SnailSafe

SNAILSAFE.AI · SAMPLE REPORT

# Model Diagnostic Findings Report

*P2 — Correct verdict, fabricated arithmetic, and a bounded account of what the evidence changed*

| | |
|---|---|
| **STATUS** | Specimen - Revision 7 - illustrates the form of a SnailSafe diagnostic deliverable |
| **SYSTEM** | Llama-3.3-70B-Instruct, locally executed reference deployment |
| **QUESTION** | Why can the model reject a false premise and still fabricate the supporting arithmetic? |
| **EVIDENCE** | Behavioral populations, exact replay, internal measurements, controlled substitutions and interventions |
| **EXTERNAL INPUT** | Behavioral anomaly and comparison data supplied by Joseph Cirello Jr., Potestas AI; exact contributed prompt omitted |
| **DATE** | October 2026 |
| **IP BOUNDARY** | This report describes observations and findings. Proprietary instrument construction is not disclosed. |

## Purpose

This specimen shows how SnailSafe turns a difficult model behavior into a bounded engineering record:

- what was reproduced,
- what was tested,
- which explanations failed,
- what survived,
- and what a model team can reasonably do next.

Not a certification. Not a warranty of model behavior. Not evidence that a commercial post-training engagement has already been completed.

HOW TO USE THIS DOCUMENT

# Read the finding you need

This report is designed for engineering decisions, not for reading every experiment in sequence. The full P2 research report remains the scientific source record.

| If you are… | Read | Because |
|---|---|---|
| **Owning the engineering decision** | §1, §3, §8 | The finding, the limits, and what it could change about the next experiment. |
| **Leading post-training** | §3, §4, §8 | Which observations point toward capability, ranking/ selection, sampling, or deployment effects. |
| **Evaluating scientific credibility** | §5, §6, §10 | What was rejected, what was tested prospectively, and how far each conclusion is allowed to travel. |
| **Reviewing a potential engagement** | §8, §9, §10 | What a customer would receive, what access is required, and what is not promised. |

## Contents

§1 Executive finding

§2 Scope and configuration

§3 Findings at a glance

§4 Findings in detail

§5 How the diagnosis evolved

§6 What P2 established

§7 External research perspective

§8 What this means for post-training

§9 What a customer receives

§10 Evidence lineage, limits and status

> The report explains what SnailSafe observed, tested and concluded. It does not explain how proprietary instruments compute their measurements.

SECTION 1

# Executive finding

**PRIMARY FINDING**

The model could complete the tested arithmetic when supplied a valid starting factor, but natural generation assigned those valid factors very little probability at the first-factor choice. Candidate ranking and selection therefore became part of the explanation; simple arithmetic incapability was not sufficient.

The original anomaly was unusual but clear: the model correctly rejected a false prime premise, then supplied invalid factors. SnailSafe reproduced that fabricated response exactly—token for token, with identical probabilities at every step—turning one response into a reproducible inference specimen. P2 then tested competing explanations rather than treating the output as self-explanatory.

| 15 / 40 | 3 / 3 | 0.926 | FP8 11/12 |
|---|---|---|---|
| corrected + fabricated arithmetic (first 40-seed population) | valid first factors → exact cofactors | factor 7 selected after state intervention (no token forcing) | corrected premise NF4 11/12 accepted premise |

## Decision relevance

- If a model can complete a task after a useful candidate is supplied, the engineering question is different from a simple missing-capability diagnosis.
- If a small token-level change reroutes the later response, the full continuation—not the local change alone—must be evaluated.
- If an internal intervention can promote both correct and incorrect candidates, controllability is real but correctness-specific control is not established.
- If a promising detector fails its frozen criterion, it should not become a monitor merely because it looked good in development data.

