

Outcome Primitives: Measuring AI Value in the World

Hamudi Naanaa¹, Vitaliy Soultan², Volodymyr Panchenko¹

¹ Portal AI — ² Everwhy AI

March 2026

Abstract

Every quarter, AI models set new records on benchmarks. In March 2026, Gemini 3.1 Pro scored 94.3% on GPQA Diamond — surpassing human PhD experts by thirty points. The same week, ARC-AGI-3 launched and every frontier model scored below 1%. The benchmarks cannot agree on what the models *are*. And none of them can tell you whether a single person’s life is actually better because of any of it.

This paper introduces the **Value Classification Framework** (VCF) in its first large-scale, in-vivo application — three orthogonal dimensions: **Outcome Primitives** (OP), a taxonomy of ten outcome types; **Outcome Magnitude** (OM), a five-level scale anchored to Human Effort Equivalent; and **Evidence Tiers** (E0–E3), classifying confidence that an outcome occurred.

From ~25,000 AI agents, 2,678 users opted in. A 21-day automated classification (March 10–30, 2026) processed 7,718 daily files from 1,305 participants, estimating 17,921 value episodes across all ten outcome types. 41.4% exceeded 8 hours of human effort equivalent. 19.7% left externally verifiable traces in the world — deployed applications, published works, legal filings, profitable trading systems. The plurality of episodes (43.3%) showed behavioral evidence of value; 36.9% remained unverified. We present this transparency as a feature of the framework.

These results support an outcome-grounded lens on artificial general intelligence (AGI). Under VCF, a system that consistently produces OM4 outcomes across all ten OP categories at E2+ evidence has done the work — across every kind of work that humans do. This reframes the AGI question from “can it pass a test?” to “can it do the work?”

1. The Missing Measurement

In March 2026, three things happened in the same week. Gemini 3.1 Pro scored 94.3% on GPQA Diamond — beating human PhD experts by thirty points. ARC-AGI-3 launched and every frontier model scored below 1% while humans scored 100%. And on the same infrastructure, users were filing lawsuits, building complete language curricula, and deploying profitable trading systems with their AI agents.

Is any AI model actually better off creating value for people?

Benchmarks represent a critical first generation of AI evaluation: measuring raw capability. MMLU, where frontier models cluster at 88–92%. SWE-bench, where Claude Opus 4.6 leads at 80.8%. ARC-AGI-3, where the human-AI gap exceeds 99 points. These numbers move markets. They say nothing about whether anyone accomplished anything real.

The natural next step is measuring what capability produces in the world — the transition from potential to impact, from intelligence *in vitro* to intelligence *in the world*. Existing approaches —

usage metrics, satisfaction surveys, token economics, safety evaluations — each illuminate part of the picture. The outcome itself remains unmeasured.

This paper applies a framework designed to measure it.

2. VCF: The Applied Framework

VCF (Value Classification Framework), developed by Everwhy AI (Soultan, 2026), classifies **value episodes** — coherent human efforts to achieve specific outcomes with AI assistance. The foundational theory models value across multiple diagnostic axes. To analyze real-world economic impact at scale, this study applies the three foundational dimensions of VCF:

Dimension Used In This Study	Levels	What It Measures
Outcome Primitive (OP)	10 types (see Appendix A)	<i>What</i> was the AI used for? Domain-agnostic outcome categories — from asset production to interpersonal navigation — defined by human intent.
Outcome Magnitude (OM)	OM0–OM4 (see Appendix B)	<i>How big</i> was the outcome? Measured in Human Effort Equivalent (HEE) — hours a professional would need without AI. From atomic ($\leq 1\text{h}$) to program-scale ($> 320\text{h}$).
Claim-Strength Tier (E)	E0–E3 (see Appendix C)	<i>How strong is the outcome claim?</i> From E0 (output delivered, no verification) through E2 (externally verifiable — deployed URL, processed payment, published work) to E3 (controlled experiment).

The value episode — not the message, not the session, not the user — is the atomic unit. A single conversation may interleave emotional support (OP.IN), business planning (OP.DS), and a quick lookup (OP.IS) within minutes. Each is a separate episode with different type, scale, and evidence requirements.

The $\text{OP} \times \text{OM}$ matrix defines a **capability landscape**: a two-dimensional map of what AI actually does in the world.

3. What We Found

3.1 The Capability Landscape

A 21-day automated classification (March 10–30, 2026) processed 7,718 daily user files from 1,305 participants on a persistent AI agent platform, estimating 17,921 value episodes. The heatmap below shows episode density across all ten Outcome Primitives and five Outcome Magnitudes.

