

Does Quantization Change What Agreement Means? A Paired Audit of Homogeneous LLM Panels

Ananto Nayan Bala

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

nayan.ananto@gmail.com

Abstract

Agreement among repeated large language model (LLM) responses is often used as a confidence signal. We test whether that signal changes after quantization by comparing five-response panels from the same checkpoints at FP16 and 4-bit NF4. We evaluate two open models on 300 TruthfulQA multiple-choice questions and 300 GSM8K problems, matching prompts, sampling settings, and seeds. Under strict parsing, quantization reduces GSM8K panel accuracy by 10.5 percentage points for Qwen2.5-3B and 6.3 for Phi-3.5-mini. The probability that all five responses are wrong rises by 9.0 and 5.0 points. Agreement also becomes less useful for selecting reliable answers, but all-wrong panels do not consistently become more likely to give one identical answer. Relaxed parsing preserves the selective-reliability result for both models, although Phi’s panel-accuracy and all-wrong differences become inconclusive. On TruthfulQA, Phi degrades while Qwen is largely stable. Quantization can therefore reduce panel reliability without consistently pushing all members toward the same wrong answer.

1 Introduction

Sampling a language model several times gives a simple uncertainty signal: if the answers agree, we are more willing to trust them. Self-consistency and multi-response uncertainty methods aggregate or compare several answers (Wang et al., 2023; Feng et al., 2025; Kruse et al., 2025), while multi-agent debate can explicitly communicate agent confidence (Yoffe et al., 2025). The signal is useful, but it assumes that agreement keeps the same meaning after the system changes.

Quantization stores model weights with fewer bits. We compare a 16-bit floating-point baseline (FP16) with a 4-bit NormalFloat copy (NF4) of the same checkpoint. Compression reduces running costs, but it can weaken calibration (the match

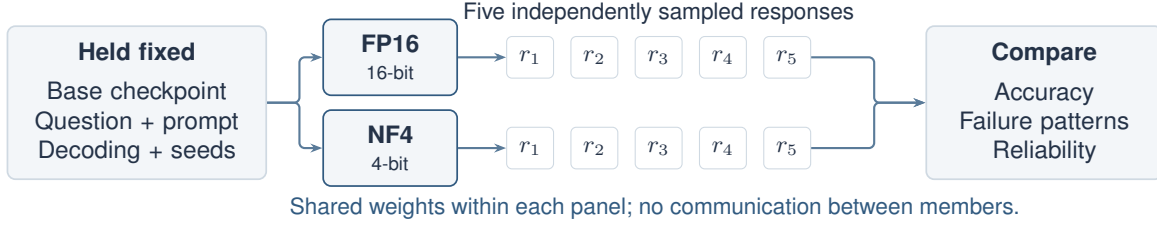
between confidence and correctness) even when accuracy changes only modestly (Zhong et al., 2025; Tong et al., 2026). Existing studies mainly examine confidence or prediction sets for individual predictions. For a panel, we also need to know how its members fail together. Similar individual accuracy can hide a higher probability that all members are wrong. More all-wrong panels, in turn, need not mean more agreement on one wrong answer.

We study this distinction in a controlled, small-scale audit. For each checkpoint, task, and question, we draw five responses at each precision. Each response is a *panel member*: members share weights and prompts but are sampled independently and never communicate. We ask whether quantization changes accuracy, how often all members are wrong, and how often those all-wrong panels give one identical answer. We also test whether vote share remains useful for ranking panel decisions by reliability. Figure 1 shows the paired design and three example panel outcomes that motivate our measurements.

The practical distinction is simple. If all five responses are wrong but disagree, the panel fails without unanimous agreement. If they repeat the same wrong answer, the panel both fails and appears maximally confident. Counting only unanimous wrong answers can miss the first kind of failure. Counting only errors can miss how confident the second kind appears.

