languages and two open-source models (Gemma3-27B and Qwen3-235B). We excluded closed-source models to ensure a fair comparison between open-box and closed-box approaches.

The best-performing method is "Self Verbalized", which achieves an average AUROC of 0.72 and demonstrates consistent performance across all languages and models. The second-ranked method is the open-box "Token Entropy" with an AUROC of 0.71, followed by "Self Certainty" at 0.70. The

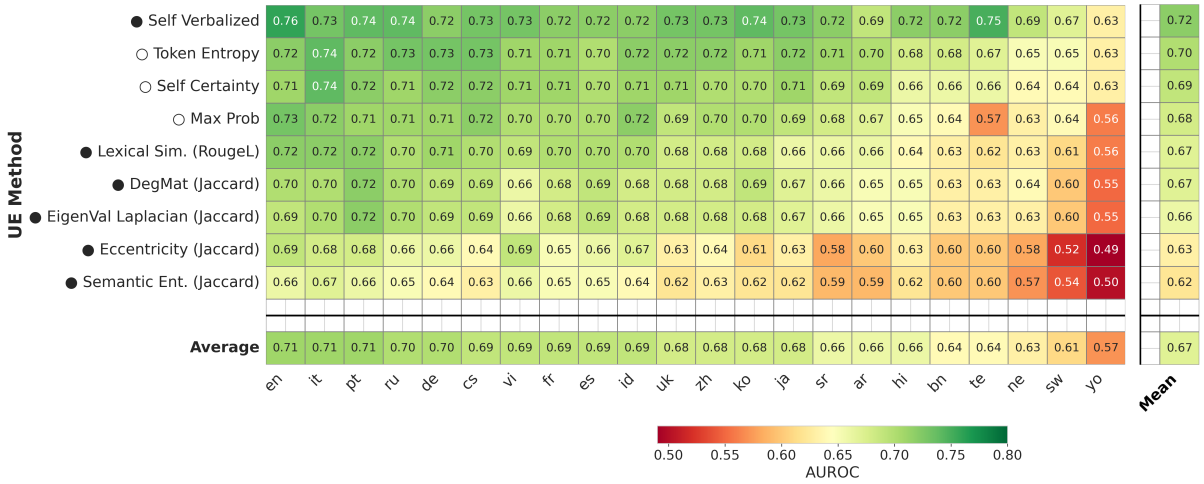


Figure 3: Per-language AUROC for nine UE methods across 22 languages on the dense Gemma3-27B and MoE Qwen3-235B-A22B models. Languages are sorted by mean AUROC (median sorting produces identical rankings). Closed-box UE methods are marked with a black circle while open-box are marked with a white circle. In the left block we average across models, in the middle block across languages, while in the right block we average across languages and models. Remarkably, the closed-box method *Self Verbalized* achieves top performance, rivaling open-box methods such as *Token Entropy* and *Self Certainty*—suggesting that effective UE is achievable without access to token-level probabilities.

open-box "Max Prob" performs reasonably well on average (0.68) but degrades substantially on low-resource languages, dropping to 0.56 on Yoruba and 0.57 on Telugu.

Sampling-based methods maintain reasonable performance on high-resource languages but fail on low-resource ones, with Eccentricity and Semantic Entropy dropping to near-random performance on Yoruba (0.49 and 0.50 AUROC). These methods generate multiple reasoning traces and use their diversity as an uncertainty signal—the intuition being that confident models produce consistent outputs while uncertain models vary. To understand why this fails, we analyze the diversity of reasoning traces across all 22 languages using pairwise Jaccard distance. For high-resource languages, the expected pattern holds: incorrect predictions produce more diverse reasoning than correct ones (English: 0.72 vs. 0.63, $p < 10^{-6}$; German: 0.73 vs. 0.65; Spanish: 0.74 vs. 0.63; French: 0.69 vs. 0.64), with gaps of 0.08–0.11. For low-resource languages, this gap collapses to $\leq$0.03: Yoruba (0.93 vs. 0.93), Swahili (0.93 vs. 0.92), Bengali (0.83 vs. 0.80). LLMs generate highly variable reasoning regardless of correctness, leaving no diversity signal to exploit. Self-verbalized approaches, by contrast, leverage meta-cognitive capabilities (Steyvers and Peters, 2026; Xu et al., 2025) that transfer more robustly across languages. The Self Verbalized signal is also robust to the elicitation wording: across three prompt phrasings (Appendix D.3), error-detection AUROC remains stable (0.72–0.76) with no statistically significant difference, confirming the ranking is not an artifact of a single prompt formulation.

RQ2: How does model scale affect multilingual UE method efficacy? We investigate the effect of model parameters scale on UE performance, we evaluate the top-2 open-box (Token Entropy, Self Certainty) and top-2 closed-box (Self Verbalized, Lexical Sim.) methods (selected for figure clarity) that scale best across all languages (see RQ1); for all other methods, see Appendix I. In Figure 4, we computed the average across all the languages and averaging across datasets. We select Qwen3 as it offers the broadest parameter range among the open-source models in our study from 4B to 235B. We observe a clear divergence in scaling behavior. Token Entropy changes only modestly across the evaluated scales, with overlapping confidence intervals. Self Verbalized, by contrast, improves substantially with scale (from 0.65$\rightarrow$0.66$\rightarrow$0.77 AUROC): the difference between 4B and 30B is not significant, while the gain at 235B is significant in 22 of 22 languages on Global-MMLU (unpaired DeLong tests, BH-FDR corrected, mean gap $+0.153$ AUROC; Appendix L). At 4B and 30B

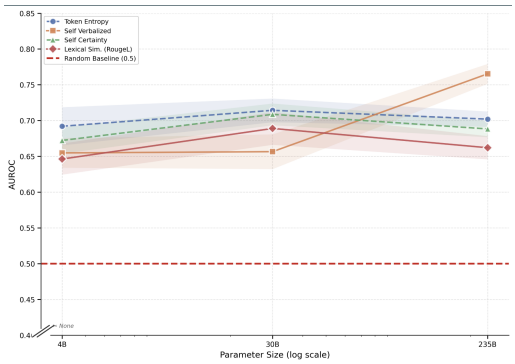


Figure 4: Effect of model scale on UE performance (Qwen3 family). Shaded regions denote 95% CIs. Dashed line: random baseline.