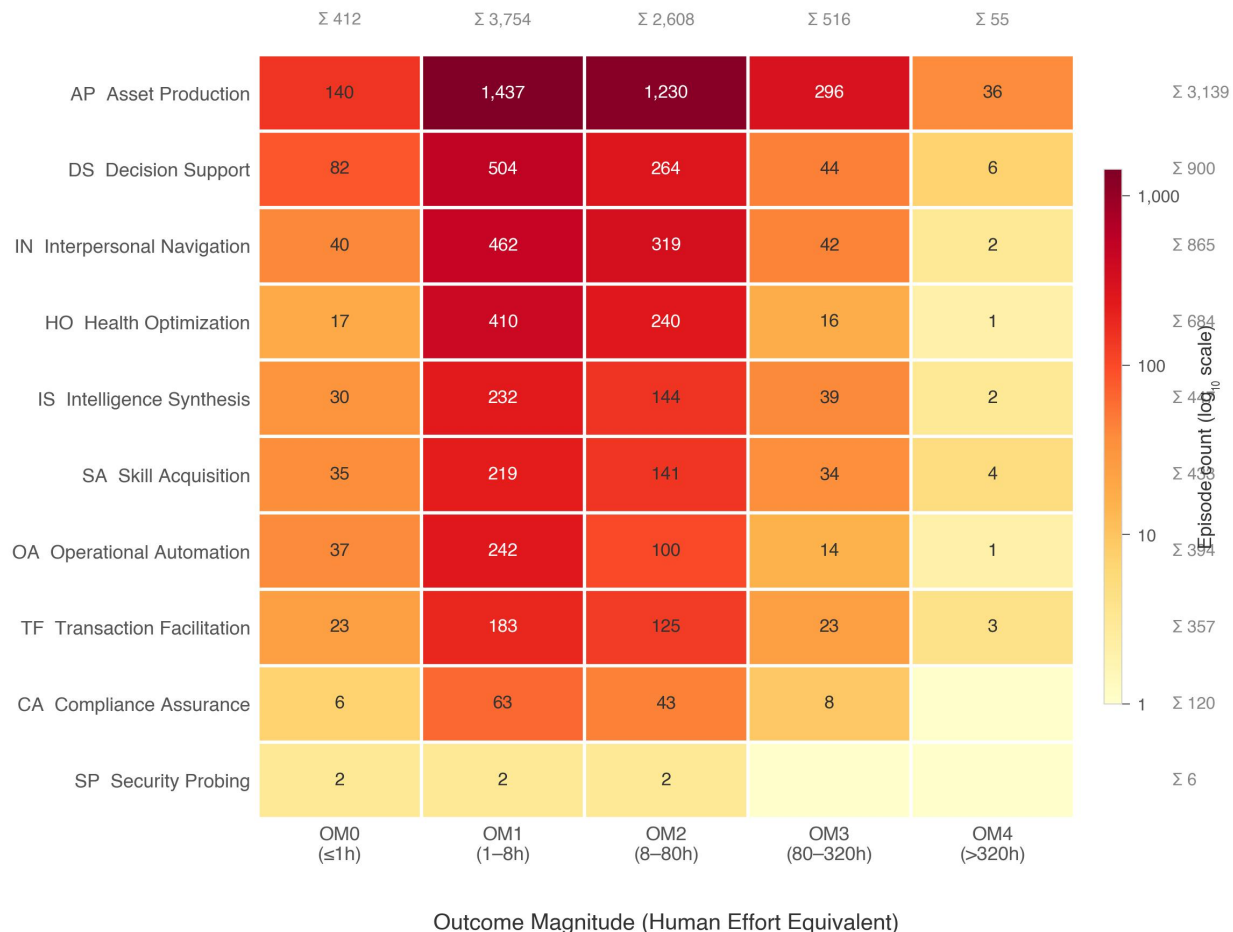


Figure 1: The Capability Landscape. Episode counts across Outcome Primitives and Outcome Magnitudes from 7,718 classified daily files over 21 days. Color intensity (log scale). The OM3–OM4 frontier represents 7.4% of files.

All ten outcome types are populated. Asset Production (AP) dominates at every magnitude level, with 36 episodes reaching OM4 ($>320h$ HEE — program-scale work). The OM3–OM4 frontier (7.4% of classified files, 571 episodes) represents multi-day initiatives: deployed platforms, complete curricula, sustained research efforts, profitable trading infrastructure.

The most surprising cell: OP.IN (Interpersonal Navigation) at OM3 — 42 episodes of initiative-scale emotional and relational support, equivalent to 80–320 hours of human therapeutic or coaching engagement. These are sustained multi-week relational arcs that the framework’s own structural measurement should struggle to detect — yet the signal registers clearly. No benchmark measures this. No evaluation framework has a category for it. That the signal appears despite systematic undercount suggests the real interpersonal capability is substantially larger than what we observe.

