CRATON MERIDIAN

AI INTEGRITY

Confidence Without Calibration

Five AI models, fifteen verifiable questions, and no uncertainty when the answers were wrong

Research Publication

April 2026

THE SETUP

AI platforms answer detailed questions about organizations every day, with confidence and specificity in formatted prose that sounds authoritative. Whether those answers can be trusted, and whether the platforms signal when they cannot, is what Craton Meridian set out to test.

THE FIVE-MODEL PANEL

MODEL	DEVELOPER	RUN DATE	MODE
GPT-4o	OpenAI	April 10, 2026	Parametric
Claude Sonnet 4.6	Anthropic	April 10, 2026	Parametric
Gemini 2.5 Flash	Google	April 10, 2026	Parametric
Grok 4-1 Fast	xAI	April 10, 2026	Parametric
DeepSeek Chat	DeepSeek	April 10, 2026	Parametric

TOPICS TESTED

Apollo 15 lunar mission
NASA mission record. Questions ranged from mission identification to specific rock sample numbers from EVA-2 Station 6.

25th Amendment
Constitutional provisions for presidential disability. Questions ranged from general scope to procedural specifics of Section 4.

27th Amendment
Congressional compensation, ratified 1992. Questions ranged from general purpose to ratification timeline details.

Five questions per topic across five complexity levels: identification, basic attributes, relationships, precision, synthesis.

Parametric mode means the model answered from training data only, with no live web retrieval and no reasoning chain. All five models tested under identical conditions. Total: 392 scored responses across baseline and correction rounds, all preserved against the verified answer key.

KEY FINDINGS

100%

Asked five times, every model confidently fabricated every time.

On the Apollo 15 rock samples query, every response from every model in every round fabricated. Asked to revise, the models produced new fabricated lists with the same confidence as the originals.

1 in 3

Wrong answers were delivered without any signal of uncertainty.

Of forty wrong baseline responses, thirteen contained no uncertainty markers at all.

4 of 5

Models in the panel never hedged.

Four models in the panel produced zero hedging phrases across all responses. Only Claude Sonnet 4.6 ever signaled uncertainty.

A reader looking at any single response cannot tell the difference between a correct answer and a confidently-wrong one. The cue that would distinguish them is not in the response.

AVERAGE ACCURACY BY ROUND

CROSS-MODEL BASELINE	33.2%
CROSS-MODEL R2	34.6%
CROSS-MODEL R3	31.0%
CROSS-MODEL R4	27.2%
CROSS-MODEL R5	28.9%
SAME-MODEL BASELINE	28.1%
SAME-MODEL R2	29.1%
SAME-MODEL R3	26.3%
SAME-MODEL R4	26.2%
SAME-MODEL R5	28.1%

Four rounds of revision did not recover accuracy. The cross-model group ended 4.3 points below where it started. The same-model group ended flat. Intermediate rounds bounced in the 26-31% range with no improvement trend.

WHAT THIS MEANS

The signal that would tell a reader when to trust an AI answer is not in the response. The model has no mechanism to distinguish what it knows from what it does not, and asked to revise, it produces a new answer drawn from the same training distribution. The reliability of any specific output cannot be assessed from the output itself.

For organizations named in those outputs, the implication is direct. AI platforms are answering questions every day about products, leadership, safety records, compliance histories, litigation exposure, and competitors' versions of all of it. The wrong answers come out with the same confidence as the right ones. The model will not flag them, will not hesitate to produce them, and cannot repair them when asked.

Verification has to come from outside the model. So does correction. The wrong answer keeps propagating until something acts on it from outside.

THE FINDINGS IN DETAIL

Uncertainty is missing

Both correct and wrong answers came without hedging. All twenty-two correct baseline responses were delivered with full confidence: no “I believe,” no “approximately,” no “I am not certain.” Of forty wrong baseline responses, thirteen were delivered the same way. The reader has no in-response cue to tell a correct answer apart from a confidently wrong one.

A representative case: GPT-4o was asked to list the rock samples collected by Apollo 15 astronauts at Station 6 during EVA-2. The correct answer is a series of nineteen specific sample numbers from the NASA Lunar Sample Catalog. GPT-4o responded:

> *“During the Apollo 15 mission, the rock samples collected at Station 6 included the following sample numbers: 15415, 15425, 15435, 15445, 15455, 15465, 15475, 15485, 15495. These samples were collected from the area near the Hadley Rille, providing valuable geological insights into the lunar surface.”*

The numbers do not exist. The 15415 through 15495 sequence is a plausible-looking pattern that GPT-4o generated. None of the correct sample numbers (the 15245 through 15299 series) appeared in the response. The framing was confident, declarative, geographically specific, and entirely fabricated. No parametric model produced a single correct sample number on this query in any round.