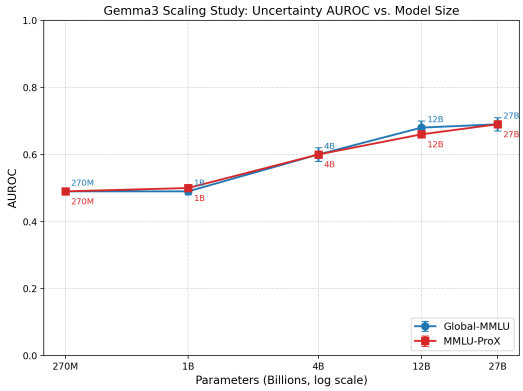


Figure 5: Uncertainty estimation performance (AUROC ± CI 95%) across varying model scales. Results are averaged across 22 languages in both datasets. Increasing model size generally yields higher AUROC scores.

scale, Token Entropy significantly outperforms Self Verbalized (19 of 22 languages at 30B, paired DeLong tests; Appendix L), but this relationship reverses at 235B where Self Verbalized achieves the highest AUROC among all methods. These findings have practical implications: at smaller scales, open-box methods such as Token Entropy provide the most reliable uncertainty signal; at larger scales, Self Verbalized surpasses all alternatives by a significant margin, suggesting that the quality of explicit self-assessment improves substantially at larger model scales, consistent with prior findings on self-knowledge in LLMs (Kadavath et al., 2022; Steyvers and Peters, 2026). However, training pipelines are not uniform across model sizes within the same family: the Qwen3 Technical Report (Yang et al., 2025) discloses that smaller Qwen3 variants are distilled from larger teacher models (Qwen3-235B-A22B), so part of the observed gap may also reflect that self-verbalization capabilities are sensitive to training procedures.

Furthermore, in Figure 5 we show results across Gemma from 270M to 27B parameters, averaged across all nine UE methods. A clear scaling trend emerges again: models at or below 1B parameters (Gemma3-0.27B, Gemma3-1B) perform at the random baseline (0.50 AUROC), indicating that UE methods fail to extract meaningful uncertainty signals from very small models. Performance increases substantially between 1B and 4B parameters, reaching ∼0.60–0.65 AUROC. Beyond 12B parameters, performance plateaus around 0.67–0.70 AUROC.

We report Expected Calibration Error in Appendix K, where larger models show a slight tendency toward overconfidence.

RQ3: How does the language of reasoning affect UE quality across languages? To investigate whether the language used for reasoning influences multilingual UE, we compare two configurations: *Multilingual*, where the model reasons in the target language (i.e., language of the question and answers), and *English Reasoning*, where reasoning is conducted in English regardless of the target language. Figure 6 reports AUROC with 95% confidence intervals across 22 languages, averaged across all datasets and the largest models (Qwen3-235B, Gemma3-27B, and Claude Sonnet 4.5), grouped by WALS language genus[1].

On average, English reasoning improves UE quality (0.72 vs. 0.68 AUROC). Crucially, this improvement is driven by low-resource languages: English reasoning effectively brings their UE performance to the level of high-resource languages. For Yoruba, AUROC improves from 0.58 to 0.68; for Swahili, from 0.58 to 0.64; for Nepali, from 0.66 to 0.71—all showing non-overlapping 95% confidence intervals. After applying English reasoning, these languages achieve AUROC scores comparable to Germanic and Romance languages (0.71–0.73), which show no significant change between conditions. These results reveal an asymmetry between natural language understanding and generation. In both conditions, the question remains in the target language—only the reasoning language changes. If comprehension were the bottleneck, changing the reasoning language would not help: the model would not know what to reason about. The fact that English reasoning closes the gap suggests that models comprehend low-resource

[1] https://wals.info/

questions adequately but struggle to generate coherent reasoning in those languages. By shifting generation to English, models leverage their stronger generative capabilities, producing stronger uncertainty signals.

Two controls confirm this interpretation: First, some sampling-based estimators may themselves exhibit language sensitivity, as they rely on similarity (e.g., Jaccard index). To isolate the reasoning-language effect from this potential measurement bias, we repeat the analysis using only Self Verbalized uncertainty, which relies directly on the model's reported confidence and involves no similarity scoring: the improvement under English reasoning remains significant (AUROC 0.665 $\rightarrow$ 0.679, paired bootstrap $p < 0.001$).

Second, the effect persists on the subset of questions answered identically under both conditions (88.4% of aligned questions), where accuracy is constant by construction (AUROC 0.670 $\rightarrow$ 0.684, $p < 0.001$); the improvement in uncertainty quality is thus independent of the accuracy gains that English reasoning also delivers. Together, these controls isolate the effect from both scoring artifacts and correctness shifts, directly supporting the generation-bottleneck interpretation.

To control for benchmark contamination, we exploit the English-reasoning condition as a memorization test: a memorized answer would be recalled regardless of the reasoning language, since question and options stay in the target language. Instead, English reasoning significantly boosts low-resource accuracy across all three model families (paired McNemar tests, FDR-corrected; Appendix M).

RQ4: How do UE methods perform under cross-lingual answer settings? Multilingual applications often involve cross-lingual scenarios where different components of an input span multiple languages—a common situation in cross-lingual retrieval-augmented generation (RAG) systems where documents retrieved from multiple languages must be integrated. We evaluate whether UE methods remain robust when this cross-lingual complexity is introduced. We construct instances where the question and reasoning are in the target language, but each answer option is drawn from a different language. Leveraging the parallel nature of our multilingual datasets, we use aligned answer IDs across translations and sample each option from a different language variant. For instance, an English question asking "What day comes after Monday?" would present answer options such as "Mardi" (French, Tuesday), "Mercoledì" (Italian, Wednesday), "Donnerstag" (German, Thursday), and "Viernes" (Spanish, Friday). Crucially, unlike RQ3, the reasoning language always follows the question language in both conditions. We evaluate using the largest models in our study: Gemma3-27B, Qwen3-235B-A22B, and Claude Sonnet 4.5. Figure 7 reports AUROC with 95% confidence intervals comparing the standard multilingual setting (solid bars) against the cross-lingual setting (hatched bars). This setting represents a more challenging evaluation scenario than standard multilingual UE. Despite this increased difficulty, the cross-lingual setting does not degrade UE performance. On average, AUROC remains stable (0.68 vs. 0.67) with overlapping confidence intervals across all language families. These results demonstrate that UE methods are robust to cross-lingual answer configurations, suggesting they can be reliably deployed in multilingual systems where language mixing is common.