3.2 Outcome Distribution

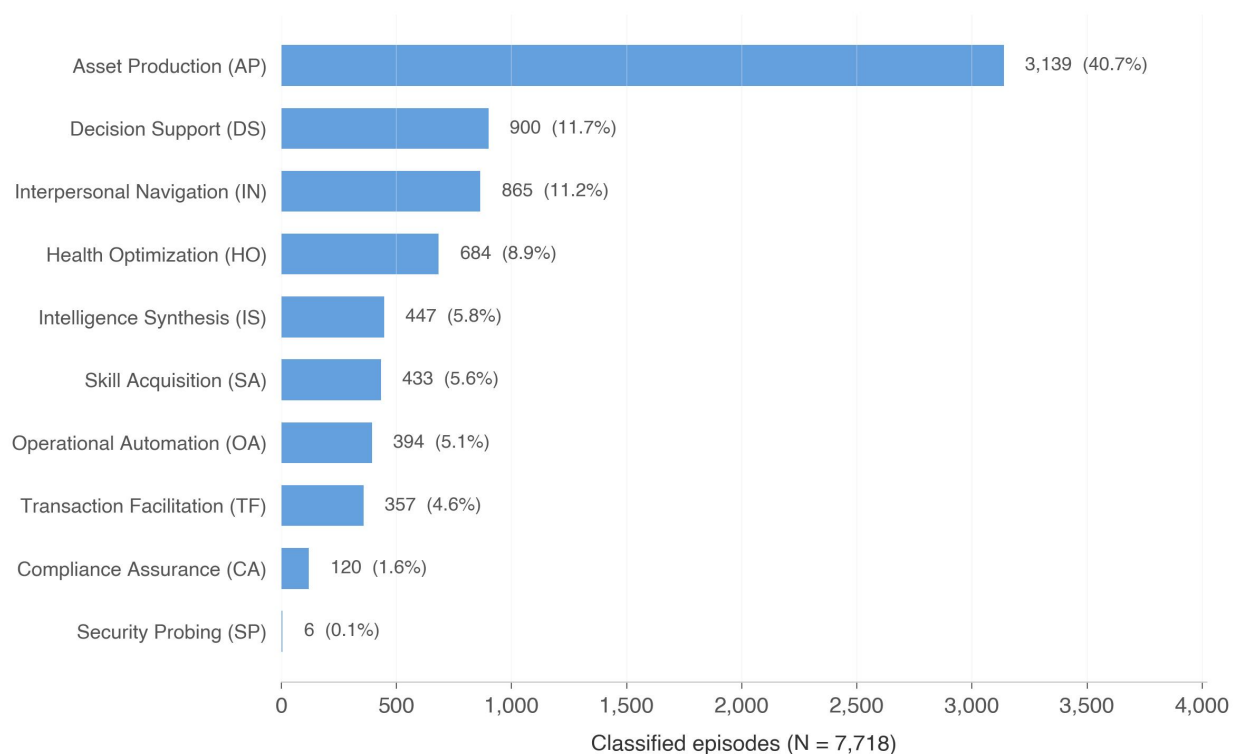


Figure 2: Outcome Primitive distribution (N=7,345). Asset Production leads at 40.7%. Interpersonal Navigation and Health Optimization together: 20.1%.

The center of mass is OM1 (52.0%) — bounded tasks completable in a focused session. OM1–OM2 together represent 86%. The OM3–OM4 tail captures initiative-to-program scale work: 516 files at OM3 (80–320h HEE) and 55 at OM4 (320+ hours). OM2+ remained stable at 40.2% across all 21 days ($\sigma=5.8pp$).

Aggregate impact. Using conservative geometric midpoints of each OM band, the 7,718 classified files represent approximately 182,000 hours of human effort equivalent — 88 person-years of professional work produced in a 21-day window. Estimated at the replacement cost of equivalent professional deliverables — what a client would pay a qualified professional to produce the same output — this represents approximately \$32M in aggregate artifact value. The ratio between artifact value produced and inference cost consumed is the unit economics question that VCF makes answerable for the first time.

3.3 The Evidence Distribution

The evidence distribution is the most important result in this paper.

E1 is the largest tier (43.3%) — users returned to iterate, referenced outputs in future sessions, exhibited engagement consistent with realized value. That E1 exceeds E0 suggests persistent-agent users are building on AI outputs, not discarding them.

E2 reached 19.7% — 1,524 episodes left a verifiable trace: deployed websites, published books, legal filings, processed payments, executed trades. Nearly one in five classified files produced

