Outcome Primitives: A Framework for Measuring AGI Value in the World

Hamudi Naanaa[1], *Vitaliy Soultan*[2], *Volodymyr Panchenko*[1]

[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.

We introduce **VCF** (Value Classification Framework) — a framework for measuring AI value in the lived experience of real people. 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** (E), a four-level framework 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.

VCF yields a natural definition of artificial general intelligence (AGI). 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 proposes a framework to measure it.

---

2. VCF: The Framework

VCF (Value Classification Framework), developed by Everwhy AI as a general-purpose value classification system, defines three orthogonal dimensions that classify a **value episode** — a coherent attempt by a human to achieve a specific outcome with AI assistance. This paper presents its first large-scale application to in-vivo AI evaluation, using data from a persistent AI agent platform operated by Portal AI.

Dimension	Levels	What It Measures
Outcome Primitive (OP)	10 types (see Appendix A)	*What* 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)	*How big* was the outcome? Measured in Human Effort Equivalent (HEE) — hours a professional would need without AI. From atomic ($\leq$1h) to program-scale (>320h).
Evidence Tier (E)	E0–E3 (see Appendix C)	*How sure* are we? 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 OP $\times$ 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.

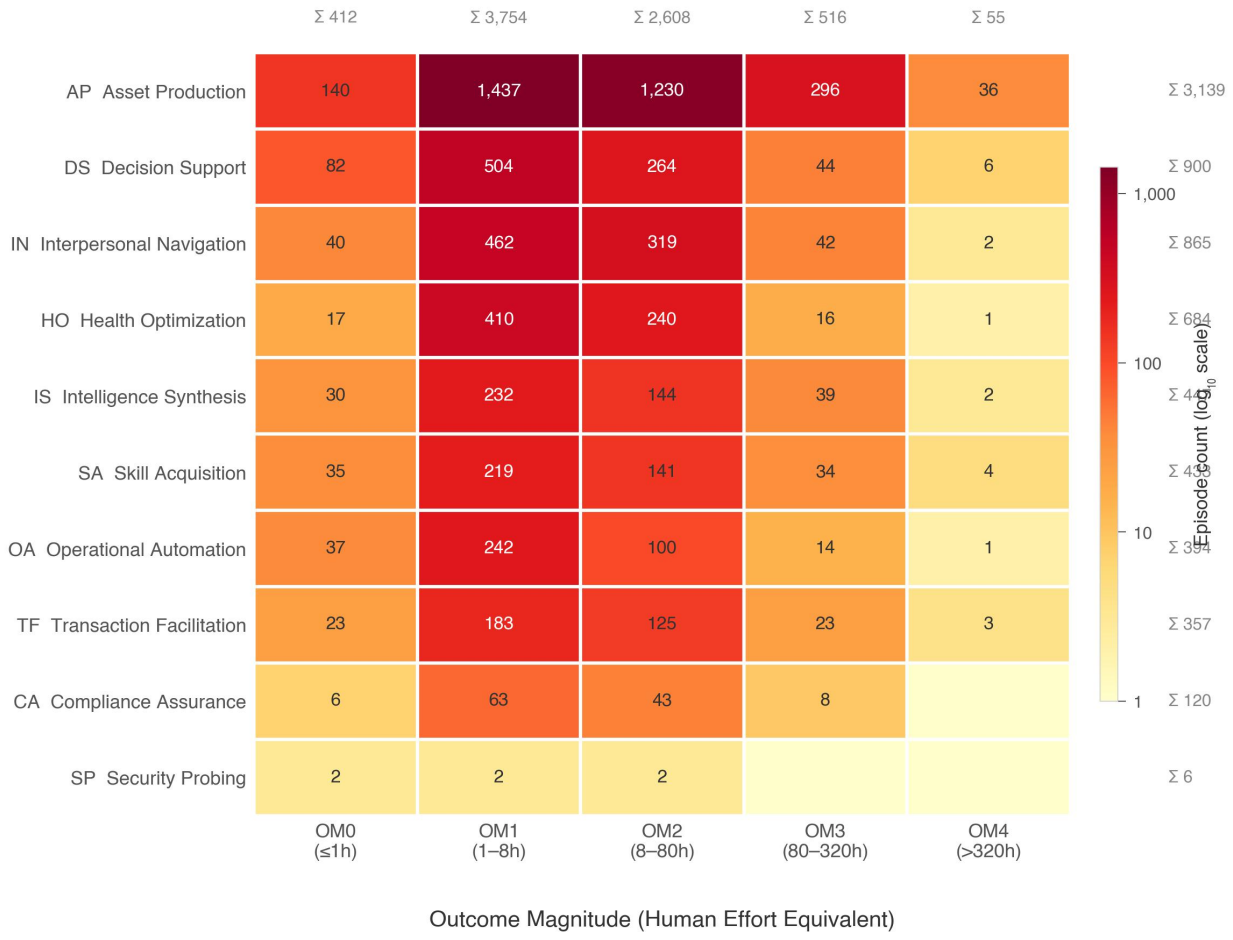

Figure 1: The Capability Landscape. Episode counts across Outcome Primitives and Outcome Magnitudes from 7,718 classified daily files over 21 days. Color intensity (log scale). The OM3–OM4 frontier represents 7.4% of files.

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

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

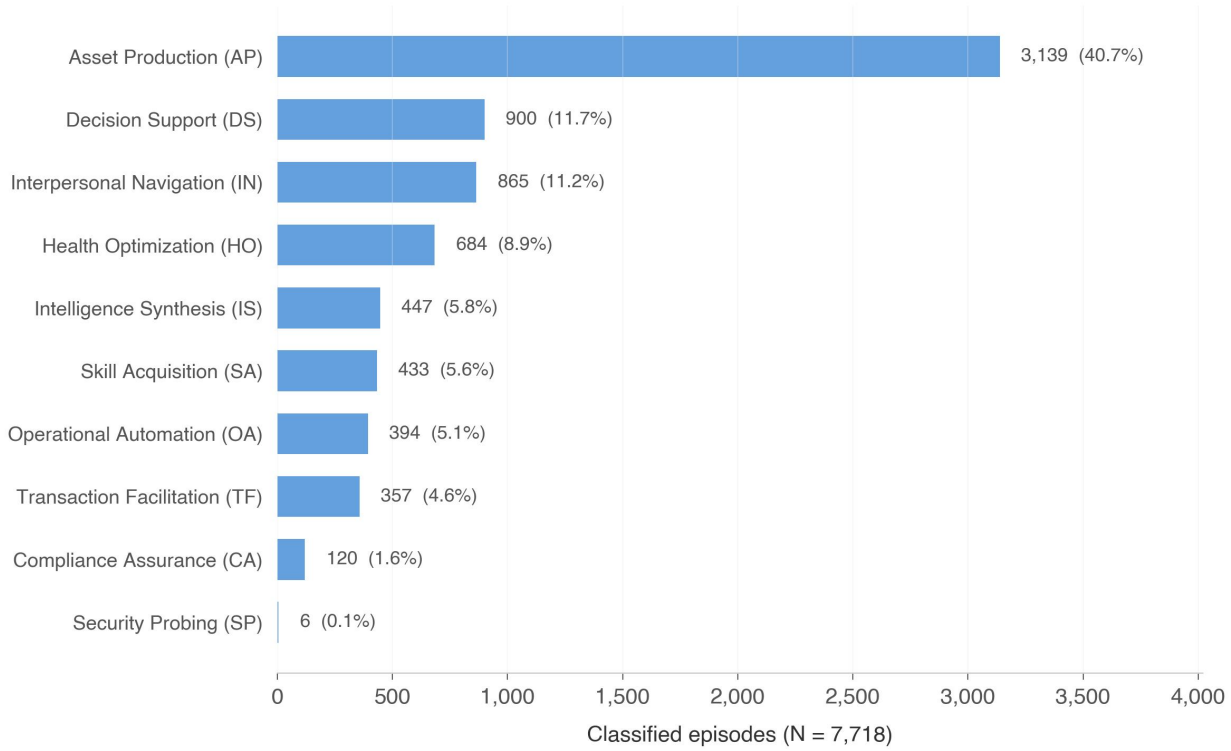

Figure 2: Outcome Primitive distribution (N=7,345). Asset Production leads at 40.7%. Interpersonal Navigation and Health Optimization together: 20.1%.