RQ5: How can we find the best UE threshold across languages? A key practical application of uncertainty estimation is selective prediction: using uncertainty scores to abstain from unreliable outputs (Madhusudhan et al., 2025), thereby improving system trustworthiness. The central challenge in multilingual settings is threshold calibration—selecting a filtering threshold on validation data for application at inference time. Given an uncertainty score u and threshold t, we accept the prediction if $u < t$, otherwise abstain (e.g., defer to human review). We select t by maximizing F1-score for error detection on validation data, comparing three calibration strategies: (1) t_{EN}, optimized on English data only and applied to all languages; (2) t_{GLOBAL}, optimized on pooled multilingual data; and (3) t_{LANG}, using language-specific thresholds transferred to corresponding target languages. We use one dataset as validation set for threshold tuning and the other as test set with Claude Sonnet 4.5, comparing against Oracle upper bounds where thresholds are tuned directly on target test data.

Table 1 reports the confusion matrix for each strategy. The selective prediction strategies reveal distinct trade-offs between error reduction, coverage, and deployment complexity. English-only calibration (t_{EN}) offers the most practical entry point, requiring minimal data collection while achieving a

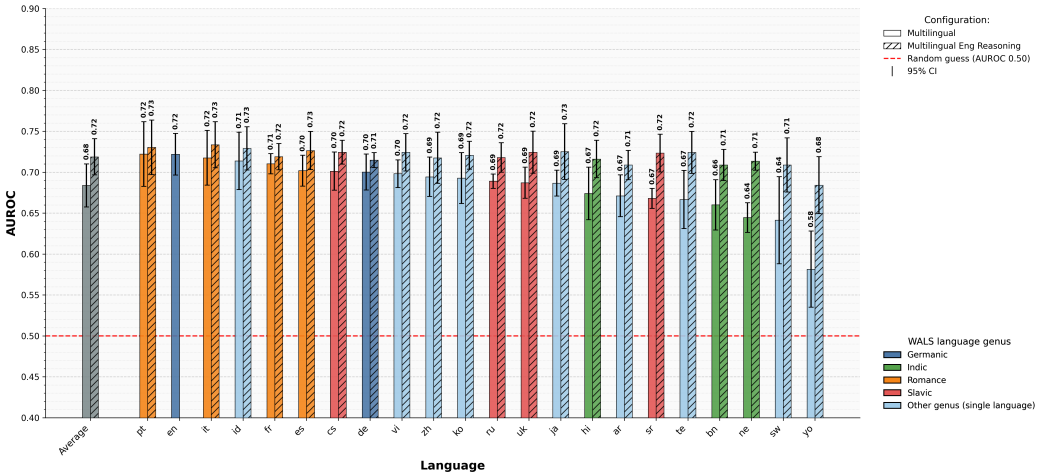


Figure 6: Effect of reasoning language on UE quality across 22 languages. Solid bars represent reasoning in target language condition; hatched bars represent the English-reasoning condition. AUROC scores are averaged across all UE methods, datasets, and the largest models (Qwen3-235B, Gemma3-27B, Claude Sonnet 4.5). Error bars denote 95% CI. English reasoning significantly narrows the performance gap for South Asian and African languages.

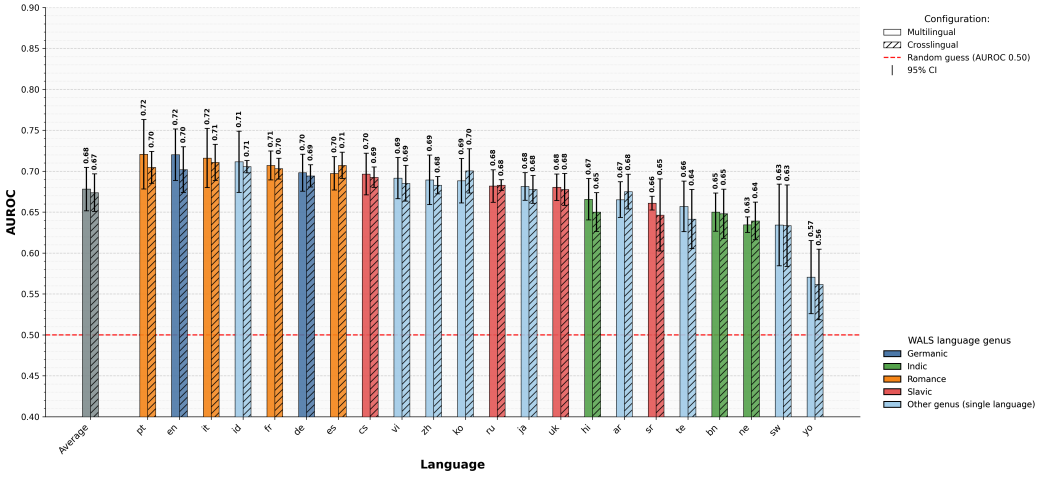


Figure 7: Effect of cross-lingual answer options on UE quality. Solid bars: multilingual setting; hatched bars: cross-lingual with each answer in a different language from the question language. Reasoning follows the question language in both conditions. AUROC scores are averaged across all UE methods, datasets, and the largest models (Qwen3-235B, Gemma3-27B, Sonnet 4.5). Error bars denote 95% CIs.