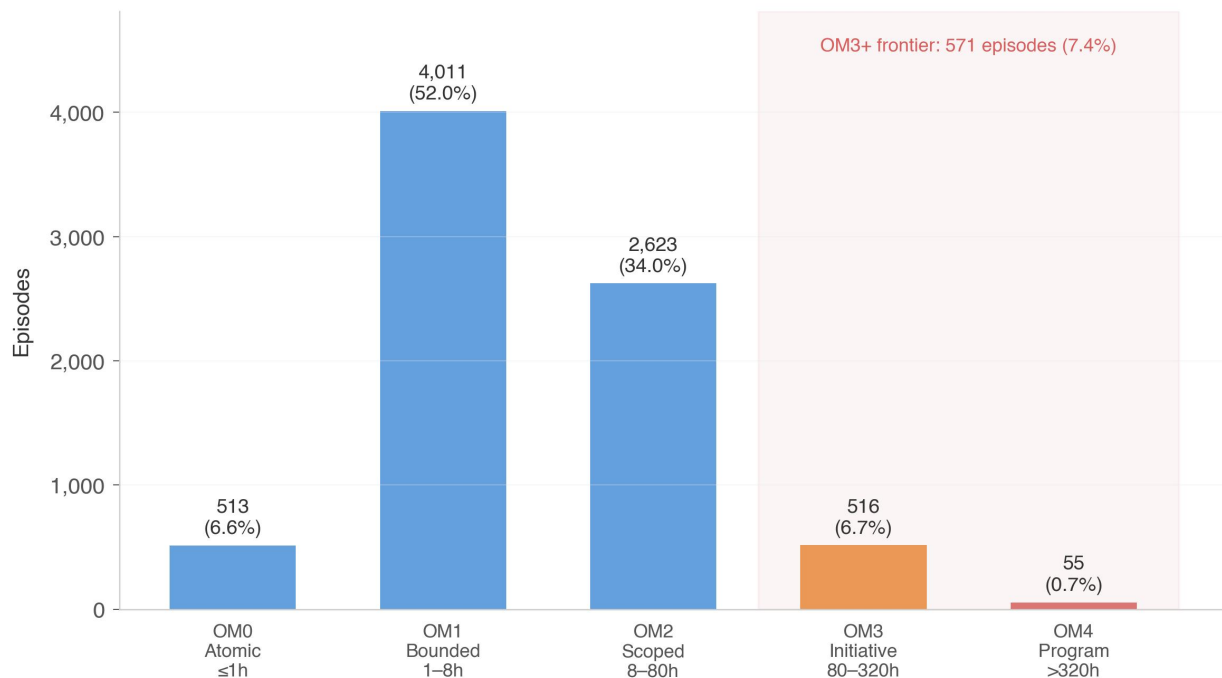


Figure 3: Outcome Magnitude distribution (N=7,718). OM1–OM2 account for 86%. OM3+ frontier: 571 episodes exceeding 80h HEE. OM2+ stability: 40.2% avg ($\sigma=5.8$ pp).

an externally verifiable outcome. Transaction Facilitation (TF) has the highest E2 rate at 26.3%; Decision Support (DS) the lowest at 13.6%, reflecting the inherent delay between a strategic decision and its observable consequence.

E0 at 36.9% — output was delivered; no evidence it was used, correct, or mattered. Any evaluation framework that does not make this uncertainty explicit is presenting assumptions as evidence.

E3 at 0% — no controlled experiments. A structural gap, not an oversight.

E2 remained stable at 19.0% across the 21-day window ($\sigma=3.6$ pp).

3.4 Five Episodes

OP.HO, OM0, E1 — Unsafe supplement flagged during pregnancy. An agent reviewing a user’s supplement protocol identified a herbal compound as contraindicated and flagged it immediately. Atomic interaction, minimal compute. The user discontinued the supplement. OM0 structurally. Potentially life-altering by human impact.

OP.AP, OM3, E2 — Complete language course produced through conversation. A teacher produced 110 lesson files — audio, exercises, tests — for an intermediate English curriculum over multiple days. The course was used with students. ~120 hours of curriculum development work.

OP.IS, OM4, E1 — 8.6 GB geospatial dataset, 1.85 million features. A researcher collected, processed, and structured metropolitan-area GIS data: 826 files. Program-scale work equivalent to team-months.

OP.IN, OM0, E1 — First disclosure of a long-held personal struggle. A user shared