days (σ=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. An independent cross-validation using U.S. Bureau of Labor Statistics median wages yields $9.0M — establishing a labor-cost floor. The 3.5× ratio falls within the standard 2–4× markup between wages and professional services billing rates. Full derivation and sensitivity analysis in Appendix E. 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 an externally verifiable outcome. Transaction Facilitation (TF) has the highest E2 rate at 26.3%;

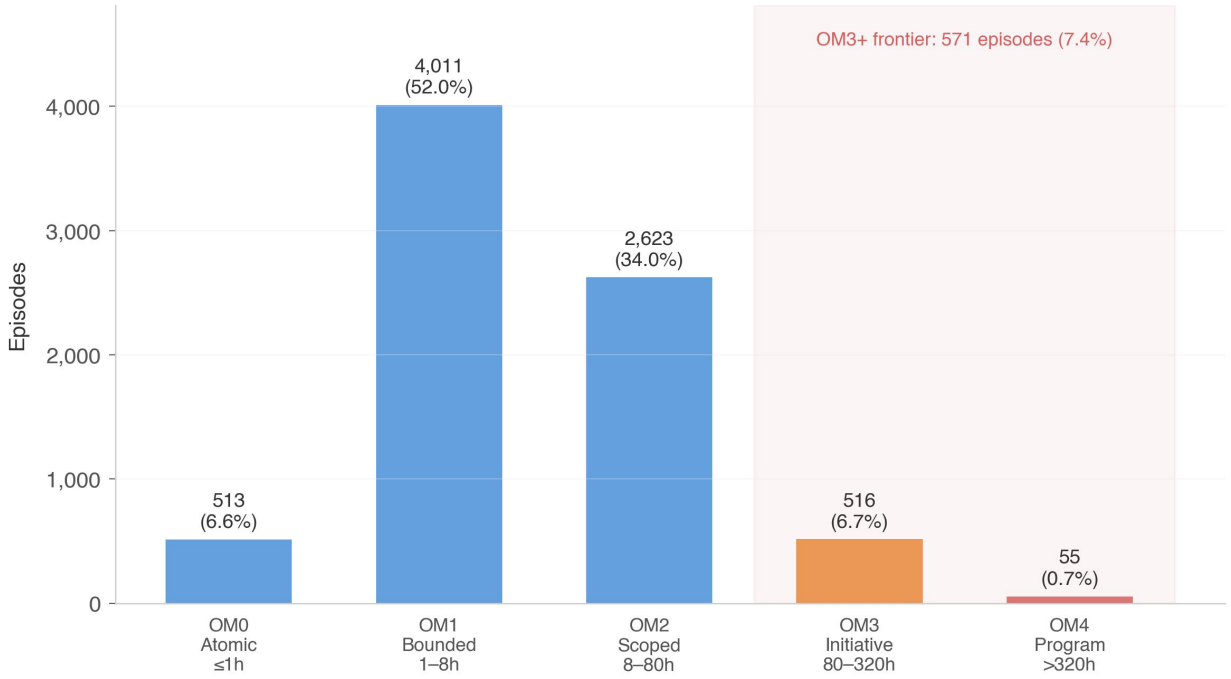

Figure 3: Outcome Magnitude distribution (N=7,718). OM1–OM2 account for 86%. OM3+ frontier: 571 episodes exceeding 80h HEE. OM2+ stability: 40.2% avg (σ=5.8 pp).

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 (σ=3.6pp).

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 something with their agent they had never shared with anyone. No artifacts. Minimal tokens.