43% relative error reduction (from 13.5% to 7.7%) and maintaining high coverage with only 6.6% of correct predictions filtered out. Pooled multilingual optimization (t_{GLOBAL}) strikes a middle ground, delivering 56% error reduction at the cost of moderate coverage loss (13.5% false negatives), making it suitable for applications where validation data spans multiple languages. Language-specific calibration (t_{LANG}) achieves the strongest practical performance with 60% error reduction and 8.1% error detection, nearly matching Oracle thresholds tuned on target data itself (93.0 vs. 93.4 accuracy), though this comes at the expense of filtering out 14.4% of correct predictions. The Oracle strategy establishes the performance ceiling at 62% error reduction but requires target-specific calibration data, limiting its practical applicability. These results demonstrate that the choice of strategy depends critically on application requirements: safety-critical systems benefit most from t_{LANG}'s superior error detection, resource-constrained deployments favor t_{EN}'s simplicity, and balanced applications achieve optimal accuracy-coverage trade-offs with t_{GLOBAL}.

5 Conclusion & Future Work

We present the largest multilingual evaluation of LLM uncertainty estimation to date, covering 9

Strategy	Acc ↑	TP ↑	FP↓	TN↑	FN ↓
No filtering	86.5	86.5	13.5	0.0	0.0
t_{EN}	91.2	79.9	7.7	5.8	6.6
t_{GLOBAL}	92.4	73.0	6.0	7.5	13.5
t_{LANG}	93.0	72.1	5.4	8.1	14.4
Oracle	93.4	72.2	5.1	8.4	14.4

Table 1: Impact of different threshold selection strategies using Self-Verbalized (Claude Sonnet 4.5). TP/FP: accepted predictions that are correct/wrong. TN/FN: filtered predictions that are wrong/correct. Acc is the accuracy over accepted predictions, $\text{TP}/(\text{TP}+\text{FP})$. All values are percentages of the 44,000 samples, averaged across the two datasets.

UE methods, 9 models (5 Gemma3, 3 Qwen3, and Claude Sonnet 4.5) spanning 270M to 235B parameters, and 22 languages from high- to low-resource settings.

Our experiments yield three actionable findings. First, generation language matters more than question language: prompting models to reason in English closes the UE performance gap for low-resource languages, indicating that the reliability bottleneck lies in generation rather than comprehension. Notably, Swahili, Yoruba, and Telugu—the lowest-performing languages in our benchmark—achieve performance on par with English when reasoning is elicited in English. Nevertheless, reasoning in the target language is important under certain scenarios (e.g., explanations in target language for non-English speakers), and closing the gap is left for future work.

Second, UE method selection should be scale-aware. All UE methods maintain stable performance across model scales, except for Self Verbalized, which improves at scale (0.65 AUROC at 4B to 0.77 at 235B), becoming the best-performing approach with larger LLMs. This trend is consistent with prior evidence that meta-cognitive capabilities emerge primarily in larger models (Tian et al., 2023; Xiong et al., 2024; Steyvers and Peters, 2026), although training-pipeline differences across sizes (e.g., distillation) may also contribute.

Third, threshold calibration transfers across datasets but benefits from language awareness. English-only calibration reduces error rates by 43%, pooled multilingual calibration by 56%, and language-specific calibration by up to 60%. We also observe that sampling-based consistency methods degrade on low-resource languages. Since these methods rely on lexical overlap measures, sensitivity to surface-level variation may explain this degradation, though further investigation is needed. Our evaluation framework provides a basis for further research on trustworthy multilingual AI. Future directions include extending to open-ended generation with reliable correctness assessment and developing UE methods explicitly designed for cross-linguistic robustness.

6 Limitations

We focus on two MCQA datasets, which, while providing reliable ground-truth labels without model-based proxies, may not fully represent the diversity of real-world NLP applications. The landscape of uncertainty estimation methods is vast, and we considered in our analysis 9 state-of-the-art methods selected to cover the main paradigms (probability-based, verbalized, and consistency-based), we did not include any training-based methods. Finally, since our benchmarks derive from MMLU, some high-resource items may have been seen during pretraining; while our English-reasoning control (Appendix M) argues against memorization on low-resource tiers, cross-lingual comparisons involving high-resource languages should be interpreted with this caveat in mind.

References

Anthropic. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–294, Toronto, Canada. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Nicola Cecere, Andrea Bacciu, Ignacio Fernández-Tobías, and Amin Mantrach. 2025. Monte Carlo temperature: a robust sampling strategy for LLM's uncertainty quantification methods. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP*

2025), pages 305–320, Albuquerque, New Mexico. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *Advances in neural information processing systems*, 38:19720–19745.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

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.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jakub Podolak and Rajeev Verma. 2025. Read your own mind: Reasoning helps surface self-confidence signals in llms. *arXiv preprint arXiv:2505.23845*.

Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. 2025. Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 743–759. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Mark Steyvers and Megan A. K. Peters. 2026. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, 35(3):131–139.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. 2023. Efficient out-of-domain detection for sequence to sequence models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1430–1454, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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*.

Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. Confronting verbalized uncertainty: Understanding how llm's verbalized uncertainty influences users in ai-assisted decision-making. *International Journal of Human-Computer Studies*, 197:103455.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, Wenxuan Zhang, and Kam-Fai Wong. 2025. MlingConf: A comprehensive study of multilingual confidence estimation on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2535–2556, Vienna, Austria. Association for Computational Linguistics.

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. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Uncertainty Estimation Methods

We classify Uncertainty Estimation (UE) methods into two primary categories based on the level of access required to the underlying Large Language Model (LLM). **Open-box methods** derive uncertainty from the model's internal probability distributions (logits), whereas **Closed-box methods** operate solely on the generated text, utilizing either self-reflection or the consistency between multiple stochastic samples.

Throughout, we denote the input prompt by x, the model's parameters by ϕ, and the generated sequence by $\mathbf{y} = (y_1, \ldots, y_T)$. The model's predictive distribution at step t is denoted as $p_\phi(y_t \mid \mathbf{y}_{<t}, x)$.

A.1 Open-Box Methods

These methods require white-box access to token-level probabilities. They are computationally efficient as they typically operate on a single greedy generation, but they are inapplicable to API-based models where logits are unavailable.

