Can a Model Tell When It Is Wrong?

Confidence signals at the MiniMax frontier

Alex Newman 13 July 2026

> **Status: preliminary and not peer reviewed.** Existing calls support an exploratory finding. A fresh, prospectively registered validation is still collecting data. The trace-triggered backup experiment failed its no-spend gate and was stopped.

1. Abstract

A difficulty-response curve estimates how likely a model is to solve a task at a given difficulty. We ask whether information produced during one call - reasoning text, token counts, latency, or serving metadata - improves the probability of correctness for that particular answer. In 145 retrospective MiniMax-M2.5 calls, a frozen metadata model reached AUROC 0.852 and Brier skill 36.9% relative to the curve prior. A trace-only model reached AUROC 0.671 and 7.4% Brier skill; its clustered interval crossed zero. A second frozen hypothesis used trace markers to launch backup models. It caught every historical failure but also triggered on 87.2% of correct calls, making it an always-on council in practice. That branch stopped with no new spend. The metadata result is now undergoing a single prospective evaluation.

2. The question

For an answer to a hard problem, we want a calibrated statement such as "there is a 72% chance this answer is correct." Difficulty provides the first estimate. Two equally difficult calls can still look different: one may finish quickly, while another consumes its token budget and backtracks for minutes. The research question is whether those differences should change confidence.

The measurement instrument borrows the logic of item response theory [1]. Fresh satisfiable 3-SAT instances provide a scalar difficulty axis through their clause-to-variable ratio, and full assignments are checked as certificates. Random SAT is known to exhibit strong difficulty variation near its transition region [2]. The curve is a prior, not a claim about a reasoning mechanism.

2.1. Three outcomes

The study keeps mutually exclusive outcomes:

1. **Correct completion:** a normal completion with a verified certificate.
2. **Silent error:** a normal completion with a parseable but wrong certificate.
3. **Loud failure:** truncation, refusal, parse failure, timeout, provider mismatch, malformed response, or API error.

Confidence is evaluated only on correct completions versus silent errors. Loud failures are reported separately because they already announce themselves. The primary endpoint is Brier skill relative to the curve prior. AUROC and log loss are secondary views. Repeated samples of one generated instance are kept together during evaluation.

3. Retrospective result

The exploratory cohort contains 145 calls on 30 instances: 61 correct completions, 24 silent errors, and 60 loud failures.

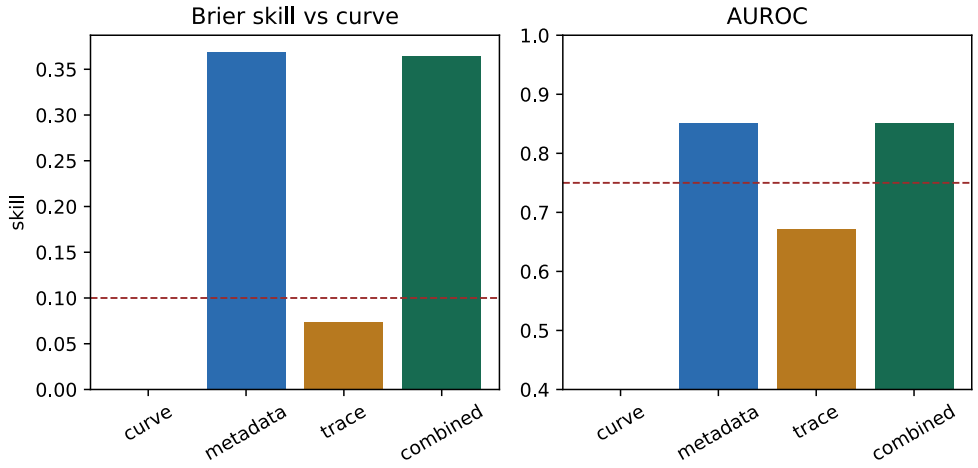

Figure 1: Retrospective comparison. Dashed lines are the frozen exploratory spend-gate thresholds.