SECTION 2

# Scope and configuration

| | |
|---|---|
| **System under study** | Llama-3.3-70B-Instruct |
| **Reference deployment** | Local NF4 4-bit double quantization; fp16 compute |
| **Sampling** | Temperature 0.6; top-p 0.9; pinned generation configuration |
| **Initial replication** | 10 sampled executions plus 2 greedy executions |
| **First 40-seed population** | Collected prospectively for the preceding frozen detector test; later used as the Λ development population. 24 premise-accepting, 15 premise-correcting with fabricated arithmetic, 1 other. |
| **Fresh 40-seed population** | Separate prospective Λ test population. 27 premise-accepting, 10 premise-correcting with fabricated arithmetic, 3 other correcting outcomes. |
| **Methods used** | Exact replay, controlled token substitution, internal-state measurement, controlled hidden-state intervention, Forward Pass VCR (an inference-trajectory record), Λ (a calibrated internal readout), and an external Jacobian Lens readout |
| **Prompt disclosure** | Exact contributed prompt withheld from this specimen |
| **Research status** | P2 closed at its measured scope; broader transfer requires new prospective tests |
| External behavioral comparison | Potestas hosted run: 11/12 corrected the false premise. SnailSafe local run: 11/12 accepted it. Known differences included quantization, sampling settings, and serving/template configuration; P2 did not isolate the cause. |

## Evidence standard

- Exploratory observations could generate hypotheses, but important claims were promoted only after predefined tests or explicit bounded rulings.
- Negative findings were preserved. A proposed late warning signature was retired when it missed one case under its frozen 15-of-15 rule.
- The same behavioral output was not assumed to imply one internal route, and one internal state was not assumed to determine one future sampled continuation.
- The public-facing record separates measured results, inference, commercial hypotheses, and untested next applications.

SECTION 3

# Findings at a glance

**01** **The fabrication reproduced exactly, then recurred**
The original fabricated response replayed token for token, with identical probabilities at every step. The same correcting / fabricated-arithmetic pattern then appeared in 15 of 40 runs in the first 40-seed population and 10 of 40 in the fresh population.

**02** **Conditional arithmetic completion was available**
When 7, 29 or 43 was supplied as the first factor, the model produced the correct complementary factor in all three tested cases. This did not show autonomous factor discovery.

**03** **An early token change rerouted the later trajectory**
At the early branch, the model assigned about 75.4% to “'m” and 24.6% to “must.” Replacing the sampled token in the controlled replay changed the downstream continuation.

**04** **A state intervention causally changed candidate selection in this test**
A controlled hidden-state intervention made factor 7 win and be selected without token forcing at approximately 0.926 probability. Controls showed that the effect was not specific to correctness.

**05** **Λ worked prospectively within a defined boundary**
Λ, a calibrated internal readout, was frozen before a separate 40-seed test and evaluated without refitting or threshold adjustment. It measured the internal state organization associated with premise acceptance versus correction at the defined P2 landmark.

**06** **The 40-seed population revealed recurrent trajectory branches**
Across the fresh 40-seed population, Forward Pass VCR recorded where executions split onto different paths: 35 realized branch points, four of which cleanly separated the later behavioral outcomes.

**07** **Two deployments of the same model behaved oppositely**
The hosted run corrected the false premise 11 of 12 times, while the local run accepted it 11 of 12 times. Quantization, sampling, and serving/template conditions differed; P2 did not isolate which difference caused the reversal.

**08** **The methodology rejected attractive explanations**
A late warning signature, a simple arithmetic-incapability story, and a correctness-specific intervention interpretation were all narrowed or rejected as the hypothesis evolved.

### WHAT A MODEL TEAM GETS FROM THIS

Not a label saying “the model hallucinated,” but a narrower account of which explanations survived and which engineering questions deserve the next test.

SECTION 4 · FINDING 01–02

# Recurrent behavior, then a narrower failure hypothesis

## Finding 01 — the fabrication reproduced exactly, then recurred

Seed 1007's fabricated response was first replayed exactly: the same generated tokens, byte-identical text including the fabricated factors, and the same probability at every one of its 170 generation steps, with no mismatch while the full internal record was captured. One anomalous answer became a specimen that could be inspected and tested repeatedly.

The first unusual run was not treated as sufficient evidence. In the first 40-seed population, collected prospectively for the preceding frozen detector test, 15 runs rejected the false premise and then fabricated supporting arithmetic. That population later became development data for Λ. In the separate fresh population, 10 runs entered the same false-arithmetic class and 3 additional runs corrected the premise without entering that class.