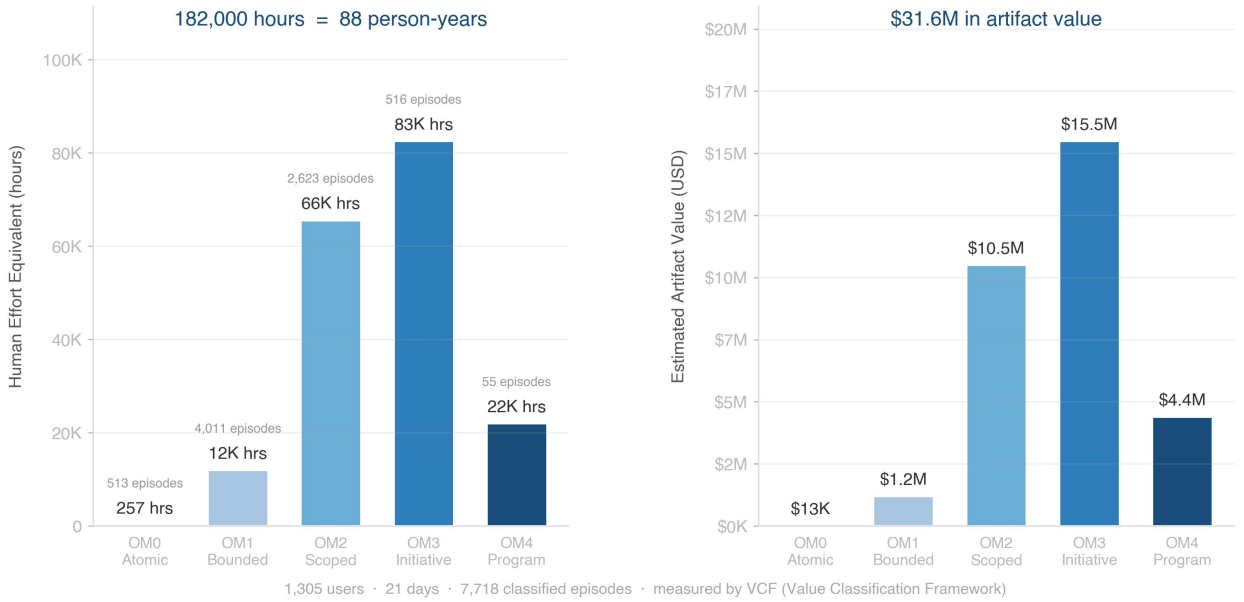

Figure 4: Aggregate impact by Outcome Magnitude. Left: Human Effort Equivalent in hours (total 182,000 hours = 88 person-years). Right: Estimated artifact replacement value (total $31.6M). 1,305 users, 21 days, 7,718 classified episodes.

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.

This mirrors how professional services measure outcomes. A therapist writes clinical notes; the

