WHAT ENGLISH-ONLY, NON-FRONTIER EVALUATION MISSES ABOUT CONSISTENCY-BASED CONFIDENCE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Confidence estimation helps users decide when to trust an AI model's answer. Most common black-box confidence estimation methods, such as self-consistency, measure agreement among answers sampled from the model given the question. Analyzing recent work, we find that research results on consistency-based methods largely come from English-only evaluations of small models, leaving other languages and recent models under-explored. We therefore evaluate consistency-based confidence estimation across nine languages and a range of small open-source to frontier models. Our results show that self-consistency performs worse outside English for all 17 models we tested. Moreover, verbalized confidence using recent Claude models outperforms self-consistency, despite using only one LLM call. For a more effective consistency-based approach across models and languages, we introduce Rosetta, which uses translation as a principled perturbation method and measures agreement between the resulting multilingual samples and the original answer to estimate confidence. Rosetta improves confidence estimation over same-language rewording in English and outside English at a matched call budget and without labeled calibration data, while narrowing the average language gap.

1 INTRODUCTION

Large language models (LLMs) have become increasingly prevalent across various applications. Confidence estimation helps users and agentic systems decide whether to trust an answer, seek verification, or abstain (Kadavath et al., 2022; Ding et al., 2025; Feng et al., 2024; Xiong et al., 2024). Confidence estimation is especially challenging in a black-box setting using only the model's input and output text, with no access to the model's internals or token probabilities. *Consistency-based* confidence estimators are the most common black-box approach (Wang & Zhang, 2026). They estimate confidence by measuring agreement between the original answer and additional sampled answers. The default form, self-consistency, generates these answers by resampling from the same question (Wang et al., 2023; Manakul et al., 2023; Xiong et al., 2024). A model can repeatedly give the same wrong answer and appear confident, making agreement misleading (Tan et al., 2025).

To see where the evidence for these methods comes from, we audited the evaluation settings of 296 recent papers; Appendix J.1 lists the venues and each paper's language coverage. Of these, 273 (92.2%) use English-only data. The evaluations are also narrow in which models they test: as shown in the left panel of Figure 1, the share of papers testing a frontier model falls from 74% in 2023 to 29% in the available 2026 proceedings, and the median open model tested has 8B parameters in 2025 and 2026. These evaluations leave the usefulness of consistency-based confidence estimation outside English and on frontier models largely underexplored. In our survey, only four papers compare the same questions across languages. The most direct of them evaluates four open models and finds that consistency-based detectors rank answers less well in low-resource languages than in English (Datta et al., 2026). We extend this comparison to 17 models across nine languages, using parallel questions covering academic knowledge and mathematical reasoning. We also evaluate on natively written exam questions (Romanou et al., 2025).

Our results show that self-consistency gives worse confidence estimates outside English compared to English for all 17 models, and rewording the question in the same language leaves a similar gap.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

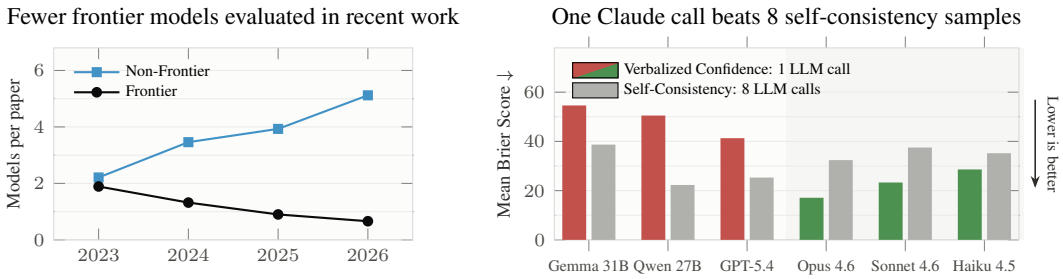


Figure 1: **Left**: Average number of frontier & non-frontier models per paper across 296 audited papers, by year. Frontier includes all closed-weight models from OpenAI, Google, and Anthropic as well as all non-distilled DeepSeek, Kimi, and Z.ai models with $\geq$200B total parameters. All other models are non-frontier. **Right**: Mean Brier score (3) over five English datasets for one verbalized-confidence call versus eight self-consistency samples. Claude models shown in green.

This gap is modest but consistent. Recent work has shown that LLMs can give different answers to the same question across languages (Qi et al., 2023; Ohmer et al., 2023). This variation motivates gathering evidence beyond the question's original language. We therefore introduce Rosetta to test whether cross-language agreement provides more useful evidence about an answer's correctness. Rosetta translates the input text into multiple languages, then samples one answer per language and uses their agreement with the original answer to estimate confidence. Rosetta closes the mean language gap and improves non-English AUROC over Paraphrase for 14 of 17 models. In English, Rosetta also improves calibration over Paraphrase at the same call budget.

Beyond language, our evaluation reveals differences across model families. Our results show that a single call – formally, *verbalized confidence* – asking the model to express its confidence in an answer outperforms eight self-consistency samples for the evaluated Claude models (Figure 1). However, this pattern does not hold for the other three recent models. Our experiments show a confidence rating from a second model can also improve on consistency-based estimation, but the benefit depends highly on the evaluator model. Furthermore, we show that one rating from a well-calibrated strong model such as Opus yields better confidence estimates than self-assessment and the tested consistency-based estimators for all six other answering models, while a rating from the strong GPT-5.6 Sol does not. Together, these findings reveal phenomena missed by English-only evaluations of non-frontier models, highlighting the importance of multilingual and frontier-model evaluation of confidence estimation. We summarize our contributions below.

Contributions.

1. Confidence Estimation Across Languages. Across nine high-resource languages, we show that self-consistency gives worse confidence estimates outside English for all 17 evaluated models, and paraphrasing leaves a similar modest yet consistent gap.

2. Rosetta for Multilingual Confidence Estimation. We introduce Rosetta, which outperforms the other consistency-based estimators on Brier in all 18 non-English comparisons and has the lowest Brier of all five estimators in 37 of 45 English comparisons.

3. Externalizing Confidence. We show that a single call to a well-calibrated frontier model, namely Opus 4.6, effectively verbalizes confidence of its own and other models' predictions competitive with Rosetta.

2 Related Work

Consistency-Based Estimators. Consistency-based estimators use answer agreement as the uncertainty signal, differing in how they measure agreement and obtain the answers. One line of work estimates the uncertainty of the model on a question from how spread out the samples are, while others estimate the confidence of one particular answer (Lin et al., 2024; Qiu & Miikkulainen, 2024). These approaches also differ in what counts as agreement. Some count identical answers (Wang et al., 2023), group answers with the same meaning (Kuhn et al., 2023; Farquhar et al., 2024), or score

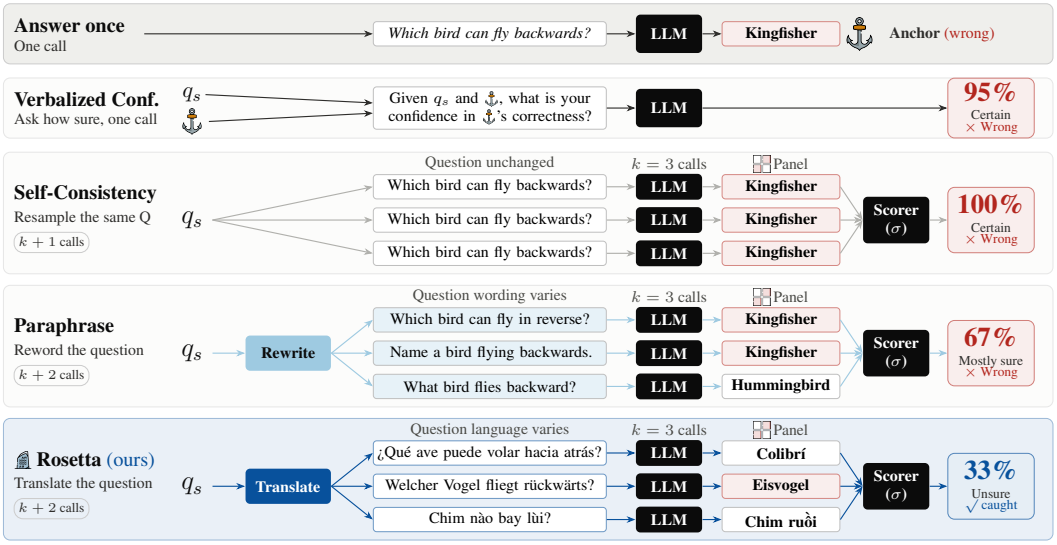


Figure 2: Self-consistency, paraphrase, and Rosetta (ours) assess uncertainty of the same anchor using different panels (size $k = 3$). Verbalized confidence directly estimates uncertainty without any panel. Each dark box represents one LLM call.

similarity between answers so that partial agreement counts (Lin et al., 2024; Nguyen et al., 2025; Nikitin et al., 2024; Qiu & Miikkulainen, 2024). The answers themselves can be generated in several ways. Self-consistency resamples a fixed input (Wang et al., 2023; Manakul et al., 2023). However, consistent errors persist as models scale up (Tan et al., 2025). To break such errors, recent work samples answers from perturbed inputs, usually by paraphrasing the question (Zhang et al., 2023; Gao et al., 2024; Song et al., 2025). Others measure agreement across models (Hamidieh et al., 2026; Kruse et al., 2025), or among several agents of the same model (Feng et al., 2025). Yet different samples, and even different models, tend to make the same mistakes, so adding more of them does not reliably help (Denisov-Blanch et al., 2026).

Verbalized Confidence. Verbalized confidence asks a model to quantify its confidence in a given answer (Lin et al., 2022; Tian et al., 2023). To reduce potential model overconfidence (Xiong et al., 2024; Wang et al., 2025a), some later work asks models to assign probabilities to several answers (Wang et al., 2025a) or to adjust their stated confidence using their confidence in alternative answers (Wang & Stengel-Eskin, 2026). Confidence can also be estimated by another model (Bouchard & Chauhan, 2025; Chen et al., 2024b). We compare verbalized confidence with consistency-based estimators to measure what additional sampling contributes across model families.

Language as an Axis. Prior work uses answers and feedback across languages to select answers (Qin et al., 2023; Khairi et al., 2025), build training data (Wang et al., 2025d; She et al., 2024), and decide when to abstain (Feng et al., 2024; Sun et al., 2025). Other studies examine how confidence estimation itself changes across languages. MlingConf compares token probabilities, P(True), and verbalized confidence separately in five languages, finding lower average calibration error in English than in most other languages for its two main models (Xue et al., 2025). Other work examines internal representations to improve multilingual calibration (Zhou et al., 2025). Datta et al. (2026) find worse ranking capability for sampling-based hallucination detectors in low-resource languages than in English. Their supervised detectors, which use internal model representations, outperform these sampling-based detectors. They also find that detection performance often declines less than task accuracy. We use agreement across languages as the confidence signal for a fixed answer and compare it with self-consistency across languages and model families.

3 Rosetta: Measuring Confidence via Multiple Languages

Problem formulation. Given a question q_s in language s, a model with response distribution p_θ generates an answer $\hat{y}_s$, which we call the *anchor*⚓. In the black-box setting, our goal is to estimate confidence in the anchor's correctness using the model's input and output. Consistency-based

estimators gather a *panel* of k additional answers and measure their agreement with the anchor. The panel contains k answers, $y_1, \ldots, y_k$, sampled from the same model. To generate answer y_j, an optional transformation g_j is applied to the original question before sampling. A scoring function σ converts agreement between the panel and the anchor into a confidence score.

$$c(q_s, \hat{y}) = \sigma(\hat{y}, y_1, \ldots, y_k) \in [0, 1], \qquad \text{where} \qquad y_j \sim p_\theta(\cdot \mid g_j(q_s)). \tag{1}$$

Rosetta. Answers sampled in one language can repeat the same mistake. We therefore introduce Rosetta, which forms the consistency panel across multiple languages. The model first generates all k translations of the original question q_s in one call. In this method, $g_j(q_s)$ is the translation function into panel language j. The same model then samples one answer y_j from each translated question to form the panel. Following Manakul et al. (2023); Zhang et al. (2023); Zhao et al. (2024), the original model rates agreement between the anchor and panel in one LLM call. The prompt asks about agreement rather than correctness and only contains the anchor and panel, omitting the original question. Figure 2 illustrates how translation changes the panel compared with Self-Consistency and Paraphrase. We also introduce **Rosetta$_{\text{AO}}$** (answer only), which keeps the question unchanged and asks the model to answer in each panel language. For Rosetta$_{\text{AO}}$, g_j adds an instruction to the question to answer in language j. Using the same panel languages and scoring procedure as Rosetta, this variant requires k answer calls and one scoring call, matching the self-consistency LLM calls.

4 Experimental Setup

Section 5's experiments focus on our first three research questions while Section 6 focuses on our fourth and fifth research questions.

Metrics. We evaluate confidence estimates using Brier Score, ECE, and AUROC. Brier measures the accuracy of confidence values; ECE measures calibration, and AUROC measures how well confidence ranks correct vs. incorrect. Appendix B.3 gives their definitions. For comparison, lower Brier and ECE are better, but higher AUROC is better. We report all metrics as percentages.

Baselines. We use Self-consistency, Paraphrase, and verbalized confidence as baselines. For *Self-consistency*, g_j in Eq. (1) is the identity transformation, $g_j(q_s) = q_s$, so stochastic sampling generates the samples (Wang et al., 2023; Manakul et al., 2023; Wang et al., 2024a). It uses k answer calls and one scoring call, matching Rosetta$_{\text{AO}}$'s number of calls. For *Paraphrase*, each $g_j(q_s)$ is a same-language perturbation, and the model samples one answer per rewrite (Zhang et al., 2023; Gao et al., 2024; Zhao et al., 2024) as shown in Figure 2. Paraphrase requires one extra call to generate the rewrites, resulting in $k + 2$ calls, matching Rosetta. For *verbalized confidence*, we ask the model to state its numeric confidence in the anchor's correctness (Tian et al., 2023; Xiong et al., 2024).

Datasets and models. Multilingual evaluation uses PolyMath (Wang et al., 2025c) and MMLU-ProX (Xuan et al., 2025), which provide the same source questions translated across our multiple languages. We evaluate 17 models ranging from 2B parameters to frontier models. For a more comprehensive comparison, we evaluate 9 models on these two multilingual datasets and five English datasets: PopQA (Mallen et al., 2023), PolyMath$_{\text{en}}$, MMLU-ProX$_{\text{en}}$, SuperGPQA (Du et al., 2025), and SimpleQA (Wei et al., 2024). Appendix B lists the employed datasets, languages, and models[1].

Evaluation protocol. We use 500 items per dataset, randomly selected with a fixed seed, and use the same sample sets across confidence estimators. For each sample, we generate an anchor⚓answer and a panel of $k = 8$ answers unless otherwise stated. The choice of panel languages is detailed in Appendix B. Following Manakul et al. (2023), anchor and panel answers are generated at temperatures 0 and 1, respectively. Paraphrase generation, translation, and scoring calls use temperature 0. The scoring prompt is in English for every anchor language, and the verbalized-confidence prompt asks for confidence in the question's own language. To evaluate the confidence estimates, a separate LLM judge, GPT-5.6 Sol (OpenAI, 2026b), grades each anchor against the reference answer. Throughout these experiments, we use no labeled calibration data and do not tune the prompts (Appendix I).

[1]We disabled reasoning to make experiments' cost manageable. Otherwise, cost increased by more than an order of magnitude.

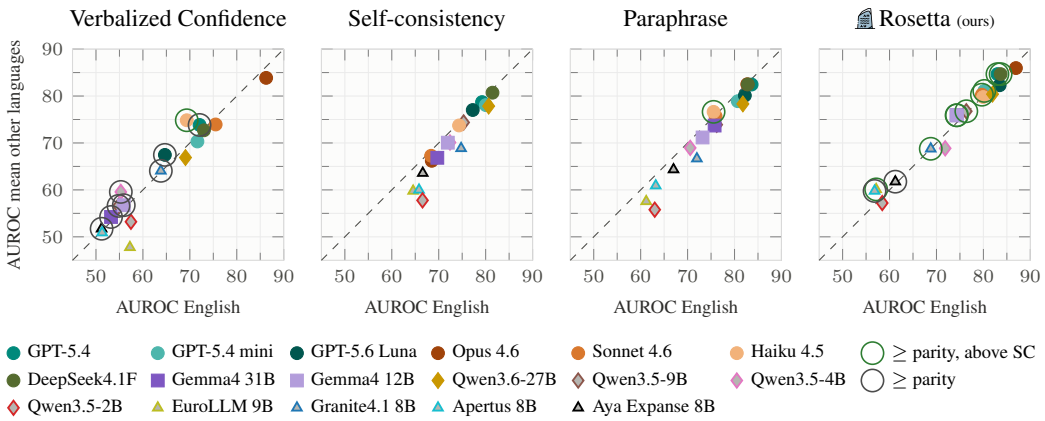


Figure 3: **Confidence Estimation in English vs. Other Languages.** Each point represents one model, with English AUROC on the x-axis and average AUROC across eight non-English languages on the y-axis. Values are averaged over PolyMath and MMLU-ProX. "≥ Parity" denotes a method's performance outside English meets or exceeds its English performance, with a green circle denoting superior English performance than self-consistency. Such points are above the dashed, diagonal line.

5 Confidence Estimation Beyond English

Each subsection below focuses on a different research question regarding self-consistency's performance beyond English. See Appendix E for a full ablation study.

5.1 **RQ1:** Do consistency-based estimators favor English?

In this experiment, we assess confidence estimation across languages using PolyMath and MMLU-ProX, which provide the same questions translated into multiple languages. On those datasets, we evaluate 17 models, ranging from small to frontier models, across nine languages: Arabic, Chinese, English, German, Japanese, Korean, Russian, Spanish, and Vietnamese. Each language serves as the anchor language in turn, with the other eight forming Rosetta's panel. Figure 3 plots each model's AUROC in English compared to its average across the other eight languages and over the two datasets. Points below the diagonal indicate worse performance outside English. We define the language gap as the non-English average minus the English value. Our results show that for all 17 models, self-consistency performs better in English with a modest but consistent gap, with the mean gap over all 17 models of -2.7 AUROC points with a 95% bootstrap interval over models $[-3.9, -1.7]$. The gap is statistically significant for 5 of the 17 models individually under 95% bootstrap intervals that resample questions jointly across languages, and the mean gap across all models is also significant. Even paraphrasing the question leaves a similar mean gap of -2.2 percentage points (pp). In contrast, with Rosetta the mean gap moves from -2.7pp to $+0.4$pp, and the number of models performing worse outside English falls from 17 to 6. By comparison, verbalized confidence shows a smaller language gap, but its AUROC is consistently lower than Rosetta. After sub-sampling to match English and non-English accuracy, self-consistency's mean AUROC gap remains -2.8 points, while Rosetta's is $+0.3$. Appendix D provides detailed per-language and per-model results, and Appendix E.5 shows details on the accuracy-matching experiment.

Takeaway: Self-consistency and Paraphrase have a modest but consistent performance gap outside English, while Rosetta (ours) closes the outside English gap.

5.2 **RQ2**: Does Rosetta Improve Confidence Estimation?

Here we examine whether Rosetta provides better confidence estimates than self-consistency and paraphrase. Table 1 compares five estimators across nine models, averaged over five English and two non-English datasets, resulting in 45 and 18 model-dataset comparisons. Our results show that asking for the answer to be in other languages with Rosetta$_{\text{AO}}$ improves confidence estimates at the

Table 1: Confidence estimation per model for the five estimators at $k = 8$. **Outside English:** Mean score over the eight non-English languages in multilingual datasets, PolyMath and MMLU-ProX. **English:** Mean English score for the two multilingual datasets as well as three additional English-only datasets. Best per row and metric in bold. The small number after each Rosetta$_{AO}$ and Rosetta is the comparison to the same call budget estimator.

	One call			Question unchanged (k+1 calls)									Question rewritten (k+2 calls)								
	Verbalized Confidence			Self-Consistency			Rosetta$_{AO}$ (ours)						Paraphrase			Rosetta (ours)					
Model	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓		AUROC↑		ECE↓		Brier↓	AUROC↑	ECE↓	Brier↓		AUROC↑		ECE↓	
Outside English (*Two datasets: eight languages, PolyMath and MMLU-ProX*)																					
GPT-5.4	37.6	73.8	37.6	23.1	78.7	25.0	19.1	−4.0	80.6	+1.9	17.9	−7.2	18.8	82.4	18.7	**16.2**	−2.6	**84.7**	+2.2	**14.4**	−4.2
GPT-5.4 mini	42.1	70.3	42.9	27.1	78.1	29.9	23.6	−3.5	78.8	+0.7	24.5	−5.3	25.1	78.9	27.4	**21.3**	−3.8	**81.1**	+2.2	**21.5**	−5.8
Sonnet 4.6	23.5	73.9	17.6	33.4	67.3	34.9	26.0	−7.4	72.3	+5.0	25.1	−9.7	23.6	75.7	22.4	**19.9**	−3.7	**80.3**	+4.6	**17.1**	−5.3
Haiku 4.5	24.5	74.8	20.8	30.9	73.7	31.8	24.1	−6.8	75.3	+1.6	22.1	−9.7	24.5	76.6	23.0	**20.6**	−3.9	**79.8**	+3.2	**17.5**	−5.5
Opus 4.6	14.4	83.9	**7.0**	25.3	66.2	26.0	17.0	−8.2	79.5	+13.3	16.0	−10.0	15.0	82.5	12.3	**13.1**	−2.0	**86.0**	+3.5	10.3	−2.0
DeepSeek V4.1F	25.3	72.7	24.4	20.8	80.7	21.1	17.0	−3.8	82.3	+1.6	13.1	−8.0	18.5	82.5	17.5	**15.1**	−3.5	**84.7**	+2.2	**10.5**	−7.0
Gemma 4 31B	51.3	54.2	51.4	40.0	66.8	43.0	33.7	−6.3	69.9	+3.1	35.3	−7.7	29.3	73.7	30.8	**26.1**	−3.2	**75.9**	+2.2	**26.0**	−4.8
Qwen3.6-27B	46.5	66.9	47.1	22.4	77.8	22.5	18.6	−3.8	78.5	+0.7	15.1	−7.4	19.5	78.3	17.2	**17.2**	−2.3	**80.4**	+2.1	**12.7**	−4.6
Gemma 4 12B	59.0	56.7	59.4	39.7	70.1	44.2	31.5	−8.2	71.9	+1.8	33.6	−10.5	29.7	71.2	31.8	**25.1**	−4.7	**75.9**	+4.7	**25.1**	−6.6
English (Five datasets: *PolyMath$_{en}$, MMLU-ProX$_{en}$ & English Only Datasets: SimpleQA, PopQA, SuperGPQA*)																					
GPT-5.4	41.3	70.2	41.3	25.3	77.2	26.9	20.7	−4.6	79.8	+2.6	19.6	−7.3	20.6	81.6	20.7	**18.0**	−2.6	**82.2**	+0.6	**14.8**	−5.9
GPT-5.4 mini	45.8	68.9	46.5	24.5	77.8	25.0	21.8	−2.7	76.3	−1.5	18.8	−6.2	23.0	78.3	22.5	**20.2**	−2.8	**78.7**	+0.4	**17.5**	−5.0
Sonnet 4.6	23.3	74.5	18.4	37.5	67.5	39.3	25.0	−12.5	73.3	+5.8	22.7	−16.6	25.7	75.2	25.3	**19.8**	−6.0	**78.8**	+3.6	**14.9**	−10.4
Haiku 4.5	28.6	68.1	27.9	35.2	70.1	38.0	24.6	−10.6	71.7	+1.6	22.7	−15.3	27.0	72.2	26.9	**21.7**	−5.4	**73.6**	+1.4	**18.2**	−8.7
Opus 4.6	**17.1**	80.9	**8.7**	32.4	66.8	33.5	21.0	−11.4	77.4	+10.6	19.9	−13.6	19.2	80.4	17.3	**17.3**	−1.9	**81.9**	+1.5	13.7	−3.6
DeepSeek V4.1F	26.5	69.8	24.5	23.9	78.2	24.5	19.2	−4.7	79.0	+0.8	13.6	−10.9	20.5	**80.3**	18.4	**18.2**	−2.3	79.1	−1.2	**10.5**	−8.0
Gemma 4 31B	54.6	54.4	54.6	38.7	70.1	43.6	28.7	−10.0	72.3	+2.2	31.6	−12.0	27.4	**75.2**	30.2	**23.5**	−3.9	74.2	−1.0	**23.0**	−7.2
Qwen3.6-27B	50.5	64.5	51.1	22.3	80.0	22.7	18.2	−4.1	78.1	−1.9	14.5	−8.2	19.1	**80.3**	17.2	**17.4**	−1.7	77.8	−2.5	**12.1**	−5.1
Gemma 4 12B	56.3	55.4	56.4	35.0	74.3	40.6	26.0	−9.0	75.9	+1.6	28.3	−12.2	26.7	74.7	30.2	**21.8**	−4.9	**76.3**	+1.6	**21.2**	−9.0

same call budget as self-consistency. Rosetta$_{AO}$ improves calibration across all nine models, with ECE falling by 11.4 points in English and 8.4 points outside English. It also lowers mean Brier and raises mean AUROC in both settings.