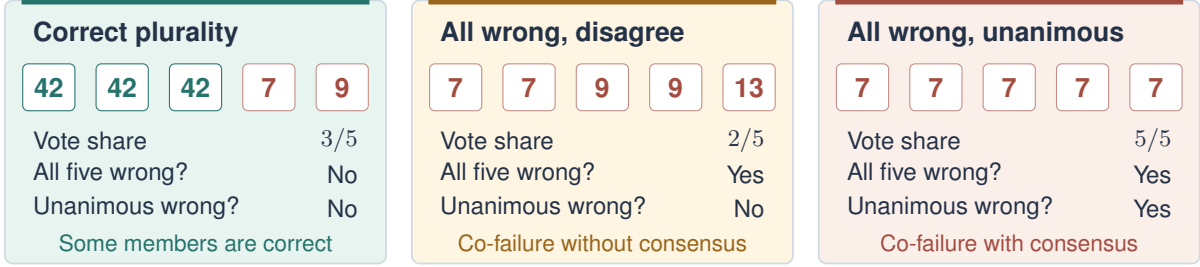
Under strict parsing, NF4 lowers accuracy and raises all-wrong and invalid-output rates on GSM8K. Selecting answers by agreement also becomes less reliable. This last result survives relaxed parsing for both models; Phi’s panel-accuracy and all-wrong differences become inconclusive. On TruthfulQA, Phi degrades while Qwen is largely stable. The tendency of all-wrong panels to give one identical answer does not change consistently.

A Compare precision on the same question



B Read the panel, not just the vote count

Example reference answer: 42



All-wrong rate β

How often are all five responses wrong?

Same-wrong concentration κ

Among all-wrong panels, how often do all five agree?

Figure 1: **One paired audit, two different failure measurements.** (A) Each question is evaluated with five responses at each precision, using the same checkpoint and matched generation settings. (B) Illustrative outcomes, not experimental results; all answers shown are valid and the reference answer is 42. The middle and right panels both contribute to β , but only the right contributes to the numerator of κ . Its vote share is maximal despite every answer being wrong.

Code and data:

<https://github.com/nayanananto/quantization-and-agreement-of-llm-models>

2 Related Work

Compression and uncertainty. Quantized LLMs can be worse calibrated; Zhong et al. (2025) show that soft prompts can help restore calibration. Conformal and broader evaluations find that accuracy and uncertainty can change differently with precision, affecting calibration, error ranking, and robustness (Tong et al., 2026; Tao et al., 2025; Ayadi et al., 2026). We instead examine how errors occur together across repeated responses from one quantized checkpoint.

Agreement and collective uncertainty. Multi-sample and multi-agent methods estimate uncertainty from response diversity or disagreement in meaning (Feng et al., 2025; Kruse et al., 2025; Jiang, 2026). Dependence between members’ correctness, adjusted for accuracy, also predicts the benefit of combining their answers (Ali, 2026). Communication can create correlated errors and false consensus (Huang et al., 2026). Pairwise cor-

relation alone does not determine how often all members are wrong (Chen, 2026), so we measure all-wrong frequency and same-wrong agreement separately. Our members do not communicate, and we hold generation settings fixed to study the effect of precision.

3 Paired Panel Audit

3.1 Models, tasks, and generation

We evaluate Qwen2.5-3B-Instruct (Qwen et al., 2024) and Phi-3.5-mini-instruct (Abdin et al., 2024). Each released checkpoint is loaded twice: as the unquantized FP16 baseline and as a locally produced bitsandbytes NF4 copy with double quantization and FP16 computation (Dettmers et al., 2023). Only the stored weights are converted after training; no retraining is performed, and prompts and decoding are unchanged.

What is paired? The same question is evaluated under both precisions, with the same prompt, decoding settings, and member seeds. Each question therefore supplies its own comparison. Matching seeds does not force the generated answers to match: changing the weights can change which

tokens are sampled. Within each panel, all five members use the same model; they differ only in their sampling seed.

We use 300 fixed questions per task. TruthfulQA MC1 requires a truthful option from a small answer set (Lin et al., 2022); GSM8K requires a free-form solution to a grade-school mathematics problem (Cobbe et al., 2021). Each question receives $K = 5$ responses per precision, with temperature 0.7, top- p 0.9, and matched seeds. TruthfulQA requests one option letter within 8 new tokens. GSM8K requests a derivation ending in `FINAL: <number>` within 256 new tokens. A response without the required parseable final marker is invalid and incorrect. Each invalid response gets its own voting label, so two invalid responses never count as agreement. Appendices B and C give runtime details and prompts.