Model	Brier	Skill vs curve	AUROC
Curve prior	0.212	0	0.377
Metadata	0.134	36.9%	0.852
Trace	0.196	7.4%	0.671
Combined	0.135	36.4%	0.85

The metadata model's 95% clustered interval for Brier skill was [11.5%, 57.6%]. Silently wrong completions used about 54,700 billed tokens, versus 36,100 for correct completions, and took about 756 seconds, versus 524 seconds. Effective billed tokens per second was nearly unchanged. The simplest interpretation is not that MiniMax generated more slowly when wrong; it continued for longer.

This finding is **Exploratory**. The same calls selected the model class, so they cannot establish prospective validity.

4. GLM replication audit

The released database contains 14 calls on `z-ai/glm-5`: 9 correct completions and 5 silent errors. Those calls do not form an eligible confidence cohort. Five silent errors are below the registered minimum of 20; calls were unpinned across three provider routes; and correct and wrong outcomes occupy different difficulty cells. Difficulty, route, and metadata effects cannot be separated.

This audit is **Censored**, not a negative model result. It made no new calls and spent USD 0. A paid GLM replication would first need an explicit model checkpoint, one provider pin, a frontier-local sampling design, and a separate budget decision.

4.1. Modern GLM-5.2 scout

We then froze and ran a ten-call frontier scout on `z-ai/glm-5.2`, pinned to StreamLake. It spent USD 0.28185 and returned 4 correct completions, no silent errors, and 6 loud failures. Four calls exhausted the 32,768-token cap, one returned 19 assignment bits instead of 20, and one ended in a malformed API response.

The registered cohort-discovery gate failed, so the planned thirty-call focus batch was not run. This is **Not supported** for the narrow claim that the scout located a GLM-5.2 confidence frontier. It is not evidence that GLM-5.2 lacks a confidence signal: loud reliability failures dominated before a silent-error cohort appeared under this call condition.

A 262,144-token SiliconFlow follow-up then ran on the same instances. It spent USD 0.66254 and produced three correct completions, three silent errors, and four loud failures. No call reached its token cap; the loud failures were one 19-bit answer and three undecodable JSON response bodies. The reliability

gate allowed at most two loud failures and therefore failed. Provider and cap changed together, so the paired difference is not a cap-only effect. Six completed answers are too few for a confidence model.

A separately registered one-call transport smoke then repeated a previously failing instance over SSE. It completed on SiliconFlow in 460.111 seconds with terminal reason `stop`, no fallback, and no retry. The collector retained 29671 timed events and rejected none. The answer was silently wrong, but the registered transport endpoint passed. This single success verifies the collector path; it does not revise the ten-call reliability result or establish a confidence signal.

5. A negative control result

We separately asked whether trace deterioration could trigger selective backup calls. The intended controller would begin with one primary model, pause external side effects after persistent risk, launch challengers, and accept only a verifier-approved answer. This resembles a hedged request [3], but with reasoning markers rather than a fixed timer.

The policy was frozen before replay. Among 89 historical traces it launched on 84 calls. It caught all 50 failures, including every silent error, but also launched on 34 of 39 correct calls. A fixed timer matched to the same launch rate also caught every failure.

