SUPERMENTE.AI: A Production Multi-Domain Outer Harness over Claude Code

Raphael Mendes da Cunha Neto[1]

[1]SUPERMENTE.AI / G3Data — Juiz de Fora, Brazil , rapha@supermente.ai · supermente.ai

June 2026

Abstract

We describe SUPERMENTE.AI, a production implementation of harness engineering: a multi-domain outer harness built on top of Claude Code (Anthropic) that has operated six real, non-coding business domains for over six months, built and run by a single non-programmer operator. The system instantiates the scenario Anthropic's Institute describes in "When AI Builds Itself" (May 2026) — humans direct, AI executes, with human effort concentrated on oversight, validation and verification — through a concrete mechanism: every error becomes a rule, every rule becomes a hook, and every hook leaves an auditable receipt. We present the architecture at a conceptual level, report verifiable operational metrics (each mapping to a re-runnable command), describe the governance layer (kill-switches, append-only audit, generator/auditor separation, cross-vendor adversarial verification), and state honestly what is proven (N=1 operation) and what remains under validation (N=2 transferability via a service-first model). The implementation itself is proprietary; this paper describes the harness, not its internals.

Keywords — *harness engineering, LLM agents, Claude Code, AI governance, cognitive amplification*

Resumo (PT-BR)

Descrevemos o SUPERMENTE.AI, uma implementação em produção de *harness engineering*: um harness externo multi-domínio construído sobre o Claude Code (Anthropic), operando seis domínios de negócio reais (não-coding) há mais de seis meses, construído e operado por um único operador não-programador. O sistema instancia o cenário descrito pelo Anthropic Institute em "When AI Builds Itself" (maio/2026) — humanos dirigem, a IA executa — através de um mecanismo concreto: cada erro vira regra, cada regra vira hook, cada hook deixa um recibo auditável. Este artigo descreve o harness em nível conceitual; a implementação é propriedade intelectual.

1 Introduction: the harness is the difference

The same language model produces dramatically different results depending on the infrastructure that surrounds it. The industry has converged on a name for this discipline: *harness engineering* — the engineering of everything around the model that is not the model: state, tool execution, feedback loops, and constraints. The model contains intelligence; the harness turns that intelligence into useful, reliable work.

Anthropic's Institute frames the background dynamics in "When AI Builds Itself" [1]: a growing share of AI development is delegated to AI itself — over 80% of Anthropic's code is now written by Claude — along a ladder of autonomy that runs from *executing specified tasks*, through *designing approaches to goals*, to *choosing which problems matter*. The scenario described as most plausible keeps humans in the director's seat, with human effort migrating toward *oversight, validation and verification*.

SUPERMENTE.AI is a production implementation of that scenario, operating since December 2025 — before the discipline received its public name in early 2026.

2 System overview

SUPERMENTE.AI is an *outer harness*: it wraps Claude Code (itself an inner harness around the Claude models) and extends it into a governed, multi-domain operating layer. The method, in one sentence:

> **Every error becomes a rule, every rule becomes a hook, every hook leaves an auditable receipt.** Text is desire; the hook is the guarantee.

The system delivers value in three layers:

Table 1: The three value layers of the harness.

Layer	What it does	Analogy
Amplify the person	Copilot shaped to the operator's cognitive profile	The suit, not the treadmill
Govern the system	Rules become deterministic interceptions (hooks)	Chassis and brakes
Orchestrate agents	Coordinates specialized agents under human direction	The conductor

Architecturally, the harness comprises: a library of specialized skills (progressive disclosure — knowledge loads only when needed); deterministic hooks across the session lifecycle (interception points that react to every action); persistent cross-session memory; a cognitive profile that shapes every interaction to the operator; and append-only audit ledgers.

3 Verifiable metrics

Each number maps to a stated, reproducible source. For external readers the canonical verification is the **public live endpoint** `supermente.ai/metrics.json`, regenerated by the system itself — the same methodological honesty Anthropic applies to its own metrics in [1].

Table 2: Operational metrics (snapshot 2026-06-11; live endpoint stays current).

Metric	Value	Verification
Specialized skills	351	`metrics.json`; locally: `find ~/.claude/commands/sc -name '*.md' | wc -l`
Hooks (interception points)	134 active / 633 installed[a]	`metrics.json` (hook-coherence ledger)
Specialized agents	28	`metrics.json`; agent registry count
Business domains in production	6 (non-coding)	revenue records under NDA
Continuous operation	6+ months	operator's public git history

[a]Definitions: *active* = hooks enabled in the running orchestrators (`HOOKS_ENABLED`); *installed* = hook files registered by the coherence ledger. A naive file count over the hooks directory returns more (∼900+) because it includes libraries, tests and archived files that are not interception points.

4 Governance and safety

The system operates "humans direct, AI executes" with concrete mechanisms, translating the *oversight, validation and verification* agenda of [1] into operations:

1. **Human direction preserved** — the system executes mandates; it does not set its own goals.
2. **Kill-switches** — every autonomous subsystem is disabled by a single flag.
3. **Append-only audit** — thousands of events in immutable ledgers.
4. **Generator/auditor separation** — the producer is never the verifier; verification measures the final state of the environment, not the agent's self-report.
5. **Cross-vendor adversarial verification** — significant verdicts are contested by models from independent providers.

5 Honest status and limitations

Proven (N=1): the architecture has sustained six paying domains for over six months with a single non-programmer operator. **Under validation (N=2):** transferability of the method to a second operator, via a service-first model in which the client receives amplified work output, never code. **Self-measured metrics:** the founder's "Symbiotic IQ" progression shown on the public site is an internal operational indicator (akin to an index, not a psychometric measure), self-measured with N=1, and labeled as such wherever it appears.

6 Related work

This system's concepts converged independently with several public formalizations, with git-timestamped records preceding publication in each case: dynamic workflow patterns [3], agent-design canon [2], and the RSI governance agenda [1]. We claim convergence, not causation; the records demonstrate that the harness operated these patterns in production before they were publicly named.

References

[1] Anthropic Institute. *When AI Builds Itself: Recursive Self-Improvement*. May 2026. https://www.anthropic.com/institute/recursive-self-improvement

[2] Anthropic. *Building Effective Agents*. December 2024. https://www.anthropic.com/research/building-effective-agents

[3] Anthropic Engineering. *A Harness for Every Task*. June 2026. https://www.anthropic.com/engineering

---

This document describes the harness at a conceptual level. The implementation (internal rule architecture, profiles, propagation machinery) is SUPERMENTE.AI intellectual property, available to partners under NDA. Built on Claude API + Claude Code; team completing the Anthropic Academy certification path within the Claude Partner Network program. Contact: `rapha@supermente.ai`.