The panel selects the answer with the most votes (plurality voting). If answers tie, we average correctness over them as if choosing uniformly at random. Accuracy therefore does not depend on sampling order, which matters especially for free response, where ties are common.

For example, if the correct answer is 42, the responses [42, 42, 7, 7, 9] tie between 42 and 7. This panel receives accuracy credit 1/2, the expected correctness of a uniform choice between those two answers. We use this expectation rather than making another random draw.

3.2 Metrics

We separate answer quality from the pattern of errors. *Individual accuracy* averages correctness across responses. *Panel accuracy* scores the plurality decision, including the tie rule above. The following measures then describe how failures and agreement relate.

All-wrong rate. Let N be the number of questions and F_i a *co-failure*: all K members are wrong on question i . Alongside mean individual and panel accuracy, we report the fraction of questions with a co-failure:

$$\beta = N^{-1} \sum_{i=1}^N \mathbf{1}[F_i]. \quad (1)$$

Here, $\mathbf{1}[F_i]$ is one for an all-wrong panel and zero otherwise. A wrong panel decision does not always count as a co-failure: one member may be correct but outvoted by the others.

Same-wrong concentration. Among all-wrong panels, what fraction give one identical answer? With member answer a_{ik} and reference answer y_i , we define

$$\kappa = \Pr(a_{i1} = \dots = a_{iK} \neq y_i \mid F_i). \quad (2)$$

The numerator requires five valid answers. Thus β measures how often everyone is wrong; κ measures agreement within those failures.

The denominators matter: β uses all questions, whereas κ uses only all-wrong panels. Their product, $\beta\kappa$, is the fraction of all panels giving one identical wrong answer. It can rise simply because all-wrong panels become more common, even if their tendency to agree stays the same.

Reliability of agreement. Confidence c_i is the largest answer count divided by K . Three matching responses out of five give $c_i = 3/5$; unanimity gives $c_i = 1$. This is a vote share, not a guarantee of correctness, as the rightmost example in Figure 1 shows.

We evaluate vote share in two ways. *Brier score* averages the squared error between confidence and correctness, including the uniform voting tie rule. It measures how closely the confidence values match outcomes. *Area under the risk-coverage curve* (AURC) evaluates using agreement to decide which answers to keep. We retain panels from highest to lowest agreement. Coverage is the fraction retained; risk is the error rate among those panels. AURC summarizes this risk as coverage increases. Lower values are better for both metrics. For AURC, we average over orderings within vote-share ties.

Uncertainty in the comparison. All NF4–FP16 intervals use 10,000 paired question-level bootstrap samples. Each sample resamples questions with replacement and uses the same selected questions for both precisions. The resulting interval describes uncertainty in the paired difference under question resampling.

4 Results

4.1 More all-wrong panels under strict parsing

Table 1 and Figure 2 show the strict-parser GSM8K results. All changes are NF4 minus FP16: lower accuracy or higher β , Brier, and AURC indicate degradation. Qwen’s individual accuracy falls by 10.8 percentage points (95% CI $[-13.7, -8.0]$),

Model	Task	Prec.	Ind.	Panel	β	Invalid	κ	Brier	AURC
Qwen2.5-3B	TruthfulQA	FP16	49.7	49.3	49.0	0.0	93.9	0.485	0.486
Qwen2.5-3B	TruthfulQA	NF4	48.3	48.0	50.7	0.0	92.8	0.496	0.499
Qwen2.5-3B	GSM8K	FP16	51.8	62.3	26.3	41.9	1.3	0.093	0.102
Qwen2.5-3B	GSM8K	NF4	41.0	51.9	35.3	54.4	0.0	0.095	0.158
Phi-3.5-mini	TruthfulQA	FP16	54.9	55.8	43.0	0.0	84.5	0.409	0.415
Phi-3.5-mini	TruthfulQA	NF4	50.2	50.0	48.0	0.0	86.8	0.455	0.466
Phi-3.5-mini	GSM8K	FP16	33.3	47.4	40.0	63.1	0.0	0.114	0.197
Phi-3.5-mini	GSM8K	NF4	26.5	41.1	45.0	70.6	0.7	0.126	0.264

Table 1: Absolute results (%). Brier and AURC are unitless; lower is better. β is the all-members-wrong rate and κ is same-wrong-answer concentration conditional on co-failure. Invalid outputs are treated as distinct abstentions, not agreement.