Translating the question with Rosetta further improves all three metrics on average in both settings. At the same call budget as Paraphrase, Rosetta lowers mean ECE and Brier and raises mean AUROC in both settings, with larger AUROC gains outside English. Outside English, Rosetta improves all three metrics over Paraphrase in all 18 model–dataset comparisons. In English, it lowers ECE in 43 of 45 comparisons and significantly lowers Brier in 31, with no significant Brier losses (Appendix Table 16). Among all five estimators, Rosetta has the lowest Brier in 37 of 45 English comparisons and the highest AUROC in 17 of 18 non-English comparisons (Appendix Tables 4, 5, and 15).

However the improvement is not uniform across models. Our results show AUROC gains grow with the model's non-English accuracy. Table 2 shows a mean gain over self-consistency of 4.6pp for models with accuracy from 30% to 45% and 9.0pp for those at or above 45%, compared with −0.6 below 30%. Rosetta improves AUROC over both self-consistency and Paraphrase for all eleven models at or above 30%, including all 9 models in Table 1. Across models, the AUROC gain correlates with non-English accuracy (Pearson's $r = 0.86$). Appendix Table 3 lists each model's accuracy, its AUROC under self-consistency and Rosetta, and the gain.

Table 2: Difference between Rosetta and self-consistency by model competence in the 8 non-English languages; wins count models where Rosetta is better.

Accuracy	Models	AUROC↑ Δ (wins)	ECE↓ Δ (wins)
<30%	6	−0.6 (1/6)	−8.6 (5/6)
[30%, 45%)	5	+4.6 (5/5)	−12.5 (5/5)
≥45%	6	+9.0 (6/6)	−13.7 (6/6)

Finally, we test whether the improvements extend to questions originally written in each language, using eight languages of INCLUDE (Romanou et al., 2025). Rosetta beats self-consistency for five of six models on both Brier and AUROC (Qwen3.6-27B ties on AUROC), and beats Paraphrase for three on Brier and five on AUROC (Table 10 in Appendix E.6).

Takeaway: Rosetta improves on confidence estimation compared to self-consistency.

5.3 **RQ3**: Why Does Rosetta Help?

To understand the source of Rosetta's improvement, we first compare the panels produced by self-consistency, Paraphrase, and Rosetta. Figure 4(b) compares answer novelty across the three methods. We measure novelty as the share of a panel's answers whose content appears in neither of the other

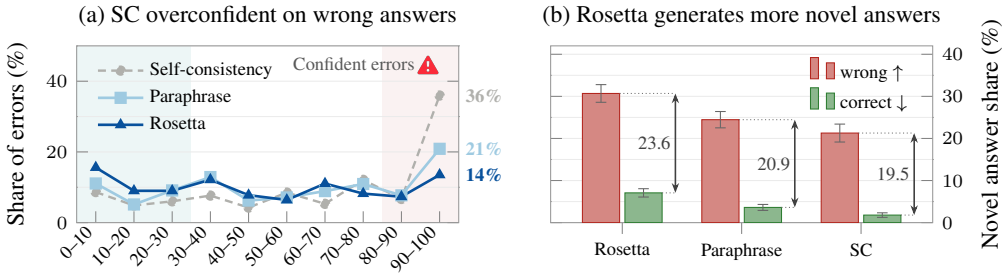


Figure 4: Confidence and answer novelty for GPT-5.4 on PolyMath and MMLU-ProX. (a) Confidence distributions on wrong English answers. (b) Novel-answer shares for wrong and correct anchors across nine languages. Arrows show the difference between these shares.

two panels. GPT-5.6 Luna judges this overlap blind across 1,800 items per model from PolyMath and MMLU-ProX in nine anchor languages. For GPT-5.4, all three methods produce more novel answers when the anchor is wrong, but Rosetta shows the largest difference. Rosetta's novel-answer share rises from 7.1% on correct anchors to 30.7% on wrong anchors, a difference of 23.6 points compared with 20.9 for Paraphrase and 19.5 for self-consistency. Rosetta's wrong–correct novelty gap is larger than Paraphrase's across all seven models in the appendix table. Details are shown in Appendix Table 11.

If other languages catch errors caused by the question's language, Rosetta should gain most on questions the model answers correctly in English. Across 17 models answering the same 500 questions in every language of the nine-language grid for PolyMath and MMLU-ProX, we call a wrong non-English answer language-specific when the model is right on the same question in English and shared when it is also wrong in English, with language-specific errors making up 18% of non-English wrong answers on average and 9% to 26% across models. Rosetta separates correct answers from both kinds of wrong answers better than every other estimator, with mean AUROC over models of 75.8 for language-specific errors and 73.9 for shared errors and gains over self-consistency of 6.9 points for language-specific errors (15 of 17 models) and 3.9 points for shared errors (13 of 17 models). Paraphrase also gains more on language-specific errors than on shared errors, with translation giving the largest gain in separating correct from wrong answers by confidence.

We next examine how the variances in the panels are reflected in confidence on wrong answers. Figure 4(a) compares the confidence distributions of self-consistency, Paraphrase, and Rosetta on GPT-5.4's wrong anchor answers on English MMLU-ProX and PolyMath. Rosetta assigns at least 90% confidence to 14% of GPT-5.4's wrong answers, compared with 21% for Paraphrase and 36% for self-consistency. The same pattern holds across five models and two datasets, where these shares average 16%, 22%, and 42%, respectively (Appendix F.2). The net Brier gain is positive against self-consistency in all ten model–dataset comparisons and against Paraphrase in eight of ten comparisons.

Takeaway: When the anchor is wrong, Rosetta provides better uncertainty estimates by producing more novel answers, and in turn assigning lower confidence.

6 Better Contextualizing Consistency-Based Uncertainty Estimation on Frontier Models

A claimed benefit of consistency-based uncertainty estimation is its applicability to black-box models, including frontier models (Chen et al., 2024a; Nguyen et al., 2025; Datta et al., 2026; Chen et al., 2026; McCabe et al., 2026). However, our survey in Section 1 shows that the consistency-based uncertainty literature increasingly de-emphasizes frontier model evaluation – likely due to the prohibitive cost. This section tests this often cited but under-tested claim more thoroughly.

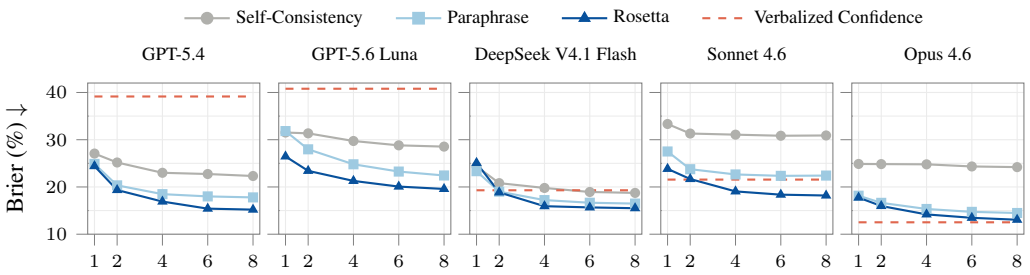


Figure 5: Mean Brier versus panel size (k) on English datasets PolyMath$_{en}$ and MMLU-ProX$_{en}$. The dashed line shows verbalized-confidence's performance with its single LLM call. Lower is better.

6.1 **RQ4**: When Does Consistency-Based Estimation Benefit Frontier Models?

Extending Section 5.2's experiments, Figure 5 sweeps panel size k from 1 to 8 for five frontier models – GPT-5.4, GPT-5.6 Luna, DeepSeek V4.1 Flash, Sonnet 4.6, and Opus 4.6 – on datasets PolyMath$_{en}$ and MMLU-ProX$_{en}$. Below, we break down the results by model family.

OpenAI Models: All three consistency-based methods substantially outperform verbalized confidence for all tested values of k, including the second newest of the models tested: GPT-5.6 Luna (released July 2026 (OpenAI, 2026b)).[2]

DeepSeek V4.1 Flash: Self-consistency achieves comparable performance to verbalized confidence at large k: mean Brier of 18.7% with 8 calls vs. 19.3% for one call, respectively. Paraphrase and Rosetta show parity with verbalized confidence at $k = 2$ and more meaningful gains at $k = 8$: mean Brier of 16.5% and 15.5%, respectively. DeepSeek V4.1 Flash is the newest of the tested models, just released in September 2026 (DeepSeek-AI, 2026).

Anthropic Models: Showing the starkest contrast, verbalized confidence using Sonnet 4.6 or Opus 4.6 consistently outperformed Self-Consistency and Paraphrase up to $k = 8$. On Sonnet 4.6, Rosetta (ours) begins to outperform verbalized confidence (21.6%) at $k = 4$ and most outperforms at $k = 8$ (18.2%). Opus 4.6's verbalized confidence outperformed all other models and methods tested (12.5%) – including Rosetta at $k = 8$ (13.1%).

We see similar patterns outside English and across the five English datasets (Table 1, detailed breakdown in Appendix H). For Sonnet 4.6, verbalized confidence has lower mean Brier than self-consistency on all five English datasets, while Rosetta outperforms verbalized confidence on three of five. For Haiku 4.5, verbalized confidence beats self-consistency on four of five datasets, while Rosetta wins on all five. Pooled over eight non-English languages and two datasets, one assessment also has lower Brier than self-consistency for Sonnet 4.6 (23.5% vs 33.4%), Haiku 4.5 (24.5% vs 30.9%), and Opus 4.6 (14.4% vs 25.3%), and for none of the other six models. Rosetta further lowers non-English Brier to 19.9% for Sonnet 4.6 and 20.6% for Haiku 4.5.

These combined results indicate a few salient trends.

Takeaway:

- Rosetta on frontier models outperforms self-consistency & paraphrase on all but one k.
- GPT models' and DeepSeek V4.1 Flash's verbalized confidence is less well-calibrated than the Claude 4.6 family of models, despite GPT-5.6 Luna and DeepSeek being newer models.
- Self-consistency and paraphrase are totally non-competitive on Claude 4.6 models. While Rosetta (ours) outperforms verbalized confidence on Sonnet 4.6, verbalized confidence provides the best uncertainty estimates with just a single LLM call.

[2]This evaluation excludes GPT-5.6 Sol since it was used as the anchor answer's correctness judge.

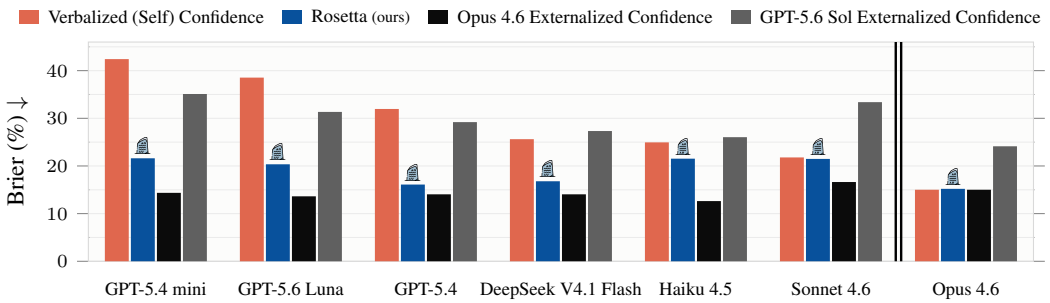


Figure 6: Mean Brier score (lower is better) across five settings: MMLU-ProX in English, Chinese, and Russian as well as English-only datasets SuperGPQA and SimpleQA. Verbalized (self) confidence denotes a model assessing confidence in its own predictions. Externalized confidence denotes a stronger model estimating confidence in a weaker model's answer. Opus 4.6 is right of the double line as a reference; its self and externalized assessments are identical.

6.2 **RQ5**: Can Strong Frontier Models Assist Weaker Models?

The previous section shows that Sonnet 4.6 and particularly Opus 4.6 can effectively estimate the confidence of their own predictions. Here, we explore frontier *externalized confidence* estimation where a stronger frontier model estimates the confidence of a weaker frontier model's prediction. Like verbalized (self) confidence where a model assesses itself, externalized confidence uses one LLM call.

Specifically, Figure 6 uses Opus 4.6 and GPT-5.6 Sol as externalized confidence quantifiers across six other frontier models and five settings: MMLU-ProX in English, Chinese and Russian as well as English-only datasets SuperGPQA and SimpleQA. Opus 4.6 estimates confidence in other models' predictions (mean Brier range of 12.6% to 16.6%) about as well as it assesses itself (15.0% mean Brier). By contrast, Sol poorly assesses confidence in other models' predictions, ranging from a mean Brier of 26.0% on Haiku 4.5 to 35.1% on GPT-5.4 mini. Sol's variance is not clearly attributable to model competence or family given its scores on Sonnet 4.6 (mean Brier 33.4%) versus Opus 4.6 (24.1%). Moreover, Sol's externalized confidence estimates consistently and substantially lagged these weaker frontier models' self estimates using Rosetta (ours). These findings compound our earlier results showing GPT models' limited effectiveness at verbalizing confidence.

Takeaway: Well-calibrated frontier models (e.g., Opus 4.6) can effectively verbalize confidence of their own and other models' predictions. When such calibrated frontier models are unavailable, Rosetta is the best performing alternative out of the box.

7 Conclusion

Our survey of recent work shows that consistency-based estimators are under-explored outside English and that evaluation increasingly focuses on non-frontier models. Rosetta (ours) leverages multilingual generation as part of uncertainty estimation to (1) improve English uncertainty estimation and (2) close the modest but consistent performance gap across languages. Our systematic evaluation of eight frontier models across three model families shows that when available, a well-calibrated frontier model like Opus 4.6 provides the best uncertainty estimates – both of its own predictions and the predictions of other models. When such frontier models are not viable (e.g., high cost, data privacy), Rosetta is a highly viable alternative.

AI Use Statement

Generative AI tools assisted with manuscript editing, figure preparation, and literature screening. Our confidence estimation methods are LLM-based, and we also used LLM-assistance for correctness grading and answer-novelty assessment, as described in the experimental setup. The literature audit

was LLM-assisted and partially human-reviewed. The authors take responsibility for the final content, including all claims, analysis, and artifacts produced with AI assistance.

REFERENCES

Anthropic. System card: Claude Haiku 4.5, 2025. URL https://www.anthropic.com/claude-haiku-4-5-system-card. System card.

Anthropic. System card: Claude Opus 4.6, 2026a. URL https://www.anthropic.com/claude-opus-4-6-system-card. System card.

Anthropic. System card: Claude Sonnet 4.6, 2026b. URL https://www.anthropic.com/claude-sonnet-4-6-system-card. System card.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, LLM judge, and ensemble scorers. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2504.19254.

Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, and Gaël Varoquaux. Reconfidencing LLMs from the grouping loss perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1567–1581, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.85. URL https://aclanthology.org/2024.findings-emnlp.85/.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=cSimKw5p6R. Featured Certification.

Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Evangelos E. Papalexakis, and Hua Wei. Uncertainty quantification of large language models through multiple uncertainty sources. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11015–11029, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.535. URL https://aclanthology.org/2026.findings-acl.535/.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. Uncertainty-aware decoding with minimum bayes risk. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hPpyUv1XyQ.

John Dang et al. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*, 2024. doi: 10.48550/arXiv.2412.04261. URL https://arxiv.org/abs/2412.04261.

Debtanu Datta, Mohan Kishore Chilukuri, Yash Kumar, Saptarshi Ghosh, and Muhammad Bilal Zafar. Do LLM hallucination detectors suffer from low-resource effect? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2959–2985, Rabat, Morocco, 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.136. URL https://aclanthology.org/2026.eacl-long.136/.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026. doi: 10.48550/arXiv.2609.19969. URL https://arxiv.org/abs/2609.19969.

Yongxin Deng, Zhen Fang, Sharon Li, and Ling Chen. Beyond in-domain detection: SpikeScore for cross-domain hallucination detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y16qXOaylp.

Yegor Denisov-Blanch, Joshua Kazdan, Jessica Chudnovsky, Rylan Schaeffer, Sheng Guan, Soji Adeshina, and Sanmi Koyejo. Truthfulness does not scale like reasoning: Why polling fails as a proxy verifier. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=g8bbEh2E2W.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in LLMs. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 12–21. ACM, 2025. doi: 10.1145/3767695.3769500. URL https://doi.org/10.1145/3767695.3769500.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a3c5af1f56fc73eef1ba0f442739f5ca-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using Semantic Entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4125–4150, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.239. URL https://aclanthology.org/2024.emnlp-main.239/.

Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishaloy Halder, Yang Li, Yassine Benajiba, and Dan Roth. Rethinking LLM uncertainty: A multi-agent approach to estimating black-box model uncertainty. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12349–12375, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.660. URL https://aclanthology.org/2025.findings-emnlp.660/.

Lorenzo Jaime Yu Flores, Ori Ernst, and Jackie CK Cheung. Improving the calibration of confidence scores in text generation using the output distribution's characteristics. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 172–182, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.15. URL https://aclanthology.org/2025.acl-short.15/.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2336–2346. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.143. URL https://aclanthology.org/2024.eacl-long.143/.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. doi: 10.48550/arXiv.2607.02770. URL https://arxiv.org/abs/2607.02770.

Granite Team, IBM. Granite 4.1, 2026. URL https://huggingface.co/ibm-granite/granite-4.1-8b. Model card.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lOoRJo8xWy.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, et al. Apertus: Democratizing open and compliant LLMs for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46877–46955. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2172. URL https://doi.org/10.18653/v1/2026.acl-long.2172.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Ammar Khairi, Daniel D'souza, Ye Shen, Julia Kreutzer, and Sara Hooker. When life gives you samples: The benefits of scaling up inference compute for multilingual LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27559–27583, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1402. URL https://aclanthology.org/2025.emnlp-main.1402/.

Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. Simple yet effective: An information-theoretic approach to multi-LLM uncertainty quantification. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30493–30504, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1551. URL https://aclanthology.org/2025.emnlp-main.1551/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2305.19187.

Zihang Liu, Jiawei Guo, Hao Zhang, Hongyang Chen, Jiajun Bu, and Haishuai Wang. Long-form hallucination detection with self-elicitation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4082–4100, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.211. URL https://aclanthology.org/2025.findings-acl.211/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Pedro Henrique Martins et al. EuroLLM-9B: Technical report. *arXiv preprint arXiv:2506.04079*, 2025. doi: 10.48550/arXiv.2506.04079. URL https://arxiv.org/abs/2506.04079.

Lucas H. McCabe, Rimon Melamed, Thomas Hartvigsen, and H. Howie Huang. Estimating semantic alphabet size for LLM uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uYK6GPVg1O.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4530–4540, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.234. URL https://aclanthology.org/2025.findings-acl.234/.

Alexander V. Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=j2wCrWmgMX.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 258–276, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.22/.

OpenAI. GPT-5.4 thinking system card, 2026a. URL https://deploymentsafety.openai.com/gpt-5-4-thinking. System card.

OpenAI. GPT-5.6 system card, 2026b. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf. System card.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10650–10666, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658/.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2695–2709, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.163. URL https://aclanthology.org/2023.emnlp-main.163/.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026a. URL https://qwen.ai/blog?id=qwen3.5. Model card.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b. Model card.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, et al. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.19799.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.539. URL https://aclanthology.org/2024.acl-long.539/.

Haoyi Song, Ruihan Ji, Naichen Shi, Fan Lai, and Raed Al Kontar. Inv-Entropy: A fully probabilistic framework for uncertainty quantification in language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KMbl8lg5Rv.

Ponhvoan Srey, Xiaobao Wu, Cong-Duy T Nguyen, and Anh Tuan Luu. Learning uncertainty from sequential internal dispersion in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40088–40106, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1862. URL https://aclanthology.org/2026.acl-long.1862/.

Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. CausalAbstain: Enhancing multilingual LLMs with causal reasoning for trustworthy abstention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14060–14076, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.723. URL https://aclanthology.org/2025.findings-acl.723/.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.238. URL https://aclanthology.org/2025.emnlp-main.238/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Tim Tomov, Dominik Fuchsgruber, and Stephan Günnemann. Task-awareness improves LLM generations and uncertainty. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Zuews4ULW6.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. CoCoA: A minimum Bayes risk framework bridging confidence and consistency for uncertainty quantification in LLMs. In *Advances in Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=H1NGlLNaVC.

Roman Vashurin, Maiya Goloburda, Preslav Nakov, and Maxim Panov. UNCERTAINTY-LINE: Length-invariant estimation of uncertainty for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7881–7908, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.400. URL https://aclanthology.org/2025.emnlp-main.400/.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. Unconditional truthfulness: Learning unconditional uncertainty of large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35673–35694, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1807. URL https://aclanthology.org/2025.emnlp-main.1807/.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Mrinmaya Sachan, Preslav Nakov, Timothy Baldwin, and Artem Shelmanov. Efficient hallucination detection for LLMs using uncertainty-aware attention heads. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=nddfPlfS09.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Lifeng Jin, Haitao Mi, Jinsong Su, and Dong Yu. Self-consistency boosts calibration for math reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6023–6029. Association for Computational

Linguistics, 2024a. doi: 10.18653/v1/2024.findings-emnlp.349. URL https://aclanthology.org/2024.findings-emnlp.349/.

Ante Wang, Weizhi Ma, and Yang Liu. Let the model distribute its doubt: Confidence estimation through verbalized probability distribution. *arXiv preprint arXiv:2511.14275*, 2025a. URL https://arxiv.org/abs/2511.14275.

Jiayi Wang and Xu-Yao Zhang. A systematic evaluation of black-box uncertainty estimation methods for large language models. *arXiv preprint arXiv:2606.19868*, 2026. URL https://arxiv.org/abs/2606.19868.

Tuo Wang, Adithya Kulkarni, Tyler Cody, Peter A. Beling, Yujun Yan, and Dawei Zhou. GENUINE: Graph enhanced multi-level uncertainty estimation for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20522–20541, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1119. URL https://aclanthology.org/2025.findings-emnlp.1119/.

Victor Wang and Elias Stengel-Eskin. Calibrating verbalized confidence with self-generated distractors. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pZs4hhemXc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Junyang Lin, Fei Huang, and Jingren Zhou. PolyMath: Evaluating mathematical reasoning in multilingual contexts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 47947–48003, 2025c. doi: 10.52202/085713-1431. URL https://arxiv.org/abs/2504.18428.