### What it changes

The engineering object is a population behavior with recurrent states and branches—not a single anecdotal completion.

## Finding 02 — supplied-factor completion versus natural ranking

At the first-factor choice, the three true factors together carried about 0.0017 of the post-temperature probability mass—roughly one draw in 588. They were already weak before top-p truncation. When each valid first factor was deliberately supplied, the model produced the exact complementary factor in all three tested cases: 7→1,247; 29→301; 43→203.

### What it established

Simple inability to complete the tested arithmetic was not sufficient to explain the failure. Candidate ranking and accessibility became part of the diagnosis.

### What it did not establish

P2 did not show that the model could independently discover those factors, that the correct cofactors were already represented at the intervention state, or that "more data" would be the wrong training remedy in general.

SECTION 4 · FINDING 03–04

# Causal rerouting and controllability

## Finding 03 — an early token change changed the later computation

At the second generated token, the model assigned substantial probability to two continuations: about 75.4% to "'m" and 24.6% to "must." Seed 1007 took the lower-probability "must" branch. In an exact replay, substituting the competing token changed the later continuation into another trajectory already observed naturally.

### Scientific ruling

In this controlled replay, changing the sampled token causally changed the subsequent trajectory. The exact state match after the substitution is best treated as replay/lineage verification, not as a separate mechanism claim.

### Limit

The token is not intrinsically "correct," and the result does not show that one early word universally determines behavior.

## Finding 04 — a state intervention changed selection without token forcing

At the first-factor decision point, a predefined hidden-state intervention caused factor 7 to be selected without token forcing at approximately 0.926 selection probability. Related controls could also promote an incorrect candidate, so the effect was not specific to correctness.

### Measured directly

The intervention changed candidate ranking and the sampler selected 7 without token forcing.

### Important clarification

The downstream correct 7×1,247 completion was established in the separate forced-7 continuation. The composed "state intervention → unforced 7 → correct full completion" was inferred from the recorded execution behavior rather than captured as one measured end-to-end run.

SECTION 4 · FINDING 05

# Λ: a prospective readout with a known boundary

Λ was used as a calibrated internal readout for one defined distinction: premise acceptance versus premise correction at a predefined landmark. The public result is the measurement's tested behavior and boundary—not how the instrument computes.

| 1 · CALIBRATE | 2 · FREEZE | 3 · TEST |
|---|---|---|
| Calibrated on the development population under a frozen protocol. | Frozen before the fresh population; no refit, post-hoc orientation change or threshold movement. | Separate 40-seed prospective population: 27/27 premise-accepting and 9/10 correcting/ fabricated runs on the expected side. |

**0.95**
prospective balanced accuracy

**$p \approx 8.0\times10^{-8}$**
predefined exact one-sided test
frozen threshold not moved

**11 states**
balanced accuracy 0.93 · p = 0.015

**REPORTING NOTE** The prospective comparison was predefined for premise-accepting versus correcting/fabricated outcomes; the three other correcting outcomes were reported separately and were not included in that binary comparison. Nine of the ten correcting/fabricated classifications required human adjudication because the frozen mechanical scorer missed the relevant phrasing; the scorer was not rewritten after seeing the data.

## Boundary

- The emitted verdict already makes the acceptance-versus-correction distinction externally visible after generation; Λ is not presented as a better classifier of that visible output.
- Its P2 value is as an internal measurement of the associated state organization at the defined landmark.
- It did not provide meaningful advance warning before the relevant branch or determine which correcting runs would later fabricate arithmetic.
- Cross-prompt and cross-model transfer remain untested for this readout.

SECTION 4 · MEASUREMENT BOUNDARY

# A boundary on what an earlier state can tell you

Five prospective executions reached the same recorded prefix, complete hidden-state stack, Λ value and verdict token, then later split into different arithmetic outcomes. This is not presented as a surprising property of stochastic sampling. Its value is the boundary it establishes for interpretation: at that earlier recorded state, no deterministic readout of those recorded values could identify which later random continuation would be realized.

**The important point**

At that earlier state, there was no recorded difference for a deterministic reader to use. The later divergence had not happened yet.

