Do Large Language Models Know When They Are Wrong? A Case Study of Confidence and Abstention Under Linguistic Corruption

Syed Muhammad Ali Abbas
Jožef Stefan Institute
Ljubljana, Slovenia
Syed.Muhammad.Ali.Abbas@ijs.si

Matjaž Gams
Jožef Stefan Institute
Ljubljana, Slovenia
matjaz.gams@ijs.si

Abstract

Self-reported confidence is useful only if it changes when model performance becomes unreliable. We study this as behavioral metacognitive monitoring using 60 matched reasoning problems presented in clean and severely corrupted form, derived from a 90-item stress-test benchmark. We compare GPT-5.6 Sol Medium and o3 in separate ChatGPT sessions using identical answer-plus-confidence prompts. Sol remained nearly invariant (100.00% clean vs. 98.33% corrupted accuracy), and its only corrupted error still received 98% confidence. o3 solved 95.00% of the clean problems but only 1.67% of the corrupted versions, a 93.33-point decrease; under corruption it produced 34 explicit "Unknown" responses and omitted 25 items. Exact paired testing showed a significant corruption-associated performance difference for o3 ($p < 0.001$) but not for Sol. The results reveal a sharp contrast between confidence saturation and abstention, and separate three aspects of LLM metacognitive behavior: robustness to degraded input, calibration of committed answers, and control over whether to answer.

Keywords

large language models, metacognition, confidence calibration, linguistic corruption, surface-form robustness, abstention, Brier score

1 Introduction

Metacognition concerns the monitoring and regulation of one's own cognitive processes [1]. Classical accounts distinguish an *object level*, where a task is performed, from a *meta level* that monitors and can regulate object-level processing [2]. Confidence judgments provide an observable form of such monitoring, but confidence and objective performance need not coincide [3].

This distinction is increasingly important for large language models (LLMs). A model may be highly accurate on average yet remain almost certain on its rare failures. Conversely, a less robust model may signal uncertainty or abstain when an input exceeds its effective capabilities. Prior work shows that LLMs can verbalize uncertainty [4], can exhibit useful self-evaluation signals [5], and can support calibration and failure prediction when confidence is explicitly elicited [6, 7]. However, confidence is sensitive to elicitation and task design, and high-confidence errors remain possible [8].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.cogni.92

A methodological problem is that many standard reasoning items are now too easy for strong models, yielding too few failures to examine confidence behavior. We therefore use composite linguistic degradation as a controlled stressor. Each stressed item has a matched clean reconstruction, so the same underlying reasoning problem can be evaluated with and without corruption.

We address two research questions:

RQ1: How does linguistic corruption affect surface-form robustness when the underlying reasoning problems and answers are held constant?

RQ2: How do self-reported confidence and response behavior differ between successful answers, committed errors, explicit abstentions, and omissions under linguistic corruption?

Our contribution is threefold: (i) a matched clean/corrupted stress-test design that holds the underlying reasoning problems constant, (ii) paired evidence of strongly different robustness profiles across model configurations, and (iii) a coverage-aware metacognitive analysis that separates committed errors from abstention. Throughout the paper, verbal confidence is treated as an observable behavioral metacognitive signal.

2 Related Work

Confidence elicitation is a common black-box approach to LLM uncertainty. Lin et al. showed that models can learn to express uncertainty in natural language [4]. Tian et al. found that verbalized confidence can be meaningfully calibrated in instruction-tuned models [6], while Xiong et al. distinguished aggregate calibration from failure prediction—the ability of confidence to distinguish likely correct from likely incorrect outputs [7]. These properties are not equivalent: mean confidence can be close to mean accuracy even when confidence barely changes on individual errors.

Robustness to noisy surface forms is a related but distinct problem. Zhao et al. evaluate 18 open-source LLMs under multilingual typographical errors and find substantial model- and task-dependent degradation [9]. Our manipulation extends beyond typos to deletion, token fusion, grammar removal, symbolic compression, and distractors, while pairing each corrupted reasoning item with the same clean problem.

Prompting and reflection can alter calibration [8], but monitoring does not automatically imply effective control; intrinsic self-correction can remain unreliable without external feedback [10]. Recent work has used metacognitive signals as control variables for tool use [11]. A recent survey frames LLM metacognition around monitoring and control [12], and related work examines how self-monitoring can guide test-time reasoning [13]. Our study focuses on response behavior under a matched linguistic stress manipulation.

Table 1: Example of the matched linguistic manipulation (Q061).