Yumeng Wang, Zhiyuan Fan, Qingyun Wang, Yi R. Fung, and Heng Ji. CALM: Unleashing the cross-lingual self-aligning ability of language model question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2809–2817, Albuquerque, New Mexico, 2025d. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.152. URL https://aclanthology.org/2025.findings-naacl.152/.

Yuxia Wang, Minghan Wang, and Preslav Nakov. Rethinking STS and NLI in large language models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 965–982, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.65. URL https://aclanthology.org/2024.findings-eacl.65/.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024. doi: 10.48550/arXiv.2411.04368. URL https://arxiv.org/abs/2411.04368. SimpleQA benchmark.

Yue Wu, Yangmin Huang, Qianyun Du, Lixian Lai, Zhiyang He, Jiaxue Hu, and Xiaodong Tao. Should I believe in what medical AI says? a Chinese benchmark for medication based on knowledge and reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1155–1164, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.91. URL https://aclanthology.org/2025.acl-short.91/.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova,

and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21381–21396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL https://aclanthology.org/2025.findings-acl.1101/.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Fan Xu, Huixuan Zhang, Zhenliang Zhang, Jiahao Wang, and Xiaojun Wan. JointCQ: Improving factual hallucination detection with joint claim and query generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1138–1159, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.58. URL https://aclanthology.org/2026.findings-acl.58/.

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, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.79. URL https://aclanthology.org/2025.emnlp-main.79/.

Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, Wenxuan Zhang, and Kam-Fai Wong. MlingConf: A comprehensive study of multilingual confidence estimation on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2535–2556, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.129. URL https://aclanthology.org/2025.findings-acl.129/.

Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 691–713, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.41. URL https://aclanthology.org/2025.findings-naacl.41/.

Bo Zhang, Cong Gao, Linkang Yang, Bingxu Han, Minghao Hu, Zhunchen Luo, Guotong Geng, Xiaoying Bai, Jun Zhang, Wen Yao, and Zhong Wang. SafeConf: A confidence-calibrated safety self-evaluation method for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3483–3495, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.186. URL https://aclanthology.org/2025.findings-emnlp.186/.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. SAC3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 2023. Association for Computational Linguistics. URL https://arxiv.org/abs/2311.01740.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what LLMs DO NOT know: A simple yet effective self-detection method. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7051–7063. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.390. URL https://aclanthology.org/2024.naacl-long.390/.

Ej Zhou, Caiqi Zhang, Tiancheng Hu, Chengzu Li, Nigel Collier, Ivan Vulić, and Anna Korhonen. Beyond the final layer: Intermediate representations for better multilingual calibration in large language models. *arXiv preprint arXiv:2510.03136*, 2025. URL https://arxiv.org/abs/2510.03136.

A LIMITATIONS

Our main evaluation covers nine relatively high-resource languages, and we provide experiments with low-resource panel languages on English questions in Appendix E.2 (Table 6). Reasoning and thinking modes were disabled due to high cost. The same model answers the question, translates it, and scores panelagreement; Appendices E.3 and E.4 replace its translations, its scoring, and the correctness grading with other sources. We match budgets by the number of model calls, which does not ensure equal token cost or latency. The survey is LLM-assisted, deatails in Appendix J.

B EVALUATION DETAILS

Detailed dataset coverage, model inventories, and metric definitions supplement Section 4.

B.1 DATASETS AND LANGUAGES

For English-anchor evaluation, we use PopQA (Mallen et al., 2023), the English subsets of PolyMath (Wang et al., 2025c) and MMLU-ProX (Xuan et al., 2025), SuperGPQA (Du et al., 2025), and SimpleQA (Wei et al., 2024). We use the MMLU-ProX-lite release as MMLU-ProX throughout. MMLU-ProX and SuperGPQA contain multiple-choice questions across academic disciplines, while PopQA and SimpleQA ask free-form questions about long-tail facts. Each model–dataset comparison uses 500 items drawn at random with a fixed seed. For these English experiments, Rosetta uses a fixed panel of eight languages: French, German, Dutch, Chinese, Japanese, Arabic, Hindi, and Vietnamese.

To compare confidence estimation across languages, we use the PolyMath and MMLU-ProX, which have the same questions across languages. The nine anchor languages are English (EN), Chinese (ZH), Arabic (AR), German (DE), Spanish (ES), Japanese (JA), Korean (KO), Russian (RU), and Vietnamese (VI). We select these shared languages to cover several scripts and language families. Each language takes a turn as the anchor language, with the other eight forming Rosetta's panel. For an anchor in language s, q_s is the dataset's own version of the question. The same 500 items appear in every anchor language, keeping question content fixed across the comparison.

In both settings, Rosetta generates its panel using the model's own translations, even when reference translations are available. This lets us evaluate the method without requiring reference translations at deployment.

B.2 MODELS AND SETTINGS

The nine headline models used in our experiments are GPT-5.4, GPT-5.4 mini (OpenAI, 2026a), Opus 4.6 (Anthropic, 2026a), Sonnet 4.6 (Anthropic, 2026b), Haiku 4.5 (Anthropic, 2025), DeepSeek V4.1 Flash (DeepSeek-AI, 2026), Gemma 4 31B and Gemma 4 12B (Gemma Team, 2026), and Qwen3.6-27B (Qwen Team, 2026b). They cover five closed API models from two vendors (OpenAI and Anthropic), DeepSeek V4.1 Flash with open weights called through an API, and three open models from two families (Gemma and Qwen) between 12B and 31B that we served ourselves using vLLM. The 17-model analysis in Section 5.1 adds GPT-5.6 Luna (OpenAI, 2026b), and seven open models from 2B to 9B: Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B (Qwen Team, 2026a), Granite 4.1 8B (Granite Team, IBM, 2026), EuroLLM 9B (Martins et al., 2025), Apertus 8B (Hernández-Cano et al., 2026), and Aya Expanse 8B (Dang et al., 2024). The last three are designed for multilingual use and are included to examine how multilingual models perform and benefit from Rosetta.

We accessed `gpt-5.4`, `gpt-5.4-mini`, and `gpt-5.6-luna` through the OpenAI API, and Haiku 4.5, Sonnet 4.6, and Opus 4.6 through the Anthropic API. We accessed `deepseek/deepseek-v4.1-flash` through OpenRouter, pinned to DeepSeek's own endpoint. We served the other open models with vLLM from their Hugging Face checkpoints.

Anchor answers use temperature 0, and panel answers use temperature 1 to obtain diverse samples. For the Claude models, the API supports a maximum temperature of 1. Reasoning and thinking modes are disabled for every model to keep the full-grid cost manageable. The scoring prompt is in English for every panel.

To obtain correctness labels, a separate judge, GPT-5.6 Sol (`gpt-5.6-sol`) (OpenAI, 2026b), grades each original answer against the reference answer once. All estimators share the original answer and its correctness label z_i for a given model and item. The comparisons are therefore paired item-for-item, with only the assigned confidence changing.

B.3 Metric Definitions

For each of the n anchor answers in one model–dataset–language combination, $c_i \in [0, 1]$ is the estimated confidence and z_i is the correctness label assigned by the grader, with 1 for correct and 0 for incorrect. We report three metrics: ECE, AUROC, and Brier score.

Expected calibration error (ECE) measures how closely confidence matches accuracy. We group answers into $M = 10$ equal-width confidence bins and average the absolute difference between confidence and accuracy, weighted by the number of answers in each bin.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{conf}(B_m) - \mathrm{acc}(B_m)\right|. \tag{2}$$

Here, $\mathrm{conf}(B_m)$ and $\mathrm{acc}(B_m)$ are the mean confidence and accuracy in bin B_m, and $|B_m|$ is its number of answers. Lower ECE is better.

The area under the receiver operating characteristic curve (AUROC) measures how often a correct anchor answer receives higher confidence than an incorrect one, counting ties as one half.[3] Higher AUROC indicates better ranking.

Brier measures the mean squared difference between confidence and correctness. It penalizes confident errors more heavily, and lower values are better. Brier reflects both the confidence estimates and the proportion of correct answers.

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} (c_i - z_i)^2. \tag{3}$$

Because accuracy varies across anchor languages, we use AUROC to compare how well confidence ranks correct above incorrect answers across languages. The formulas above give values on a 0–1 scale; tables and figures report these values multiplied by 100.

[3] See the scikit-learn documentation for `roc_auc_score`.

C MULTILINGUAL COMPETENCE BY MODEL

To understand which models benefit from multilingual sampling, we compare Rosetta's AUROC improvement over self-consistency with each model's accuracy on English and non-English questions. Table 3 reports English and non-English accuracy alongside Rosetta's non-English AUROC gain, averaged over PolyMath and MMLU-ProX. Our results show that across the 17 models, Rosetta's AUROC gain correlates with non-English accuracy (Pearson's $r = 0.86$). Multilingual competence also tracks the gains on English questions. All eleven models with non-English accuracy of at least 30% also improve in English AUROC under Rosetta relative to self-consistency. Below 30%, five of the six models lose more than two points of English AUROC, while their non-English AUROC changes by at most 1.8 points in either direction.

We next ask whether the gains outside English relate more closely to accuracy outside English or to how much English accuracy the model retains in other languages. Accuracy retention, defined as non-English accuracy divided by English accuracy, has a weaker correlation with the AUROC gain (Pearson's $r = 0.57$). For example, Aya Expanse 8B retains 96% of its English accuracy, close to the 97% of Opus 4.6, but has only 21% non-English accuracy and has lower AUROC under Rosetta than under self-consistency.

Table 3: Multilingual competence and Rosetta's gains for all 17 models, averaged over PolyMath and MMLU-ProX. English accuracy is measured on the English anchor⚓; non-English accuracy and AUROC are means over the eight non-English anchor⚓ languages. Rows are ordered by non-English accuracy; rules separate the descriptive accuracy groups used in Table 2. Gain is Rosetta minus self-consistency (SC) AUROC, so positive values favor Rosetta.

	Accuracy (%)		Non-English AUROC		
Model	English	Non-English	SC	Rosetta	Gain
EuroLLM 9B	18.6	16.0	59.8	60.1	+0.3
Apertus 8B	20.8	16.5	60.0	59.8	−0.2
Qwen3.5-2B	21.8	18.2	57.8	57.2	−0.6
Aya Expanse 8B	21.7	20.8	63.6	61.8	−1.8
Granite 4.1 8B	30.9	25.8	68.8	68.8	−0.1
Qwen3.5-4B	33.4	28.4	70.1	68.9	−1.2
Qwen3.5-9B	36.1	31.7	74.4	76.8	+2.4
Gemma 4 12B	36.4	33.1	70.1	75.9	+5.8
GPT-5.4 mini	44.1	40.4	78.1	81.1	+3.0
Qwen3.6-27B	46.0	43.2	77.8	80.4	+2.6
Gemma 4 31B	48.8	44.0	66.8	75.9	+9.1
Haiku 4.5	46.7	45.1	73.7	79.8	+6.1
GPT-5.6 Luna	47.7	45.2	77.0	82.3	+5.3
DeepSeek V4.1 Flash	55.8	50.4	80.7	84.7	+4.0
GPT-5.4	54.7	52.8	78.7	84.7	+6.0
Sonnet 4.6	58.3	54.1	67.3	80.3	+13.0
Opus 4.6	70.2	68.0	66.2	86.0	+19.8

D PER-LANGUAGE RESULTS

To see how the overall language gap varies across models and datasets, we examine the results for each language separately. Tables 4 and 5 report Brier, AUROC, and ECE for each model, estimator, and anchor language. The self-consistency language gap appears on both datasets. AUROC is lower outside English for all 17 models on MMLU-ProX and 10 of 17 on PolyMath, counting ties at the displayed precision. With Rosetta, mean non-English AUROC rises from 70.1 to 74.4, improving over self-consistency for 12 of 17 models. The language gap also shifts from -2.7 to $+0.4$ points, with fewer models performing worse outside English. Appendix C examines how these gains relate to multilingual competence.

Table 4: MMLU-ProX results by anchor language. Each cell gives Brier / AUROC / ECE, with ECE to the nearest point. All values are percentages. Lower Brier and lower ECE are better, higher AUROC is better. Bold values are best among the five estimators for that model and language, with ties at the displayed precision. Brier and ECE carry a base-rate term, so they compare estimators within a language rather than across languages. SC denotes self-consistency.

Method	EN	ZH	AR	DE	ES	JA	KO	RU	VI
GPT-5.4									
Verbalized	21.1/69.5/20	22.4/72.0/21	21.7/74.2/20	25.3/70.9/25	24.1/70.6/23	24.5/70.5/23	23.4/73.7/22	21.3/72.5/20	20.4/73.1/19
SC	15.9/76.5/15	18.4/74.2/16	16.9/75.4/16	18.1/76.2/17	19.4/73.7/17	19.5/71.1/17	19.8/70.5/18	18.0/72.4/16	17.1/73.7/16
Rosetta$_{AO}$	13.9/77.3/ **8**	15.0/81.3/10	15.0/79.1/10	17.3/74.1/12	16.5/78.7/12	18.7/73.1/15	16.8/78.3/12	15.9/77.6/11	15.6/74.4/**10**
Paraphrase	13.6/**80.4**/11	14.8/82.5/11	16.0/80.6/13	15.9/**81.5**/13	16.8/80.0/13	17.5/75.6/14	17.0/78.7/14	15.6/76.4/12	15.6/76.0/11
Rosetta	**12.9**/79.9/ **8**	**12.7/84.3/ 8**	**13.4/84.6/ 9**	**15.1/81.5/10**	**15.0/81.1/10**	**16.1/81.2/12**	**14.4/84.3/10**	**13.5/82.9/ 8**	**14.3/78.8/10**
GPT-5.4 mini									
Verbalized	30.5/68.2/30	35.1/69.6/35	34.1/66.6/34	34.9/64.2/34	36.4/69.6/36	37.6/72.0/38	40.8/68.0/41	34.8/64.4/34	35.3/67.0/35
SC	21.5/77.6/19	25.8/73.9/23	25.6/74.6/24	25.3/75.8/24	27.0/72.0/25	25.7/74.7/24	27.3/70.8/24	25.8/73.5/24	26.3/72.5/24
Rosetta$_{AO}$	20.7/75.8/15	22.5/76.9/19	23.3/73.8/19	22.4/75.4/18	23.7/74.8/20	22.5/79.2/20	24.5/75.5/21	23.4/**75.9**/21	25.1/71.4/21
Paraphrase	20.6/78.5/18	23.4/76.4/21	22.6/77.3/20	25.0/74.1/23	25.0/74.0/23	25.0/76.2/23	24.4/76.9/23	25.1/73.1/22	23.7/73.4/20
Rosetta	**19.4/79.3/14**	**20.2/80.1/15**	**21.3/78.8/17**	**20.7/78.9/16**	**20.8/80.2/18**	**21.5/80.9/19**	**21.4/79.0/17**	**22.8**/74.8/**18**	**22.8/75.4/18**
GPT-5.6 Luna									
Verbalized	25.2/65.9/25	32.5/61.5/32	30.7/62.3/30	26.7/64.8/26	28.5/67.2/29	28.0/64.9/27	27.6/66.4/27	26.6/66.7/25	25.4/66.6/24
SC	20.7/73.3/20	26.2/72.4/26	25.8/74.7/26	23.9/74.0/24	24.5/71.3/24	24.7/71.0/23	22.4/74.4/22	23.5/72.2/23	24.6/71.3/24
Rosetta$_{AO}$	17.9/77.8/14	21.7/78.5/19	20.5/79.9/19	20.7/76.2/18	19.6/79.2/17	21.1/75.8/18	20.6/75.3/16	19.2/79.4/16	20.5/77.3/16
Paraphrase	17.4/82.1/15	21.9/76.9/19	22.4/76.4/20	20.1/79.3/18	21.4/76.0/19	20.3/78.0/16	19.3/77.5/16	21.4/74.7/19	19.2/**80.0**/16
Rosetta	**15.2/83.7/10**	**19.3/79.6/16**	**17.8/83.8/15**	**18.4/82.0/16**	**18.1/81.6/14**	**18.5/79.4/15**	**18.0/80.2/13**	**17.9/80.1/13**	**18.1**/79.8/**13**
Opus 4.6									
Verbalized	10.9/85.0/ 8	**11.4/84.6/ 4**	11.5/84.6/ **4**	**10.0/87.2/ 4**	**11.1/84.5/ 3**	11.1/**83.2/ 4**	**12.3/81.9/ 5**	11.6/81.5/ **4**	**11.2/83.1/ 3**
SC	15.2/65.3/15	17.7/60.8/17	17.4/62.5/17	16.9/61.6/16	17.2/60.3/17	16.0/58.5/15	16.5/58.5/16	16.8/57.2/16	16.1/63.6/15
Rosetta$_{AO}$	12.5/74.1/ 9	14.9/77.1/12	13.6/76.4/11	13.5/76.5/11	14.2/75.6/12	13.8/76.9/11	13.7/74.1/11	14.5/70.2/12	13.3/73.4/10
Paraphrase	12.1/76.8/10	13.4/82.1/10	11.8/81.4/ 9	11.1/84.0/ 8	12.5/80.2/10	13.0/76.0/10	14.2/75.3/12	12.0/78.4/ 9	11.3/79.5/ 8
Rosetta	**10.3/85.1/ 6**	12.4/82.7/ 9	**9.9/86.2/** 5	10.4/84.1/ 7	11.5/84.4/ 8	**10.9**/82.3/ 6	12.4/80.4/ 8	**10.8/83.8/** 7	12.0/80.4/ 8
Sonnet 4.6									
Verbalized	**19.0**/72.6/ **8**	**20.4**/73.7/**11**	22.4/69.5/**14**	**21.1**/70.0/**10**	**21.0**/71.2/**11**	24.3/69.6/**18**	**20.4**/73.4/**11**	**21.9**/67.1/**12**	22.9/69.8/**14**
SC	27.8/60.5/28	30.6/60.8/30	32.9/55.5/32	30.9/58.6/30	32.5/57.8/32	35.4/57.3/35	30.3/60.3/30	30.2/54.4/29	34.3/56.4/34
Rosetta$_{AO}$	23.8/62.6/20	25.5/67.6/23	28.0/61.5/25	28.0/60.3/25	27.6/63.8/25	30.1/66.5/29	24.4/67.8/22	26.9/59.5/25	28.3/66.2/26
Paraphrase	21.8/70.7/20	25.4/68.2/23	27.0/61.7/24	24.3/70.6/22	27.8/62.4/26	29.4/64.0/27	22.6/**75.0**/21	25.5/66.0/23	27.7/70.1/27
Rosetta	19.7/**73.3**/14	21.0/**75.3**/18	**20.8/74.6**/15	22.0/**72.6**/17	22.9/**72.3**/18	**23.8/75.1**/19	20.6/74.2/15	22.2/**71.0**/17	**21.7/75.5**/18
Haiku 4.5									
Verbalized	22.1/72.2/15	20.2/**75.9**/12	23.8/72.0/17	22.1/73.2/15	21.2/75.6/14	21.0/73.5/13	19.1/77.5/11	20.8/76.4/13	19.5/77.0/11
SC	21.6/73.8/18	21.6/69.9/15	28.4/68.1/25	24.4/71.6/20	25.2/72.3/23	22.0/71.3/16	20.2/76.3/15	24.1/72.4/21	23.9/71.7/19
Rosetta$_{AO}$	19.5/76.0/13	22.3/68.7/14	22.9/75.1/17	21.6/74.8/15	20.4/79.6/15	22.1/70.1/14	19.0/75.6/10	20.6/76.1/15	21.8/72.5/14
Paraphrase	19.0/76.8/12	21.0/71.2/12	24.4/72.3/18	21.9/74.6/15	21.3/76.8/16	19.7/75.8/13	17.4/**80.8**/11	20.1/75.9/14	22.7/74.7/17
Rosetta	**18.6/77.3/ 9**	**19.7**/75.1/**11**	**18.1/81.5/ 9**	**19.3/79.0/12**	**17.8/82.6/10**	**17.7/78.5/ 7**	**17.0**/80.6/ **8**	**17.8/79.3/ 8**	**19.0/79.0/10**
DeepSeek V4.1 Flash									
Verbalized	18.0/69.8/16	22.3/72.2/20	22.8/67.2/20	21.2/72.9/19	21.7/72.9/20	22.7/71.0/21	21.8/71.4/19	20.2/74.2/17	24.1/68.2/23
SC	15.1/77.8/13	18.8/75.9/14	18.0/76.5/13	18.3/77.1/14	18.3/75.9/14	20.1/73.7/15	17.9/79.0/15	17.6/76.1/12	18.7/76.2/15
Rosetta$_{AO}$	**12.6**/83.1/ **5**	16.8/78.5/ 8	15.3/81.1/ 7	15.7/80.7/ 7	15.6/80.7/ 8	16.8/80.4/10	16.2/79.0/ 7	15.8/79.0/ 6	15.7/81.1/ 7
Paraphrase	13.3/82.3/ 8	18.2/76.4/12	16.7/78.7/10	16.1/81.8/10	15.6/82.0/10	18.0/77.9/12	16.4/79.5/11	15.9/79.5/ 9	16.4/79.1/12
Rosetta	13.0/**83.6/ 5**	**14.8/81.9/ 6**	**14.6/82.7/ 4**	**14.4/82.4/ 5**	**14.4/82.8/ 6**	**15.0/82.8/ 6**	**13.8/85.0/ 6**	**13.6/83.6/ 3**	**14.4/83.6/ 5**
Gemma 4 31B									
Verbalized	31.6/54.2/32	38.1/61.5/38	32.5/58.4/32	38.3/57.3/39	36.4/55.3/36	43.2/56.6/44	36.7/55.3/37	35.7/56.8/36	34.5/56.5/35
SC	30.1/60.7/31	37.7/57.2/38	29.8/61.9/30	36.1/59.0/36	33.8/61.8/35	42.8/57.2/43	34.3/60.9/35	33.3/60.1/33	31.8/57.9/32
Rosetta$_{AO}$	26.4/65.6/25	32.0/65.8/32	26.5/66.4/25	31.4/63.9/30	30.2/64.2/29	36.8/63.0/36	28.5/66.9/28	29.4/64.4/28	28.9/58.5/26
Paraphrase	**22.4/71.9**/20	29.7/68.3/28	23.0/**74.4**/21	28.8/**69.8**/27	28.5/**68.8**/27	33.7/65.8/32	29.1/65.7/28	26.3/71.6/26	26.3/66.8/23
Rosetta	22.9/69.4/**18**	**24.5/74.2/23**	**22.1**/73.3/**17**	**25.9**/69.5/**21**	**26.0**/68.1/**22**	**28.2/71.2/26**	**22.1/76.6/20**	**23.6/72.6/20**	**24.0/68.9/18**
Qwen3.6-27B									
Verbalized	27.5/65.3/27	30.7/65.5/30	31.0/63.5/31	28.3/65.8/28	28.9/66.0/28	34.1/60.5/34	28.1/61.7/27	28.9/66.6/29	29.3/64.2/29
SC	17.4/81.4/14	22.2/74.5/18	21.1/74.5/17	20.4/76.7/15	20.5/77.3/17	22.3/76.0/18	19.7/76.5/15	21.3/75.6/17	21.0/76.4/18
Rosetta$_{AO}$	16.5/80.9/10	19.3/78.5/14	20.3/**76.9**/14	18.4/77.9/10	18.4/78.7/11	20.4/78.0/15	**18.3/78.6**/12	19.2/78.4/13	19.1/78.3/13
Paraphrase	16.3/82.7/11	20.0/78.4/15	**19.5**/76.8/**12**	18.2/78.1/11	20.0/76.1/13	20.9/76.6/15	18.9/76.8/**11**	17.9/**80.6**/12	19.2/76.7/12
Rosetta	**15.4/82.9/ 6**	**18.5/79.6/10**	19.9/75.9/**12**	**18.1/78.2/ 8**	**16.3/83.0/ 7**	**19.5/79.0/13**	19.1/77.0/12	**17.5/80.6/ 9**	**17.4/79.7/ 9**

Continued on next page

Table 4 continued. MMLU-ProX. Brier / AUROC / ECE.