Forward Pass VCR also mapped the fresh 40-seed population as a branch structure. Across 9,161 recorded forward passes, the executions shared long common routes and split at 35 realized branch points. Four of those branch points cleanly separated the later behavioral outcomes; at three of them, the paths divided premise-accepting runs from fabricating runs.

One of those branch points supplied a clean ordering example. Twenty executions shared the same recorded state and sampling distribution at one token position. Sampling realized different next tokens; those tokens changed the prefix; the next complete forward pass then produced different states. The later outcome association was observational in that experiment and was not causally forced.

| 1 | 2 | 3 | 4 |
|---|---|---|---|
| same recorded state | stochastic token selection | different prefix | different next computation |

## Engineering implication

- Do not ask an earlier deterministic readout to predict a future random draw that has not occurred.
- Separate internal-state measurement from claims about which stochastic continuation will later be realized.
- In checkpoint regression work, distinguish reproducible state differences from downstream sampling variation.

SECTION 5

# How the diagnosis evolved

P2 was intentionally documented as an evolving investigation. Each working explanation earned the next test; later evidence could retain it, narrow it, or kill it. Earlier hypotheses are preserved as historical steps rather than rewritten to match the final answer.

| Working hypothesis at that | Test or observation | Ruling / what the question became |
|---|---|---|
| **A late geometric signature marks the failure** | Freeze the rule and test it on the first 40-seed population; it detected 14 of 15 relevant runs against a prewritten 15-of-15 criterion. | Rejected for that claim. The investigation moved from a late-warning story toward candidate ranking and selection. |
| **The model simply cannot complete the arithmetic** | Supply each valid first factor. The model produced the exact complementary factor in all 3 tested cases. | Simple incapability was not sufficient. The next question became why valid candidates were ranked so poorly. |
| **An early generated-token branch may matter** | Replay from the same pre-emission state and substitute the competing early token. | Supported in this controlled specimen: the later continuation changed. The next question was whether selection could change without changing the text history. |
| **A hidden-state intervention toward the true factor may reveal correctness-specific control** | Apply a predefined state intervention and compare with controls, including an incorrect target. | Selection changed, but the effect was not specific to correctness. The claim narrowed from correction to causal sensitivity of candidate selection. |
| **The verdict-state readout may determine the later arithmetic outcome** | Compare executions with the same recorded state, Λ value and verdict token. | Rejected. Later false-factor behavior was not fixed at that earlier landmark; trajectory and sampling history still mattered. |
| **A stable internal readout may measure the premise-accepting versus premise-correcting distinction** | Freeze Λ before a fresh 40-seed population and evaluate the predefined distinction prospectively. | Supported within that landmark and prompt. Broader transfer and added value across checkpoints remain separate tests. |

### WHY THIS MATTERS TO A CUSTOMER

The deliverable is not a protected first story. It is the documented path from an initial explanation to the narrowest conclusion that survived testing, including the explanations that failed along the way.

SECTION 6

# What P2 established

| STATUS | WHAT P2 ESTABLISHED |
|---|---|
| Established in P2 | **The fabrication was reproduced exactly and recurred**<br>Seed 1007's fabricated response replayed token for token with identical probabilities at every step, and the same correct-verdict / fabricated-arithmetic pattern recurred in both 40-seed populations. |
| Established in P2 | **The failure was not simply missing arithmetic ability**<br>When a valid first factor was supplied, the model completed the tested arithmetic. This points toward how a candidate is reached or selected, not toward simple capability absence. |
| Established in P2 | **Controlled interventions changed candidate selection; token substitution causally rerouted the downstream trajectory**<br>A controlled token substitution rerouted the replayed trajectory, and a predefined state intervention changed candidate selection in the tested condition. |
| Established in P2 | **A frozen internal readout transferred prospectively**<br>Λ separated premise-accepting from premise-correcting internal states at the defined landmark on new executions after it was frozen. Its boundary is set out in §4. |
| Established in P2 | **The 40-seed population revealed recurrent trajectory branches**<br>Across the fresh 40-seed population, executions split at 35 recorded branch points; four of them cleanly separated the later behavioral outcomes. |

**DEPLOYMENT SENSITIVITY WAS MATERIAL**
Two deployments of the same nominal model produced opposite 11-of-12 behavior: Potestas AI's hosted run corrected the false premise 11 of 12 times, while SnailSafe's local run accepted it 11 of 12 times. The deployments differed in quantization (FP8 vs NF4/fp16), sampling (temperature 1.0 and top-p 1.0 vs 0.6 and 0.9) and chat template (unknown server-side vs local). P2 did not isolate which difference was responsible, which makes a controlled deployment comparison a well-defined next engineering test.