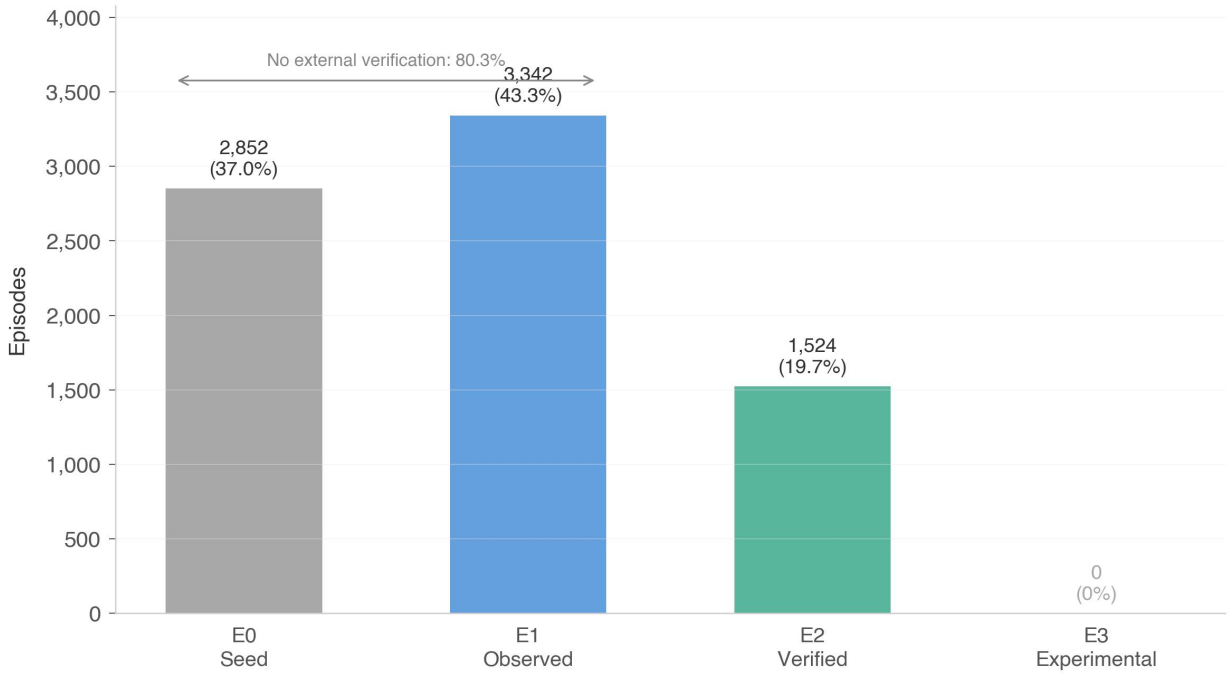

Figure 5: Evidence Tier distribution (N=7,718). E1 is the plurality at 43.3%. E2 reached 19.7%: 1,524 episodes with externally verifiable outcomes. E2 stability: 19.0% avg (σ=3.6 pp).

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

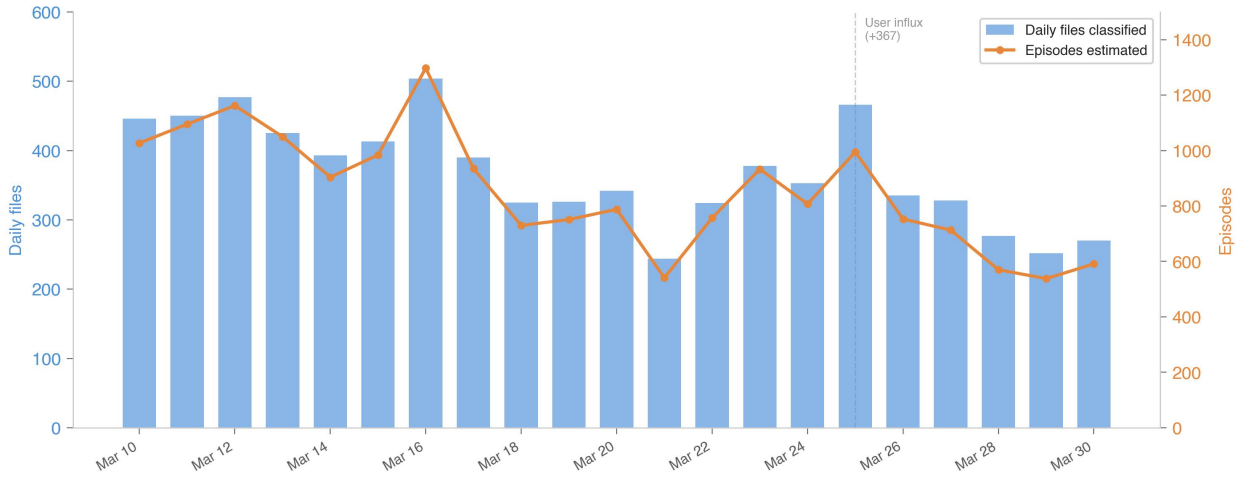

Figure 6: Daily classification activity over the 21-day window. Bars: classified files. Line: estimated episodes. User influx on March 25 temporarily diluted OM2+ and E2; both recovered within 3 days.