Method	EN	ZH	AR	DE	ES	JA	KO	RU	VI
Gemma 4 12B									
Verbalized	43.1/53.7/43	46.1/55.4/46	42.0/55.9/42	43.1/57.5/43	43.5/59.9/44	62.0/57.0/63	48.3/56.4/49	43.1/57.3/43	41.3/58.1/41
SC	34.2/66.2/34	38.0/63.8/38	33.1/65.4/32	35.6/62.4/35	36.9/63.1/37	52.8/55.9/55	39.7/65.3/41	34.9/65.4/34	35.1/63.0/34
Rosetta$_{AO}$	27.4/**71.9**/24	30.6/66.3/26	27.9/67.4/23	29.0/67.1/25	29.9/65.8/26	42.9/64.2/43	29.7/72.0/29	29.5/66.2/27	29.6/63.6/25
Paraphrase	27.4/71.3/25	31.3/68.3/30	28.4/70.7/25	30.2/66.2/26	30.6/66.4/27	42.8/60.5/44	34.9/64.1/33	26.7/72.7/24	28.7/67.0/24
Rosetta	**25.2**/71.6/**19**	**25.6/73.1/21**	**23.3/74.1/16**	**25.8/71.9/21**	**26.4/70.9/21**	**32.1/72.2/32**	**23.4/77.5/18**	**23.1/75.3/18**	**24.4/73.2/17**
Qwen3.5-9B									
Verbalized	38.0/57.3/38	42.9/56.2/43	46.3/59.7/47	44.6/58.1/45	43.6/55.8/44	47.2/57.1/47	45.3/57.1/45	42.7/60.4/43	43.9/59.4/44
SC	24.8/75.0/21	26.8/72.3/21	28.9/72.7/25	27.9/71.3/24	27.2/75.0/25	28.2/71.6/24	26.7/73.3/22	25.9/73.7/21	28.4/70.7/24
Rosetta$_{AO}$	**21.3**/77.5/15	**23.3**/74.6/16	23.9/75.4/18	23.4/75.2/18	23.8/76.0/19	24.1/75.5/19	23.0/77.0/17	23.0/75.9/17	**22.2/76.2**/15
Paraphrase	22.2/**78.0**/18	24.0/**76.0**/18	27.3/71.1/21	24.8/74.5/19	26.5/73.5/22	27.5/71.9/22	25.0/74.5/20	25.6/72.7/19	26.5/72.0/21
Rosetta	21.6/75.6/**12**	24.1/72.0/**15**	**22.7/76.8/17**	**22.2/75.5/14**	**21.8/76.7/14**	**22.1/77.4/15**	**21.4/78.6/15**	**20.1/78.8/13**	22.7/75.0/**14**
EuroLLM 9B									
Verbalized	52.6/54.6/54	60.6/49.9/62	64.4/50.4/66	65.4/49.0/67	61.5/50.2/63	67.3/42.8/69	66.5/46.9/68	62.6/50.0/64	73.0/47.8/74
SC	37.8/66.2/42	37.7/63.7/42	41.6/58.1/45	36.8/**68.5**/42	38.7/64.3/42	40.2/**65.0**/45	37.6/**68.4**/44	39.3/**67.4**/43	39.1/56.7/45
Rosetta$_{AO}$	**35.3/66.4/39**	**34.9/64.6/36**	40.4/58.4/43	37.0/68.2/42	38.7/64.2/42	41.3/61.2/45	37.6/67.4/43	39.0/63.5/42	37.9/58.3/45
Paraphrase	35.6/65.6/40	35.9/59.0/**36**	**38.4**/57.6/**41**	**35.5**/67.8/**40**	**34.6/67.1/38**	38.7/63.0/**42**	37.8/60.0/**41**	38.6/63.5/**41**	38.6/54.5/**44**
Rosetta	40.8/55.3/42	36.3/61.7/38	39.0/**59.2**/42	38.1/63.9/43	38.9/63.4/42	**37.4**/63.8/**42**	**37.0**/65.4/**41**	**37.9**/64.2/**41**	**37.7/59.1/44**
Granite 4.1 8B									
Verbalized	41.4/63.0/42	46.6/63.1/47	50.1/58.4/51	48.6/57.9/49	45.6/58.6/46	49.7/64.0/51	51.4/59.9/53	45.5/60.4/46	49.9/62.3/52
SC	37.3/**69.9**/39	41.6/65.8/43	42.2/63.2/44	40.6/67.9/42	39.5/67.9/41	42.3/63.5/44	41.6/65.5/43	40.9/64.9/42	46.0/60.0/48
Rosetta$_{AO}$	32.5/69.6/32	36.8/**67.8**/38	37.5/66.3/38	34.9/**69.8**/34	**34.4**/66.9/34	37.8/64.1/38	38.4/65.2/38	35.2/68.8/36	**36.6/68.1/37**
Paraphrase	35.6/69.2/36	38.1/65.1/39	38.0/66.0/39	39.8/63.9/41	35.6/**70.1**/37	37.9/66.0/40	38.0/63.6/40	38.6/65.0/39	39.7/63.8/41
Rosetta	**31.4**/67.5/**29**	**35.2**/64.0/**34**	**33.8/66.8/33**	**34.0**/65.2/**33**	34.6/63.9/**33**	**33.0/70.4/33**	**36.2/68.4/36**	**31.8/71.4/31**	37.5/63.8/38
Aya Expanse 8B									
Verbalized	62.5/50.9/63	66.0/49.3/66	70.3/50.9/71	63.2/53.3/64	63.4/49.7/64	68.1/51.0/68	66.1/52.0/66	61.8/50.7/62	63.1/53.7/64
SC	57.6/60.4/59	60.2/**60.2**/62	66.1/55.4/68	60.4/56.8/62	59.7/59.7/61	61.8/**61.0**/64	60.2/58.3/62	58.3/58.2/60	59.5/63.0/61
Rosetta$_{AO}$	52.4/**63.1**/55	**56.0**/59.2/**59**	59.8/**61.5**/63	53.9/58.1/56	54.3/**62.7**/57	**55.9**/59.6/**59**	**54.6**/60.0/**57**	53.1/60.2/55	53.8/59.9/**56**
Paraphrase	**50.6**/62.8/**53**	56.8/58.6/**59**	61.3/58.1/64	54.1/56.9/56	**53.7**/59.7/**56**	57.7/56.4/60	**54.6**/60.1/**57**	53.8/60.5/56	**53.3/64.6/56**
Rosetta	**50.6**/60.9/**53**	58.1/52.7/60	**59.6**/57.7/**62**	**52.8/58.3/55**	54.4/60.8/57	56.8/59.0/**59**	54.9/**61.5**/58	**51.1/61.2/53**	53.6/60.1/**56**
Apertus 8B									
Verbalized	50.2/54.1/52	57.1/55.3/62	54.6/50.0/58	53.4/58.1/57	50.4/55.4/53	58.4/50.6/62	67.3/50.1/68	49.0/57.7/52	60.1/52.8/63
SC	50.6/61.5/52	54.6/58.2/56	51.9/56.8/53	54.9/56.8/57	53.2/57.7/55	53.5/55.8/55	51.4/58.7/53	54.5/56.9/56	53.8/53.4/55
Rosetta$_{AO}$	43.9/57.9/**45**	**42.1/60.3/44**	**41.3/57.8/43**	43.7/61.2/45	**42.5**/61.5/**44**	43.9/59.7/46	**39.6/63.9/41**	**41.2/62.3/43**	**42.0/60.4/43**
Paraphrase	46.6/**61.9**/48	52.8/59.3/55	50.2/56.6/52	53.6/54.7/55	49.3/**62.6**/51	50.3/58.8/52	47.2/59.9/49	48.8/61.7/51	51.0/57.3/53
Rosetta	**43.7**/55.6/**45**	43.3/57.8/45	43.6/**57.8**/45	**41.3/62.5/43**	45.6/56.2/47	**42.7/62.1/44**	44.3/56.9/46	41.8/60.4/**43**	42.4/59.5/44
Qwen3.5-4B									
Verbalized	42.2/53.6/42	46.1/56.9/46	54.5/55.0/55	49.7/59.1/50	44.5/56.7/45	47.6/58.3/48	50.4/56.9/51	48.9/54.9/49	49.4/58.7/50
SC	26.8/73.6/25	30.0/70.2/28	32.1/**69.7**/31	32.1/68.6/30	29.7/**70.9**/27	30.6/67.9/28	29.7/**70.8**/28	28.7/**70.5**/26	33.5/68.2/33
Rosetta$_{AO}$	**22.1/74.9/15**	**25.1**/70.6/**18**	28.8/66.6/24	27.2/69.1/22	26.2/68.1/19	28.2/66.2/**22**	27.9/67.9/22	**25.9**/68.6/**18**	**25.6/71.7/20**
Paraphrase	27.6/71.2/26	28.3/**71.1**/26	31.4/68.2/29	29.5/**71.2**/28	30.4/68.0/28	29.3/**68.5**/26	30.4/67.6/27	28.2/69.4/24	29.9/69.5/28
Rosetta	24.4/71.3/17	29.1/64.3/23	**27.8**/68.2/**22**	**26.5**/70.3/**21**	**24.2**/70.7/**17**	**28.0**/67.0/**22**	**27.0**/68.6/**21**	27.1/67.6/21	28.1/66.5/22
Qwen3.5-2B									
Verbalized	55.6/54.0/57	45.1/56.2/46	64.6/49.5/66	54.8/54.6/57	53.0/56.8/55	59.0/49.4/62	54.5/54.7/58	58.8/52.5/61	55.3/51.0/56
SC	32.7/**69.3**/33	31.8/**70.1**/32	37.7/**64.1**/41	37.4/**62.0**/38	36.9/**64.5**/38	37.6/54.0/36	43.0/56.2/43	37.5/61.1/38	35.5/65.0/37
Rosetta$_{AO}$	32.6/58.3/25	31.6/57.8/23	36.7/52.9/30	36.3/54.7/28	34.3/56.5/**25**	32.1/58.6/**26**	**33.0**/55.7/**25**	31.3/59.3/**24**	**30.6**/62.4/24
Paraphrase	30.7/67.8/28	35.0/63.0/35	38.0/57.6/39	38.2/59.5/38	37.1/62.9/38	34.3/61.1/35	40.5/**60.1**/42	34.3/62.7/34	33.9/**66.3**/35
Rosetta	**29.6**/59.4/**20**	**29.3**/61.4/**20**	**32.9**/56.9/**28**	**31.0**/60.3/**26**	**30.9**/61.4/**25**	**29.7/61.5/26**	**33.0**/55.7/27	**28.3/65.5**/25	30.7/59.5/**23**

Table 5: PolyMath results by anchor⚓ language. Each cell gives Brier / AUROC / ECE, with ECE to the nearest point. All values are percentages. Lower Brier and lower ECE are better, higher AUROC is better. Bold values are best among the five estimators for that model and language, with ties at the displayed precision. Brier and ECE carry a base-rate term, so they compare estimators within a language rather than across languages. SC denotes self-consistency.

Method	EN	ZH	AR	DE	ES	JA	KO	RU	VI
GPT-5.4									
Verbalized	57.2/74.6/59	49.6/74.0/51	51.2/72.3/52	55.7/77.4/58	53.8/77.4/55	53.3/75.1/54	46.1/73.4/47	50.1/78.8/52	58.4/75.7/60
SC	28.7/81.9/34	26.8/85.0/33	28.8/83.6/34	30.9/82.1/36	28.9/85.6/36	27.0/82.8/32	25.6/83.5/30	28.0/84.7/34	26.9/84.7/32
Rosetta$_{AO}$	21.5/85.7/25	22.2/81.1/24	22.8/83.2/25	23.4/83.6/27	21.9/87.8/27	20.7/84.6/22	20.9/82.0/20	22.2/85.9/25	20.8/85.3/23
Paraphrase	22.0/**86.9**/26	19.4/**87.8**/23	22.4/84.2/25	25.3/82.4/28	22.2/89.6/29	18.4/**87.1**/21	19.9/83.4/19	21.4/87.4/26	21.9/85.9/25
Rosetta	**18.9**/86.1/**20**	**17.8**/85.7/**19**	**18.3/87.0/20**	**19.6/86.7/22**	**17.1/90.5/22**	**17.3**/86.8/**17**	**18.2/84.0/16**	**17.4/88.5/20**	**18.4/86.7/20**
GPT-5.4 mini									
Verbalized	49.5/75.0/51	43.8/72.9/47	39.0/74.6/40	47.8/74.6/50	43.2/77.7/45	52.1/65.7/53	48.7/72.6/51	54.4/67.4/56	55.1/78.1/58
SC	25.4/82.4/32	27.0/83.0/35	28.0/84.7/36	30.2/82.5/38	28.1/82.1/35	28.5/81.8/37	26.8/**83.8**/35	27.3/**80.7**/33	28.3/82.9/36
Rosetta$_{AO}$	22.4/82.1/27	24.0/80.6/29	22.9/85.0/29	24.2/80.7/29	24.0/82.3/29	23.4/83.6/30	23.8/81.7/29	24.0/80.1/28	23.7/83.4/30
Paraphrase	23.7/**82.9**/29	25.3/**83.1**/32	25.8/82.6/33	26.5/**84.0**/34	25.8/83.3/33	27.0/82.6/36	23.9/83.1/30	26.9/79.0/32	25.8/83.4/33
Rosetta	**20.4**/81.1/**23**	**22.8**/81.2/**28**	**21.0/85.2/26**	**21.6**/82.8/**27**	**20.8/83.7/25**	**20.2/86.2/27**	**19.8/83.8/23**	**22.4**/80.2/**26**	**20.4/86.3/26**
GPT-5.6 Luna									
Verbalized	56.4/63.5/56	47.4/71.7/49	48.1/65.7/49	46.3/72.2/48	47.0/67.3/47	46.1/70.2/48	38.7/74.2/39	54.9/70.7/57	46.1/67.1/47
SC	36.4/81.2/45	32.9/81.7/41	37.7/83.0/46	39.3/81.7/48	38.5/79.0/46	35.3/83.1/44	34.9/78.9/42	36.6/83.0/45	39.6/80.6/48
Rosetta$_{AO}$	27.6/81.6/33	26.5/80.4/32	27.6/83.6/35	28.2/82.9/35	27.7/80.2/35	24.0/84.9/30	27.3/77.6/31	27.8/**83.4**/34	28.4/80.3/34
Paraphrase	27.5/82.3/34	25.2/**84.0**/31	27.6/83.7/35	27.9/83.0/34	29.3/80.9/36	24.1/84.3/32	25.6/80.7/30	28.1/82.5/35	28.3/84.0/36
Rosetta	**23.2/83.2/29**	**24.6**/80.5/**28**	**23.8/86.1/31**	**22.9/85.8/28**	**23.9/83.2/29**	**20.4/86.4/26**	**22.3/81.9/26**	**24.9**/82.1/**31**	**23.7/84.3/29**
Opus 4.6									
Verbalized	**14.1**/87.5/ **3**	19.1/81.2/**11**	16.2/86.0/ **9**	17.8/85.0/**14**	16.3/84.9/ **6**	17.9/83.0/**10**	17.4/82.8/ **7**	17.8/84.1/**11**	18.4/84.1/**13**
SC	33.2/71.8/35	33.2/71.2/35	34.2/72.1/36	33.8/73.4/36	32.9/69.6/35	34.7/75.0/37	34.4/70.2/36	32.3/74.0/35	34.1/70.7/37
Rosetta$_{AO}$	18.1/84.5/18	20.7/82.0/20	19.9/84.0/20	20.6/85.2/22	20.4/82.8/21	20.8/83.7/22	20.3/83.9/20	18.6/87.0/20	19.4/83.6/20
Paraphrase	14.8/88.6/11	19.0/81.8/15	17.2/87.1/17	17.9/85.3/16	20.0/82.7/17	16.5/87.4/14	15.8/87.4/14	17.3/84.9/14	17.3/86.5/15
Rosetta	14.4/**88.7**/12	**14.5/87.9**/12	**13.8/90.5**/13	**15.7/89.2**/15	**16.0/88.5**/15	**15.0/88.4**/12	**13.8/89.6**/12	**15.5/87.2**/13	**14.1/89.8**/14
Sonnet 4.6									
Verbalized	24.2/78.4/21	25.8/76.6/23	24.1/75.5/19	26.0/75.2/22	23.3/78.9/21	26.6/75.9/24	24.4/78.1/22	24.8/78.9/23	26.3/79.5/26
SC	34.0/76.2/36	34.0/77.9/38	34.7/75.1/38	35.1/76.1/38	33.9/77.5/38	33.9/78.7/38	36.6/77.4/40	34.5/76.6/37	34.5/76.1/37
Rosetta$_{AO}$	23.6/80.1/23	23.5/81.1/24	25.0/78.9/25	25.4/80.5/27	23.4/83.0/25	25.0/80.2/26	26.7/78.2/27	24.2/80.3/24	24.5/81.7/25
Paraphrase	23.0/81.2/22	20.6/83.9/20	21.3/84.5/22	21.7/83.6/21	20.7/84.2/20	20.2/84.1/19	20.2/**86.2**/21	22.1/82.1/21	21.2/84.8/21
Rosetta	**17.5/86.1/16**	**17.7/87.2/17**	**18.4/86.3/17**	**18.2/87.1/18**	**18.3/86.7/18**	**16.1/88.6/15**	**18.5**/85.9/**16**	**18.1/86.5/17**	**18.6/85.6/18**
Haiku 4.5									
Verbalized	31.4/66.4/31	26.9/73.0/26	28.0/74.6/28	30.6/71.6/31	26.6/73.4/**25**	26.3/74.7/26	27.3/78.2/28	29.5/76.5/32	28.9/74.4/32
SC	40.6/74.9/47	37.2/75.5/44	39.7/73.1/46	41.8/75.0/49	41.4/73.6/48	33.5/76.7/39	29.9/**82.2**/36	39.2/75.6/45	41.3/74.2/48
Rosetta$_{AO}$	25.0/79.8/29	25.6/76.5/28	27.9/77.2/32	27.2/80.0/33	27.2/78.0/32	26.4/73.9/27	26.9/68.7/23	27.1/78.2/33	26.6/80.6/33
Paraphrase	28.6/74.3/31	25.7/**79.2**/30	28.4/76.0/31	29.4/77.8/34	29.3/77.1/32	25.7/**80.8**/30	25.2/80.9/27	30.0/75.3/34	30.5/76.9/36
Rosetta	**20.7/82.5/24**	**22.6**/78.2/**25**	**23.2/82.4/27**	**22.4/84.3/28**	**23.2/80.3**/27	**22.5**/78.7/**22**	**22.4**/75.5/**20**	**23.2/80.7/27**	**24.2/80.9/29**
DeepSeek V4.1 Flash									
Verbalized	20.6/76.2/20	28.0/69.6/28	23.1/64.2/23	24.9/77.8/26	25.2/79.9/26	29.3/73.1/30	27.2/76.1/26	28.2/78.7/29	41.9/73.8/45
SC	22.3/**85.2**/26	21.6/**83.8**/24	23.8/84.8/29	24.0/87.2/31	23.6/86.0/29	22.3/85.5/29	21.9/83.9/25	24.0/83.4/29	23.9/**86.5**/30
Rosetta$_{AO}$	18.5/82.3/16	18.3/82.8/17	17.2/85.3/18	17.6/86.5/21	19.8/82.5/20	18.1/84.9/21	18.3/82.2/17	17.1/85.1/17	17.1/**86.5**/18
Paraphrase	19.7/83.4/20	19.8/83.6/21	21.6/85.9/25	21.1/86.5/26	19.9/**87.6**/25	19.3/**86.9**/24	20.1/83.8/23	20.0/84.2/23	21.1/86.3/27
Rosetta	**15.9**/83.7/**12**	**16.6**/83.5/**15**	**15.7/87.6/16**	**15.5/87.3/17**	**15.9**/85.6/**16**	**15.1**/86.8/**18**	**15.3/86.5/13**	**15.4/87.7/16**	**16.6**/85.2/**17**
Gemma 4 31B									
Verbalized	63.2/52.2/63	66.3/49.6/66	64.6/51.6/65	65.7/52.1/66	60.6/55.5/61	67.9/50.2/68	66.6/49.8/66	65.6/51.3/66	68.0/49.9/68
SC	40.9/78.7/47	45.9/70.8/50	44.4/76.1/51	45.9/74.5/52	41.9/75.7/48	46.2/74.0/53	43.9/73.6/50	46.4/72.8/52	45.2/75.8/51
Rosetta$_{AO}$	32.6/78.7/37	35.5/74.0/40	36.3/77.3/42	38.2/76.6/44	35.6/77.6/40	38.5/74.5/42	36.4/74.1/40	37.1/74.3/41	37.5/77.5/43
Paraphrase	28.7/**79.6**/32	30.5/**79.0**/35	31.5/77.4/37	29.5/79.8/35	30.1/77.8/33	31.7/76.5/36	29.3/77.4/34	30.7/**79.7**/36	29.5/80.6/35
Rosetta	**26.1**/79.2/**28**	**27.3**/78.5/**30**	**28.1/80.7/32**	**28.1/80.6/33**	**27.3/80.3/30**	**27.1/80.2/31**	**26.9/79.3/28**	**28.7**/78.8/**32**	**27.7/81.5/33**
Qwen3.6-27B									
Verbalized	58.9/72.8/60	68.6/72.1/71	65.2/67.3/67	61.6/69.3/64	58.5/67.4/60	62.8/60.3/63	58.5/73.9/60	62.6/73.6/65	66.6/72.5/69
SC	26.6/79.9/31	24.8/80.2/31	23.9/75.2/26	23.9/81.3/29	26.0/79.4/31	22.9/78.4/27	20.6/**81.8**/24	23.3/81.2/28	24.3/80.6/29
Rosetta$_{AO}$	17.6/80.9/16	18.6/78.0/20	18.2/77.1/18	18.7/79.0/19	19.3/77.1/18	17.3/80.2/16	**16.1**/81.3/15	17.1/80.1/17	18.3/78.4/18
Paraphrase	20.2/80.8/23	18.8/**82.7**/23	19.6/76.1/22	20.1/79.5/23	21.8/77.9/25	19.1/78.5/20	18.5/79.6/19	18.7/80.3/19	20.2/78.7/23
Rosetta	**16.5/81.0/13**	**16.1**/81.6/**15**	**16.5/79.8/15**	**15.8/82.6/16**	**17.0/79.6/16**	**15.4/83.2/15**	**16.1**/81.1/**14**	**15.6/82.1/14**	**15.8/83.5/16**

Continued on next page

Table 5 continued. PolyMath. Brier / AUROC / ECE.