Maximum Sequence Probability Following the implementation in LM-Polygraph, we estimate sequence-level uncertainty as the negative log-probability of the generated sequence:

$$\tilde{u}_{\text{MSP}}(x) = -\sum_{t=1}^{T} \log p_\phi(y_t \mid \mathbf{y}_{<t}, x).$$

We then apply dataset-level min–max normalization:

$$u_{\text{MSP}}(x) = \text{Norm}\left(\tilde{u}_{\text{MSP}}(x)\right).$$

Higher values indicate lower sequence probability and therefore greater uncertainty.

Self Certainty Following Kang et al. (2025), we measure how far each token's predictive distribution is from a uniform distribution over the vocabulary $\mathcal{V}$, via the KL divergence $\text{KL}(\mathcal{U} \,\|\, p_\phi)$. Averaging over the generated tokens and negating so that larger values denote higher uncertainty:

$$u_{\text{SC}}(x) = \text{Norm}\left(\frac{1}{T}\sum_{t=1}^{T} s_t\right),$$
$$s_t = \log|\mathcal{V}| + \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \log p_\phi(v),$$

where $p_\phi(v)$ is shorthand for $p_\phi(v \mid \mathbf{y}_{<t}, x)$. A flatter (more uniform) predictive distribution yields higher uncertainty.

A.2 Closed-Box Methods

These methods treat the LLM as a black box. We further divide them into *Verbalized* methods (single generation, self-reflection) and *Consistency* methods (multiple generations, aggregation).

A.2.1 Self-Verbalize

This approach prompts the model to explicitly report its confidence, relying on the assumption that instruction-tuned models possess internal calibration capabilities (Kadavath et al., 2022; Tian et al., 2023). We append a follow-up query asking the model to report the probability $s \in [0, 1]$ that its answer is correct, and define the uncertainty as the complement of the reported confidence:

$$u_{\text{SelfVerb}}(x) = 1 - s_\phi(x).$$

If the model fails to output a valid numerical score, we assign maximum uncertainty ($u = 1$).

A.2.2 Consistency and Graph-based Uncertainty

These methods approximate uncertainty by measuring the semantic dispersion of k stochastic samples $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$ generated at temperature $T = 1$.

Semantic Entropy (Jaccard) We use a lexical variant of Semantic Entropy (Kuhn et al., 2023) in which semantic-equivalence clustering is replaced by Jaccard-based clustering. Given k stochastic generations $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$, we group generations into clusters $C_1, \ldots, C_m$ according to their Jaccard similarity. We estimate the probability mass of each cluster empirically as

$$\hat{p}(C_j) = \frac{|C_j|}{k}.$$

Uncertainty is then computed using the Semantic Entropy formulation over the resulting cluster distribution:

$$u_{\text{SemEnt}}(x) = \text{Norm}\left(-\sum_{j=1}^{m} \hat{p}(C_j)\log\hat{p}(C_j)\right).$$

Thus, the entropy formulation follows Semantic Entropy, while Jaccard similarity replaces entailment-based semantic equivalence when constructing the clusters.

Lexical Similarity. We compute the average pairwise similarity using the ROUGE-L (Lin, 2004) metric as a proxy for consistency. Higher disagreement implies higher uncertainty:

$$S_{\text{LexSim}}(x) = \frac{1}{\binom{k}{2}} \sum_{i<j} \text{RougeL}\big(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}\big).$$

$$u_{\text{LexSim}}(x) = \text{Norm}\big(1 - S_{\text{LexSim}}(x)\big).$$

Graph-Theoretic Measures. Following Lin et al. (2024) and the LM-Polygraph implementation, we construct a weighted similarity matrix $W \in \mathbb{R}^{k\times k}$ over the k sampled generations. For the Jaccard variants used in our experiments, W_{ij} is the pairwise Jaccard similarity between generations $\mathbf{y}^{(i)}$ and $\mathbf{y}^{(j)}$.

We define the degree matrix as

$$D = \text{diag}\left(\sum_j W_{1j}, \ldots, \sum_j W_{kj}\right),$$

and the normalized graph Laplacian as

$$L = I - D^{-1/2} W D^{-1/2}.$$

1. **Laplacian Eigenvalue (EigValLaplacian).** Let $\lambda_1, \ldots, \lambda_k$ denote the eigenvalues of L. Following LM-Polygraph, the raw uncertainty score is

$$\tilde{u}_{\text{Eig}}(x) = \sum_{r=1}^{k} \max\left(0,\, 1 - \lambda_r\right).$$

   We report its normalized value

$$u_{\text{Eig}}(x) = \text{Norm}\left(\tilde{u}_{\text{Eig}}(x)\right).$$

2. **Degree Matrix (DegMat).** Following LM-Polygraph, the raw uncertainty score is

$$\tilde{u}_{\text{Deg}}(x) = \frac{\text{tr}\left(kI - D\right)}{k^2}.$$

   We report

$$u_{\text{Deg}}(x) = \text{Norm}\left(\tilde{u}_{\text{Deg}}(x)\right).$$

3. **Eccentricity.** Let $(\lambda_r, \mathbf{v}_r)$ denote the eigenvalue–eigenvector pairs of the normalized graph Laplacian L. Following the default LM-Polygraph implementation, we retain eigenvectors associated with eigenvalues below the threshold $\tau_{\text{eig}} = 0.9$:

$$\mathcal{I} = \{r : \lambda_r < 0.9\}.$$

   For each retained eigenvector, we subtract its mean component:

$$\bar{v}_r = \frac{1}{k} \sum_{i=1}^{k} (\mathbf{v}_r)_i,$$

$$\tilde{\mathbf{v}}_r = \mathbf{v}_r - \bar{v}_r \mathbf{1}.$$

   LM-Polygraph computes

$$c_r = -\left\|\tilde{\mathbf{v}}_r\right\|_2,$$

   and aggregates these components as

$$\tilde{u}_{\text{Ecc}}(x) = \left(\sum_{r\in\mathcal{I}} c_r^2\right)^{1/2}.$$

   We report the normalized uncertainty

$$u_{\text{Ecc}}(x) = \text{Norm}\left(\tilde{u}_{\text{Ecc}}(x)\right).$$

B Why not open-ended QA?

