AI Native Business Benchmark

Vendor neutral measurement protocol and validation plan

Consultation edition 0.51 | 6 November 2026 | Founding sponsor: Alien Cortex | aliencortex.com

At a glance

The AI Native Business Benchmark measures how much useful value a business produces for every hour of human involvement, and whether that holds or improves over time. It is open and vendor neutral: any business can run it with any AI tools. aBOS is one implementation, not a requirement. aBOS (Alien Business Operating System) is the commercial product of Alien Cortex (aliencortex.com), the founding sponsor: it installs a private team of AI agents for a business, with people approving consequential actions and finance work kept on local hardware.

The headline: Value per Human Hour

Measure	Definition
Value per Human Hour	Profile-defined realized value divided by all attributable human hours, for the same scope and period. For most businesses, value is gross profit (section 6).
Human minutes per $1,000	The same result inverted: 60,000 x human hours / value, published only when value is positive. It falls toward zero as people move above the loop.
Trajectory	The change against the business's own declared baseline, in dollars per hour and, where the baseline is positive, in percent (sections 10 and 11).

Human hours count everyone: owners, staff, contractors and outsourced services, including time spent prompting, reviewing, approving and fixing AI work (sections 2 and 5).

No headline without its guardrails

- Quality status: qualified, breached, insufficient evidence or provisional (section 7). A rising headline with a breached quality condition is not an improvement claim.
- AI cost: full AI and program cost shown beside the headline, so a gain bought with runaway spend stays visible (section 6).
- AI operating coverage: the share of reference workload completed by qualified AI execution (section 3).
- Evidence: completeness, uncertainty and method version (sections 8 and 13). Observed change is never presented as AI-caused without the evidence in section 9.

The headline is a communication aid, not a composite maturity score. It is always reported with this context and is never used to rank unlike businesses (section 9).

Two ways to take part

Track	Who it suits	What it takes
Lite	Businesses under about 50 people without a data team.	A frozen one-page boundary, three numbers a quarter, a coverage count for the top three workflows and a quality check. Estimated at about an hour to set up and 30 minutes each quarter. See the Lite self-assessment below.
Full	Larger organizations, assured reports and published comparisons.	The complete measurement contract, ledgers, reproduction package and assurance disclosure in sections 2 to 13.

Pilot participation, the starter kit and updates to this consultation are coordinated through aliencortex.com. Comments are welcome from any business, provider or reviewer, and every comment is considered under the same open process (section 16).

Name, status and licence

- Name: AI Native Business Benchmark. The working titles AI Leverage Benchmark and Open Business AI Benchmark used in earlier drafts are retired.
- Status: consultation edition 0.51. The metrics, thresholds and comparisons still require independent validation (section 14).
- Licence: not yet granted. Proposed: a Creative Commons Attribution licence for the specification and an open-source licence for the calculator, chosen with patent counsel before public release (section 15). Until then, “open” describes the intent, not a licence grant.
- Founding sponsor: Alien Cortex (aliencortex.com), credited without preferential scoring or control of the rules.
- Starter kit: a local reference calculator computes the Lite headline, coverage, leverage and a two-point trajectory. It uploads nothing.

Lite self-assessment

The Lite track uses the same definitions as the full protocol with proportionate evidence. Its results are self-reported and not assured, and they are for a business’s own trend. Do not compare Lite results between companies or publish rankings from them.

Step 1. Freeze the boundary before you start. Write down the entity; the period (one full quarter); what counts as value, normally gross profit (revenue minus direct cost of goods and delivery); and whose time counts, which is everyone in section 2, including contractors and agencies. Change it only with a dated amendment.

Step 2. Record three numbers for the period. Value V; all human hours H, from timesheets, calendars or supplier invoices, with the method written down; and full AI cost, including subscriptions, APIs and a share of local hardware.

Step 3. Calculate the headline. Value per Human Hour = V / H. Human minutes per $1,000 = 60,000 x H / V, only when V is positive.

Step 4. Count AI coverage for your top three workflows. For each workflow, record the cases due, the cases completed by AI without substantive human repair and accepted, and an agreed reference time per case. Coverage = reference hours of qualified AI cases / reference hours of all due cases (section 3).

Step 5. Check quality. Count reopened, refunded or corrected cases and any incident, then record one status: qualified, breached, insufficient evidence or provisional (section 7).