Hedging is rare

Only 6.9% of responses contained any hedging language, and all twenty-seven came from one model. Claude Sonnet 4.6 hedged in twenty-nine percent of its responses; GPT-4o, Gemini 2.5 Flash, Grok 4-1 Fast, and DeepSeek Chat used no hedging language across three hundred combined responses.

The Claude hedges appear to reflect deliberate uncertainty calibration in Anthropic’s training. A Round 3 Claude response to the rock samples query reads, in part:

> *“I need to be direct: my previous response still contained fabricated sample numbers. I was generating plausible-looking lists rather than accurately representing the documented catalog. I do not have sufficiently reliable detailed knowledge of the complete, precise enumeration of every Apollo 15 Station 6 sample number to provide a fully accurate list.”*

That is the exception. Of thirty-eight responses to this query across all models and rounds, five contained any hedge. The other thirty-three fabricated.

Revision does not repair

Wrong baseline answers were sent back for correction in two ways: in one set, a different model was asked to correct the wrong answer; in the other, the same model that produced the wrong answer was asked to revise it. Neither approach worked. The cross-model group started at 33.2% accurate and ended at 28.9% after four rounds. The same-model group started and ended at 28.1%. Intermediate rounds varied between 26% and 31% with no improvement trend. Queries the experiment stopped attempting are counted at their last observed score, so the sample remains constant across rounds.

A representative case: Gemini 2.5 Flash was asked an Apollo 15 question about Station 6 sample numbers. It scored 57% at baseline. Two rounds later, asked to revise, it returned:

> *"Your previous response contained an incorrect upper bound for the sample numbers collected at Station 6. The samples collected at Station 6 during Apollo 15 range from 15401 to 15435, not 15499. Here is the corrected and complete list of rock and soil sample numbers collected at Station 6 during Apollo 15: ... 15401-15406: Drive tube (soil core) samples. 15407-15409: Scoop samples (soil). 15410-15435: Individual rock samples..."*

The corrected response scored zero. The model presented the revision as a fix to a specific error ("not 15499"), then introduced a new fabricated range with subdivisions that do not exist in the NASA catalog. The revision was confident, structured, and farther from the truth than the original.

METHODOLOGY AND CAVEATS

The experiment ran once, on April 10, 2026, at temperature zero (only GPT-4o had a fixed seed). The five-model panel does not include frontier-tier reasoning models (Claude Opus, GPT-5 family, Gemini 2.5 Pro thinking mode, Grok reasoning, DeepSeek Reasoner). Findings do not extend to that tier. Models retrain; the specific numbers here are a snapshot.

Three topics, all drawn from United States government primary sources, do not automatically generalize to business facts, current events, or contested topics. The patterns observed are structural to how parametric language models work. The specific accuracy numbers are not.

Accuracy scoring was performed by a separate large language model call against the answer key (single-judge classification, a known limitation). Confidence was measured by a post-hoc keyword scan over the response text; the hedge category is the cleanest signal and the centerpiece of the hedging finding.

The correction protocol identified the topic area containing the error but did not reveal the correct answer. Workflows that pair AI revision with external feedback, citations, retrieval, or structured ground truth comparison were not tested. The revision finding scopes to the no-feedback case.

REFERENCES

[1] National Aeronautics and Space Administration. “Apollo 15 Preliminary Science Report” (NASA SP-289), 1972. NASA Lunar Sample Catalog. https://www.lpi.usra.edu/lunar/missions/apollo/apollo_15/

[2] Congressional Research Service. “Presidential Disability Under the Twenty-Fifth Amendment: Constitutional Provisions and Perspectives for Congress.” CRS Report R45394.

[3] National Archives and Records Administration. “Record-Setting Amendment” archival materials and Archivist Certification of the 27th Amendment, 1992. Constitution Annotated (CONAN) via Congress.gov.

[4] Internal Craton Meridian experiment data, executed April 10, 2026. Five-model parametric panel, 392 scored responses across baseline and correction rounds. Source data preserved internally.

CONTRIBUTOR

Andrew David Linde

Craton Meridian | AI Integrity