Clean	Arrange A–G in positions 1–7 once each. D is immediately after B; A is immediately after D; F is before B; $A < G < C < E$; F is not adjacent to A; G is not first. Find the unique order.
Corrupted	`7obj A..G /p1-7 uniq. D=B+1p; A=D+1p; F«B; A<G<C<E; |pF-pA|!=1; G!=p1. full seq?`

3 Method

3.1 Benchmark and Matched Control

The full benchmark contains 90 independently scored items. Q001–Q030 are clean baseline problems. Q031–Q060 combine harder reasoning with irregular linguistic degradation, including character deletion, token fusion and splitting, omitted grammatical words, abbreviations, symbolic relations, and distractor fragments. Q061–Q090 intensify the degradation and use multi-step tasks such as constraint satisfaction, Bayesian inference, counterfactual reasoning, temporal dependencies, dynamic state tracking, graph reasoning, combinatorics, and symbolic transformations.

For every Q031–Q090 item, a separately maintained clean natural-language reconstruction and predetermined answer were created. The main experiment therefore contains 60 paired problems in two conditions: *clean* and *corrupted*. Logical content and answer key are held constant while the surface representation changes. Table 1 illustrates the manipulation.

3.2 Models and Procedure

Pilot testing showed contrasting response profiles under linguistic corruption. We therefore compare GPT-5.6 Sol Medium, which exhibited high robustness, with o3, which showed substantially greater sensitivity to the corrupted items. This contrast enables direct analysis of how confidence and abstention behave when two capable reasoning models respond differently to the same matched stressor. The label "Medium" denotes the reasoning-effort setting displayed in ChatGPT at test time.

All reported runs were conducted in the ChatGPT interface on 26 August 2026; model labels are reported exactly as displayed at test time. Clean and corrupted conditions were run in separate sessions with the same minimal prompt requiring one answer and an integer confidence score from 0–100 for every ID. Visible explanations were suppressed to keep the output format identical across models and make confidence scoring directly comparable. Models were not told that the clean items reconstructed previously corrupted questions.

Scoring followed the predetermined answer key. Numerically equivalent answers and reasonable rounding differences were accepted. Extra continuation information was accepted when it contained the requested answer. Explicit "Unknown" responses and omissions were incorrect for required-answer accuracy, but were tracked separately rather than treated as ordinary wrong attempts.

3.3 Metrics

For item i, correctness is $y_i \in \{0, 1\}$ and confidence is $c_i \in [0, 1]$. Accuracy is

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{1}$$

Table 2: Matched Q031–Q090 results. Confidence is averaged over returned rows; Brier score is calculated only on concrete attempts.

Model	Cond.	Acc.	Out.	Commit	Conf.	BS (n)
Sol	Clean	100.00	100.00	100.00	100.00	0.0000 (60)
Sol	Corr.	98.33	100.00	100.00	98.50	0.0165 (60)
o3	Clean	95.00	96.67	96.67	90.86	0.0218 (58)
o3	Corr.	1.67	58.33	1.67	11.71	0.0900 (1)

Acc., Out., Commit, and Conf. are percentages. Out.=any returned row; Commit=concrete-answer coverage. The corrupted o3 Brier estimate is based on one concrete answer and is not stable. Wilson 95% CIs for accuracy are Sol clean [94.0,100], Sol corr. [91.1,99.7], o3 clean [86.3,98.3], and o3 corr. [0.3,8.9].

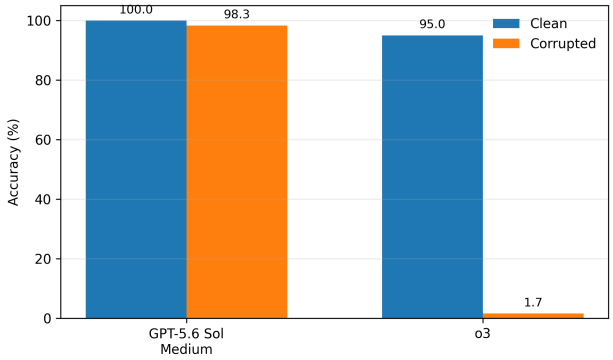

Figure 1: Required-answer accuracy on the same 60 reasoning problems in clean and corrupted form.

For concrete answer attempts, we compute the Brier score [14],

$$\text{BS} = \frac{1}{M} \sum_{i=1}^{M} (c_i - y_i)^2, \tag{2}$$

where lower is better. "Unknown" and omitted items are excluded from the primary Brier score because they do not contain a concrete proposed solution. We additionally report output coverage, concrete-answer coverage, and response type counts. Because clean and corrupted conditions contain the same 60 items, we use the exact McNemar test [15] for paired changes in correctness.