Step 6. Repeat after 90 days under the same rules. Report the change in dollars per hour and, if the baseline is positive, in percent. A Lite report states observed change only, never an AI-caused effect (section 9). If the rules changed, start a new series.

Worked example (illustrative). A six-person agency records $180,000 of gross profit and 2,400 human hours in a quarter: $75 per human hour, or 800 human minutes per $1,000. AI cost was $3,000. The next quarter it records $189,000 and 2,100 hours: $90 per human hour, up $15 or 20%, with quality status qualified. That is an observed change; on its own it does not show that AI caused it.

The starter kit includes lite_template.json and a --lite mode in the reference calculator that produces these figures. It is available through aliencortex.com.

Purpose and recommendation

Establish a common, auditable way to measure useful economic or mission value relative to all human involvement, and how reliably that relationship is maintained or improved over time. The standard must work for businesses using aBOS, competing platforms, internally built systems or mixed technologies.

Develop an open measurement standard that any business can run and any provider can implement. Report current performance and its trajectory separately. aBOS will be one built-in implementation. Independent validation and shared governance must precede claims of an accepted industry standard.

Two complementary questions

Question	Required result
1 What useful value does the business produce for all human involvement?	Current profile-defined value and human hours, their ratio and underlying totals; operational leverage and supervision burden; matched quality and full-cost evidence.
2 How reliably is that relationship maintained or improved over time?	Comparable performance series, absolute and permitted relative changes, elapsed interval, uncertainty, quality and cost changes, confirmation evidence and attribution class.

Evidence that gives both answers meaning

Report AI operating coverage, business scope, completeness, quality and control status, resource cost, observation maturity, method and profile versions, and verification procedures alongside both answers. These are required context, not ingredients in an arbitrary weighted total.

The two questions do not by themselves establish AI nativeness or an AI-caused gain. Pricing, demand, conventional automation and capital investment can change performance. Keep the role of AI and causal evidence separately visible.

Meaning of AI native

For this working draft, AI nativeness describes the extent to which material business processes are organized around substantive AI execution, connected to authoritative business information, operated within an explicit authority policy, and evaluated against accepted outcomes. This is a proposed operational definition to validate with businesses and independent reviewers.

Higher autonomy and faster percentage growth are not automatically preferable. A business may retain human judgment or maintain an already efficient process. A new business may have no comparable history; report its trajectory as unavailable, not zero.

Status: consolidated consultation and pilot edition. The proposed metrics, thresholds and comparisons still require independent validation. Credit the founding sponsor without granting preferential scoring or sole control of the rules.

1 Standard structure and permitted claims

Three parts with distinct purposes

Part	What it standardizes	What it does not imply
Common measurement protocol	Definitions, scope, identities, human time, resource cost, accepted outcomes, evidence and calculation rules.	A universal ranking of unlike businesses.
Sector and workflow profiles	Outcome units, complexity strata, quality floors, observation horizons, reference effort and allowable comparisons.	Freedom to redefine success after observing results.
Independent assurance and benchmark population	Verification procedures, challenge handling, reference cohorts and transparent statistical comparisons.	That measurement conformance proves excellent performance or guarantees future results.

A standard can be widely applicable without one universal score

An invoice review, a completed construction milestone, a resolved customer problem and a retained learning gain are different outcomes. Their measurement contracts can follow the same rules without treating their economic or social value as interchangeable. Use within-business trends and profile-matched peer comparisons; do not rank a hospital, a retailer and a software company on one unqualified number.

Claim grammar

Every report must identify its entity or process, dates, profile and method versions, scope and completeness, current level, trajectory or reason it is unavailable, uncertainty, quality status, attribution class and assurance disclosure. A narrow process result must not silently become an enterprise claim.

Illustrative form: For scope S and period T, value per all human hours was X under profile P and method V. Compared with the declared baseline, the change was D and trajectory status R. AI coverage was C%; quality status was Z. Attach the underlying totals, cost bridge and evidence limitations. These are placeholders.

Terminology used in this proposal

Must indicates a condition for conformance to this proposed protocol. Should indicates a recommended practice with a documented reason for departure. May indicates an optional disclosure. These terms do not assert endorsement or approval by a standards organization.