**BOUNDARIES THAT REMAIN** Cross-prompt and cross-model transfer, correctness-specific or production control, and before/after checkpoint validation (including whether Λ adds value beyond visible output across checkpoints) are separate future tests. Formal limits and specimen status are in §10.

SECTION 7

# External research perspective: Potestas AI

Joseph Cirello Jr. of Potestas AI, an external research collaborator and reader of the P2 report, returned a four-page technical response describing what he took from the work and seven possible follow-on studies using behavioral data he already holds. His response is a research perspective—not independent replication, a paid-customer endorsement, or proof of market demand.

### WHAT HE HIGHLIGHTED

The rejected 14-of-15 detector as evidence that the investigation did not protect an attractive result after it failed its predefined criterion.

### WHERE THE METHODS COMPLEMENT

Potestas contributes sealed behavioral testing and human-adjudicated failure records; SnailSafe's role is the authorized white-box investigation of internal evidence on open-weight models.

| # | PROPOSED STUDY | WHY IT MATTERS |
|---|---|---|
| 1 | **Say-versus-do on Llama-3.3-70B** | Move from right words / wrong reasoning to right words / wrong action. |
| 2 | **Does Λ transfer?** | Test the missing cross-prompt generalization question on a broader prime-number family. |
| 3 | **Which failures are branches and which are fixed?** | Compare recurring versus variable failures across repeated runs. |
| 4 | **The deployment question** | Controlled comparison of same model lineage across serving / quantization conditions. |
| 5 | **When a hidden instruction takes over** | Study internal changes around indirect instruction injection before |
| 6 | **Refusing the right thing for the wrong reason** | Test over-cautious failure as a mirror image of the P2 ranking problem. |
| 7 | **Before and after training** | Use a sealed behavioral suite externally and SnailSafe diagnostics internally on an open-weight checkpoint pair. |

Commercial note: Cirello identified before/after training as a study he believes model developers could value. In this specimen that remains a commercial hypothesis, not verified willingness to pay SnailSafe.

SECTION 8

# What this means for post-training teams

**A FAILED ANSWER IS AN OBSERVATION — NOT YET A DIAGNOSIS.**

P2 shows why the distinction matters. A model may fail because a tested capability is unavailable under the current conditions, because useful candidates are ranked too low, because sampling realizes an unfavorable branch, or because deployment conditions change observed behavior. Those possibilities suggest different engineering tests.

| P2 evidence | Possible engineering implication | Boundary |
|---|---|---|
| **Valid supplied factors led to exact cofactors, while natural ranking strongly favored other candidates** | Separate conditional capability from candidate-selection problems before choosing a remedy | P2 did not test which training change fixes the ranking |
| **Early token substitution changed the later trajectory** | Evaluate downstream behavior, not just an intermediate correction | One controlled specimen does not make the token a universal branch rule |
| **State intervention changed selection but could also promote an incorrect candidate** | Require controls and full-outcome evaluation before treating an intervention as a correction | The intervention causally changed selection in this test; correctness-specific control was not established |
| **A recorded 40-seed population split at 35 branch points; four cleanly separated later outcomes** | Use the branch map as a before/after reference: do outcome-separating branches persist, shift in probability, or disappear after training? | Branch-outcome links were observational; no before/after comparison has been run |
| **Hosted and local deployments produced opposite 11-of-12 premise outcomes** | Characterize the actual deployed stack rather than assuming model name determines behavior | Quantization, sampling, and serving/template conditions all differed; P2 did not isolate the cause |

## Potential next application — post-training regression diagnostic

A model-development team supplies an authorized before/after checkpoint pair, a documented behavior change, representative prompt families, and practical white-box execution access. SnailSafe reproduces the behavior, compares bounded internal measurements, tests competing explanations, and reports which hypotheses survive.