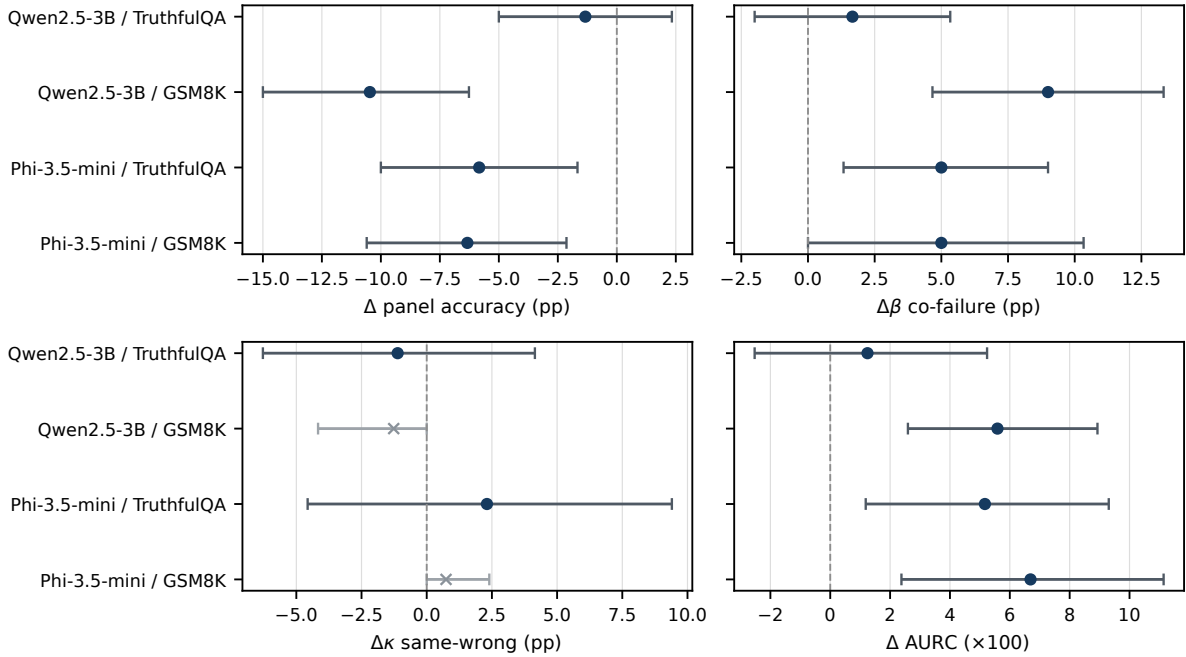


Figure 2: Paired NF4 minus FP16 effects with 95% question-bootstrap intervals. Negative accuracy and positive β /AURC values indicate degradation. The $\Delta\kappa$ panel shows no consistent change in conditional same-wrong concentration; gray crosses mark the sparse GSM8K estimates, where exact same-wrong events occur at most once per condition.

panel accuracy by 10.5 $[-15.0, -6.3]$, and β rises by 9.0 $[4.7, 13.3]$. For Phi, the changes are $-6.9 [-9.7, -4.1]$, $-6.3 [-10.6, -2.1]$, and $+5.0 [0.0, 10.3]$, respectively.

TruthfulQA results differ by model. Phi loses 4.7 points of individual accuracy and 5.8 of panel accuracy; β rises by 5.0 points. All three intervals exclude zero. Qwen changes by less than two points, with intervals including zero. The strict-parser GSM8K pattern therefore appears in both models, but does not hold across both task formats.

Reading these changes. Individual accuracy counts each response, while panel accuracy asks whether voting recovers the answer. A rise in β

means more questions have no correct response available to the vote. It does not tell us whether those wrong responses agree; that requires κ . An interval that includes zero leaves the direction of a difference unresolved, rather than establishing that the two precisions are equivalent.