Method	EN	ZH	AR	DE	ES	JA	KO	RU	VI
Gemma 4 12B									
Verbalized	65.0/58.2/65	73.3/56.2/74	77.1/55.7/78	69.9/60.6/71	69.7/55.0/70	69.7/56.9/70	67.4/57.0/67	76.0/53.3/76	72.1/55.7/73
SC	39.7/**77.7**/47	42.0/76.9/51	42.2/74.6/51	40.3/**79.6**/50	42.4/76.5/51	40.8/79.5/50	41.5/77.9/51	40.6/76.7/49	39.9/**74.8**/47
$\text{Rosetta}_{\text{AO}}$	30.6/**77.7**/37	31.9/75.8/39	32.3/77.9/40	30.2/79.5/38	32.4/**76.9**/40	30.9/**79.9**/39	32.1/76.5/39	32.8/76.5/40	32.3/**74.8**/38
Paraphrase	27.9/75.0/33	28.6/77.7/36	27.3/75.8/34	26.0/74.3/32	29.2/72.0/35	28.5/77.6/35	26.7/78.7/35	27.4/73.9/35	28.2/73.1/34
Rosetta	**24.1**/77.1/**28**	**24.4/79.2/29**	**23.2/82.3/30**	**24.5**/75.8/**29**	**26.1**/75.2/**31**	**24.3**/78.7/**29**	**23.6/81.5/30**	**25.0/79.4/31**	**25.6**/74.2/**30**
Qwen3.5-9B									
Verbalized	75.6/52.3/76	80.8/53.0/81	75.2/56.7/75	78.5/54.4/79	77.7/53.1/78	66.7/58.8/67	69.9/55.5/70	74.2/57.3/75	80.6/52.8/81
SC	21.2/75.6/27	21.6/73.6/29	24.7/73.4/33	20.6/**78.2**/29	21.4/76.9/28	20.2/74.7/28	18.4/81.8/26	21.7/73.4/28	21.4/77.2/29
$\text{Rosetta}_{\text{AO}}$	15.1/76.6/20	15.5/**77.6**/22	15.7/76.5/24	16.0/76.5/22	16.4/**78.3**/23	15.8/76.0/24	14.1/**82.6**/22	16.6/72.4/21	16.0/79.7/24
Paraphrase	18.0/74.5/21	18.0/73.5/24	20.1/71.9/28	18.5/74.7/24	19.8/73.9/26	17.3/**77.0**/24	17.2/79.4/25	16.5/74.3/21	19.8/71.4/26
Rosetta	**14.3/77.1/18**	**14.4**/74.9/**20**	**13.4/79.1/20**	**14.8**/75.9/**21**	**15.8**/76.1/**21**	**14.0**/76.4/**21**	**13.7**/78.8/**20**	**15.7/76.3/19**	**14.5/79.9/21**
EuroLLM 9B									
Verbalized	40.7/59.9/**42**	92.5/49.1/94	85.8/47.8/89	85.8/45.5/90	78.2/55.5/84	91.3/47.0/94	90.7/38.1/93	87.2/43.3/91	90.6/50.6/91
SC	35.3/**62.9**/52	30.1/58.7/50	31.9/**58.1**/50	34.9/51.7/53	34.6/**58.1**/53	31.7/53.8/**51**	31.7/58.5/50	33.6/56.3/52	28.9/49.1/48
$\text{Rosetta}_{\text{AO}}$	34.1/53.2/50	34.2/59.8/53	37.1/57.6/54	34.4/61.4/53	35.4/47.3/53	34.5/56.7/53	35.9/60.1/54	36.4/54.9/54	31.8/**54.5**/51
Paraphrase	**33.4**/56.8/50	**29.4/63.7/49**	**30.1**/50.1/**48**	**31.5**/59.1/**51**	**32.7**/49.0/**51**	**31.6**/52.1/**51**	**29.7**/58.3/**48**	**31.6**/48.7/**51**	**27.6**/47.4/**47**
Rosetta	34.1/59.0/51	35.0/58.3/54	36.1/**58.1**/53	33.5/**67.4**/53	38.3/44.6/55	32.6/**58.5**/52	34.2/**65.8**/53	35.5/**61.4**/54	36.3/47.0/54
Granite 4.1 8B									
Verbalized	50.1/64.7/53	62.8/66.6/68	62.2/66.7/68	55.4/69.6/61	63.6/56.6/67	70.4/69.3/74	71.3/73.3/76	52.5/66.6/56	57.5/71.6/65
SC	43.8/**79.6**/54	42.9/74.1/54	42.0/**72.0**/53	43.7/**78.7**/56	46.7/70.9/58	36.8/67.6/47	38.2/**78.0**/50	44.8/68.6/55	44.2/**72.7**/56
$\text{Rosetta}_{\text{AO}}$	32.0/68.5/40	32.2/69.5/40	29.7/71.0/37	31.8/70.2/40	32.6/70.0/41	28.7/67.0/36	28.7/72.3/36	31.2/**73.7**/40	31.5/70.9/39
Paraphrase	33.9/74.7/43	33.8/70.3/44	35.8/63.9/45	35.7/69.3/46	35.9/**74.1**/46	29.7/64.7/37	30.6/61.8/39	34.7/71.1/46	34.9/68.0/44
Rosetta	**24.1**/70.0/**30**	**23.6/74.7/30**	**25.5**/69.7/**31**	**25.1**/71.3/**32**	**26.4**/68.4/**32**	**24.3/70.8/29**	**24.5**/71.6/**31**	**26.8**/69.4/**34**	**26.5**/70.4/**33**
Aya Expanse 8B									
Verbalized	91.9/51.5/92	88.0/55.9/89	90.8/54.0/92	89.3/50.9/90	90.1/52.3/91	90.8/50.2/91	90.2/50.9/90	92.7/47.7/93	88.9/55.1/90
SC	68.6/**72.8**/78	68.4/**71.8**/77	70.9/**69.5**/79	73.8/62.7/80	72.3/68.9/79	69.1/**72.4**/78	72.2/67.4/79	74.8/59.9/82	73.8/**72.3**/81
$\text{Rosetta}_{\text{AO}}$	61.8/61.6/72	**59.0**/67.1/**69**	62.4/68.6/72	60.2/66.6/70	61.4/**72.5**/72	**59.2**/66.5/**69**	**61.9**/59.8/**71**	62.6/66.5/73	**60.8**/70.0/**71**
Paraphrase	**57.1**/71.1/**69**	63.5/71.4/73	66.8/61.6/76	**54.7/73.5/67**	**56.6**/66.7/**67**	62.8/71.9/73	68.9/**73.7**/77	**60.8**/68.4/**72**	64.4/67.3/74
Rosetta	59.4/61.6/70	60.6/69.9/70	**58.1**/61.4/**69**	58.9/61.7/69	59.3/65.2/70	63.7/63.8/73	66.3/54.3/74	62.0/**72.5/72**	62.1/67.9/72
Apertus 8B									
Verbalized	80.2/48.5/85	77.8/48.5/86	84.8/41.4/89	84.0/44.9/87	66.4/54.2/72	84.3/49.3/88	94.1/52.1/94	69.6/46.3/78	89.0/47.8/91
SC	32.6/**70.1**/35	28.8/59.2/30	30.2/58.8/32	34.7/58.9/36	34.8/67.1/37	29.8/56.9/31	25.6/**68.6**/27	34.1/65.9/36	31.4/70.3/33
$\text{Rosetta}_{\text{AO}}$	23.5/60.3/25	**20.0/62.5/22**	**16.9/64.4/18**	24.7/60.5/27	21.2/56.6/23	**20.1/64.2/21**	24.2/55.5/26	**19.0**/63.9/**21**	**19.9/70.4/22**
Paraphrase	29.9/64.5/32	26.4/62.4/28	27.7/52.6/29	31.4/**62.1**/33	29.4/**68.5**/31	24.7/58.9/26	23.8/64.9/25	30.0/**71.0**/32	30.8/63.0/32
Rosetta	**21.8**/58.0/**23**	20.8/56.7/23	19.2/58.7/21	**20.9**/61.1/**22**	**19.6**/61.4/**21**	20.6/57.3/22	**19.6**/61.3/**21**	19.2/70.0/**21**	20.8/56.5/23
Qwen3.5-4B									
Verbalized	79.4/57.0/80	69.9/67.5/72	80.2/59.9/82	80.1/67.5/82	75.5/59.4/77	75.9/52.5/76	69.6/61.4/70	79.0/57.9/80	79.1/70.5/82
SC	29.3/71.2/31	28.4/71.1/30	25.7/68.6/29	29.3/71.3/33	31.3/66.3/34	22.6/**72.2**/27	23.8/**72.2**/26	24.7/72.8/26	29.2/70.4/33
$\text{Rosetta}_{\text{AO}}$	21.6/**73.7**/21	21.7/**71.3/19**	21.0/66.9/21	**21.2/73.9/23**	21.6/69.0/19	22.3/68.4/24	22.9/64.2/23	21.0/67.9/20	19.8/72.5/21
Paraphrase	26.7/69.9/26	25.2/71.0/27	20.3/71.6/21	27.8/64.4/29	29.1/67.5/30	21.4/68.6/25	21.9/67.0/22	25.0/68.4/25	26.1/71.1/28
Rosetta	**17.5**/72.4/**17**	**20.8**/65.1/20	**19.2/71.8/20**	22.1/68.2/**23**	**20.9/72.2/18**	**20.7**/66.8/**22**	**21.0**/65.9/**19**	**19.7/73.1/17**	**15.9/75.3/18**
Qwen3.5-2B									
Verbalized	87.0/60.9/90	71.0/56.1/77	92.9/55.6/94	85.9/**57.7**/90	86.2/48.8/88	88.3/55.9/91	79.7/39.4/84	90.0/54.1/92	81.3/59.1/86
SC	**33.9/63.8**/43	**27.7/56.7/36**	48.5/46.4/58	42.8/49.8/53	40.8/52.8/52	**27.6/60.0/37**	53.8/48.8/61	37.0/52.9/46	41.7/**59.7**/51
$\text{Rosetta}_{\text{AO}}$	36.0/54.7/**41**	35.3/52.8/40	**36.5**/43.0/**44**	42.2/46.6/50	39.7/**56.2**/45	34.7/47.4/39	37.5/45.0/42	36.2/49.1/41	37.5/47.2/**41**
Paraphrase	36.5/58.2/44	30.6/51.8/38	48.2/44.4/58	43.1/53.0/54	**37.6**/52.7/48	28.4/50.3/38	52.2/43.1/58	41.8/53.7/51	40.9/50.3/50
Rosetta	38.5/57.5/43	34.5/51.7/39	37.5/**62.9**/46	**32.4**/53.3/**40**	38.0/47.0/**44**	36.0/47.7/44	**34.5/55.7/41**	**33.6/59.0/39**	**37.3**/55.6/43

E Ablation Study

These experiments examine how panel size, panel languages, translation source, and scoring affect confidence estimation. We also check whether the language gap persists after matching accuracy and whether Rosetta helps on natively written questions. Together, these tests supplement the matched-budget comparisons in Section 5.2.

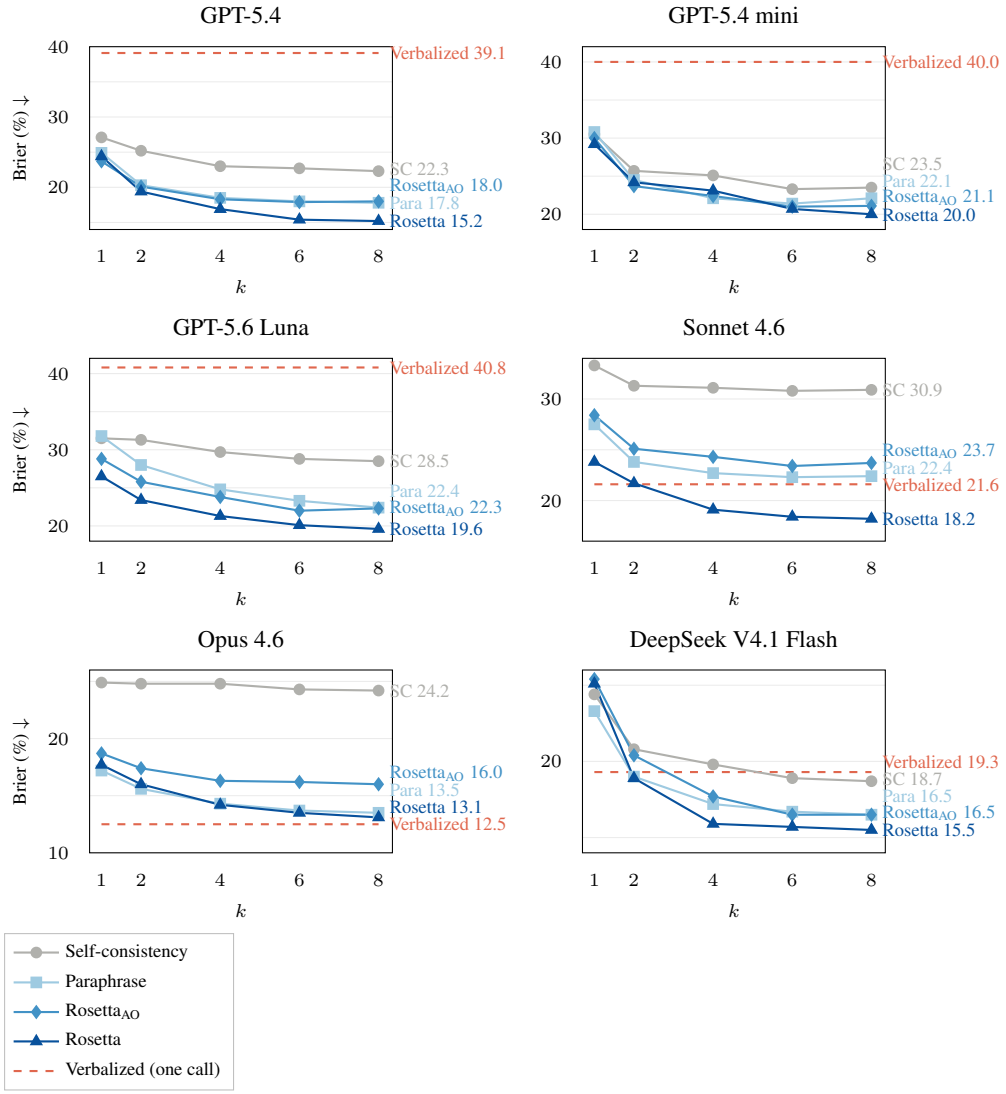


Figure 7: Brier score on a sweep of panel sizes on English MMLU-ProX and PolyMath for GPT-5.4, GPT-5.4 mini, GPT-5.6 Luna, DeepSeek V4.1 Flash, Sonnet 4.6, and Opus 4.6. Each point is the mean Brier over the two datasets and rescores subsets of the same eight-answer panels over 500 items per dataset; lower is better. The dashed line shows one verbalized-confidence call and does not depend on k.

E.1 Panel Size

To measure the benefit of additional answers, we vary panel size for Self-consistency, Paraphrase, Rosetta$_{AO}$, and Rosetta. We evaluate GPT-5.4, GPT-5.4 mini, GPT-5.6 Luna, DeepSeek V4.1 Flash, Sonnet 4.6, and Opus 4.6 on English MMLU-ProX and PolyMath, with 500 items per dataset. Each method rescores subsets of its existing eight-answer panel at $k \in \{1, 2, 4, 6, 8\}$, keeping the original answer fixed and generating no new answers. For Rosetta and Rosetta$_{AO}$, each subset uses the

Table 6: Rosetta with three different panel language sets on the five English datasets, 500 items per cell. Default: French, German, Dutch, Chinese, Japanese, Arabic, Hindi, Vietnamese. Alternative: Spanish, Chinese, Arabic, Indonesian, Japanese, Hindi, Vietnamese, Telugu. Low-resource: Swahili, Telugu, Urdu, Yoruba, Hindi, Vietnamese, Indonesian, Arabic. Rosetta$_{AO}$ uses the default languages only. Each method has two sub-columns, Brier (lower is better) and AUROC (higher is better). Best per row and metric in bold, second best underlined.

Model	Dataset	SC		Para		Rosetta$_{AO}$		Default		Alt.		Low-res.	
		Brier↓	AUROC↑	Brier↓	AUROC↑	Brier↓	AUROC↑	Brier↓	AUROC↑	Brier↓	AUROC↑	Brier↓	AUROC↑
GPT-5.4	PopQA	26.0	77.0	19.6	83.0	19.8	83.0	18.0	82.7	**16.8**	**84.0**	18.0	82.2
	PolyMath	28.7	81.9	22.0	86.9	22.2	84.8	17.6	87.2	**17.0**	**87.7**	18.6	85.9
	MMLU-ProX	15.9	76.5	13.6	80.4	13.8	78.4	**12.8**	**81.6**	12.8	80.1	13.3	80.9
	SuperGPQA	31.3	66.6	26.9	72.3	27.3	70.2	**23.1**	**76.2**	23.5	73.8	23.3	74.4
	SimpleQA	24.8	84.1	20.8	**85.2**	20.6	82.6	18.4	83.3	**18.1**	82.6	18.3	82.0
	Mean	25.3	77.2	20.6	81.5	20.7	79.8	18.0	**82.2**	**17.7**	81.6	18.3	81.1
GPT-5.6 Luna	PopQA	26.9	77.4	20.8	81.1	19.9	80.9	18.2	81.8	19.3	80.1	**17.7**	**82.4**
	PolyMath	36.4	81.2	27.5	82.3	27.0	82.6	23.1	82.9	22.3	**84.5**	**22.2**	84.3
	MMLU-ProX	20.7	73.3	17.4	**82.1**	17.6	80.0	**16.1**	82.0	16.9	79.4	16.5	80.0
	SuperGPQA	34.8	69.4	31.4	68.3	30.0	70.4	**26.6**	**72.6**	27.2	71.2	28.0	67.7
	SimpleQA	35.1	75.6	29.9	**76.9**	25.0	74.1	21.8	76.4	21.5	74.1	**20.9**	75.2
	Mean	30.8	75.4	25.4	78.1	23.9	77.6	21.1	**79.1**	21.4	77.8	**21.1**	77.9
Sonnet 4.6	PopQA	35.6	68.1	22.3	80.7	20.4	80.8	17.6	**82.6**	17.9	81.6	**17.5**	82.3
	PolyMath	34.0	76.2	23.0	81.2	24.8	80.1	17.0	86.5	16.1	**87.4**	**15.9**	86.7
	MMLU-ProX	27.8	60.5	21.8	70.7	22.6	67.0	19.4	73.9	**18.4**	**75.3**	18.6	71.8
	SuperGPQA	43.3	54.3	32.9	62.9	33.5	60.6	27.8	**67.3**	**26.8**	66.5	27.0	65.0
	SimpleQA	46.7	78.1	28.7	80.4	23.6	77.9	**17.0**	**83.5**	18.6	79.9	17.2	80.8
	Mean	37.5	67.4	25.8	75.2	25.0	73.3	19.8	**78.8**	19.6	78.1	**19.2**	77.3
Gemma 4 31B	PopQA	36.4	80.6	20.9	87.2	19.0	86.9	**14.7**	**87.9**	14.8	86.6	15.0	85.6
	PolyMath	40.9	78.7	28.7	79.6	32.9	78.9	26.8	78.7	25.1	**80.1**	**24.2**	78.8
	MMLU-ProX	30.1	60.7	22.4	71.9	25.4	67.0	22.7	70.5	22.3	70.9	**21.9**	**72.1**
	SuperGPQA	46.8	57.4	36.9	**64.1**	41.2	62.1	**34.9**	63.5	36.8	59.4	35.6	60.6
	SimpleQA	39.4	73.2	28.1	**73.4**	25.2	66.8	18.2	70.4	18.4	67.9	**17.3**	67.8
	Mean	38.7	70.1	27.4	**75.2**	28.8	72.3	23.4	74.2	23.5	73.0	**22.8**	73.0

first k languages in the fixed order French, German, Dutch, Chinese, Japanese, Arabic, Hindi, and Vietnamese. Each point in Figure 7 reports the mean Brier over the two datasets.

Larger panels lower Brier for both self-consistency and Rosetta, but Rosetta gains more from the additional answers for every tested model. Self-consistency improves least for Opus 4.6, whose confidence estimates change little as the panel grows. For all six models, Rosetta with four answers has lower Brier than self-consistency with eight answers.

We also compare these panels with one verbalized-confidence assessment. All four consistency-based methods have lower Brier at every tested panel size for GPT-5.4, GPT-5.4-mini, and GPT-5.6 Luna. However, the pattern is different for Opus and Sonnet. For DeepSeek V4.1 Flash, Paraphrase and Rosetta beat one assessment from $k = 2$, Rosetta$_{AO}$ from $k = 4$, and self-consistency from $k = 6$. For Sonnet 4.6, Rosetta has the lowest Brier of the five methods at $k = 8$. For Opus 4.6, one assessment has a lower Brier score than all four methods at every tested size. All models use temperature 1 for panel answers and temperature 0 for anchors (Appendix B.2). For Opus and Sonnet, temperature 1 is also the API maximum, so their panels cannot be made more diverse by raising the temperature.

E.2 Panel Languages

To test whether Rosetta's gains depend on the panel languages, we compare the default, alternative, and low-resource sets listed in Table 6 across four models and five English datasets. Across all three sets, Rosetta lowers Brier relative to Paraphrase in most model–dataset comparisons. Specifically, the low-resource set achieves the lowest mean Brier for Sonnet 4.6 and Gemma 4 31B. For ranking, however, the default set achieves higher mean AUROC than the low-resource set for every tested model. Thus, Rosetta's Brier gains hold across different language sets, while language choice affects how well it ranks answers.

E.3 Translation Source

To test whether the model's own translations limit Rosetta's performance, we replace them with the dataset's reference translations. Table 7 reports this comparison for three models on PolyMath, using

Table 7: Component ablation on PolyMath at the English anchor⚓ of the nine-language grid ($n = 500$, $k = 8$; panel: the other eight grid languages), Brier; lower is better. *Reference translations* replaces the model's own translations with the dataset's reference translations. Lowest per column in bold.

Arm	GPT-5.4	GPT-5.6 Luna	Gemma 4 31B
PolyMath			
Self-consistency	28.7	36.4	40.9
Translated self-consistency	27.7	34.3	40.0
Rosetta$_{AO}$	21.5	27.6	32.6
Paraphrase	22.0	27.5	28.7
Translate, answer in English	20.5	25.9	28.7
Rosetta	18.9	**23.2**	26.1
Reference translations	**17.3**	**23.2**	**25.6**

500 English original answers and panels in the other eight grid languages. *Translated self-consistency* translates the self-consistency panel answers into Rosetta's panel languages for the answering model to score with the same agreement prompt, testing whether multilingual text alone explains Rosetta's gain. *Translate, answer in English* translates the question into each panel language but answers in English, the anchor language, to separate question translation from answering in another language. Reference translations slightly lower Brier for GPT-5.4 and Gemma 4 31B and leave it unchanged for GPT-5.6 Luna. Most of Rosetta's Brier improvement over self-consistency therefore remains when the model generates its own translations.

E.4 Grader and Scorer Controls

We test whether the language gap depends on which model grades correctness or scores answer agreement. We replace each model separately, keeping the original answers and panels fixed.

Correctness grader. GPT-5.6 Sol labels every answer correct or wrong. If its labels were less reliable outside English, grading could create the language gap. To check this, Sonnet 5 regrades answers from GPT-5.4 and Gemma 4 31B on PolyMath and MMLU-ProX in all nine languages. The graders agree on 97.8% to 99.5% of labels, and accuracy changes by at most 0.5 points. With these new labels, self-consistency's language gap keeps its sign in all four model–dataset pairs and changes by at most 1.2 AUROC points.

Panel scorer. In our setting, each model scores agreement between its panel and anchor. To test whether the findings depend on this choice, we replace the scorer with GPT-5.6 Luna while keeping the answers fixed. We rescore self-consistency, Paraphrase, and Rosetta for four open models: Gemma 4 12B, Gemma 4 31B, Qwen3.5-9B, and Qwen3.6-27B. The comparison covers the nine-language evaluation and all five English datasets. With this shared scorer, self-consistency still performs worse outside English for all four models. Table 8 shows that Rosetta retains its mean AUROC gains over both baselines outside English and its mean Brier gains in English.

Table 8: Rosetta's gains with the answering model or GPT-5.6 Luna scoring the same panels for four open models. Positive values indicate lower Brier or higher AUROC than the baseline, in points. English results average five datasets; non-English results average two datasets and eight anchor languages. Metrics are computed separately for each model, dataset, and language before averaging.

	English Brier gain		Non-English AUROC gain	
Scorer	Over SC	Over Paraphrase	Over SC	Over Paraphrase
Answering model	9.6	3.1	5.0	3.0
GPT-5.6 Luna	10.8	5.0	4.6	2.6

E.5 Matching Accuracy Across Languages

Accuracy is lower outside English for almost every model. To check whether the language gap persists at comparable accuracy, we randomly remove correctly answered English items until English accuracy approximately matches the average over the eight other languages. We do this separately for each model and dataset, repeating the selection 50 times.