4 Results

4.1 Matched Clean vs. Corrupted Performance

Table 2 summarizes the matched-condition metrics. Figure 1 shows a strongly model-specific corruption effect. Sol solved all 60 clean items and 59 of 60 corrupted items, a decrease of only 1.67 percentage points. o3 solved 57 of 60 clean items but only one of 60 corrupted items, a 93.33-point decrease. The same underlying reasoning problems were therefore highly solvable by o3 when written clearly, but not when their surface form was heavily degraded.

The paired analysis supports the same conclusion. For Sol, only one item changed from correct when clean to incorrect when corrupted, with no changes in the opposite direction (exact McNemar $p = 1.0$). For o3, 56 items changed from correct to incorrect and none changed from incorrect to correct ($p < 0.001$). Thus the matched data show a large corruption-associated performance difference for o3 but not for Sol.

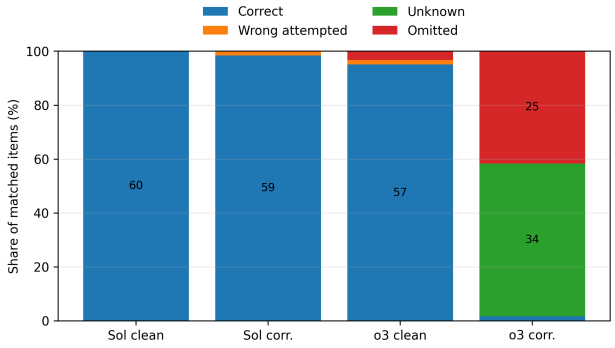

Figure 2: Response outcomes on the matched clean and corrupted conditions. Numbers inside large segments are item counts.

Table 3: Confidence by response type. Values in parentheses are item counts.

Model	Cond.	Correct	Wrong attempt	Unknown
Sol	Corr.	98.51 (59)	98.00 (1)	–
o3	Clean	90.96 (57)	85.00 (1)	–
o3	Corr.	70.00 (1)	–	10.00 (34)

4.2 Response Policy and Confidence

The two models also differed in how they failed. Sol continued to provide concrete answers in both conditions. Its only corrupted failure was an incomplete reconstruction of a personal name and received 98% confidence. o3 produced 57 correct answers, one concrete error, and two omissions in the clean condition. Under corruption it produced one correct concrete answer, 34 explicit "Unknown" responses, and 25 omissions (Fig. 2). Its concrete-answer coverage therefore fell from 96.67% to 1.67%.

Sol's aggregate calibration remained strong: its Brier score rose from 0.0000 clean to 0.0165 corrupted. However, Table 3 shows why aggregate calibration alone is insufficient for metacognitive interpretation: its rare error received almost the same confidence as its correct answers.

For o3, the dominant metacognitive signal under corruption appears in *response control*: heavy corruption shifted the model from answering to low-confidence abstention and omission. Its only clean committed error received 85% confidence, while the corrupted condition contained too few committed wrong answers to characterize error-level discrimination. Mean confidence across returned rows fell to 11.71%, reflecting the large number of explicit "Unknown" responses and the accompanying drop in answer coverage.

4.3 Illustrative Error Cases

The few committed errors help clarify what the aggregate metrics do and do not measure. Sol's only corrupted error occurred on Q046, a working-memory item in which the name *Saniya* appeared in degraded form. Sol correctly tracked the relevant person but returned the incomplete reconstruction "Sniya" with 98% confidence. The clean version of the same item was answered as "Saniya" with 100% confidence. This is primarily a linguistic reconstruction failure, not a failure of the ordering operation itself.

o3's only concrete clean error occurred on Q082, a multi-step symbolic string transformation. It returned "HNBWQU" rather than "HNWBQU" with 85% confidence. Under corruption, however, o3 did not produce a corresponding pool of confidently wrong transformed answers: after Q031 it largely switched to "Unknown" and later omissions. The comparison therefore supports a distinction between *confidence on a committed solution* and *willingness to commit*.

5 Discussion

The matched control changes the scientific interpretation compared with a simple difficulty-tier experiment. o3 can solve the underlying Q031–Q090 problems at 95% accuracy when they are clean, yet required-answer accuracy falls to 1.67% when the same logical content is heavily corrupted. Sol is nearly invariant to the manipulation. The composite linguistic corruption therefore produces a large, model-specific robustness effect rather than a uniform increase in reasoning difficulty.