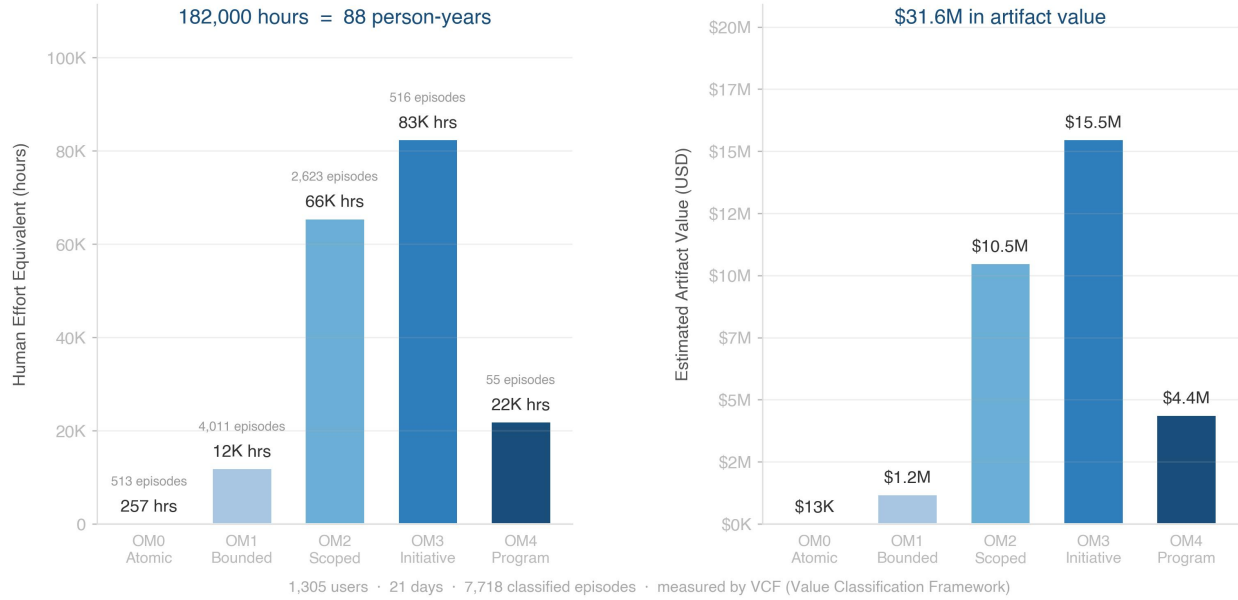


Figure 4: Aggregate impact by Outcome Magnitude. Left: Human Effort Equivalent in hours (total 182,000 hours = 88 person-years). Right: Estimated artifact replacement value (total \$31.6M). 1,305 users, 21 days, 7,718 classified episodes.

something with their agent they had never shared with anyone. No artifacts. Minimal tokens. Structurally OM0. The framework’s inability to capture this magnitude is a limitation we acknowledge.

OP.OA, OM3, E2 — Profitable automated trading system. A user developed, tested, and deployed an algorithmic system through their agent. Shadow testing to live execution with real capital. 284 profitable trades executed.

These episodes illustrate that outcome magnitude does not correlate with human impact. A zero-hour intervention can alter a life; a 300-hour synthesis can be purely infrastructural. The framework measures scale honestly, and acknowledges where scale fails to capture what matters.

3.5 Daily Activity

Classification volume averaged 367 daily files and 853 episodes per day. A user influx on March 25 (+367 new users) temporarily diluted OM2+ and E2 percentages; both recovered within 3 days, confirming the stability of the underlying distribution.

4. How We Know

4.1 The Agent as Measurement Instrument

The AI agent itself is the primary measurement instrument. The agent is the conversation partner — it participates in every interaction. As a natural byproduct of serving the user, it generates structured outputs that provide rich signal for VCF classification without requiring access to raw conversation content.