An alternative to MCQA would be the use of open-ended QA datasets such as TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), or their multilingual counterpart MKQA (Longpre et al., 2021). We do not adopt this setup, because correctness on these datasets must still be assessed on the open-ended string, which requires regex matching, an LLM-as-a-judge, or embedding similarity, reintroducing exactly the biases in the correctness signal that we want to avoid. Open-ended answers are also extremely short factoids (e.g. "Albert Einstein"): the median is 1 word for TriviaQA and 2 words for Natural Questions, so they do not provide longer text for UE than MCQA labels do, and reasoning elicitation would be needed in either case.

MCQA with elicited reasoning is therefore the only configuration that simultaneously satisfies all three requirements: (1) multilingual coverage including low-resource languages, (2) sufficiently long generations for UE, (3) and unbiased model-free correctness.

C Languages

Our benchmark covers 22 languages given by the intersection of Global-MMLU and MMLU-ProX, so that every instance has a strictly parallel multiple-choice question and shared answer key across all languages.

The languages and their ISO codes are: Arabic (ar), Bengali (bn), Czech (cs), German (de), English (en), Spanish (es), French (fr), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Korean (ko), Nepali (ne), Portuguese (pt), Russian (ru), Serbian (sr), Swahili (sw), Telugu (te), Ukrainian (uk), Vietnamese (vi), Yoruba (yo), and Chinese (zh).

D Prompts

This section details the prompts used in our experiments. All uncertainty estimation methods share the same base prompt for answer generation. The *self-verbalized* method additionally employs a follow-up prompt to elicit a confidence score from the model.

D.1 Answer Generation Prompt

```
Provide your best guess for the
    following question about {subject}
    selecting one of the options. Before
     giving your answer, provide a step-
    by-step explanation of your thought
    process. Then on a new line give the
     guess with no other words or
    explanation.

For example:

Explanation: <one sentence step-by-step
    explanation of your thought process>
Guess: <most likely guess, as short as
    possible; not a complete sentence,
    just the guess!>

Q: {question}
A. {choice_A}
B. {choice_B}
C. {choice_C}
D. {choice_D}
```

D.2 Self-Verbalized Confidence Prompt

```
Provide the probability that your guess
    is correct. Give ONLY the
    probability, no other words or
    explanation.

For example:

Probability: <the probability between
    0.0 and 1.0 that your guess is
    correct, without any extra
    commentary whatsoever; just the
    probability!>
```

To transform the probability into the uncertainty score we computed the uncertainty as $1-$probability score.

D.3 Prompt Phrasing Ablation for Self-Verbalized Confidence

We re-ran the Self-Verbalized method with Qwen3-235B on 200 English questions (100 from Global-MMLU, 100 from MMLU-ProX) under three elicitation phrasings. Error-detection AUROC was stable across wordings (Variant 1: 0.72; Variant 2: 0.74; Variant 3: 0.76), with no statistically significant difference between them.

D.3.1 Variant 1 — Baseline.

```
The following is a multiple-choice
    question. Think step by step, then
    give your final answer and your
    confidence.

Follow this format exactly:
<rationale> [your step-by-step reasoning
    ] </rationale>
<final> [the answer letter only, e.g. A]
     </final>
<verbalized_confidence> [confidence
    score between 0 and 100] </
    verbalized_confidence>
```

D.3.2 Variant 2 — "How sure are you".

```
Answer the following multiple-choice
    question. First reason through it,
    then state which option is correct
    and how sure you are on a 0 to 100
    scale.

Respond strictly in this format:
<rationale> [your reasoning] </rationale
    >
<final> [single answer letter] </final>
<verbalized_confidence> [an integer from
     0 to 100 for how confident you are]
     </verbalized_confidence>
```

D.3.3 Variant 3 — "Probability correct".

```
You are answering a multiple-choice
    question. Work through it, choose
    the best option, and estimate the
    probability that your answer is
    correct.

Use exactly this format:
<rationale> [reasoning steps] </
    rationale>
<final> [answer letter] </final>
<verbalized_confidence> [a number 0-100
    giving the percent chance your
    answer is correct] </
    verbalized_confidence>
```

E Question Answering Categories

We report the list of categories in both datasets used. The categories of Global-MMLU are [business, humanities, medical, other, stem, social sciences].

The categories of MMLU-ProX are [biology, business, chemistry, computer science, economics, engineering, health, history, law, math, other, philosophy, physics, psychology]

F Cross-lingual Q&A Accuracy comparison

This section provides task accuracy results for the cross-lingual evaluation discussed in RQ4 (Section 4). In the cross-lingual setting, the question and reasoning are in the target language, but each answer option is drawn from a different language.

We compared the task accuracy between the standard multilingual setting and the cross-lingual setting. As expected, the cross-lingual setting is more challenging: accuracy drops by 4–6 percentage points across all models across all the scales. This decrease reflects the additional difficulty of matching reasoning to answer options expressed in different languages.

Crucially, while task accuracy decreases, uncertainty estimation performance (AUROC) remains stable (see Figure 7). This indicates that UE methods successfully capture model confidence regardless of the cross-lingual configuration, and do not rely solely on lexical matching between reasoning and answer options.

Beyond uncertainty estimation, we also investigate whether English reasoning improves task accuracy. Table 2 compares language-specific and English reasoning in terms of Q&A accuracy, showing the relative improvement ($\Delta_{\text{rel}}\%$) and McNemar significance across the 21 non-English languages, pooled across Gemma3-27B, Qwen3-30B, and Claude Sonnet 4.5 on Global-MMLU and MMLU-ProX.

The results indicate that English reasoning consistently improves task accuracy, with an average relative improvement of $+3.3\%$ and statistically significant gains in 19 of 21 languages. The improvement is most pronounced for lower-resource languages such as Yoruba (YO, $+12.8\%$), Swahili (SW, $+8.8\%$), and Telugu (TE, $+7.8\%$), while higher-resource languages such as Italian (IT), Vietnamese (VI), and Portuguese (PT) exhibit smaller gains. This suggests that English reasoning not only helps models better recognize their uncertainty but also enhances their ability to answer correctly, particularly in languages where the model's native reasoning capabilities are weaker.