4.2 All-wrong frequency and wrong agreement differ

TruthfulQA’s small answer set makes identical wrong answers common. Yet NF4 changes κ little: Qwen moves from 93.9% to 92.8% ($\Delta\kappa = -1.1$, CI $[-6.3, 4.1]$), and Phi from 84.5% to 86.8% ($+2.3$, $[-4.6, 9.4]$). The directions differ and both intervals cross zero. For Phi, the fraction of *all* pan-

els giving one identical wrong answer ($\beta\kappa$) rises from 36.3% to 41.7%, a 5.3-point change before rounding. This mainly reflects more all-wrong panels, rather than a clear change in agreement within those panels. Appendix A reaches the same conclusion after adjusting for chance agreement.

On GSM8K, five identical wrong answers occur only once for Qwen FP16, never for Qwen NF4 or Phi FP16, and once for Phi NF4. A zero here means no observed event, not that wrong consensus is impossible. Each event requires five wrong, validly parsed, identical answers. The unrestricted numeric answer space and many invalid outputs make this rare. The strict-parser invalid rate is already 41.9–63.1% at FP16; NF4 raises it by 12.5 points for Qwen and 7.5 for Phi, with both intervals excluding zero. Too few panels have five valid answers for a useful κ sensitivity estimate restricted to that subset. The primary analysis does not guess answers from unfinished derivations.

4.3 Agreement becomes less useful for selecting answers

AURC rises by 5.6 points for Qwen GSM8K and 6.7 for Phi GSM8K; both intervals exclude zero. Phi TruthfulQA also worsens by 5.2 points, while Qwen remains stable. Brier score significantly worsens only for Phi TruthfulQA (+0.046, CI [0.009, 0.085]). The more consistent change is therefore in selecting reliable answers using agreement, rather than in the average squared error of vote-share confidence.

Many GSM8K responses reach the 256-token limit without a parseable `FINAL: <number>` answer. We therefore check a permissive parser that keeps explicit final answers and uses the last numeric string only for otherwise invalid responses. AURC still worsens by 7.0 points for Qwen (CI [3.7, 10.7]) and 3.8 for Phi ([0.5, 7.1]). Qwen’s panel-accuracy and β effects persist; Phi’s intervals cross zero. Selecting answers by agreement is therefore less reliable under both parsers, while Phi’s collective-failure result depends on the parsing rule. Appendix A gives all estimates.

Across both parsers, agreement is less useful for selecting reliable GSM8K panel answers after quantization. All-wrong panels also become more frequent, most clearly for Qwen, but do not consistently become more likely to give one identical wrong answer.

5 Conclusion

In this two-model, two-task study, quantization changes panel reliability without consistently increasing agreement on one wrong answer. The most robust GSM8K result is that agreement becomes less useful for selecting reliable answers under both parsers and for both models. Lower accuracy and more all-wrong panels are clearest for Qwen; Phi’s panel and co-failure effects depend on parsing. Measuring how often everyone is wrong separately from how often those failures agree makes the effect of compression easier to interpret.

6 Limitations

This case study covers two small model families, one 4-bit method, two tasks, and 300 fixed questions per task. Members are independent samples from shared weights, without specialized roles or communication. Exact numeric matching suits GSM8K, but the final-marker requirement and 256-token cap produce many invalid outputs. The results therefore reflect both model performance and adherence to the output format. Relaxed parsing removes invalid parses but may select an intermediate number from unfinished reasoning; it is a stress test, not verified recovery of intended answers. Multiple choice restricts the answer set, so its high κ should not be compared directly with free response. Vote share also has few possible values, so we emphasize Brier, AURC, and paired intervals rather than expected calibration error (ECE).

References

- Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Junade Ali. 2026. Quantifying diversity of thought: A predictive law of weighted LLM ensemble lift. *arXiv preprint arXiv:2607.17384*.
- Sirine Ayadi, Sándor Daróczi, Stephan Günnemann, and Bertrand Charpentier. 2026. Reliability scaling laws for quantized large language models. *arXiv preprint arXiv:2607.10855*.
- Josef Chen. 2026. When does combining language models help? a co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv preprint arXiv:2606.27288*.