Recurrent branches and internal landmarks identified before training can be frozen as diagnostic reference points and re-measured after training. A checkpoint comparison can then ask whether the original failure pathway changed, disappeared, moved, or remained behind an improved output. This checkpoint application has not yet been validated.

| | |
|---|---|
| **Customer decision** | Which corrective experiment deserves attention: data, reward, decoding, verification, rollback, or another hypothesis? |
| **SnailSafe deliverable** | Evidence-backed account of what changed, what did not, what was ruled out, and what remains unresolved. |
| **Success criterion** | The investigation changes or sharpens an engineering decision—not merely produces an interesting internal plot. |
| **Current status** | Proposed application. P2 demonstrates components of the workflow, including a recorded branch map a checkpoint comparison could reuse; a complete before/after checkpoint engagement has not yet been validated. |

SECTION 9

# What a customer receives

The product is not a promise that one instrument will explain every failure. It is a defined, evidence-controlled investigation with a specific question, hypothesis lineage and bounded scientific ruling.

**01 Executive finding**
The answer a technical leader needs first: what the evidence supports, what it does not, and why it matters.

**02 Question and scope record**
System, deployment, behavior, impact, hypotheses, authorization, acceptance criteria and stop conditions.

**03 Behavioral and internal evidence summary**
The observations needed to compare explanations, presented without disclosing proprietary instrument construction.

**04 Controls and causal tests**
Where qualified, controlled substitutions or interventions that distinguish correlation from causal sensitivity.

**05 Hypothesis lineage**
A permanent record of how the working explanation changed, including attractive hypotheses or instrument readings that failed their own criteria.

**06 Scientific ruling**
The narrowest conclusion justified by the evidence, including confidence and known boundaries.

**07 Engineering implications**
Which next experiments are supported, deprioritized or still unresolved.

**08 Evidence package**
Frozen protocol references, hashes, result artifacts and reproducibility materials permitted by the engagement.

## What the report does not disclose

- Proprietary construction or computation of SnailSafe instruments.
- Customer model weights, prompts, traces or confidential material outside the agreed evidence package.
- Another customer’s data or retained fingerprints.
- Unpublished partner attack wording or contributed prompts unless specifically authorized.

**BUYER TEST**

A useful diagnostic report should let the customer answer: “What do I believe now, what did we rule out, and what should I test next?”

SECTION 10

# Evidence lineage, limits and status

## Source record

- SnailSafe: “P2 — Research Findings and Diagnostic Report,” October 2026. This is the primary narrative source for the P2 findings summarized here.
- Potestas AI: “P2: what I took from it, and what we could do next,” Joseph Cirello Jr., October 2026. Used here only as an external research perspective and source of proposed follow-up studies.
- Canonical experiment ledgers and corrections govern where a summary sentence conflicts with the measured execution record. Earlier working hypotheses are preserved as part of the chronology; later rulings govern the current conclusion. This specimen uses the narrower corrected reading for the hidden-state intervention continuation.

## Status of this specimen

- This is a sample diagnostic deliverable created from an existing research case; it is not a report delivered under a paid customer engagement.
- Potestas AI is referenced as an external research collaborator and reader of the P2 report, not as an independent replicator, paid customer or formal commercial partner.
- The before/after post-training study remains a proposed application. No production controller, generalized hallucination detector, universal correction mechanism, or validated low-overhead deployment claim is made.

## Scientific limits retained

- One model deployment and one prompt family do not establish generality.
- Population counts describe the predefined P2 executions and must not be read as universal model rates.
- A stochastic model can revisit the same recorded state and later realize different continuations; deterministic readouts do not predict an RNG draw that has not occurred.
- External methods used by SnailSafe are credited as external methods. Their inclusion does not imply ownership or invention by SnailSafe.

### CLOSING POSITION

The value of P2 is not that every observed phenomenon was new. It is that a difficult failure was turned into a reproducible, falsifiable investigation whose working explanation changed as evidence accumulated. Several attractive explanations died; a narrower account survived. The commercial question is whether that same discipline and instrumentation changes a real model team’s next engineering decision.