

May 10, 2026 Skynet Production Smoke-Test: Bounded Status Report

Exzil Calanza

Independent Researcher, Iloilo City, Philippines
Independent research preprint. IEEE-style formatting only; this work has not been accepted by, published by, certified by, or peer reviewed by IEEE.

ABSTRACT This report documents the May 10, 2026 production verification of Skynet port 8420 following a backend binary reload. Using controlled exact-token tasks, direct execution was confirmed for Gemini (gemini-3.1-pro-preview) and Codex (gpt-5.4), while Claude Opus and Sonnet task types completed via explicit fallback due to provider quota limits. Independent cross-validation confirmed operational status without overstating native provider availability.

INDEX TERMS AI swarm verification; smoke testing; fallback routing; exact-token validation; model registry; production health monitoring

I. INTRODUCTION

STUDY TYPE: Empirical field report. **RESEARCH QUESTION:** Did production port 8420 run the patched backend after a cooperative reload, and did registered model lanes return controlled results under a repeatable test design? This preprint formalizes the existing public evidence record as a empirical field report. No new experiment, dataset, measurement, or validation is claimed beyond the source article and its cited evidence.

A bounded production-status report on the May 10, 2026 Skynet backend reload, direct Gemini and Codex smoke tests, and the Claude task-type fallback boundary. The claim is deliberately narrow: measured routing worked for the sampled lanes, while native Claude execution remained quota-blocked during the observation window.

This bounded status report replaces the earlier short correction with an evidence-first smoke-test account of what was observed on May 10, 2026. The research question is narrow: did production port 8420 run the patched backend after a cooperative reload, and did the registered model lanes return controlled results under a repeatable test design? The answer is bounded. During the May 10, 2026 observation window, the patched production backend was observed live, the direct Gemini and Codex lanes returned exact tokens, and the Claude Opus/Sonnet task types returned exact tokens through an explicit Codex/GPT-5.4 fallback. Native Claude execution itself remained quota-blocked. [1]

The page deliberately avoids the language that made the prior version weak. It does not claim a flawless system, permanent

autonomy, or unqualified native Claude availability. It reports observations, methods, limitations, and reproducibility criteria. If a value was not measured, it is not promoted to a fact. If a result depends on fallback routing, the fallback is named. That is the minimum standard for a public technical post about an autonomous AI swarm verification event. [4]

II. MAY 10 PRODUCTION PROBE - ROUTING REALITY VS CLAUDE BOUNDARY

III. RESEARCH QUESTION AND SCOPE

The useful scientific question is not whether Skynet is "perfect." Perfect is not an engineering measurement. The useful question is whether a specific production system, at a specific time, produced observable outputs that match a defined test plan. This report uses that framing. It treats the backend reload, worker registry, smoke-task output, independent review, and provider-limit boundary as separate variables rather than blending them into a marketing claim. The scope is deliberately limited. A smoke test is a fast verification that the most important path is alive after a change. It is not a proof of long-run stability, security, throughput under sustained load, or future model availability. The experiment therefore supports one bounded conclusion: the patched production backend was live and cross-lane task routing worked for the sampled Gemini, Codex, Claude Opus task-type, and Claude Sonnet task-type routes. It does not prove that all future tasks will succeed or that direct Claude execution was available during the observation window.

XVIII. LIMITATIONS AND CLAIM BOUNDARY

The verification covers a bounded observation window. Direct Claude execution was quota-blocked, and the report does not evaluate long-run throughput, stress conditions, security completeness, or guaranteed future uptime.

Claim boundary - Present Claude task-type successes explicitly as operating through Codex/GPT-5.4 fallback rather than native Claude execution.

XIX. REFERENCES AND SOURCE RECORD

The method used controlled token-return tasks because exact-token output is easy to falsify. Each lane received a task whose success condition was not subjective prose quality but the presence of a specific string in the result payload. That design matters because model systems can produce fluent explanations even when the underlying routing path is wrong. An exact token, a task ID, a return code, a duration, and a fallback marker create a tighter evidence chain.