Table 9 shows that the AUROC gaps barely change after matching. Self-consistency still performs worse outside English for all 17 models, while Rosetta's average gap remains near zero. Paraphrase and verbalized confidence also retain their original patterns.

Table 9: The AUROC language gap with accuracy held equal. For each model, English items answered correctly are dropped at random until English accuracy matches the model's accuracy outside English, then the gap is recomputed (mean over 50 draws, 17 models, each dataset scored separately and then averaged, as in Figure 3). Each cell gives the mean gap, with the number of the 17 models below parity in parentheses. All values are percentages.

Estimator	All items	Accuracy matched
Self-consistency	−2.7 (17)	−2.8 (17)
Paraphrase	−2.2 (16)	−2.2 (16)
Rosetta	+0.4 (6)	+0.3 (6)
Verbalized confidence	−0.2 (8)	−0.2 (8)

E.6 Natively Written Questions

To test whether multilingual sampling also helps on natively written questions, we evaluate the five estimators on INCLUDE (Romanou et al., 2025) across eight languages and six models. Table 10 reports each model's results averaged over the eight languages.

Our results show that changing only the answer language with Rosetta$_{AO}$ improves Brier, AUROC, and ECE over self-consistency for all six models. Rosetta also improves AUROC and ECE over Paraphrase for five of the six models, although its Brier gains are less consistent. For Sonnet 4.6, verbalized confidence again achieves the best results on all three metrics. These results show that multilingual sampling also improves confidence estimates on questions originally written in other languages.

Table 10: INCLUDE-44, natively written exam questions in eight languages (zh, ar, de, es, ja, ko, ru, vi), for six models: each metric is averaged over the eight anchor languages, with about 500 items per cell and 139 for German. Rosetta translates the question only. Accuracy is that of the original answers. The small number after each Rosetta$_{AO}$ and Rosetta value is its change relative to the estimator with the same call budget, self-consistency and Paraphrase, respectively. Best per row and metric in bold, second best underlined. All values are percentages.

		Verbalized Confidence			Self-Consistency			Rosetta$_{AO}$			Paraphrase			Rosetta		
Model	Acc. (%)	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓
GPT-5.4	79.8	18.0	**75.3**	16.9	17.1	65.4	16.5	16.0 −1.0	68.1 +2.7	13.6 −2.9	16.1	69.2	15.0	**15.6** −0.5	70.9 +1.7	**12.6** −2.4
GPT-5.6 Luna	80.3	17.4	71.8	16.8	16.2	67.3	15.8	**14.9** −1.3	71.6 +4.3	12.4 −3.4	**14.9**	71.7	13.2	15.4 +0.5	**72.7** +1.0	**11.8** −1.3
Sonnet 4.6	81.4	**11.8**	**82.0**	**6.8**	16.3	63.3	15.6	13.6 −2.7	73.5 +10.2	10.7 −4.9	13.8	72.7	11.5	14.1 +0.3	74.0 +1.3	9.7 −1.8
Gemma 4 31B	73.8	24.6	56.3	24.6	24.2	57.1	24.3	21.0 −3.2	64.3 +7.2	19.8 −4.5	20.5	66.4	19.4	**19.4** −1.1	**67.8** +1.4	**15.6** −3.8
Gemma 4 12B	64.5	33.0	56.3	32.7	30.8	60.5	30.3	26.3 −4.5	67.2 +6.7	23.7 −6.6	26.5	66.8	24.0	**23.4** −3.1	**70.3** +3.5	**18.3** −5.7
Qwen3.6-27B	76.0	21.8	64.2	21.6	17.0	72.7	14.8	16.6 −0.4	73.9 +1.2	13.5 −1.3	**16.5**	**74.8**	**12.5**	18.5 +2.0	72.7 −2.0	14.7 +2.2

F Answer Novelty and Confident Errors

This appendix examines whether Rosetta produces different answers, reduces confident errors, and improves Brier overall, supporting Section 5.3. All methods assess the same anchor answer for each item, so its correctness stays fixed across methods.

F.1 Answer Novelty

To test whether translation produces answers absent from same-language sampling, we compare the panels from Rosetta, Paraphrase, and self-consistency. For each panel, GPT-5.6 Luna checks whether each of its eight answers appears in either of the other two panels, comparing content rather than language or phrasing. An answer counts as novel only when its content appears in neither panel. The judge uses the same prompt in one call per panel and does not see the question, reference answer, or anchor correctness. We then separate the results by anchor correctness to examine whether novel answers occur more often when the anchor is wrong. The table reports seven models, with 1,800 items per model across nine anchor languages and two datasets, PolyMath and MMLU-ProX.

Table 11 reports each method's novel-answer share on wrong and correct anchors and the difference between them. Opus 4.6 shows the largest margin over Paraphrase. Rosetta's novel-answer share rises from 4.1% on correct anchors to 20.8% on wrong anchors, a 16.7-point gap compared with 10.5 points for Paraphrase. All three methods produce more novel answers when the anchor is wrong. Rosetta's wrong–correct difference exceeds Paraphrase's for all seven models, by 0.8 to 6.2 percentage points. The 95% bootstrap intervals over items exclude zero for five models and include zero for Gemma 4 31B and Qwen3.6-27B.

To check whether the judge recognizes the same content across languages, we insert a translated copy of a panel answer into a comparison panel and test whether it finds the match. This check uses about 1,800 answers per model. The judge misses at most 1.5% of these matches for every model, reported as the planted-copy miss rate in the table.

F.2 Robustness to the Confidence Threshold

To check whether the reduction in confident errors depends on the chosen cutoff, we vary the confidence threshold from 50% to 95%. Figure 8 shows the share of wrong answers assigned confidence at or above each threshold, averaged over five models for each of the two English datasets, PolyMath and MMLU-ProX. Rosetta has the lowest share and self-consistency the highest among the consistency-based methods at every threshold on both datasets. Thus, the ordering holds across the tested thresholds rather than only at the 90% cutoff used in the main text.

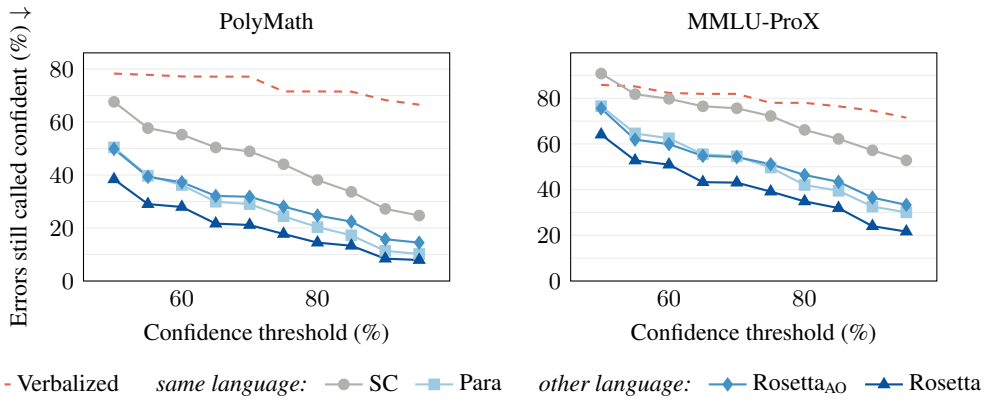


Figure 8: The confident-error ordering does not depend on the 90% threshold used in the main text. Each curve is the share of wrong answers scored at or above the threshold, macro-averaged over the five models; lower is better. Among the consistency-based methods, self-consistency is worst and Rosetta best at every threshold from 50% to 95% on both datasets.

Table 11: Novelty control. For each source, the share of its panel answers whose content the judge finds in neither of the other two panels, on wrong and on correct anchors, pooled over nine anchor languages and two datasets, 100 items per cell. Gap is wrong minus correct, in points. Excess is Rosetta's gap minus the same-language source's gap, with a 95% bootstrap interval over items. Planted miss is how often the judge failed to find a planted translated copy of a panel answer. Judge: GPT-5.6 Luna.

Model	Source	Wrong (%)	Correct (%)	Gap	Excess of Rosetta [95% interval]	Planted miss
GPT-5.4	Rosetta	30.7	7.1	23.6		1.5%
	Paraphrase	24.5	3.6	20.9	+2.8 [+1.0, +4.6]	
	SC	21.3	1.8	19.5	+4.1 [+2.3, +6.0]	
GPT-5.4 mini	Rosetta	34.9	9.9	25.0		1.4%
	Paraphrase	27.4	4.6	22.8	+2.2 [+0.3, +4.0]	
	SC	27.0	3.1	23.9	+1.0 [−0.8, +2.9]	
Opus 4.6	Rosetta	20.8	4.1	16.7		0.6%
	Paraphrase	16.0	5.5	10.5	+6.2 [+3.7, +8.6]	
	SC	9.2	0.7	8.4	+8.2 [+5.5, +10.8]	
Sonnet 4.6	Rosetta	22.9	7.4	15.5		0.8%
	Paraphrase	15.8	5.7	10.1	+5.5 [+3.2, +7.6]	
	SC	11.7	1.5	10.2	+5.4 [+3.2, +7.7]	
Haiku 4.5	Rosetta	28.3	9.0	19.2		0.5%
	Paraphrase	21.5	7.2	14.2	+5.0 [+3.0, +7.0]	
	SC	17.0	3.6	13.4	+5.8 [+3.8, +7.9]	
Gemma 4 31B	Rosetta	21.5	7.1	14.4		0.6%
	Paraphrase	17.6	4.6	13.0	+1.4 [−0.3, +3.3]	
	SC	11.9	0.9	10.9	+3.5 [+1.5, +5.4]	
Qwen3.6-27B	Rosetta	35.3	9.3	26.0		0.7%
	Paraphrase	34.7	9.5	25.2	+0.8 [−1.2, +2.9]	
	SC	30.2	5.6	24.6	+1.4 [−0.4, +3.3]	

F.3 Brier Contributions from Wrong and Correct Answers

Lower confidence helps when the anchor is wrong but can hurt when it is correct. We separate these two contributions to determine which accounts for Rosetta's overall Brier improvement. Table 12 compares Rosetta with self-consistency and Paraphrase on five models and two English datasets, with 500 items per model–dataset pair. The wrong-answer and correct-answer contributions are each weighted by their share of the items, so they sum to the overall Brier gain. Positive values indicate lower Brier under Rosetta.

Rosetta reduces Brier on wrong answers in every comparison, showing that it assigns less confidence to errors. It generally increases Brier on correct answers, however, so the overall gain depends on the balance between these changes. For GPT-5.4 on PolyMath, the improvement on wrong answers contributes 5.0 points over Paraphrase, while the loss on correct answers costs 0.6 points, leaving a net gain of 4.4 points.

The benefit on wrong answers outweighs the cost on correct answers in all ten comparisons with self-consistency and eight of ten with Paraphrase. The two exceptions are Gemma 4 31B and Qwen3.6-27B on MMLU-ProX, where the cost on correct answers is larger. Rosetta's overall Brier gains therefore come from reducing confident errors enough to offset the loss on correct answers.

Table 12: Brier improvement of translated-question evidence over each monolingual baseline, split by anchor⚓ correctness. Positive values favor translated-question evidence. The contributions of wrong and correct answers sum exactly to the net column. The wrong-answer contribution is positive in every cell against both baselines. Each cell gives the improvement over SC / over Paraphrase.

Model	Dataset	Wrong	Correct	Net
GPT-5.4	PolyMath	+11.7 / +5.0	−0.6 / −0.6	+11.1 / +4.4
GPT-5.4	MMLU-ProX	+5.5 / +2.2	−2.4 / −1.4	+3.1 / +0.8
Sonnet 4.6	PolyMath	+19.5 / +6.1	−2.4 / 0.0	+17.1 / +6.1
Sonnet 4.6	MMLU-ProX	+11.8 / +4.5	−3.5 / −2.1	+8.3 / +2.4
Gemma 4 31B	PolyMath	+16.6 / +2.5	−2.4 / −0.6	+14.2 / +1.9
Gemma 4 31B	MMLU-ProX	+11.2 / +2.4	−3.9 / −2.7	+7.3 / −0.3
Qwen3.6-27B	PolyMath	+13.8 / +6.4	−3.8 / −2.8	+10.1 / +3.7
Qwen3.6-27B	MMLU-ProX	+5.0 / +2.7	−4.6 / −3.5	+0.4 / −0.8
Gemma 4 12B	PolyMath	+16.9 / +3.1	−2.5 / −0.5	+14.4 / +2.6
Gemma 4 12B	MMLU-ProX	+12.5 / +3.2	−3.3 / −0.8	+9.2 / +2.4

G Additional Frontier Comparisons

G.1 External Confidence Assessment

This experiment compares confidence estimators on 500 items per answering model for MMLU-ProX, SuperGPQA, and SimpleQA; Figure 6 averages these three together with MMLU-ProX in Chinese and Russian. The original answers and labels remain fixed within each model's comparison. The answering model produces its self-assessment and Rosetta panel, while Opus 4.6 provides a separate direct assessment of the original answer. Table 13 reports the results used in Figure 6.

Table 13: Every method on each answering model, for the three English benchmarks included in Figure 6's average. The last two columns are outside assessors applying the verbalized-confidence prompt to the model's fixed answer; the others are run by the answering model on its own answer. Each cell gives Brier↓ / AUROC↑. All values are percentages. GPT-5.6 Sol also grades correctness everywhere in this paper, so its columns partly measure agreement with its own labels.

Model	Benchmark	Acc. (%)	Self-assessment	Self-consistency	Paraphrase	Rosetta	Opus 4.6 assessment	GPT-5.6 Sol assessment
GPT-5.4 mini	MMLU-ProX	64.4	30.5 / 68.2	21.5 / 77.6	20.6 / 78.5	20.1 / 76.7	10.2 / 92.3	22.7 / 70.4
	SuperGPQA	43.6	48.8 / 61.1	32.9 / 70.9	31.5 / 70.2	26.8 / 74.2	18.3 / 80.7	40.1 / 61.1
	SimpleQA	21.2	61.7 / 65.2	22.3 / 77.7	20.7 / 76.6	18.2 / 75.5	18.9 / 82.6	58.6 / 68.8
GPT-5.6 Luna	MMLU-ProX	70.8	25.2 / 65.9	20.7 / 73.3	17.4 / 82.1	16.1 / 82.0	8.9 / 93.6	19.5 / 74.8
	SuperGPQA	49.2	42.8 / 58.7	34.8 / 69.4	31.4 / 68.3	26.6 / 72.6	19.0 / 78.6	35.7 / 64.5
	SimpleQA	27.4	62.7 / 62.5	35.1 / 75.6	29.9 / 76.9	21.8 / 76.4	18.9 / 81.2	56.9 / 67.2
GPT-5.4	MMLU-ProX	78.0	21.1 / 69.5	15.9 / 76.5	13.6 / 80.4	12.8 / 81.6	10.3 / 88.2	16.3 / 70.5
	SuperGPQA	55.4	38.3 / 64.6	31.3 / 66.6	26.9 / 72.3	23.1 / 76.2	18.0 / 79.7	36.1 / 55.2
	SimpleQA	34.2	55.5 / 65.1	24.8 / 84.1	20.8 / 85.2	18.4 / 83.3	19.8 / 79.5	53.8 / 65.2
Haiku 4.5	MMLU-ProX	66.4	22.1 / 72.2	21.6 / 73.8	19.0 / 76.8	18.5 / 76.8	11.2 / 91.1	19.2 / 74.6
	SuperGPQA	45.0	28.9 / 61.7	34.8 / 64.3	27.8 / 68.6	26.0 / 68.7	18.1 / 80.5	32.0 / 67.7
	SimpleQA	11.2	33.1 / 57.5	47.4 / 56.6	35.6 / 61.2	25.7 / 58.7	12.5 / 91.1	42.4 / 77.7
Sonnet 4.6	MMLU-ProX	69.4	19.0 / 72.6	27.8 / 60.5	21.8 / 70.7	19.4 / 73.9	13.2 / 84.8	22.5 / 66.8
	SuperGPQA	51.4	23.5 / 67.5	43.3 / 54.4	32.9 / 63.0	27.8 / 67.4	19.0 / 78.3	35.0 / 59.1
	SimpleQA	24.4	26.6 / 72.9	46.7 / 78.1	28.7 / 80.4	17.0 / 83.5	20.1 / 86.3	58.2 / 66.5
Opus 4.6	MMLU-ProX	83.0	10.9 / 85.0	15.2 / 65.3	12.1 / 76.8	10.8 / 83.4	–	13.8 / 70.6
	SuperGPQA	60.6	19.7 / 74.6	35.7 / 59.1	26.2 / 69.9	23.3 / 72.3	–	29.8 / 62.2
	SimpleQA	42.8	21.1 / 75.7	42.5 / 71.8	22.4 / 86.2	18.7 / 84.8	–	47.4 / 64.7
DeepSeek V4.1 Flash	MMLU-ProX	78.6	18.0 / 69.8	15.1 / 77.8	13.3 / 82.3	13.6 / 81.7	10.1 / 89.2	15.5 / 69.8
	SuperGPQA	52.8	34.6 / 61.6	29.4 / 70.0	26.8 / 72.9	22.7 / 73.3	18.4 / 79.7	36.3 / 59.7
	SimpleQA	35.2	33.0 / 69.6	28.0 / 80.6	22.7 / 82.2	19.2 / 77.5	19.9 / 79.7	48.3 / 70.2

G.2 Every Assessor Against Every Answering Model

Table 14 extends the external assessment to a full grid: each assessor applies the verbalized-confidence prompt to every answering model's fixed answer on three benchmarks that span the difficulty range: MMLU-ProX, SuperGPQA, and SimpleQA.

Table 14: Every assessor against every answering model: Brier of one verbalized-confidence call on the other model's fixed answer, lower is better and darker is worse. Rows and columns are in the same order, so a starred diagonal cell is the model's own verbalized confidence on its answer, as in Table 13. GPT-5.6 Sol assesses but never answers, so it has no column. Under each benchmark: the answering model's Rosetta panel from Table 13, and the base rate, which is what always predicting that model's accuracy scores without reading the answer. Brier is comparable down a column but not across benchmarks, so the base-rate row marks how much of a cell is skill. Best per column in bold and second best underlined, over the assessors and Rosetta. GPT-5.6 Sol also grades correctness everywhere in this paper, so its row partly measures agreement with its own labels; Sonnet 4.6, Haiku 4.5 and Opus 4.6 rating one another are same-family pairs.

Assessor	Opus 4.6	Sonnet 4.6	Haiku 4.5	GPT-5.4	GPT-5.4 mini	GPT-5.6 Luna	DeepSeek V4.1 Flash
MMLU-ProX, 500 items, answering models at 64–83% accuracy							
Opus 4.6	10.9*	**13.2**	**11.2**	**10.3**	**10.2**	**8.9**	**10.1**
Sonnet 4.6	13.1	19.0*	14.9	13.8	14.9	13.3	14.6
GPT-5.6 Sol	13.8	22.5	19.2	16.3	22.7	19.5	15.5
Haiku 4.5	18.3	23.4	22.1*	18.2	21.1	19.2	19.6
GPT-5.4	16.5	26.0	24.3	21.1*	27.7	23.9	18.2
GPT-5.4 mini	16.8	27.1	24.1	21.0	30.5*	24.4	20.6
GPT-5.6 Luna	18.5	29.1	26.9	20.0	30.6	25.2*	21.8
DeepSeek V4.1 Flash	12.6	21.9	21.7	17.0	22.2	18.5	18.0*
Rosetta	**10.8**	19.4	18.5	12.8	20.1	16.1	13.6
Base rate	14.1	21.2	22.3	17.2	22.9	20.7	16.8
SuperGPQA, 500 items, answering models at 44–61% accuracy							
Opus 4.6	**19.7***	**19.0**	**18.1**	**18.0**	**18.3**	**19.0**	**18.4**
Sonnet 4.6	21.2	23.5*	23.2	22.1	21.7	21.8	23.0
GPT-5.6 Sol	29.8	35.0	32.0	36.1	40.1	35.7	36.3
Haiku 4.5	27.0	28.8	28.9*	29.0	27.0	25.8	28.5
GPT-5.4	28.6	34.9	33.5	38.3*	41.0	35.0	37.1
GPT-5.4 mini	30.3	35.0	40.1	36.9	48.8*	41.6	40.0
GPT-5.6 Luna	30.4	35.2	37.7	38.0	43.5	42.8*	40.6
DeepSeek V4.1 Flash	26.2	29.9	30.2	29.5	31.5	31.9	34.6*
Rosetta	23.3	27.8	26.0	23.1	26.8	26.6	22.7
Base rate	23.9	25.0	24.8	24.7	24.6	25.0	24.9
SimpleQA, 500 items, answering models at 11–43% accuracy							
Opus 4.6	21.1*	20.1	**12.5**	19.8	18.9	**18.9**	19.9
Sonnet 4.6	23.9	26.6*	16.9	22.8	23.8	23.1	23.4
GPT-5.6 Sol	47.4	58.2	42.4	53.8	58.6	56.9	48.3
Haiku 4.5	27.7	25.5	33.1*	24.9	23.6	24.0	22.7
GPT-5.4	45.5	56.9	59.1	55.5*	57.0	55.0	47.0
GPT-5.4 mini	46.4	57.5	51.8	53.6	61.7*	55.7	49.9
GPT-5.6 Luna	50.6	63.2	69.1	56.5	64.3	62.7*	55.4
DeepSeek V4.1 Flash	29.9	32.8	33.6	31.8	29.8	30.8	33.0*
Rosetta	**18.7**	**17.0**	25.7	**18.4**	**18.2**	21.8	**19.2**
Base rate	24.5	18.4	9.9	22.5	16.7	19.9	22.8

H PER-CELL ENGLISH RESULTS

Table 15 gives AUROC, Brier score, and ECE for every model, dataset, and estimator. Tables 16 and 17 give Brier and AUROC with percentile bootstrap 95% intervals over each cell's 500 items, using 1,000 resamples.

Table 15: English-anchor⚓ results for the 9 models on the five datasets. Each cell gives AUROC↑ / Brier↓ / ECE↓. Best per row and metric in bold, second best underlined.