The title question receives a differentiated answer. Sol combines very high robustness with confidence saturation: its aggregate Brier score is excellent, but its sole failure remains at 98% confidence. o3 expresses uncertainty in a different way: under corruption it shifts strongly from answering to explicit abstention and omission. These contrasting profiles show that metacognitive behavior can appear both in confidence assigned to a solution and in control over whether to commit to one. They motivate separating three properties: *surface-form robustness*, *calibration of committed answers*, and *abstention/coverage behavior*.

This separation matters methodologically. Collapsing "Unknown", omission, and wrong attempts into one incorrect category would make o3 appear to have excellent error-level confidence sensitivity, because most failures carry 10% confidence. That conclusion would be misleading: the low values accompany abstentions, not a large sample of committed wrong solutions. Conversely, a small mean confidence–accuracy gap for Sol could hide a near-certain individual error. Confidence metrics should therefore be interpreted jointly with response coverage and response type.

The findings also show that benchmark difficulty is model-dependent. The same linguistic stressor is nearly harmless to Sol but produces a dramatic change in o3's accuracy and response policy. Matched clean/corrupted controls therefore provide a useful way to distinguish the difficulty of the underlying reasoning problem from sensitivity to its surface representation.

Beyond this benchmark, the result has a practical implication for LLM evaluation: accuracy alone cannot distinguish a model that confidently commits to an answer from one that recognizes difficulty by withholding a response. Jointly reporting correctness, confidence, and coverage provides a richer view of model behavior under degraded inputs and makes qualitatively different failure modes visible.

6 Limitations and Future Work

This focused study examines two contrasting model configurations under a composite linguistic stressor. The paired within-item design establishes the reported robustness contrast on this benchmark; broader model families and repeated runs will test how widely the pattern generalizes. Future work will also separate individual corruption types and add human readability calibration. Because the experiments used ChatGPT, backend decoding parameters were not directly observable. The matched question pairs, prompt, answer key, and scoring protocol are archived with the study to support reproduction and extension.

7 Conclusion

On 60 matched reasoning problems, linguistic corruption had almost no effect on GPT-5.6 Sol Medium but reduced o3 required-answer accuracy by 93.33 percentage points despite 95% clean accuracy on the same items. Sol remained highly robust and almost maximally confident, including on its rare failure, whereas o3 shifted primarily toward low-confidence abstention and omission. These contrasting profiles support a three-part view of behavioral LLM metacognition: robustness to degraded input, calibration of committed answers, and control over whether to answer. Evaluating these dimensions jointly reveals behavior that accuracy or mean confidence alone would miss.

References

[1] John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34, 10, 906–911. https://doi.org/10.1037/0003-066X.34.10.906

[2] Thomas O. Nelson and Louis Narens. 1990. Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation* 26, 125–173.

[3] Stephen M. Fleming. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology* 75, 241–268. https://doi.org/10.1146/annurev-psych-022423-032425

[4] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

[5] Saurav Kadavath et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

[6] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of EMNLP 2023*, 5433–5442. https://doi.org/10.18653/v1/2023.emnlp-main.330

[7] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *ICLR 2024*.

[8] Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models. In *Findings of ACL 2024*, 8702–8718. https://doi.org/10.18653/v1/2024.findings-acl.515

[9] Raoyuan Zhao, Yihong Liu, Lena Altinger, Hinrich Schuetze, and Michael A. Hedderich. 2026. Evaluating Robustness of Large Language Models Against Multilingual Typographical Errors. In *Proceedings of ACL 2026*, 16059–16078. https://doi.org/10.18653/v1/2026.acl-long.729

[10] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models Cannot Self-Correct Reasoning Yet. In *ICLR 2024*.

[11] Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. Adaptive Tool Use in Large Language Models with Meta-Cognition Trigger. In *Proceedings of ACL 2025*, 13346–13370. https://doi.org/10.18653/v1/2025.acl-long.655

[12] Gabrielle Kaili-May Liu, Areeb Gani, Jacqueline Lu, Jordan Thomas, Mark Steyvers, and Arman Cohan. 2026. Metacognition in LLMs: Foundations, Progress, and Opportunities. *arXiv preprint arXiv:2607.11881*.

[13] Qi Cao, Yufan Wang, Peijia Qin, Shuhao Zhang, and Pengtao Xie. 2026. LLMs Know When They Know, but Do Not Act on It: A Metacognitive Harness for Test-time Scaling. *arXiv preprint arXiv:2605.14186*.

[14] Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78, 1, 1–3.

[15] Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* 12, 153–157. https://doi.org/10.1007/BF02295996