The first data source was the production health and status probe on port 8420. It reported backend health ok, timestamp 2026-05-10T16:31:00+08:00, 22 registered agents, 0 unknown model records, 22 dead workers, 0 workers currently working, and connection saturation false. The second data source was the post-reload smoke report generated after the binary swap. The third source was independent model review: Gemini accepted the verdict and Codex/GPT-5.4 rated the evidence strong, while direct Claude Opus and Sonnet validation attempts hit a provider quota limit. [1] [2] [3]

V. PRODUCTION RUNTIME OBSERVATIONS

- Backend health: ok
- Backend timestamp: 2026-05-10T16:31:00+08:00
- Production PID after reload: 22336
- Backend uptime at report: 102.770946 seconds
- Registered agents: 22
- Unknown model records: 0
- Dead workers: 22
- Workers currently working: 0
- Workers with reported errors: 0
- Connection saturated: false
- Patched binary size: 11485184 bytes
- Candidate binary size: 11485184 bytes

The runtime observation matters because the May 9 version of this page reported stale numbers: fewer workers, unknown

model identities, and no completed task proof. Those figures are historical correction data. The May 10 probe showed a resolved model registry and a non-saturated connection state. That does not mean the system is unconstrained; it means the sampled control surface was not reporting the earlier degraded state during the May 10 verification window. [1]

VI. MODEL REGISTRY SNAPSHOT

- claude-opus-4-7-prime : 2
- claude-sonnet-4.7 : 1
- gemini-3.1-pro-preview : 18
- gpt-5.4 : 1

The registry snapshot is important because worker names alone do not prove model identity. A worker called gemini-1 is not evidence unless the status surface reports the model binding. During this verification, the Gemini worker group reported gemini-3.1-pro-preview, the Prophet lane reported gpt-5.4, and the Claude task-type lanes reported their Claude model labels while also relying on the fallback marker when native Claude quota blocked execution. [1] [4]

VII. WHAT WAS FIXED SINCE THE RETRACTION

- The prior public status page was a stale May 9 correction snapshot and no longer reflected the production state.
- The production binary was safely swapped to the patched backend and reloaded through /reload .
- Claude quota failures now produce a truthful fallback path instead of silently blocking Claude task types.
- Claude fallback outputs include the explicit marker `FALLBACK_USED: claude_quota_to_codex_gpt-5.4` .

The important repair was not cosmetic. The public claim changed only after production 8420 was observed running the patched binary and after the sampled task routes returned exact-token outputs. The fallback design also changed the truth surface. Instead of pretending Claude is natively available, the system exposes a fallback marker so the reader can distinguish "Claude task type completed" from "Claude provider completed natively." That distinction is the difference between an audit-grade report and an ambiguous success banner. [2] [4]

VIII. POST-RELOAD SMOKE PROOF

TABLE 1

Lane	Worker	Status	Expected token	Fallback	Duration	Task ID	Latest results
Claude Opus task type via fallback	claude-opus	success	skynet-prod-opus-fallback-ok	yes	10853.148 ms	task_1778396122655455100	yes
Claude Sonnet task type via fallback	claude-sonnet	success	skynet-prod-sonnet-fallback-ok	yes	12464.190 ms	task_1778396122655997400	yes
Gemini direct lane	gemini-1	success	skynet-prod-gemini-ok	no	14528.384 ms	task_1778396122653566100	yes
Codex/GPT-5.4 direct lane	prophet	success	skynet-prod-codex-ok	no	7555.961 ms	task_1778396122654370000	yes

The table is the core result. A language model can sound convincing while still failing an integration test; therefore the pass condition here was a token match, not rhetorical quality. The Gemini lane returned skynet-prod-gemini-ok. The Codex/GPT-5.4 lane returned skynet-prod-codex-ok. The Claude Opus task type returned skynet-prod-opus-fallback-ok and the Claude Sonnet task type returned skynet-prod-sonnet-fallback-ok. Both Claude task-type successes must be read with the fallback column, because the native provider was not available during the direct validation attempt. [2] [4]

IX. EXECUTION PATH EVIDENCE STRENGTH

X. INDEPENDENT CROSS-VALIDATION

- Gemini gemini-3.1-pro-preview accepted the verdict production_8420_reloaded_and_cross_lane_smoke_passed_with_claude_quota_fallback .
- Codex/GPT-5.4 judged the evidence strong because the post-reload PID was observed, the model registry had 0 unknown records, and 4/4 smoke tasks returned exact tokens.
- Direct Claude Opus and direct Claude Sonnet validation did not run natively because both returned the quota message: You've hit your limit - resets May 13, 8am (Asia/Manila) . Cross-validation is useful only when it can disagree. Here it did. Gemini and Codex accepted the bounded production verdict, but the direct Claude attempts supplied a negative

Question	Observation
Was production reloaded?	New PID observed after /reload
Did direct Gemini work?	Exact Gemini token returned
Did direct Codex work?	Exact Codex token returned
Did Claude task types work?	Exact fallback tokens returned
Did native Claude work?	Provider quota message returned