Boundary of the claim

The protocol can verify that a number was measured consistently. It cannot make all causal assumptions incontestable, prove the value of every human purpose, or guarantee future performance. Objections should be resolved through published definitions, evidence, replication and an independent appeal process.

2 Measurement contract and organizational boundary

Create and freeze a measurement contract before the evaluation period. Preserve every amendment with its effective date. The same contract must govern the baseline, intervention and comparison cohorts.

Required field	Minimum content
Entity and scope	Legal or operating entity; included functions, sites and value-chain boundary; all material exclusions and the reason.
Process inventory	Stable top-level business processes and outcomes, including manual and outsourced work; source used to establish completeness.
Outcome and acceptance	What counts once; who or what accepts it; quality, timeliness, substantive completion and reopening rules.
Cohort	Assignment or original due-date rule; treatment of retries, cancellations, rejected cases, fraud and genuine duplicates.
AI contribution	Material AI role, permitted execution modes, authority policy and evidence distinguishing substantive AI from a trivial add-on.
Reference effort	Process and complexity strata, source of reference hours, uncertainty, baseline dates and approved version.
Human and cost scope	Time categories, service-provider boundary, shared-resource allocation, missingness and sensitivity rules.
Quality and impact	Predeclared error, severity, service and affected-segment criteria; critical incidents; discovery horizon.
Evaluation design	Independent acceptance, sampling, holdouts and causal method; trajectory baseline, materiality, guardrails, confirmation horizon and time basis.
Reporting and assurance	Cadence, observation maturity, rounding, confidentiality, verification procedures and dispute owner.

Comparable denominators

The primary boundary includes internal employees, founders and the attributable human services purchased to execute the measured business processes. It does not extend automatically to every employee in every upstream supplier. A direct service provider performing an outsourced business function is included; the general staff of a utility or model provider is not counted as the client's labor input.

Disclose purchased automation and purchased human services separately. A firm cannot improve its human leverage simply by moving the same work to an uncounted agency or contractor. Material unknown outsourcing effort prevents a fully verified all-human-hours claim.

Establish intake completeness against an upstream source independent of the evaluated AI queue. Otherwise difficult cases could disappear before measurement. Process redesign may eliminate tasks, but it must preserve the underlying business outcome and its denominator.

Include project, creative, research and exception work through appropriate profiles. Long-duration work may use independently accepted milestones, with final outcomes reported later; unfinished work must not be relabeled as complete. Disclose material effort shifted to customers or other parties outside the labor boundary.

3 Qualified AI operating coverage

Use top-level business outcomes, weighted by frozen reference human effort. Agent counts, task counts, seats, tokens and model calls are not measures of organizational AI nativeness.

For each process and complexity stratum s: b is reference human hours per case; N is every in-scope case due in the period under the original cohort rule; A is the subset completed through substantive AI execution within its authority envelope and meeting the acceptance, quality and timing requirements. A cannot exceed N.

What qualifies

- The AI must make a material contribution to the declared outcome. A token AI step attached to a conventional workflow does not qualify the entire process.
- The defined execution scope must be completed without substantive human construction or repair. A required decision-ready authorization may remain, and its time must be counted.
- Acceptance must have sufficient independence from the executor for the risk involved: maintained tests, reconciliation, customer confirmation or authorized judgment. A profile may permit sampling with an explicit estimate; sampled assurance is not verification of every case.
- All failures, abandoned attempts and human rerouting remain in the due-case denominator unless a predeclared exclusion rule applies. Retried work is one business case.

Mutually exclusive case classes

Completed class	Interpretation
Human only	Accepted outcome with no substantive AI execution.
AI assisted	Accepted outcome with material AI contribution and substantive human execution or repair.
Qualified AI with authorization	AI completes execution without substantive human repair; case-specific human authorization remains.
Qualified AI without case touch	AI completes execution without case-specific human execution or release authorization.

Classify failed, incomplete and unknown cases separately. A comprises only the two qualified AI classes. Report every class share, including valuable AI assistance; these classes are not a universal ladder of desirability. Retrospective assurance still counts as human effort but is not case execution or release authorization.

A qualification count may be observed or estimated under a profile-approved sampling design. Estimates must disclose the sampling frame, selection method, estimator, uncertainty and unobserved cases. Do not describe extrapolated qualification as directly observed success.