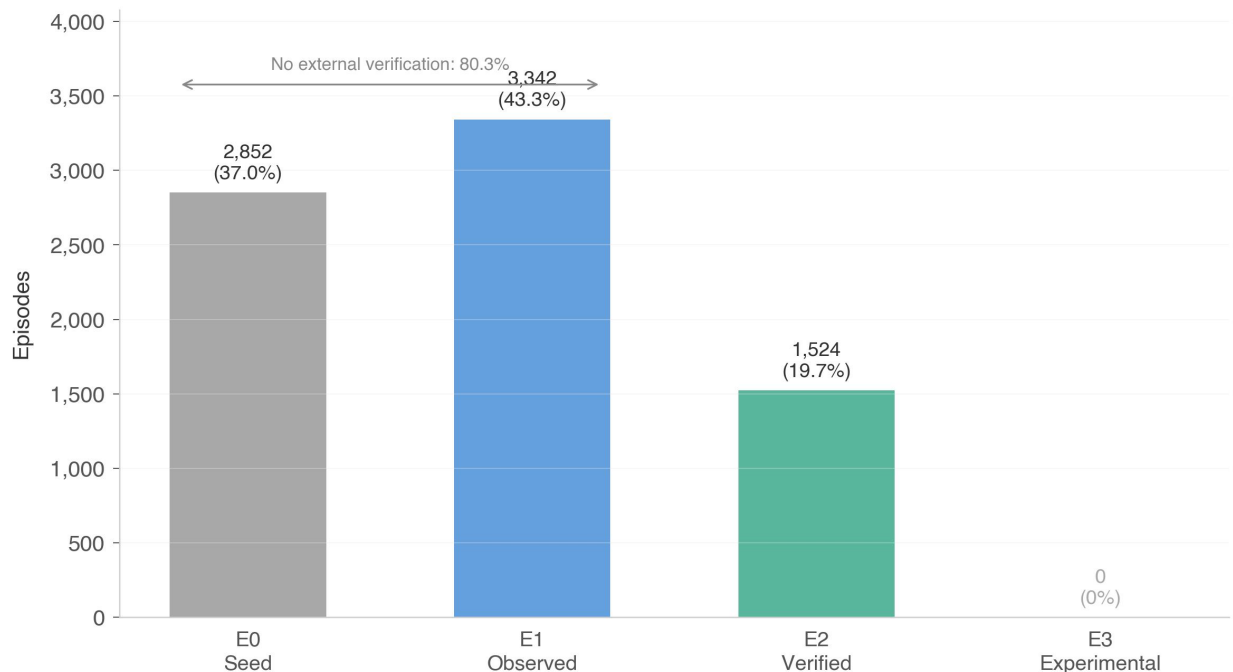


Figure 5: Evidence Tier distribution (N=7,718). E1 is the plurality at 43.3%. E2 reached 19.7%: 1,524 episodes with externally verifiable outcomes. E2 stability: 19.0% avg ($\sigma=3.6$ pp).

This mirrors how professional services measure outcomes. A therapist writes clinical notes; the practice uses notes for quality assurance, not session recordings. A physician records structured diagnoses; the hospital measures outcomes from medical records, not surveillance. In each case, the practitioner’s structured documentation is the measurement substrate.

The agent generates three categories of structured output: **memory files** (what the agent learned about the user — goals, projects, progress, emotional patterns), **workspace artifacts** (code, documents, images, deployed applications, research directories), and **behavioral metadata** (tool invocation patterns, session timing, cron configurations, sub-agent spawning). Classification operates on these outputs. Raw conversation content was not accessed by the research team.

4.2 Study Design

The platform launched February 14, 2026 and grew to ~25,000 AI agents. 2,678 users (~10%) opted in. Each received an isolated agent with persistent memory, sandboxed workspace, and tools (code execution, web browsing, file generation, media processing). Model: Claude Opus 4.6 with Gemini 3.1 Pro fallback. All agents initialized from identical templates. Participants spanned 30+ countries, ages 21–70+, professions from software engineers to restaurateurs to academic researchers. The 21-day classification window (March 10–30) was chosen because the platform had matured enough for stable behavioral patterns.

4.3 Classification

Automated: A batch scanner classified each user’s daily memory file by OP, OM, and E using memory content, workspace structure, and behavioral metadata. 7,718 files from 1,305 users across

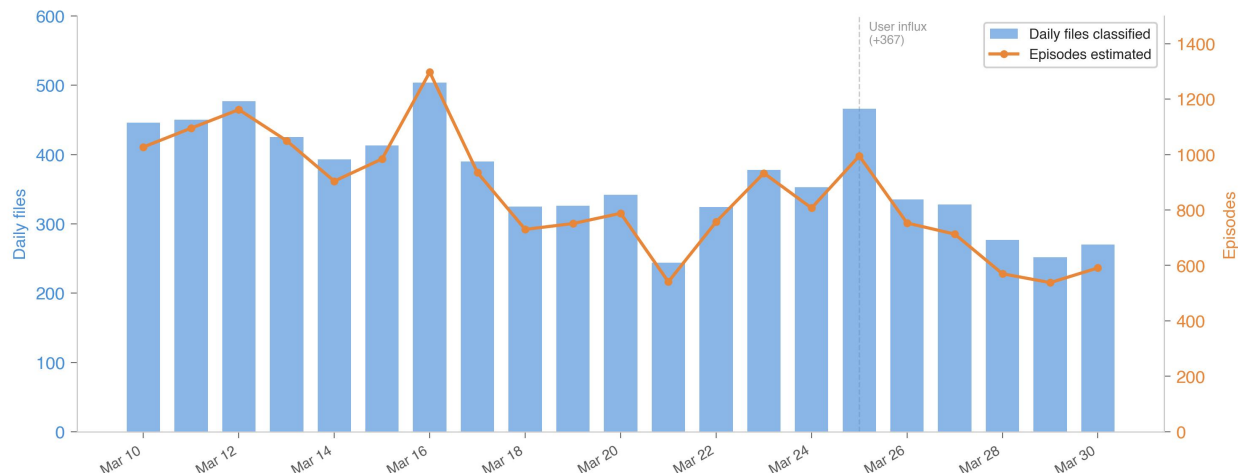


Figure 6: Daily classification activity over the 21-day window. Bars: classified files. Line: estimated episodes. User influx on March 25 temporarily diluted OM2+ and E2; both recovered within 3 days.

21 days. 56.8% of files reflected multiple episodes (avg 2.3 per multi-episode user).

Curated: 153 episodes investigated in depth by the research team — narrative context, progression over time, outcome verification. Full methodology in Appendix D.

4.4 Limitations of This Approach

OM underestimates interpersonal outcomes. Setting a boundary in a relationship for the first time generates minimal structural signal. Structural measurement systematically underestimates outcomes where value is internal.