Lang.	Lang-Spec.	Eng. Reason.	$\Delta_{\text{rel}}\%$	Sig
YO	0.457	0.516	+12.8%	***
SW	0.588	0.639	+8.8%	***
TE	0.691	0.745	+7.8%	***
NE	0.705	0.750	+6.4%	***
BN	0.697	0.738	+5.8%	***
HI	0.709	0.749	+5.5%	***
AR	0.719	0.741	+3.1%	***
KO	0.725	0.748	+3.1%	***
UK	0.745	0.765	+2.6%	***
SR	0.737	0.753	+2.2%	***
CS	0.749	0.765	+2.1%	***
ES	0.753	0.768	+2.0%	***
JA	0.733	0.746	+1.8%	**
RU	0.747	0.761	+1.8%	**
FR	0.750	0.762	+1.7%	**
ZH	0.740	0.751	+1.5%	**
DE	0.747	0.758	+1.5%	**
ID	0.751	0.761	+1.3%	*
PT	0.759	0.769	+1.3%	**
VI	0.750	0.755	+0.7%	ns
IT	0.765	0.768	+0.4%	ns
Avg	**0.715**	**0.739**	**+3.3%**	

Table 2: Q&A accuracy under language-specific vs. English reasoning, pooled across Gemma3-27B, Qwen3-30B, and Claude Sonnet 4.5 on Global-MMLU and MMLU-ProX. Paired McNemar tests per language, BH-FDR adjusted across the 21 non-English languages. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

G Does the quantization impact multilingual UE performance?

In Figure 8, we compare the performance of full-precision Qwen3-30B against its 8-bit and 4-bit quantizations. We selected Qwen3-30B as it is the largest model in our study that fits in memory at full precision, making it the most relevant candidate for quantization. As can be noticed, there is no statistically significant impact from quantizing the model.

H Hardware Infrastructure

All experiments were conducted on an Amazon EC2 g6e.48xlarge instance with an AMD EPYC 7R13 processor (192 vCPUs), 1.5 TiB of system memory, and 8 NVIDIA L40S GPUs (48GB VRAM each, 384GB total).

I Complete Multilingual UE Method Scaling Effect

In Figure 9, we report the model size scaling factor with all the UE methods.

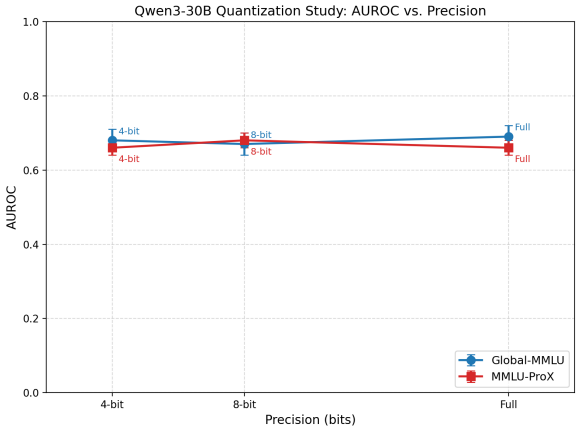


Figure 8: Uncertainty estimation performance for Qwen3-30B with quantization to 8-bit and 4-bit. Quantization has minimal impact on both accuracy and uncertainty estimation at this model scale. (± indicates 95% CI).

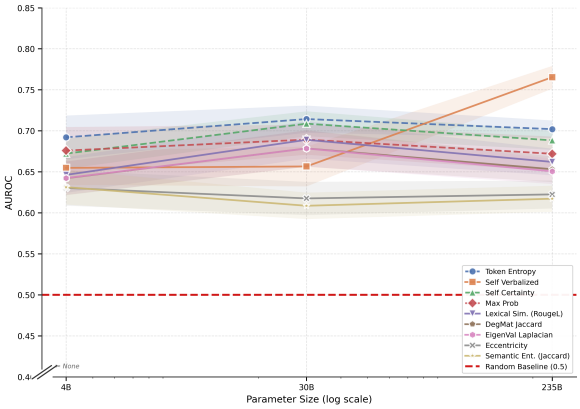


Figure 9: Effect of model scale on UE performance (Qwen3 family). Shaded regions denote 95% CIs. Self Verbalized improves with statistical significance at 235B. Dashed line: random baseline. This figure complements Figure 4 presented in the main where for simplicity we show only 4 methods.

J Label Distribution Analysis for Positional Bias

To rule out positional bias in our evaluation, we analyzed the distribution of correct answer labels across the two MCQA benchmarks used in our study. Tables 3 and 4 show that, for both Global MMLU (4 options, expected uniform $= 25\%$) and MMLU-ProX (10 options, expected uniform $= 10\%$), all labels fall within approximately ± 2 percentage points of the uniform expectation. We therefore conclude that label distributions are sufficiently balanced to exclude positional bias as a confounding factor in the results reported in the main paper.

Label	A	B	C	D
Freq. (%)	22.96	24.65	25.50	26.90

Table 3: Label distribution for Global MMLU.

Label	**Freq. (%)**
A	11.57
B	11.14
C	10.93
D	11.11
E	9.58
F	9.42
G	9.78
H	9.32
I	9.22
J	7.92

Table 4: Label distribution for MMLU-ProX.

K Calibration Analysis

Figure 10 extends the AUROC results from the main paper (Figure 5) with Expected Calibration Error (ECE), which measures the alignment between predicted confidence and actual accuracy. Lower ECE indicates better calibration.

L Statistical Significance Analysis

To complement the 95% confidence intervals reported in the main paper, we perform formal hypothesis testing with the DeLong test for ROC curves, applying Benjamini–Hochberg (BH) FDR correction across all 22 languages, complementing the confidence intervals shown in Figures 4 and 9. We report per-language results for two analyses on Global-MMLU; note that per-language values on a single dataset differ from the cross-dataset averages shown in Figure 4.