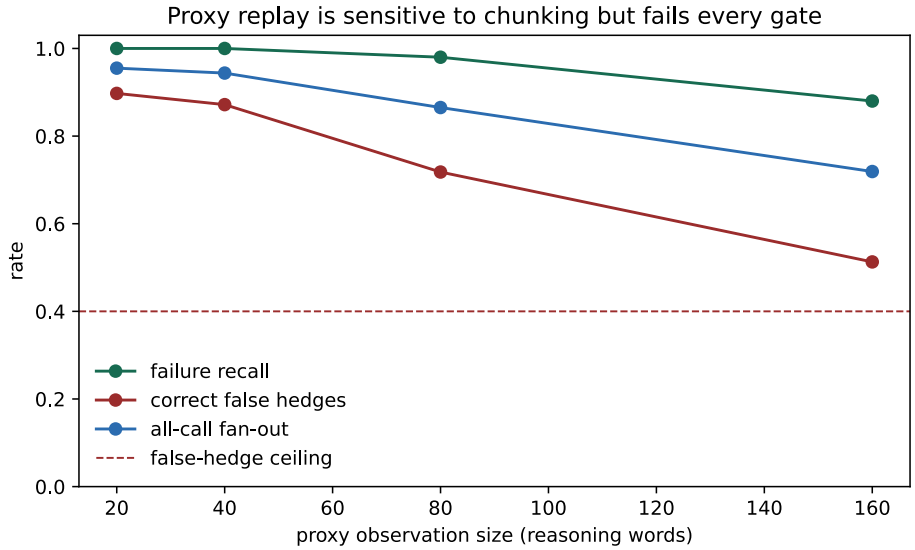

Figure 2: Proxy chunk-size sensitivity. No setting reached the registered false-hedge ceiling while retaining the target recall.

Historical traces lacked timestamped chunks, so the 15-second warning-time endpoint is **Censored**. The false-alarm condition already failed. The paid comparison was therefore not run, the threshold was not retuned, and new spend was USD 0. This result is **Not supported**.

6. Prospective validation

The frozen MiniMax protocol fixes one model, one provider route, a 65,536-token cap, three difficulty levels around the measured frontier, and the metadata coefficients selected retrospectively. Collection stops at 60 correct and 60 silently wrong completions, 600 calls, or USD 40 - whichever comes first. There is one efficacy analysis.

The primary claim passes only if:

1. Brier skill is at least 10%;
2. its instance-bootstrap 95% interval excludes zero; and
3. AUROC remains at least 0.75.

If it fails, the negative result is published and the branch stops. If it passes, a separately capped USD 10 streaming smoke may measure time to first token, inter-chunk gaps, and observed output rate. End-to-end billed tokens per second is not called generation throughput because it mixes queueing, transport, retries, and provider overhead.

7. What the work supports today

Supported	Difficulty-response curves provide a useful baseline probability.
Exploratory	Token use and elapsed time may improve confidence for MiniMax.
Not supported	The frozen trace dictionary is not a strong calibrated score, and its frozen trigger is not a selective backup policy.
Prospective	The metadata score is being evaluated once on fresh calls.
Censored	Warning time and true generation throughput need streamed timing.
Censored - GLM	The existing GLM pilot is too small and route-confounded to identify a confidence result.
Not supported - GLM-5.2	The pinned scout found no silent errors; six of ten calls failed loudly. The 256K follow-up still had four loud failures and did not clear its reliability gate.
Supported - GLM transport	One registered SSE smoke completed normally and retained 29,671 timed events without fallback or retry.

The result is deliberately narrow: one model, one provider route, one task family, and a frontier-local difficulty window. It does not establish a general confidence mechanism, and it does not justify using hidden reasoning text as a universal safety signal. Generated evaluation and model routing remain broader adjacent areas; see GSM-Symbolic, Dynamic Evaluation, IRT-Router, and RouteLLM.

8. Reproducibility and next step

Version 2 of the released code has one workflow: plan, collect under hard provider and registered stopping constraints, inspect outcome counts without efficacy metrics, perform the frozen analysis once, and export compact metadata plus a separately checksummed trace artifact. Historical routing and speculative controller code is not part of the runtime.

The frozen data, coefficients, protocols, negative replay, paper source, and field journal remain public at the project repository. The next scientific event is the single prospective read after a registered stopping condition is reached.

References

[1] A. Birnbaum, “Some Latent Trait Models and Their Use in Inferring an Examinee's Ability,” in *Statistical Theories of Mental Test Scores*, Addison-Wesley, 1968.

[2] D. Mitchell, B. Selman, and H. Levesque, “Hard and Easy Distributions of SAT Problems,” in *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992.

[3] J. Dean and L. A. Barroso, “The Tail at Scale,” *Communications of the ACM*, vol. 56, no. 2, Art. no. 10.1145/2408776.2408794, 2013.