XIII. REPRODUCIBILITY CHECKLIST

A future verification run should reproduce the result without relying on the prose in this article. The minimum repeatable checklist is: capture /health, capture /status, record the production PID, run exact-token tasks against Gemini and Codex direct lanes, run Claude Opus and Sonnet task-type probes, preserve result task IDs, record whether the fallback marker appears, and independently review the evidence with at least one model that did not produce the implementation. This is the difference between a status update and a scientific report. [5] [6]

The next stronger experiment is a generational validation round after the Claude quota resets. That round should include a native Claude smoke, the same Codex/Gemini controls, repeated runs across time, and an explicit comparison between fallback and native Claude behavior. Until that data exists, the correct public statement remains the bounded one used here. [4]

XIV. KEY TAKEAWAYS

control: they did not prove native Claude execution because they failed at the provider quota boundary. That disagreement improves the report because it prevents a stronger claim than the evidence supports. The public conclusion therefore separates "task type operational through fallback" from "native provider operational." [3] [4]

XI. TRUTH BOUNDARY AND NON-CLAIMS

The bounded May 10, 2026 observation is: the patched production backend was live during the verification window. Gemini and Codex executed directly. Claude Opus/Sonnet task types were operational through explicit Codex/GPT-5.4 fallback while native Claude execution remained quota-blocked. This claim is supported by the observed post-reload PID, the resolved status registry, the exact-token smoke tests, and the fallback marker. [1] [2]

The truthful current non-claim is equally important: native Claude Opus/Sonnet execution is not verified live right now. It must be re-smoked after the quota resets on **May 13, 2026 at 8:00 AM Asia/Manila**. The report also does not claim long-duration stability, security completeness, model-quality superiority, or guaranteed future uptime. Those would require different experiments: load testing, fault injection, security review, repeated provider probes, and longitudinal monitoring. [4]

XII. SCOPE OF CLAIMS

TABLE 2

Interpretation	Status
Patched backend is live	Supported
Sampled direct Gemini lane passed	Supported
Sampled direct Codex lane passed	Supported
Claude task types passed through Codex fallback	Fallback-supported
Native Claude was not verified	Not supported

- Use evidence nouns, not status adjectives: "4 of 4 smoke tasks returned exact tokens" is stronger than "works perfectly."
- Keep fallback visible: Claude task-type success is valid only when the Codex/GPT-5.4 fallback marker remains attached.
- Separate provider health from task health: native Claude was quota-blocked even though Claude task types completed through fallback.
- Treat stale public pages as incidents: the May 9 snapshot was accurate for its time but wrong as a current status page after the May 10 reload.
- Make the next test harder: after the quota reset, rerun native Claude, fallback, Gemini, and Codex in the same controlled harness.

XV. EMBEDDED EVIDENCE RECORD

The local evidence bundle used to build this public article is preserved under the ScreenMemory workspace. The file names are included because the backend evidence is local-first and not a public API surface. The public WordPress REST

record is also linked in the references so readers and crawlers can verify the current title, excerpt, and rendered content. [5] [6]

- data/truth_reports/skynet_production_8420_post_reload_cross_validation_20260510T1455.json
- data/truth_reports/skynet_production_8420_post_reload_cross_validation_20260510T1455.md
- data/public_evidence/skynet-prophet-autonomous-swarm-live-2026-05-07_20260510T163104+0800_production_truth.json

XVI. RELATED READING

- AI WP Manager Blog
- AI WP Manager Platform
- More AI agent analysis

XVII. REFERENCES

AUTHOR BIOGRAPHY

EXZIL CALANZA is an independent researcher and software builder based in Iloilo City, Philippines. His work focuses on browser and desktop automation, reproducible evaluation, agent reliability, workflow verification, and AI-assisted systems engineering.

- [1] Current production probe section , generated from live Skynet /health and /status observations, accessed May 10, 2026.
- [2] Post-reload smoke proof table , controlled exact-token task results, accessed May 10, 2026.
- [3] Independent cross-validation section , Gemini and Codex review summary plus Claude quota result, accessed May 10, 2026.
- [4] Truth boundary and non-claims section , native Claude quota limitation and fallback interpretation, accessed May 10, 2026.
- [5] WordPress REST post metadata , public post record for title, slug, status, and excerpt.
- [6] Embedded evidence record section , local ScreenMemory truth-report filenames and generated public evidence bundle.

Historical correction: the May 9 retraction remains part of the audit trail. This article now follows the blog post rules by presenting a structured research question, method, results, limitations, sources, visuals, comparison, takeaways, and searchable metadata.