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 A Definition of General Intelligence

The OP × OM matrix yields an outcome-empirical definition:

A system that consistently produces OM4 outcomes across all ten OP categories at E2+ evidence is generally intelligent — it has done the work.

No current system achieves this. OM4 concentrates in AP and IS. OP.SP (6 episodes) is too sparse for analysis. The matrix is filling in, but consistent OM4 across all categories at E2+ remains 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.

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

The question of how to measure intelligence has a long history. Chollet's "On the Measure of Intelligence" (2019) argued that benchmarks measure skill, not intelligence, and proposed skill-acquisition efficiency over novel tasks as the true metric. VCF shares this dissatisfaction with static benchmarks but takes a different path: rather than designing better tests, we measure what happens when AI systems operate in the wild with real people pursuing real goals. The evidence dimension (E0–E3) addresses something Chollet's controlled setting does not require — the honest acknowledgment that most real-world outcomes cannot be verified.

This paper also builds on "Meaning Is All You Lose" (Naanaa & Panchenko, 2026), which modeled communication as lossy compression between individual meaning manifolds and showed that persistent AI agents learn individual meaning spaces over time. That work explains the mechanism — *why* persistent agents create value through accumulated understanding. VCF provides the measurement — *whether* that value materializes, at what scale, and with what confidence.

Shannon (1948) solved the transport of information. Gärdenfors (2000) gave meaning a geometry. VCF attempts something adjacent: giving *outcome* a coordinate system — one where the position of a data point tells you what was accomplished, how large it was, and how sure we are it happened.

8. Closing

In 21 days, 1,305 people produced 182,000 hours of human effort equivalent — 88 person-years of professional work — through AI agents that learned who they are and what they need. Approximately $32M in artifact value, classified across ten outcome types, five magnitude levels, and three evidence tiers. 1,524 of those episodes left externally verifiable traces: deployed applications, published curricula, processed payments, executed trades, accepted legal filings.

These numbers exist because we built a framework to count them. Before VCF, they were invisible — not because they weren't happening, but because nobody had a coordinate system for outcomes. Benchmarks count capability. Usage metrics count engagement. VCF counts what changed.

The framework is young. The taxonomy will evolve. The evidence gaps are real — 37% of outcomes remain unverified, and zero controlled experiments have been conducted. What we have is a starting point: the first large-scale classification of what AI actually does for real people, measured honestly, with the limitations stated in the open.

The capability landscape is open. Any AI system can be plotted on it — any model, any product, any deployment. The taxonomy will be challenged, the evidence methods will improve, the E3 gap will close. What fills this map will define progress in artificial intelligence — not by what models score on a test, but by what people accomplish in the world.

References

- 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) — Full Taxonomy

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.

Appendix B: Outcome Magnitude (OM) — Scale Definitions

Band	HEE Hours	Label	Definition
OM0	$\leq$1h	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	>320h	Program	Team-month+ of human work

HEE decouples magnitude from compute cost. OM operates at two levels: **om_step** (event-local, for routing) and **om_goal** (episode-level, for capability analysis). A single OM3 initiative consists of hundreds of OM0–OM1 steps.

Appendix C: Evidence Tiers (E) — Detailed Criteria

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.

Appendix D: Classification Methodology

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 $\geq$2 keyword hits each, yielding an average of 2.3 episodes per multi-episode user.

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

Appendix E: Aggregate Impact Derivation

E.1 Human Effort Equivalent — Method

HEE hours are computed per OM band using **geometric midpoints** of each range. Geometric means are standard for log-spaced intervals and yield conservative estimates (below arithmetic midpoints) for right-skewed distributions within bands. OM4 is open-ended; we cap at 400h (1.25× the lower bound), well below known OM4 episodes (e.g., 826-file GIS dataset likely exceeding 1,000h).