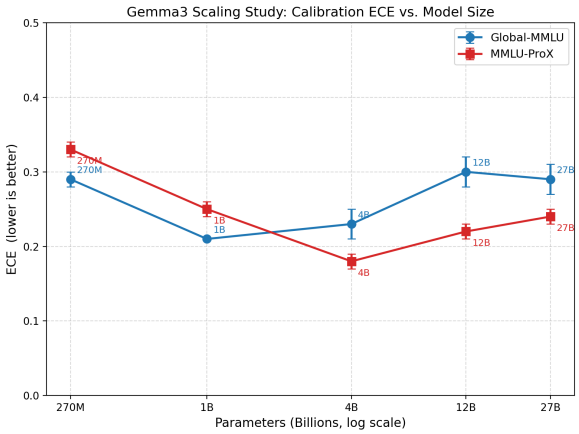


Figure 10: Effect of model scale on ECE. (Mean across langs and 95% CIs).

Model scaling (RQ2). We test Self-Verbalized AUROC on Qwen3-235B vs. Qwen3-30B. Because the two models produce distinct correctness labels, their ROC curves are independent and we use an unpaired two-sided DeLong test per language. The scaling advantage is significant in 22 of 22 languages after FDR correction (mean gap: $+0.153$ AUROC).

Lang	235B	30B	Δ	p (BH)	Sig
ar	0.693	0.558	+0.135	3.3e-05	***
bn	0.710	0.582	+0.129	2.2e-05	***
cs	0.731	0.623	+0.108	1.3e-03	**
de	0.725	0.585	+0.139	1.3e-04	***
en	0.805	0.571	+0.234	5.0e-09	***
es	0.728	0.579	+0.149	6.1e-05	***
fr	0.711	0.566	+0.145	4.6e-05	***
hi	0.746	0.621	+0.125	4.6e-05	***
id	0.742	0.585	+0.157	6.4e-06	***
it	0.748	0.578	+0.170	5.0e-06	***
ja	0.748	0.556	+0.192	3.6e-10	***
ko	0.746	0.542	+0.204	3.3e-09	***
ne	0.700	0.593	+0.106	4.4e-04	***
pt	0.742	0.567	+0.175	2.3e-06	***
ru	0.771	0.554	+0.218	3.6e-10	***
sr	0.744	0.551	+0.192	3.8e-09	***
sw	0.670	0.575	+0.096	1.5e-04	***
te	0.770	0.608	+0.162	1.8e-07	***
uk	0.771	0.613	+0.159	4.0e-06	***
vi	0.715	0.584	+0.130	6.7e-05	***
yo	0.655	0.549	+0.105	4.6e-05	***
zh	0.732	0.592	+0.140	3.3e-05	***

Table 5: Scaling within the Qwen3 family (235B vs. 30B), Self-Verbalized AUROC on Global-MMLU, all 22 languages. Unpaired DeLong test, BH-FDR adjusted. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

Method ranking (RQ1/RQ2). We compare Self-Verbalized vs. Token Entropy on Qwen3-30B (our largest unquantized open-box baseline) with a paired DeLong test for correlated ROC curves, as both methods are scored on identical samples under identical correctness labels. The difference is significant in 19 of 22 languages after correction, uniformly favoring Token Entropy across all significant outcomes (mean gap: 0.144), consistent with the scale-dependent reversal discussed in RQ2.

Lang	TE	SV	Δ (SV$-$TE)	p (BH)	Sig
ar	0.725	0.558	-0.167	7.2e-08	***
bn	0.645	0.582	-0.064	3.0e-02	*
cs	0.780	0.623	-0.158	5.6e-08	***
de	0.726	0.585	-0.141	1.2e-05	***
en	0.773	0.571	-0.202	2.7e-07	***
es	0.760	0.579	-0.181	2.8e-07	***
fr	0.774	0.566	-0.208	4.5e-12	***
hi	0.664	0.621	-0.043	1.5e-01	ns
id	0.789	0.585	-0.204	4.5e-12	***
it	0.811	0.578	-0.232	2.6e-11	***
ja	0.754	0.556	-0.198	1.8e-10	***
ko	0.696	0.542	-0.153	1.6e-06	***
ne	0.673	0.593	-0.080	1.2e-02	*
pt	0.795	0.567	-0.228	1.7e-13	***
ru	0.767	0.554	-0.214	1.6e-11	***
sr	0.732	0.551	-0.180	1.1e-09	***
sw	0.636	0.575	-0.061	1.6e-02	*
te	0.631	0.608	-0.022	4.7e-01	ns
uk	0.746	0.613	-0.134	4.7e-05	***
vi	0.744	0.584	-0.160	1.9e-07	***
yo	0.514	0.549	$+0.035$	2.3e-01	ns
zh	0.754	0.592	-0.162	1.6e-06	***

Table 6: Method ranking (Token Entropy (TE) vs. Self-Verbalized) on Qwen3-30B, AUROC on Global-MMLU, all 22 languages. Paired DeLong test, BH-FDR adjusted. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

Scale-dependent reversal (RQ2). The per-language DeLong test above compares methods within each language ($\sim$600 items). To test the scale-dependent *ranking* reversal visible in Figure 4, which concerns the mean across languages, we use a paired Wilcoxon signed-rank test over the 22 per-language AUROC values on Global-MMLU. At 30B, Token Entropy significantly outperforms Self-Verbalized (mean 0.722 vs. 0.579; $W = 2.0$, $p < 10^{-5}$; cf. Table 6). At 235B the ranking reverses: Self-Verbalized significantly outperforms Token Entropy (mean 0.732 vs. 0.696; higher in 17/22 languages; $W = 31.0$, $p = 1.1 \times 10^{-3}$), confirming the crossover.

M Contamination Control: English Reasoning as a Memorization Test

If a model had memorized an answer during pretraining, it should recall it regardless of the

reasoning language, since the question and options remain in the target language in both conditions. We test this with paired McNemar tests (BH-FDR corrected across the 21 non-English languages) on the predictions pooled across Gemma3-27B, Qwen3-30B, and Claude Sonnet 4.5 on both datasets. Table 2 reports the per-language accuracies with the corresponding significance levels: English reasoning significantly improves accuracy in 19 of 21 languages, with the largest gains on the lowest-resource languages (Yoruba $+12.8\%$, Swahili $+8.8\%$, Telugu $+7.8\%$); on these languages the improvement is also significant within each model individually ($p < 0.05$).