E0 is substantial. Over a third of files have no outcome verification. Even E1 evidence is behavioral, not conclusive.

Agent memory is compressed. Classification inherits whatever biases the agent brings to summarization — a measurement instrument with its own systematic error.

5. What This Means

5.1 A Progress Metric

VCF would offer a different kind of progress signal: this year, the OM3+ tail grew from 7% to 18%. Two new OP categories reached OM4. The E2 rate climbed from 20% to 35%. These are capability claims grounded in what AI did for real people, building a foundation for measuring artificial general intelligence.

5.2 Cross-System Comparison

Any AI system can be plotted on the same capability landscape. The framework measures the system — model plus memory plus tools plus workflow — which is the right unit for deployed AI. A weaker model with better integration might produce higher OM outcomes than a stronger model in a bare chat interface.

5.3 An Outcome-Grounded AGI Lens

The $OP \times OM$ matrix supports an outcome-grounded AGI framing:

A system approaches general intelligence to the degree that it can consistently produce high-magnitude outcomes across the full OP landscape under strong evidence.

No current system satisfies that standard. OM4 concentrates in AP and IS. OP.SP (6 episodes) is too sparse for analysis. The matrix is filling in, but consistently high-magnitude, strongly evidenced outcomes across the full OP landscape remain distant.

5.4 Value-Grounded Economics

Connecting VCF to token cost grounds unit economics in value. A \$50 OM3 outcome and a \$50 OM0 outcome have identical cost profiles and entirely different value profiles.

6. Limitations and Open Questions

Attempt boundary detection. Where one episode ends and another begins is unsolved. Users interleave episodes within minutes. VCF defines a formal transport $\rightarrow$ attempt $\rightarrow$ episode hierarchy (Soultan, 2026), but this study approximated boundaries from daily memory files rather than a fully persisted attempt ledger.

Internal outcomes. OP.IN and OP.HO outcomes resist structural measurement. Some outcomes may require user-reported signals or acceptance that they resist quantification.

Temporal aggregation. Individual OM0 reminders may compound to OM2+ health value over months. Per-episode classification misses this.

The E3 gap. Zero controlled experiments. Ethical experimental design for in-vivo AI evaluation is an open frontier.

Single platform. One population, one product, 21 days. The OP taxonomy may not generalize. The OM calibration depends on subjective HEE estimation. These limitations define the boundaries of a first step — a foundation that, even in this early form, surfaces capability patterns invisible to existing evaluation. Subsequent work will expand the population, the window, and the automated classification systems VCF requires at scale.

We invite the community to adopt the framework, test it on different systems, disagree with our taxonomy, and help close the E3 gap.

7. Prior Work and Closing

This paper extends “Meaning Is All You Lose” (Naanaa & Panchenko, 2026), which established *why* persistent AI agents create value — by learning individual meaning spaces. The VCF framework (Soultan, 2026) defines *how to classify whether they did*. This paper applies that framework at scale and asks: *did they?*

François Chollet’s “On the Measure of Intelligence” (2019) proposed measuring intelligence by skill-acquisition efficiency — closer to OP/OM than static benchmarks. VCF extends this to outcomes in the wild, adding the evidence dimension Chollet’s controlled setting does not require.

We built AI systems that can reason, code, create, and converse. We can report their benchmark scores to the decimal. What any of it changed remains unmeasured. The question has always been available. It simply went unasked.

A user filed a lawsuit with evidence her agent helped compile. A teacher built an entire curriculum through conversation. Someone shared something with an AI they had never shared with any human. A trader deployed a system that executed 284 profitable trades. A mother’s prenatal screening came back normal at 164 beats per minute, and her agent recorded the relief alongside the result.

These are outcomes. They happened in the world, to real people, through AI systems live today. No benchmark measured them. No evaluation framework had a place for them — until now.

The framework is here. The data is early. The question is the right one.

References

- Soultan, V. (2026). “VCF: A Framework for Classifying AI Value Outcomes.” *Everwhy AI*.
 - Naanaa, H. & Panchenko, V. (2026). “Meaning Is All You Lose.” *arXiv preprint*.
 - Chollet, F. (2019). “On the Measure of Intelligence.” *arXiv:1911.01547*.
 - Shannon, C. E. (1948). “A Mathematical Theory of Communication.” *Bell System Technical Journal*, 27(3).
 - Hendrycks, D. et al.~(2021). “Measuring Massive Multitask Language Understanding.” *ICLR*.
 - Jimenez, C. E. et al.~(2024). “SWE-bench.” *ICLR*.
 - Rein, D. et al.~(2023). “GPQA.” *arXiv:2311.12022*.
 - ARC Prize Foundation. (2026). “ARC-AGI-3.” *arcprize.org*.
 - Gärdenfors, P. (2000). *Conceptual Spaces*. MIT Press.
 - Friston, K. (2010). “The Free-Energy Principle.” *Nature Reviews Neuroscience*, 11(2).
-