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36.
- Yu Feng, Phu Mon Htut, Zheng Qi, Wei Xiao, Manuel Mager, Nikolaos Pappas, Kishalay Halder, Yang Li, Yassine Benajiba, and Dan Roth. 2025. [Rethinking llm uncertainty: A multi-agent approach to estimating black-box model uncertainty](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12349–12375.
- Jiatan Huang, Mingchen Li, Ziming Li, Sunjae Kwon, Hong Yu, and Chuxu Zhang. 2026. Counterfactual graph for multi-agent llm calibration. *arXiv preprint arXiv:2605.30653*.
- Bo Jiang. 2026. DiscoUQ: Structured disagreement analysis for uncertainty quantification in LLM agent ensembles. *arXiv preprint arXiv:2603.20975*.
- Maya Kruse, Majid Afshar, Saksham Khatwani, Anoop Mayampurath, Guanhua Chen, and Yanjun Gao. 2025. [Simple yet effective: An information-theoretic approach to multi-llm uncertainty quantification](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30493–30504.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.
- Qwen, An Yang, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Linwei Tao, Yi-Fan Yeh, Minjing Dong, Tao Huang, Philip Torr, and Chang Xu. 2025. Revisiting uncertainty estimation and calibration of large language models. *arXiv preprint arXiv:2505.23854*.
- Yujia Tong, Yuxi Wang, Yunyang Wan, Tian Zhang, Junhao Dong, and Jingling Yuan. 2026. Does compression preserve uncertainty? a unified benchmark for quantized and sparse llms via conformal prediction. *arXiv preprint arXiv:2606.01850*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.
- Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2025. [DebUnc: Improving large language model agent communication with uncertainty metrics](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23299–23315.
- Mingyu Zhong, Guanchu Wang, Yu-Neng Chuang, and Na Zou. 2025. [Quantized can still be calibrated: A unified framework to calibration in quantized large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30503–30517.

A Sensitivity Analysis

Chance-adjusted concentration. For TruthfulQA, we also subtract the agreement expected under an independence baseline. Questions are grouped by the correct option, and the baseline preserves each member’s wrong-answer frequencies within each group. The NF4–FP16 change in this excess κ is +1.2 points for Qwen (95% CI $[-4.6, 7.3]$) and +2.9 for Phi ($[-4.4, 10.6]$). As in the raw analysis, neither interval establishes a change in agreement beyond this baseline.

Relaxed GSM8K parsing. We preserve every explicit FINAL: answer and use the last numeric string only for a response that is invalid under the primary parser. This changes the invalid rate from 41.9–70.6% to zero, but can mistake an intermediate number for an intended answer. Table 2 therefore reports a sensitivity analysis rather than a replacement evaluation. The values are paired NF4–FP16 changes in percentage points, with 95% question-bootstrap intervals.

Metric	Qwen	Phi
Individual accuracy	−10.3 $[-13.3, -7.5]$	−4.1 $[-6.7, -1.5]$
Panel accuracy	−11.5 $[-16.1, -7.2]$	−3.4 $[-7.4, 0.6]$
β	+7.0 $[3.0, 11.0]$	+3.7 $[-0.7, 8.3]$
Brier	+1.3 $[-0.5, 3.0]$	+2.3 $[0.4, 4.2]$
AURC	+7.0 $[3.7, 10.7]$	+3.8 $[0.5, 7.1]$
κ	−2.9 $[-7.6, 0.0]$	+0.8 $[-4.3, 6.0]$

Table 2: Relaxed-parser GSM8K sensitivity. Accuracy and β are probability differences; Brier and AURC are multiplied by 100.

The main result for selecting answers by agreement is unchanged: AURC worsens for both models. Qwen’s individual accuracy, panel accuracy, and all-wrong effects remain clear. Phi’s individual accuracy still falls, but its panel-accuracy and β intervals now include zero. Same-wrong concentration among all-wrong panels remains small and changes inconsistently across models.

B Reproducibility Details

The experiment uses five samples per question and base seed 20260809, with member seed $20260809 + 5i + k$. The final runtime uses Transformers 4.57.6, bitsandbytes 0.50.0, and FP16 compute for NF4. Phi remote model code is pinned to revision `2fe1924`. Raw JSONL generations, configuration files, environment records, analysis code, and 10,000-replicate bootstrap outputs are

provided through the repository linked after the Introduction.

C Prompt Templates

TruthfulQA uses: “Answer the multiple-choice question by returning only the single option letter. Do not explain your answer.” GSM8K uses: “Solve the problem step by step. End with exactly FINAL: <number>. The final marker is required.” No demonstrations or communication between members are used.