Model	Dataset	Verbalized	SC	Rosetta$_{AO}$	Para	Rosetta
GPT-5.4	PopQA	77.1 / 34.5 / 33.0	77.0 / 26.0 / 25.9	**83.0** / 19.8 / 17.8	**83.0** / 19.6 / 17.0	82.7 / **18.0** / **11.3**
	PolyMath	74.6 / 57.2 / 58.8	81.9 / 28.7 / 33.9	84.8 / 22.2 / 24.6	86.9 / 22.0 / 26.5	**87.2** / **17.6** / **19.5**
	MMLU-ProX	69.5 / 21.1 / 20.3	76.5 / 15.9 / 14.9	78.4 / 13.8 / 10.0	80.4 / 13.6 / 11.0	**81.6** / **12.8** / **7.8**
	SuperGPQA	64.6 / 38.3 / 37.6	66.6 / 31.3 / 30.9	70.2 / 27.3 / 25.1	72.3 / 26.9 / 25.4	**76.2** / **23.1** / **18.9**
	SimpleQA	65.1 / 55.5 / 56.7	84.1 / 24.8 / 29.0	82.6 / 20.6 / 20.5	**85.2** / 20.8 / 23.6	83.3 / **18.4** / **16.3**
GPT-5.4 mini	PopQA	74.8 / 38.4 / 38.4	80.3 / 20.6 / 17.2	80.5 / 19.1 / 11.7	83.3 / 18.6 / 12.4	**84.6** / **16.3** / **9.6**
	PolyMath	75.0 / 49.5 / 50.6	82.4 / 25.4 / 31.6	81.2 / 21.6 / 24.7	**82.9** / 23.7 / 29.1	82.3 / **19.8** / **23.3**
	MMLU-ProX	68.2 / 30.5 / 29.8	77.6 / 21.5 / 18.7	76.2 / 20.5 / **15.0**	**78.5** / 20.6 / 17.7	76.7 / **20.1** / 15.4
	SuperGPQA	61.1 / 48.8 / 49.3	70.9 / 32.9 / 33.3	70.7 / 28.4 / 25.5	70.2 / 31.5 / 31.0	**74.2** / **26.8** / **24.3**
	SimpleQA	65.2 / 61.7 / 64.2	**77.7** / 22.3 / 24.1	72.8 / 19.6 / 16.9	76.6 / 20.7 / 22.4	75.5 / **18.2** / **14.8**
Opus 4.6	PopQA	**81.8** / 19.5 / 13.7	66.1 / 35.2 / 36.0	80.3 / 20.8 / 18.7	80.3 / 20.6 / 17.5	81.0 / **18.3** / **12.4**
	PolyMath	87.5 / **14.1** / **2.8**	71.8 / 33.2 / 34.9	85.4 / 19.5 / 19.9	**88.6** / 14.8 / 11.0	88.0 / 15.3 / 13.4
	MMLU-ProX	**85.0** / 10.9 / 8.0	65.3 / 15.2 / 15.0	75.1 / 12.5 / 9.8	76.8 / 12.1 / 9.7	83.4 / **10.8** / **7.1**
	SuperGPQA	**74.6** / **19.7** / **6.1**	59.1 / 35.7 / 35.5	64.5 / 29.0 / 27.0	69.9 / 26.2 / 22.5	72.3 / 23.3 / 17.8
	SimpleQA	75.7 / 21.1 / **12.8**	71.8 / 42.5 / 46.3	81.7 / 23.1 / 24.3	**86.2** / 22.4 / 25.8	84.8 / **18.7** / 17.7
Sonnet 4.6	PopQA	81.2 / 23.1 / 21.7	68.1 / 35.6 / 36.6	80.8 / 20.4 / 15.1	80.7 / 22.3 / 20.8	**82.6** / **17.6** / **8.5**
	PolyMath	78.4 / 24.2 / 20.6	76.2 / 34.0 / 35.9	80.1 / 24.8 / 24.9	81.2 / 23.0 / 22.3	**86.5** / **17.0** / **14.5**
	MMLU-ProX	72.6 / **19.0** / **8.4**	60.5 / 27.8 / 27.6	67.0 / 22.6 / 18.7	70.7 / 21.8 / 19.8	**73.9** / 19.4 / 13.5
	SuperGPQA	**67.5** / **23.5** / **9.3**	54.4 / 43.3 / 43.2	60.7 / 33.5 / 30.3	63.0 / 32.9 / 29.4	67.4 / 27.8 / 21.6
	SimpleQA	72.9 / 26.6 / 31.9	78.1 / 46.7 / 53.2	77.9 / 23.6 / 24.4	80.4 / 28.7 / 34.0	**83.5** / **17.0** / **16.3**
Haiku 4.5	PopQA	82.6 / 27.3 / 29.7	80.9 / 31.8 / 36.0	82.2 / 19.9 / 16.3	80.2 / 24.2 / 23.6	**82.9** / **17.3** / **9.5**
	PolyMath	66.4 / 31.4 / 31.4	74.9 / 40.6 / 47.0	80.8 / 24.1 / 28.2	74.3 / 28.6 / 30.8	**80.9** / **20.9** / **23.4**
	MMLU-ProX	72.2 / 22.1 / 15.3	73.8 / 21.6 / 17.5	76.1 / 19.1 / 10.1	**76.8** / 19.0 / 12.1	**76.8** / **18.5** / **7.8**
	SuperGPQA	61.7 / 28.9 / 21.1	64.3 / 34.8 / 33.1	65.4 / 29.1 / 22.6	68.6 / 27.8 / 23.5	**68.7** / **26.0** / **18.7**
	SimpleQA	57.5 / 33.1 / 41.8	56.6 / 47.4 / 56.5	54.2 / 30.9 / 36.2	**61.2** / 35.6 / 44.4	58.7 / **25.7** / **31.4**
DeepSeek V4.1 Flash	PopQA	71.7 / 26.4 / 23.1	77.3 / 24.8 / 25.1	80.2 / 18.7 / 11.3	80.6 / 20.0 / 16.0	**81.1** / **18.1** / **9.5**
	PolyMath	76.2 / 20.6 / 20.4	**85.2** / 22.3 / 25.7	82.0 / 19.5 / 17.5	83.4 / 19.7 / 19.8	82.0 / **17.4** / **14.1**
	MMLU-ProX	69.8 / 18.0 / 16.3	77.8 / 15.1 / 12.6	79.7 / 13.4 / **5.0**	**82.3** / **13.3** / 8.4	81.7 / 13.6 / 5.8
	SuperGPQA	61.6 / 34.6 / 32.3	70.0 / 29.4 / 27.6	**74.2** / 24.0 / 19.1	72.9 / 26.8 / 23.9	73.3 / **22.7** / **14.2**
	SimpleQA	69.6 / 33.0 / 30.6	80.6 / 28.0 / 31.3	79.0 / 20.6 / 15.1	**82.2** / 22.7 / 24.1	77.5 / **19.2** / **8.7**
Gemma 4 31B	PopQA	59.8 / 52.0 / 52.2	80.6 / 36.4 / 43.5	86.9 / 19.0 / 21.0	87.2 / 20.9 / 25.2	**87.9** / **14.7** / **12.3**
	PolyMath	52.2 / 63.2 / 63.2	78.7 / 40.9 / 47.2	78.9 / 32.9 / 37.8	**79.6** / 28.7 / 32.4	78.7 / **26.8** / **28.2**
	MMLU-ProX	54.2 / 31.6 / 31.6	60.7 / 30.1 / 30.6	67.0 / 25.4 / 23.8	**71.9** / **22.4** / 19.7	70.5 / 22.7 / **17.2**
	SuperGPQA	52.0 / 51.3 / 51.2	57.4 / 46.8 / 47.3	62.1 / 41.2 / 41.9	**64.1** / 36.9 / 35.8	63.5 / **34.9** / **32.1**
	SimpleQA	53.7 / 74.8 / 74.8	73.2 / 39.4 / 49.2	66.8 / 25.2 / 33.4	**73.4** / 28.1 / 37.8	70.4 / **18.2** / **25.2**
Gemma 4 12B	PopQA	63.3 / 49.9 / 50.0	86.4 / 30.1 / 38.5	90.5 / 16.4 / 20.3	**90.6** / 19.0 / 25.4	87.4 / **14.0** / **12.6**
	PolyMath	58.2 / 65.0 / 65.0	**77.7** / 39.7 / 47.4	77.6 / 30.2 / 35.9	75.0 / 27.9 / 33.4	75.1 / **25.3** / **30.3**
	MMLU-ProX	53.7 / 43.1 / 43.0	66.2 / 34.2 / 34.1	71.3 / 27.7 / 24.0	71.3 / 27.4 / 25.0	**72.2** / **25.0** / **18.5**
	SuperGPQA	50.0 / 59.2 / 59.3	63.5 / 42.6 / 44.5	64.2 / 38.5 / 38.6	62.3 / 37.4 / 36.4	**64.6** / **32.7** / **27.7**
	SimpleQA	51.7 / 64.5 / 64.8	77.7 / 28.4 / 38.3	76.0 / 17.0 / 22.9	74.4 / 21.6 / 30.9	**82.1** / **11.8** / **17.1**
Qwen3.6-27B	PopQA	62.6 / 52.7 / 53.2	86.7 / 17.2 / 16.4	84.4 / 16.0 / 13.5	**87.1** / **14.7** / **10.0**	83.0 / 16.2 / 12.1
	PolyMath	72.8 / 58.9 / 60.3	79.9 / 26.6 / 30.8	78.6 / 18.4 / 16.2	**80.8** / 20.2 / 22.8	78.5 / **16.5** / **13.4**
	MMLU-ProX	65.3 / 27.5 / 27.2	81.4 / 17.4 / 13.6	79.5 / 17.4 / 11.4	**82.7** / **16.3** / 11.0	80.2 / 17.0 / **8.2**
	SuperGPQA	62.6 / 41.8 / 41.3	71.5 / 29.1 / 27.4	**72.1** / 25.7 / 20.7	70.2 / 27.8 / 23.3	71.8 / **25.4** / **20.4**
	SimpleQA	59.0 / 71.8 / 73.5	80.4 / 21.0 / 25.4	76.0 / 13.5 / 10.8	**80.7** / 16.3 / 18.7	75.5 / **11.8** / **6.3**

Table 16: Brier score for every model, dataset, and estimator behind Table 15, with percentile bootstrap 95% intervals over the 500 items of each cell (lower is better). The last two columns give the paired difference on the same items, Rosetta minus Paraphrase and Rosetta minus self-consistency, each with its own paired percentile bootstrap 95% interval from 1,000 resamples. A difference is in bold when its interval excludes zero, which the per-estimator intervals to its left cannot show because they overlap.

Model	Dataset	Verbalized	SC	Rosetta$_{AO}$	Para	Rosetta	Δ Para	Δ SC
GPT-5.4	PopQA	34.5 [30.7, 38.4]	26.0 [23.0, 29.4]	19.8 [17.2, 22.8]	19.6 [17.1, 22.3]	18.0 [15.6, 20.5]	−1.7 [−3.4, +0.0]	**−8.1** [−10.9, −5.5]
	PolyMath	57.2 [53.2, 61.4]	28.7 [25.5, 32.0]	22.2 [19.6, 25.1]	22.0 [19.4, 24.7]	17.6 [15.4, 20.3]	**−4.4** [−6.1, −2.6]	**−11.1** [−13.2, −8.8]
	MMLU-ProX	21.1 [17.8, 24.6]	15.9 [13.2, 18.9]	13.8 [11.3, 16.4]	13.6 [11.1, 16.2]	12.8 [10.6, 15.2]	−0.8 [−2.0, +0.4]	**−3.1** [−4.9, −1.4]
	SuperGPQA	38.3 [34.3, 41.9]	31.3 [27.9, 34.7]	27.3 [24.3, 30.5]	26.9 [24.0, 30.1]	23.1 [20.6, 25.8]	**−3.9** [−5.6, −2.1]	**−8.2** [−10.5, −5.9]
	SimpleQA	55.5 [51.7, 59.4]	24.8 [22.1, 27.7]	20.6 [18.3, 23.2]	20.8 [18.5, 23.4]	18.4 [16.5, 20.6]	**−2.4** [−4.2, −0.6]	**−6.4** [−8.6, −4.2]
GPT-5.4 mini	PopQA	38.4 [34.7, 42.3]	20.6 [17.8, 23.4]	19.1 [16.6, 21.6]	18.6 [16.3, 21.0]	16.3 [14.0, 18.6]	**−2.3** [−4.0, −0.4]	**−4.2** [−6.5, −1.9]
	PolyMath	49.5 [45.6, 53.7]	25.4 [22.9, 28.0]	21.6 [19.1, 24.0]	23.7 [21.3, 26.0]	19.8 [17.6, 22.1]	**−3.9** [−5.7, −1.9]	**−5.6** [−7.8, −3.5]
	MMLU-ProX	30.5 [26.7, 34.0]	21.5 [18.7, 24.1]	20.5 [17.9, 23.2]	20.6 [18.0, 23.4]	20.1 [17.6, 22.7]	−0.5 [−2.7, +1.5]	−1.4 [−3.5, +0.8]
	SuperGPQA	48.8 [45.1, 52.6]	32.9 [29.9, 36.0]	28.4 [25.7, 31.2]	31.5 [28.6, 34.4]	26.8 [24.3, 29.5]	**−4.7** [−6.8, −2.7]	**−6.1** [−8.4, −3.8]
	SimpleQA	61.7 [57.9, 65.3]	22.3 [19.8, 24.9]	19.6 [17.4, 21.9]	20.7 [18.6, 22.9]	18.2 [16.0, 20.3]	**−2.5** [−4.4, −0.5]	**−4.2** [−6.4, −2.2]
Opus 4.6	PopQA	19.5 [17.5, 21.6]	35.2 [31.4, 38.9]	20.8 [18.0, 23.7]	20.6 [17.8, 23.5]	18.3 [15.7, 21.0]	**−2.3** [−4.1, −0.6]	**−16.8** [−19.9, −13.7]
	PolyMath	14.1 [12.4, 15.8]	33.2 [29.6, 37.3]	19.5 [16.7, 22.3]	14.8 [12.8, 17.2]	15.3 [12.8, 17.6]	+0.5 [−1.3, +2.4]	**−17.8** [−20.9, −14.8]
	MMLU-ProX	10.9 [9.5, 12.5]	15.2 [12.4, 18.4]	12.5 [10.1, 14.9]	12.1 [9.7, 14.7]	10.8 [8.7, 13.1]	−1.3 [−2.8, +0.1]	**−4.4** [−6.6, −2.5]
	SuperGPQA	19.7 [18.0, 21.5]	35.7 [31.5, 39.5]	29.0 [25.6, 32.3]	26.2 [22.9, 29.3]	23.3 [20.4, 26.1]	**−2.9** [−4.6, −1.1]	**−12.3** [−15.4, −9.5]
	SimpleQA	21.1 [19.5, 22.6]	42.5 [38.6, 46.4]	23.1 [20.3, 25.9]	22.4 [19.8, 25.0]	18.7 [16.3, 21.1]	**−3.7** [−5.5, −1.9]	**−23.8** [−27.3, −20.6]
Sonnet 4.6	PopQA	23.1 [20.7, 25.6]	35.6 [31.7, 39.2]	20.4 [17.8, 23.3]	22.3 [19.3, 25.2]	17.6 [15.3, 20.2]	**−4.8** [−7.0, −2.8]	**−18.1** [−21.2, −14.7]
	PolyMath	24.2 [21.6, 26.7]	34.0 [30.5, 37.5]	24.8 [22.0, 27.7]	23.0 [20.4, 25.6]	17.0 [14.8, 19.3]	**−6.1** [−8.1, −4.0]	**−17.1** [−20.2, −14.1]
	MMLU-ProX	19.0 [16.6, 21.3]	27.8 [24.1, 31.3]	22.6 [19.6, 25.8]	21.8 [18.7, 25.0]	19.4 [16.5, 22.1]	**−2.4** [−4.1, −0.6]	**−8.3** [−10.8, −5.9]
	SuperGPQA	23.5 [21.6, 25.3]	43.3 [39.3, 47.7]	33.5 [30.4, 37.0]	32.9 [29.5, 36.0]	27.8 [25.0, 30.6]	**−5.1** [−7.5, −2.9]	**−15.5** [−18.6, −12.5]
	SimpleQA	26.6 [24.8, 28.5]	46.7 [43.3, 50.3]	23.6 [21.0, 26.3]	28.7 [25.8, 31.6]	17.0 [14.9, 19.2]	**−11.7** [−14.0, −9.3]	**−29.7** [−33.2, −26.4]
Haiku 4.5	PopQA	27.3 [24.7, 30.0]	31.8 [28.9, 34.7]	19.9 [17.4, 22.3]	24.2 [21.6, 26.8]	17.3 [15.1, 19.3]	**−6.9** [−9.0, −4.6]	**−14.5** [−17.4, −11.6]
	PolyMath	31.4 [28.4, 34.5]	40.6 [37.7, 43.6]	24.1 [21.7, 26.5]	28.6 [26.0, 31.3]	20.9 [18.8, 23.1]	**−7.7** [−9.8, −5.5]	**−19.7** [−22.5, −16.9]
	MMLU-ProX	22.1 [19.4, 24.6]	21.6 [18.8, 24.3]	19.1 [16.6, 21.6]	19.0 [16.6, 21.6]	18.5 [16.0, 20.9]	−0.5 [−2.4, +1.3]	**−3.0** [−5.3, −0.7]
	SuperGPQA	28.9 [26.4, 31.5]	34.8 [31.4, 38.0]	29.1 [26.7, 31.7]	27.8 [25.3, 30.6]	26.0 [23.7, 28.5]	−1.8 [−3.8, +0.2]	**−8.8** [−11.2, −5.9]
	SimpleQA	33.1 [30.8, 35.6]	47.4 [44.5, 50.4]	30.9 [28.4, 33.9]	35.6 [33.0, 38.3]	25.7 [23.5, 28.2]	**−9.9** [−12.1, −7.7]	**−21.7** [−24.5, −18.8]
DeepSeek V4.1 Flash	PopQA	26.4 [23.0, 29.6]	24.8 [21.7, 27.9]	18.7 [16.1, 21.5]	20.0 [17.5, 22.7]	18.1 [15.9, 20.4]	−1.8 [−3.7, +0.3]	**−6.7** [−9.5, −3.9]
	PolyMath	20.6 [17.4, 24.0]	22.3 [19.9, 24.8]	19.5 [17.4, 21.8]	19.7 [17.5, 22.1]	17.4 [15.5, 19.5]	**−2.3** [−4.1, −0.2]	**−4.9** [−7.2, −2.5]
	MMLU-ProX	18.0 [14.8, 21.0]	15.1 [12.4, 17.7]	13.4 [11.2, 15.4]	13.3 [10.9, 15.4]	13.6 [11.5, 15.6]	+0.3 [−1.3, +1.9]	−1.6 [−3.5, +0.5]
	SuperGPQA	34.6 [31.0, 38.3]	29.4 [26.2, 32.6]	24.0 [21.5, 26.7]	26.8 [23.9, 29.5]	22.7 [20.3, 25.2]	**−4.1** [−5.9, −2.3]	**−6.7** [−9.0, −4.5]
	SimpleQA	33.0 [29.7, 36.2]	28.0 [25.3, 30.8]	20.6 [18.2, 22.9]	22.7 [20.4, 25.1]	19.2 [17.3, 21.3]	**−3.4** [−5.6, −1.3]	**−8.7** [−11.4, −6.0]
Gemma 4 31B	PopQA	52.0 [47.7, 56.4]	36.4 [33.1, 39.8]	19.0 [16.5, 21.6]	20.9 [18.4, 23.4]	14.7 [12.7, 17.0]	**−6.2** [−8.1, −4.1]	**−21.7** [−24.9, −18.5]
	PolyMath	63.2 [58.8, 67.4]	40.9 [37.5, 44.4]	32.9 [29.7, 36.2]	28.7 [25.5, 31.5]	26.8 [23.8, 30.0]	−1.9 [−3.8, +0.1]	**−14.1** [−16.9, −11.4]
	MMLU-ProX	31.6 [27.3, 35.9]	30.1 [26.1, 33.8]	25.4 [22.1, 29.0]	22.4 [19.2, 25.6]	22.7 [19.8, 25.8]	+0.3 [−1.6, +2.3]	**−7.3** [−10.1, −4.3]
	SuperGPQA	51.3 [47.1, 55.9]	46.8 [42.8, 51.1]	41.2 [37.6, 45.2]	36.9 [33.6, 40.5]	34.9 [31.5, 38.2]	−2.1 [−4.3, +0.1]	**−12.0** [−15.0, −8.9]
	SimpleQA	74.8 [71.0, 78.4]	39.4 [35.9, 42.7]	25.2 [22.3, 27.9]	28.1 [25.4, 30.9]	18.2 [15.9, 20.5]	**−9.9** [−12.1, −7.9]	**−21.2** [−24.1, −18.3]
Gemma 4 12B	PopQA	49.9 [45.5, 54.3]	30.1 [27.2, 33.3]	16.4 [14.0, 18.9]	19.0 [16.6, 21.6]	14.0 [12.1, 16.1]	**−5.0** [−7.1, −3.0]	**−16.1** [−18.8, −13.1]
	PolyMath	65.0 [60.6, 69.0]	39.7 [36.4, 43.0]	30.2 [27.0, 33.5]	27.9 [25.2, 30.7]	25.3 [22.7, 28.2]	**−2.6** [−5.0, −0.2]	**−14.4** [−17.1, −11.8]
	MMLU-ProX	43.1 [38.6, 47.3]	34.2 [30.6, 37.6]	27.7 [24.5, 30.8]	27.4 [24.2, 30.2]	25.0 [22.0, 27.7]	**−2.4** [−4.5, −0.3]	**−9.2** [−12.0, −6.4]
	SuperGPQA	59.2 [55.2, 63.2]	42.6 [39.1, 46.3]	38.5 [35.1, 42.0]	37.4 [33.9, 40.7]	32.7 [29.6, 35.7]	**−4.7** [−7.1, −2.0]	**−9.9** [−12.9, −6.8]
	SimpleQA	64.5 [60.4, 68.7]	28.4 [25.5, 31.5]	17.0 [14.5, 19.3]	21.6 [19.2, 24.1]	11.8 [10.1, 13.5]	**−9.8** [−11.7, −7.9]	**−16.6** [−19.1, −14.0]
Qwen3.6-27B	PopQA	52.7 [48.6, 57.0]	17.2 [14.7, 19.8]	16.0 [13.4, 18.6]	14.7 [12.4, 17.1]	16.2 [13.8, 18.9]	+1.5 [−0.8, +4.0]	−1.0 [−3.8, +1.7]
	PolyMath	58.9 [55.2, 62.9]	26.6 [23.8, 29.2]	18.4 [16.2, 20.6]	20.2 [18.0, 22.6]	16.5 [14.5, 18.8]	**−3.7** [−5.9, −1.5]	**−10.1** [−12.7, −7.3]
	MMLU-ProX	27.5 [23.5, 31.1]	17.4 [15.0, 20.1]	17.4 [15.2, 20.0]	16.3 [14.2, 18.8]	17.0 [14.9, 19.5]	+0.8 [−1.4, +2.9]	−0.4 [−2.7, +1.9]
	SuperGPQA	41.8 [37.5, 45.8]	29.1 [26.0, 31.8]	25.7 [23.0, 28.4]	27.8 [25.2, 30.4]	25.4 [22.6, 28.0]	**−2.4** [−4.7, −0.1]	**−3.6** [−6.2, −0.7]
	SimpleQA	71.8 [68.1, 75.5]	21.0 [18.4, 23.4]	13.5 [11.4, 15.6]	16.3 [14.3, 18.5]	11.8 [10.0, 13.8]	**−4.5** [−6.7, −2.4]	**−9.2** [−12.1, −6.2]

Table 17: AUROC for every model, dataset, and estimator behind Table 15, with percentile bootstrap 95% intervals over the 500 items of each cell (higher is better). The last two columns give the paired difference on the same items, Rosetta minus Paraphrase and Rosetta minus self-consistency, each with its own paired percentile bootstrap 95% interval from 1,000 resamples. A difference is in bold when its interval excludes zero, which the per-estimator intervals to its left cannot show because they overlap.