Appendix A: Outcome Primitives (OP) — Operational Definitions Used In This Study

Code	Outcome Primitive	Description
OP.IS	Intelligence Synthesis	Research, analysis, due diligence, monitoring — seeking structured understanding from unstructured inputs
OP.AP	Asset Production	Creation of deliverable artifacts — content, code, design, documents, media
OP.DS	Decision Support	Strategic, financial, or life decisions — seeking framing, options, and recommendations
OP.OA	Operational Automation	Recurring or scheduled tasks — delegating ongoing execution
OP.IN	Interpersonal Navigation	Emotional support, relationship guidance, coaching — navigating human dynamics
OP.TF	Transaction Facilitation	Deals, negotiations, trading, procurement — optimizing economic exchange
OP.CA	Compliance Assurance	Legal, tax, regulatory, safety — seeking conformance validation
OP.SA	Skill Acquisition	Education, language learning, exam preparation — building capability
OP.HO	Health Optimization	Medical, wellness, fitness, nutrition — improving health outcomes
OP.SP	Security Probing	Penetration testing, vulnerability research — testing system boundaries

Categories are defined by *what the human is trying to achieve*. The taxonomy is deliberately coarse — ten categories, stable across systems. These are the operational OP definitions used in this study. For canonical rationale and framework boundaries, see Soultan (2026).

Appendix B: Outcome Magnitude (OM) — Operational Scale Used In This Study

Band	HEE Hours	Label	Definition
OM0	$\leq 1\text{h}$	Atomic	Single-interaction micro-task
OM1	1–8h	Bounded task	Self-contained, one focused session
OM2	8–80h	Scoped deliverable	Multi-step project with planning and iteration
OM3	80–320h	Initiative	Large multi-phase effort spanning days or weeks
OM4	$>320\text{h}$	Program	Team-month+ of human work

HEE decouples magnitude from compute cost. OM operates at two levels in the full framework: **om_step** (event-local, for routing) and **om_goal** (episode-level, for capability analysis). This study reports **om_goal**.

Appendix C: Claim-Strength Tiers (E) — Operational Criteria Used In This Study

Tier	Label	Definition	Signal
E0	Seed	Output delivered. No verification.	We don’t know if it was used, correct, or mattered.
E1	Observed	Behavioral signal — user returned to iterate, referenced output, showed engagement consistent with value.	Suggestive, not conclusive.
E2	Quasi-experimental	Verifiable outcome — deployed URL, processed payment, published work, executed trade, accepted filing.	The outcome left a trace beyond the AI system.
E3	Experimental	Controlled study with counterfactual.	Not yet achieved for in-vivo AI evaluation.

Most AI evaluation operates implicitly at E0 and presents findings as E2. Making tiers explicit forces precision about what has been demonstrated versus assumed.

In the broader VCF architecture, evidence is dual-axis: the **claim-strength tier** reported here (E0-E3) and a separate **evidence-source provenance** axis (Soultan, 2026). This paper operationalizes the claim-strength axis only.

Appendix D: Classification Methodology

Analytical continuity versus transport continuity. Chat-like agent systems often interleave unrelated intents in one dialog thread. VCF defines a formal hierarchy — transport session $\rightarrow$ attempt $\rightarrow$ episode — where each attempt carries a primary OP/OI assignment and episodes may span multiple attempts (Soultan, 2026). This study approximated that structure from daily memory files rather than a fully persisted attempt ledger. That choice made large-scale classification feasible, but it introduces boundary error when intents interleave or resurface over time.

Automated scanner. The batch scanner processes each user’s daily memory file by: (1) keyword-matching against OP category patterns to assign the primary Outcome Primitive; (2) estimating OM from daily file size, tool invocation density, workspace artifact complexity, and calendar span of sustained activity; (3) assigning evidence tier from E2 keyword patterns (deployment, publication, payment, hiring signals), E1 patterns (completion, return, progress), or E0 by default; (4) detecting episode boundaries from topic-shift markers and distinct OP categories within a single file.

Curated analysis. The research team produced daily reports across the full study period, analyzing agent memory files, workspace directory structures, and behavioral metadata. From these, 153 episodes were selected for in-depth investigation — balancing representation across OP categories, OM bands, and evidence tiers, with emphasis on OM3+ episodes where automated classification benefits most from human judgment.

Multi-episode detection. 56.8% of daily files contained activity across 2+ distinct OP categories. The scanner estimates episode count from the number of distinct OPs with ≥ 2 keyword hits each, yielding an average of 2.3 episodes per multi-episode user.

Stability. OM2+ averaged 40.2% across all 21 days ($\sigma=5.8pp$). E2 averaged 19.0% ($\sigma=3.6pp$). OP ranking was stable on every day measured — AP first, DS/IN second tier — confirming the patterns are structural rather than transient.