OM Band	HEE Range	Geometric Midpoint	Files (N)	HEE Subtotal
OM0	≤1h	0.5h	513	257
OM1	1–8h	3.0h ($\sqrt{8} \approx 2.83$, rounded)	4,011	12,033
OM2	8–80h	25.0h ($\sqrt{640} \approx 25.3$)	2,623	65,575
OM3	80–320h	160.0h ($\sqrt{25{,}600} = 160$)	516	82,560
OM4	>320h	400.0h (capped)	55	22,000
Total			**7,718**	**182,425**

Sensitivity analysis. Using band minimums (lower bound): 84,000h = 41 person-years. Using band maximums (upper bound, OM4 capped at 640h): 443,000h = 213 person-years. The primary estimate of 182,000h sits at the 27th percentile of this range — conservative by construction.

E.2 Model A: Artifact Replacement Cost

Rates reflect what a client would pay a qualified professional to produce the equivalent deliverable. Rates scale with OM because higher-magnitude work requires more specialized expertise and carries higher market rates.

OM Band	HEE Hours	Rate ($/h)	Rate Basis	Value
OM0	257	$50	Virtual assistant / admin task rate	$13K
OM1	12,033	$100	Professional freelancer (Upwork/Toptal median)	$1.2M
OM2	65,575	$160	Senior consultant / boutique agency rate	$10.5M
OM3	82,560	$188	Specialist consulting (multi-phase initiatives)	$15.5M
OM4	22,000	$200	Program management (conservative vs. Big 4 at $300–600/h)	$4.4M
Total	**182,425**			**$31.6M**

E.3 Model B: BLS Occupational Equivalence (Cross-Validation)

Independent validation using a different decomposition axis and rate source. Instead of pricing by complexity tier (OM), we price by professional domain (OP) using U.S. Bureau of Labor Statistics Occupational Employment and Wage Statistics (May 2024). Each OP is mapped to its closest

Standard Occupational Classification. Rates are median hourly wages — what the professional earns, not what the client pays.

HEE hours per OP are computed from the OP × OM cross-tabulation (Figure 1) using the same geometric midpoints from E.1.

OP	BLS Occupation (SOC)	Median $/h	HEE Hours	Value
IS	Research Analysts (13-1161)	$50	11,351	$568K
AP	Software Dev / Content (15-1252, 27-3043)	$55	96,891	$5,329K
DS	Management Analysts (13-1111)	$50	17,593	$880K
OA	Computer Systems Analysts (15-1211)	$50	5,885	$294K
IN	Mental Health Counselors (21-1014)	$30	16,901	$507K
TF	Financial Analysts (13-2051)	$55	8,566	$471K
CA	Compliance Officers (13-1041)	$45	2,547	$115K
SA	Postsecondary Educators (25-1000)	$40	11,240	$450K
HO	Health Educators / Dietitians (21-1091)	$35	10,199	$357K
SP	Info Security Analysts (15-1212)	$55	57	$3K
Total		**wt. avg $50/h**	**181,230**	**$8,974K**

E.4 Comparison

	Model A (Replacement Cost)	Model B (BLS Wages)
Decomposition axis	OM (complexity tier)	OP (professional domain)
Rate source	Market replacement rates	BLS median hourly wages
Total	**$31.6M**	**$9.0M**
Ratio	3.5×	1.0× (baseline)

The 3.5× ratio between Models A and B corresponds to the well-documented gap between what professionals earn and what clients pay for their output. Professional services firms typically bill at 2–4× the practitioner's salary to cover overhead, profit margin, and the value premium of finished deliverables over raw labor time. The ratio falls squarely within this range.

Model B establishes a **labor-cost floor**: even at bare median wages with zero overhead, zero profit margin, and zero value premium, the work represents $9.0M. Model A estimates the **market value of the deliverables**: what a buyer would pay for the finished output. The true economic value likely sits between the two, depending on context.

Both models — using different decomposition axes, different rate sources, and different pricing philosophies — converge on the same order of magnitude, validating the aggregate impact estimate.