Model	Dataset	Verbalized	SC	Rosetta$_{AO}$	Para	Rosetta	Δ Para	Δ SC
GPT-5.4	PopQA	77.1 [73.4, 81.0]	77.0 [72.9, 80.8]	83.0 [79.3, 86.5]	83.0 [79.1, 86.6]	82.7 [79.0, 86.2]	−0.2 [−2.7, +2.3]	**+5.8** [+2.3, +9.3]
	PolyMath	74.6 [69.8, 79.4]	81.9 [77.7, 85.7]	84.8 [80.9, 88.1]	86.9 [83.1, 90.2]	87.2 [83.8, 90.2]	+0.3 [−2.4, +3.0]	**+5.3** [+2.1, +8.5]
	MMLU-ProX	69.5 [64.2, 74.7]	76.5 [71.3, 81.9]	78.4 [73.4, 83.2]	80.4 [75.0, 85.0]	81.6 [76.9, 86.5]	+1.2 [−2.7, +5.3]	+5.1 [−0.3, +10.4]
	SuperGPQA	64.6 [59.9, 69.3]	66.6 [61.7, 71.1]	70.2 [65.4, 74.7]	72.3 [67.4, 76.7]	76.2 [71.7, 80.1]	**+3.9** [+0.5, +7.7]	**+9.6** [+5.3, +13.8]
	SimpleQA	65.1 [60.2, 69.2]	84.1 [80.2, 87.2]	82.6 [78.3, 86.0]	85.2 [81.4, 88.5]	83.3 [79.3, 86.8]	−1.9 [−4.9, +0.9]	−0.8 [−4.0, +2.5]
GPT-5.4 mini	PopQA	74.8 [70.2, 79.0]	80.3 [76.2, 84.1]	80.5 [76.6, 84.2]	83.3 [80.0, 86.5]	84.6 [81.3, 88.0]	+1.3 [−1.0, +3.9]	**+4.2** [+1.4, +7.3]
	PolyMath	75.0 [69.9, 79.8]	82.4 [77.8, 86.8]	81.2 [76.3, 86.0]	82.9 [78.4, 87.3]	82.3 [77.5, 86.7]	−0.5 [−3.8, +2.8]	−0.1 [−3.7, +3.9]
	MMLU-ProX	68.2 [63.0, 72.8]	77.6 [73.2, 82.1]	76.2 [71.7, 80.6]	78.5 [74.0, 82.7]	76.7 [72.3, 81.2]	−1.8 [−5.8, +2.7]	−0.9 [−5.5, +3.7]
	SuperGPQA	61.1 [56.5, 65.5]	70.9 [66.1, 75.3]	70.7 [66.2, 75.1]	70.2 [65.6, 74.6]	74.2 [69.6, 78.4]	**+4.0** [+0.2, +7.9]	+3.4 [−0.5, +7.7]
	SimpleQA	65.2 [59.6, 70.9]	77.7 [72.7, 82.3]	72.8 [67.5, 78.2]	76.6 [71.7, 81.5]	75.5 [70.4, 80.2]	−1.1 [−5.3, +3.0]	−2.2 [−6.6, +2.4]
Opus 4.6	PopQA	81.8 [78.1, 85.3]	66.1 [61.5, 70.6]	80.3 [76.3, 84.6]	80.3 [76.1, 84.2]	81.0 [77.0, 85.0]	+0.7 [−2.1, +3.8]	**+14.9** [+9.9, +20.0]
	PolyMath	87.5 [84.6, 90.5]	71.8 [67.2, 76.3]	85.4 [81.5, 88.8]	88.6 [85.4, 91.5]	88.0 [84.8, 91.0]	−0.6 [−3.4, +2.2]	**+16.2** [+12.0, +20.8]
	MMLU-ProX	85.0 [80.6, 89.1]	65.3 [58.8, 71.9]	75.1 [69.0, 81.3]	76.8 [70.8, 82.4]	83.4 [78.5, 88.3]	**+6.6** [+1.8, +11.8]	**+18.1** [+10.9, +25.7]
	SuperGPQA	74.6 [70.0, 79.1]	59.1 [54.3, 64.0]	64.5 [59.5, 69.8]	69.9 [65.3, 74.6]	72.3 [67.6, 76.9]	+2.4 [−1.8, +6.4]	**+13.2** [+7.5, +18.9]
	SimpleQA	75.7 [71.2, 80.1]	71.8 [68.1, 75.4]	81.7 [78.1, 85.3]	86.2 [82.9, 89.2]	84.8 [81.4, 88.1]	−1.4 [−3.9, +1.4]	**+13.0** [+8.9, +17.0]
Sonnet 4.6	PopQA	81.2 [77.6, 84.9]	68.1 [63.8, 72.4]	80.8 [76.8, 84.6]	80.7 [76.9, 84.4]	82.6 [78.8, 86.0]	+1.9 [−0.8, +4.9]	**+14.5** [+10.6, +18.7]
	PolyMath	78.4 [74.4, 82.0]	76.2 [72.2, 80.3]	80.1 [76.2, 83.9]	81.2 [77.5, 85.0]	86.5 [83.2, 89.6]	**+5.3** [+2.4, +7.9]	**+10.2** [+6.6, +14.3]
	MMLU-ProX	72.6 [67.8, 77.4]	60.5 [54.9, 66.0]	67.0 [61.5, 71.8]	70.7 [65.9, 75.7]	73.9 [69.0, 78.6]	+3.1 [−1.3, +7.8]	**+13.3** [+7.4, +20.1]
	SuperGPQA	67.5 [63.2, 71.9]	54.4 [49.5, 59.2]	60.7 [55.6, 65.1]	63.0 [57.9, 67.6]	67.4 [62.8, 71.7]	**+4.4** [+0.4, +8.6]	**+13.0** [+7.7, +18.5]
	SimpleQA	72.9 [67.7, 78.3]	78.1 [73.7, 82.4]	77.9 [73.0, 82.6]	80.4 [75.8, 84.6]	83.5 [79.3, 87.0]	+3.2 [−0.7, +6.8]	**+5.5** [+1.1, +9.5]
Haiku 4.5	PopQA	82.6 [78.6, 86.3]	80.9 [77.0, 85.0]	82.2 [78.0, 85.9]	80.2 [75.9, 84.1]	82.9 [79.0, 86.6]	+2.7 [−0.6, +5.9]	+2.0 [−1.6, +5.6]
	PolyMath	66.4 [60.3, 72.5]	74.9 [70.0, 79.4]	80.8 [76.7, 85.0]	74.3 [69.6, 78.9]	80.9 [76.5, 85.2]	**+6.6** [+2.6, +10.7]	**+6.0** [+1.8, +10.5]
	MMLU-ProX	72.2 [67.9, 76.7]	73.8 [68.7, 78.5]	76.1 [71.5, 80.4]	76.8 [71.9, 81.4]	76.8 [72.2, 81.3]	+0.0 [−4.1, +3.7]	+3.0 [−1.6, +7.8]
	SuperGPQA	61.7 [57.1, 66.4]	64.3 [59.1, 68.9]	65.4 [60.3, 69.8]	68.6 [63.5, 72.9]	68.7 [63.8, 73.1]	+0.2 [−3.7, +4.1]	+4.4 [−0.6, +9.0]
	SimpleQA	57.5 [49.8, 64.7]	56.6 [48.8, 64.2]	54.2 [46.7, 60.9]	61.2 [54.0, 68.4]	58.7 [51.8, 65.2]	−2.5 [−9.1, +4.5]	+2.1 [−5.9, +9.9]
DeepSeek V4.1 Flash	PopQA	71.7 [67.2, 76.1]	77.3 [73.4, 81.2]	80.2 [75.8, 83.9]	80.6 [76.5, 84.4]	81.1 [77.3, 84.6]	+0.5 [−2.1, +3.2]	**+3.7** [+0.5, +7.1]
	PolyMath	76.2 [71.1, 81.2]	85.2 [81.2, 88.7]	82.0 [78.2, 85.7]	83.4 [79.1, 87.3]	82.0 [77.8, 85.9]	−1.4 [−4.5, +1.4]	−3.2 [−7.0, +0.3]
	MMLU-ProX	69.8 [64.1, 75.6]	77.8 [72.8, 82.4]	79.7 [74.9, 84.4]	82.3 [77.8, 86.5]	81.7 [76.9, 85.8]	−0.6 [−4.3, +3.4]	+3.9 [−0.5, +8.2]
	SuperGPQA	61.6 [56.7, 66.6]	70.0 [65.1, 74.5]	74.2 [69.7, 78.4]	72.9 [68.6, 77.2]	73.3 [68.7, 77.9]	+0.4 [−3.4, +4.4]	+3.4 [−1.0, +7.9]
	SimpleQA	69.6 [64.8, 74.5]	80.6 [76.5, 84.4]	79.0 [74.9, 83.0]	82.2 [78.2, 85.8]	77.5 [73.4, 81.5]	**−4.7** [−8.2, −1.2]	−3.0 [−6.8, +1.1]
Gemma 4 31B	PopQA	59.8 [56.0, 63.2]	80.6 [77.4, 83.5]	86.9 [83.7, 89.8]	87.2 [84.3, 90.1]	87.9 [84.7, 90.6]	+0.7 [−1.9, +3.4]	**+7.3** [+3.8, +10.7]
	PolyMath	52.2 [49.8, 55.1]	78.7 [75.5, 82.0]	78.9 [75.2, 82.7]	79.6 [75.7, 83.7]	78.7 [74.2, 82.9]	−0.9 [−4.7, +2.4]	−0.1 [−4.0, +3.5]
	MMLU-ProX	54.2 [51.6, 56.7]	60.7 [56.8, 64.6]	67.0 [62.5, 71.8]	71.9 [67.1, 76.3]	70.5 [65.4, 75.1]	−1.4 [−5.3, +2.5]	**+9.9** [+4.8, +14.9]
	SuperGPQA	52.0 [49.4, 54.5]	57.4 [53.7, 61.1]	62.1 [57.6, 66.7]	64.1 [59.3, 68.3]	63.5 [58.9, 68.3]	−0.6 [−4.5, +3.0]	**+6.0** [+1.3, +10.3]
	SimpleQA	53.7 [47.8, 58.8]	73.2 [65.6, 79.9]	66.8 [58.0, 75.7]	73.4 [65.6, 80.2]	70.4 [61.8, 78.7]	−3.0 [−9.1, +3.1]	−2.8 [−10.7, +4.5]
Gemma 4 12B	PopQA	63.3 [58.8, 67.1]	86.4 [83.1, 89.5]	90.5 [87.3, 93.2]	90.6 [87.5, 93.2]	87.4 [83.8, 90.5]	**−3.2** [−5.9, −0.6]	+0.9 [−2.5, +4.2]
	PolyMath	58.2 [55.4, 61.0]	77.7 [73.4, 81.8]	77.6 [73.0, 82.4]	75.0 [69.6, 80.1]	75.1 [69.4, 80.4]	+0.0 [−5.1, +5.0]	−2.6 [−7.1, +1.7]
	MMLU-ProX	53.7 [50.9, 56.4]	66.2 [62.3, 70.2]	71.3 [66.8, 75.7]	71.3 [66.8, 76.0]	72.2 [67.7, 76.6]	+0.8 [−2.8, +4.8]	**+5.9** [+1.5, +10.4]
	SuperGPQA	50.0 [47.1, 53.0]	63.5 [58.8, 68.3]	64.2 [59.4, 68.8]	62.3 [57.6, 67.3]	64.6 [59.7, 69.6]	+2.3 [−2.2, +6.4]	+1.1 [−4.2, +6.1]
	SimpleQA	51.7 [42.2, 60.2]	77.7 [71.0, 83.7]	76.0 [67.9, 83.0]	74.4 [66.1, 81.9]	82.1 [75.6, 87.9]	**+7.7** [+1.3, +14.8]	+4.4 [−1.9, +10.7]
Qwen3.6-27B	PopQA	62.6 [58.8, 66.8]	86.7 [83.0, 89.9]	84.4 [80.8, 88.1]	87.1 [83.9, 90.3]	83.0 [79.2, 86.6]	**−4.1** [−7.3, −1.5]	**−3.7** [−6.8, −0.9]
	PolyMath	72.8 [68.9, 76.6]	79.9 [75.3, 84.5]	78.6 [73.7, 83.3]	80.8 [76.1, 85.2]	78.5 [73.6, 83.2]	−2.3 [−6.4, +1.6]	−1.4 [−5.4, +2.5]
	MMLU-ProX	65.3 [60.6, 69.6]	81.4 [77.3, 85.1]	79.5 [75.1, 83.5]	82.7 [78.5, 86.1]	80.2 [76.0, 84.0]	−2.5 [−6.4, +1.5]	−1.1 [−5.3, +2.9]
	SuperGPQA	62.6 [58.7, 67.0]	71.5 [67.1, 75.9]	72.1 [67.7, 76.2]	70.2 [65.8, 74.2]	71.8 [67.4, 76.0]	+1.7 [−2.1, +5.4]	+0.3 [−3.8, +4.2]
	SimpleQA	59.0 [53.7, 64.4]	80.4 [75.6, 85.2]	76.0 [70.4, 81.4]	80.7 [76.0, 85.5]	75.5 [69.9, 80.9]	**−5.2** [−9.8, −0.4]	−4.8 [−10.2, +0.6]

I PROMPTS

The ten boxes below contain the prompts as sent, extracted from the code. They show the English form; for a question in another language, the answer and rewrite instructions drop the word "English", the translation instruction names the source language for every anchor⚓ language, English included, while the verbalized-confidence instruction is given in the question's own language. Placeholders in braces are filled per item with the question, language list, or numbered statements. Every scoring, confidence, translation, and rewrite call requests a JSON object with the fields named in the prompt. Consistency scoring and confidence calls each return one integer from 0 to 100, and answering, translation, and paraphrase calls return text fields.

Answering prompt, system message

```
Answer the supplied English question in the one requested language. Give one short,
complete sentence rather than only a bare word, name, number, or sentence fragment.
Do not translate or restate the question, discuss language, or add any text beyond
your answer.
```

Answering prompt, user message template

```
Question ({LANGNAME[q_lang]}):
{question}

Answer language: {LANGNAME[a_lang]}
Give one short, complete sentence, not only a bare word or phrase.
```

Verbalized confidence (Verbalized), system message

```
You are given a question and a candidate answer. State your confidence (0-100) that
the
candidate answer is correct. Answer directly with NO reasoning or explanation.
```

Verbalized confidence (Verbalized), user message

```
{item.prompt(main)}

Candidate answer: {anchor}

Confidence (0-100)?
```

Consistency scoring, system message

```
You measure AGREEMENT within a set of statements. Do not assess factual correctness.

The statements were produced independently under varied conditions. Statement [0] is
the
target; statements [1] onward are the comparison set.

Report 0-100 for how consistently the comparison set expresses the same core claim as
statement [0] (the same entity, quantity, or proposition). Treat different languages,
scripts, transliterations, spellings, name orders, phrasings, and levels of detail as
the
same claim.

Scale: every comparison states [0]'s claim -> 90-100. Most do -> 70-89. A bare
majority
does -> 50-69. The set contains several different claims -> 20-49. [0] stands alone or
nearly every comparison contradicts it -> 0-19.

Measure agreement only. A set that unanimously states the same false claim is
perfectly
consistent and must score near 100. A set containing many different claims is
inconsistent
```

```
and must score near 0 even if [0] is factually true. Do not use outside knowledge or
infer
missing context.

If the comparison set is empty, report exactly 50.
```

Consistency scoring, user message

```
Statements:
{listing}

Agreement (0-100) of statement [0] with the comparison set?
```

Translation of the anchor⚓ question into the panel languages (every anchor⚓ language, English included)

```
You are a professional translator. Translate the supplied {LANGNAME[source]} question
into each of the following languages: {target_list}. Do not answer or solve the
question. Preserve all option letters and their order, numbers, units, mathematical
notation, LaTeX, chemical formulas, and proper nouns in every translation. Translate
each language faithfully and completely, independently of the others. Output one
field per target language, keyed by its language code (shown in parentheses above),
containing only that language's translation.
```

Paraphrase, batched rewrite

```
You are an expert technical editor. Rewrite the supplied English question in English,
{k} separate times, preserving the exact factual meaning, all numbers, units,
notation,
and proper nouns in every rewrite. Make the {k} rewrites diverse from one another —
vary wording, sentence structure, and register across them. Do not add, remove,
reorder, or solve anything in any rewrite. Output one field per rewrite, keyed
{key_list},
each containing only that rewrite.
```

Correctness judge, system message

```
You are grading whether an answer to a question is factually correct, given a set of
reference correct answers.

Reason in this order, then give the verdict:
1. Identify the DIRECT answer the candidate gives — its core claim. Ignore wording,
length, tone, and hedging.
2. Correct if that core claim agrees in meaning with at least one correct reference.
Extra accurate detail, added caveats, and expressed uncertainty do NOT make it
incorrect.
3. Incorrect only if the candidate ASSERTS something false. Naming a misconception in
order to reject it is NOT asserting it.
4. Incorrect if it is evasive or non-responsive.
```

Correctness judge, user message

```
Question: {item.question}

Reference correct answers:
- {ref 1}
- {ref 2}
...

Candidate answer: {answer}

Is the candidate correct?
```

Original answers, rewrites, translations, confidence, and scoring calls use temperature 0, and panel answers use temperature 1. The answer budget is 256 tokens. The scoring budget is 1,200 tokens. The correctness judge is GPT-5.6 Sol for every model.

J SURVEY OF EVALUATION SETTINGS

We surveyed recent papers to identify which languages and models are used to evaluate consistency-based confidence estimators. This appendix describes how we selected the papers, classified their evaluations, and verified the findings.

J.1 CORPUS

We searched 63,659 papers published in 2023–2026 across ACL, NAACL, EACL, and EMNLP, including Findings, as well as ICLR, NeurIPS, and ICML. First, the filter keeps a paper when its title or abstract names a consistency method (such as self-consistency, semantic entropy, SelfCheckGPT, or answer agreement), or pairs an uncertainty term (uncertainty, confidence, calibration, hallucination, abstention, or selective prediction) with a term for large language models. Alongside these matches, the filter also keeps papers that cite core uncertainty-estimation work. The filter retained 6,576 papers, of which we obtained full text for 6,246. Full text was unavailable for the remaining 330. An LLM reader examined the available papers in full. A paper enters the corpus when it evaluates a consistency-based estimator against whether each answer is correct, yielding 296 papers.

Table 18 summarizes the evaluation settings of the 296 corpus papers. They comprise 19 papers from 2023, 81 from 2024, 86 from 2025, and 110 from 2026. The 2026 count is incomplete because the EMNLP and NeurIPS proceedings for that year were not yet published. The most-used benchmarks are TriviaQA (107 papers), GSM8K (74), and Natural Questions (58).

Table 18: Evaluation settings of the 296 surveyed papers from 2023–2026 that check a consistency-based estimator against correctness. Table 19 breaks down their language coverage.

Regime	Quantity	Value
Language coverage	English data only	273/296
	Translation is the task	11/296
	Same questions compared across languages	4/296
Subject models	Papers with no frontier model	160/296
	Papers with a 5–10B model	219/296
	Median open size, 2025–2026	8B
	Frontier models per paper, 2023 $\rightarrow$ 2026	1.9 $\rightarrow$ 0.7
Most-used frontier models	GPT-3.5-turbo (incl. ChatGPT)	68
	GPT-4	41
	GPT-4o	21
	text-davinci-003	15
	GPT-4o-mini	15
Benchmarks	TriviaQA	107
	GSM8K	74
	Natural Questions	58
	TruthfulQA	42
	CoQA	30

J.2 LANGUAGE COVERAGE

English-only data account for 273 of the 296 corpus papers (92.2%). Table 19 classifies the remaining 23 papers by how they use non-English data and evaluate the estimators. Of these, 11 evaluate translation, mostly into or from English on Workshop on Machine Translation (WMT) benchmarks. Another 2 pool languages into one result, 3 use no English data, and 2 ask different questions in each language, so their per-language results cannot be compared directly. One further paper includes multiple languages without tying them to reported results. These evaluations do not compare estimator performance on the same questions across languages.

Four papers compare the same questions across languages. Xue et al. (2025) evaluate a sampling estimator on GPT-3.5 Turbo, reporting AUROC and ECE per language on two datasets with about 200 items each. They report that their English-calibration finding holds for sampling across languages,

Table 19: Language coverage of the 296 surveyed papers that evaluate a consistency-based estimator. The other 273 use English only. Rows give how languages enter the evaluation; columns give how the estimator is checked: with a calibration measure (ECE, Brier, NLL or a reliability diagram), by ranking only (AUROC, AUPRC, PRR or a coverage curve), or with other measures only (for example the F1 score of flagged errors, or abstain accuracy at one threshold).

Languages	Calibration	Ranking only	Other measures only
Translation is the task	Wang et al. (2025b)	Hamidieh et al. (2026); Daheim et al. (2025); Nguyen et al. (2025); Tomov et al. (2026); Vashurin et al. (2025a;b); Vazhentsev et al. (2025; 2026); Xia et al. (2025)	Flores et al. (2025)
Languages pooled into one number	–	Deng et al. (2026); Srey et al. (2026)	–
No English	–	–	Sun et al. (2025); Feng et al. (2024); Wu et al. (2025)
Different items per language	Wang et al. (2024b)	–	Xu et al. (2026)
Same items, one dataset	–	Liu et al. (2025); Yaldiz et al. (2025)	–
Same items, several datasets, conclusion drawn	Xue et al. (2025)	Datta et al. (2026)	–
Languages never tied to a number	Zhang et al. (2025)	–	–

and theirs is the only paper with a calibration measure in these rows. Datta et al. (2026) evaluate two agreement-based detectors on four open models with AUROC only and conclude that the detectors are relatively robust because their drop is smaller than the drop in task accuracy. Yaldiz et al. (2025) report per-language numbers on translated TriviaQA without drawing a conclusion about language. Similarly, Liu et al. (2025) report per-language AUROC for a sampling baseline on a Chinese medical dataset and its English translation, without drawing a conclusion about language. None uses agreement across languages as the confidence score.

J.3 Models

For the survey counts, frontier is a model-family category rather than a measure of capability at the time of publication. We count every closed model from OpenAI, Google, and Anthropic, and every non-distilled model of 200 billion parameters or more from DeepSeek, Kimi, and Z.ai (GLM) as frontier. All other models count as non-frontier. For example, GPT-3.5 counts as frontier despite its age, whereas ChatGLM3-6B and Gemma count as non-frontier. Under this definition, the share of papers testing frontier models falls from 74% in 2023 to 68% in 2024, 41% in 2025, and 29% in the available 2026 proceedings.

Figure 1, left, shows how the number of models tested per paper changes over time. Frontier models per paper fall from 1.9 to 0.7 between 2023 and 2026. Over the same period, non-frontier models per paper rise from 2.2 to 5.1. An open model with 5–10 billion parameters appears in 219 papers. The median open model has 8 billion parameters in both 2025 and 2026.

Table 20 shows that most frontier-model evaluations are in English: 129 of the 136 papers that test a frontier model use English-only data. Among the 23 papers with any non-English evaluation, only 7 test a frontier model. For the four papers that compare the same questions across languages, the consistency-based estimator runs on GPT-3.5 in MlingConf (Xue et al., 2025) or on open models with up to 70 billion parameters.

Table 20: Across the 296 corpus papers, the table crosses language coverage with frontier-model coverage.

	Tests a frontier model	Tests no frontier model
English only	129	144
Any non-English evaluation	7	16

J.4 Extraction and verification

For the full reading, an LLM reader, Muse Spark (1.3-contributor), read each of the 6,246 papers in full and completed a structured form. For each paper, the form asks which estimators appear, whether agreement among answers is the confidence score, and, for each reported number, its measure, dataset, languages, and models. Every value in the form carries a verbatim quote and its location. Code applies the inclusion rule to the forms and computes the counts, keeping the recorded evidence separate from the counting rules. An automated check matched the opening text of 96.5% of 5,044 quotes to the papers after removing formatting differences. Another 3.2% had partial text matches, and 0.4% did not meet either matching criterion (percentages rounded).

Next, a separate Claude Opus agent reread every paper the first reader included at any stage, covering 368 papers, using the full text and PDF with the same definitions. The two readers agreed on corpus membership for 324 papers and on the language category for 363. For each of the 45 disagreements, a human evaluator read the papers and resolved the disagreement. In this review, the Opus reader was right in 39 cases, and the Muse in 6.

As a further check, we manually examined 50 papers. This set included every multilingual comparison and known false positives, with the remaining papers selected at random. For filter coverage, the keyword terms alone, applied to the title and abstract, catch 280 of the 296 corpus papers (94.6%); the others entered through their citations of core work. Separately, a citation chase following the references and citing papers of the multilingual corpus papers found no qualifying paper in the covered venues that the filter had missed. Per-paper readings, quotes, and code are released with the paper.