4 Reference effort and scope coverage

Two different reference systems

Reference	Permitted use	Requirement
Company baseline	Estimate change within the same business.	Representative measured work under the prior operating model, with quality, mix and labor boundary held comparable.
Common profile reference	Compare matched businesses or systems.	Independent, versioned timing study or reference dataset for the same tasks, complexity strata and acceptance standards.

Company-specific baseline hours are not automatically comparable between companies. A slow historical process can produce a large improvement multiple. That may demonstrate local improvement but cannot establish superiority to a leaner competitor.

If the company was AI-native from birth, use a documented reference study or a contemporaneous comparison. Do not invent a pre-AI baseline. If an acceptable reference cannot be established, publish raw outcome and effort measures and withhold the normalized multiplier.

Evidence gaps must remain visible

A small process pilot cannot support an enterprise-wide percentage. Report intake-case reconciliation, outcome-status coverage, human-effort measurement coverage, cost reconciliation and the share of reference workload with established weights. Define each denominator; an unknown denominator is not an omnibus completeness percentage. A missing material process inventory prevents an enterprise coverage claim.

Where the inventory and reference workload are known but qualification evidence is missing, report bounds. The lower bound credits only demonstrated qualified work; the upper bound additionally treats every unknown case as qualified. Do not present the lower bound as an exact point score, and do not remove unknown work from the denominator.

For the optional fixed-mix index, strata with no due cases have unknown observed performance. Do not assign them zero or 100%, and do not renormalize the remaining weights. Report an interval or withhold the index. A capability test is separate evidence, not proof that production work occurred.

Eligibility and physical work

Publish total-scope coverage first. A separate automation-eligible view may be useful, but it must expose the excluded workload and reasons. Do not make the eligibility definition a way to hide difficult or legally restricted work.

For a business combining knowledge and physical work, measure each declared process boundary honestly. Automating scheduling does not mean AI performed the surgery, construction or delivery. A lower AI-operated share does not imply an inferior business.

Weight controls

Reference weights must be approved before scoring, estimated with uncertainty and tested for sensitivity. Preserve counts and strata. Changes require a new version and series bridge. Publish actual workload mix; any fixed-mix trend index must be labeled standardized coverage.

Example: AI completes 95 easy cases at one reference minute each; humans complete five complex cases at 100 minutes each. Case-count coverage is 95%, but reference-workload coverage is 95 / (95 + 500) = 16.0%. Freeze complexity rules so hard cases cannot be relabeled.

5 Human leverage and supervision

The central operational question is how much accepted work the business achieves per unit of human involvement. Separate all human labor from the narrower supervision burden.

Q is accepted work meeting the quality and timing policy. W uses reference human hours: leverage of 4.0 means four reference hours of accepted work per actual human hour. This is operational leverage, not a measure of economic or social value; value must be reconciled separately.

Qualified AI oversight leverage credits only outcomes A using the same reference weights. Its denominator includes oversight of every assigned AI case, including failures and manual rescues. Manual fallback can count toward total accepted work, but cannot inflate successful AI supervision.

Denominators that cannot hide work

Measure	Include
All attributable human hours	Execution, prompting, briefing, review, required authorization, corrections, escalation, monitoring, recovery, management and allocated implementation and maintenance.
Oversight hours	The mutually exclusive subset spent reviewing, authorizing, correcting, monitoring and handling exceptions for the same AI-supervised cohort.
Reserved supervision capacity	Staffing or on-call time that must remain available, even when active intervention is low. Report separately from active attention.
Failed work	All related time and cost, even when no accepted outcome is credited. Late defects revise the original cohort under the restatement rule.

For homogeneous work, report human minutes per accepted outcome and oversight minutes per qualified AI outcome. Required authorization is not an error; substantive repair is not merely an approval.

Implementation and period accounting

If setup is material, publish operating and full-lifecycle leverage separately, with the setup allocation horizon and sensitivity. Cohort effort and period staffing are distinct views. Never divide outcomes from one cohort by time from another.

Do not double count simultaneous work or infer attention from an open browser page. Include unpaid founders and external human services. Time studies and sampled estimates are acceptable when direct capture is disproportionate; disclose method and uncertainty.

Zero measured hours require investigation and a not-estimable result. Sustainable supervision capacity requires workload stress tests with quality, backlog and response constraints. Agent count is not a substitute.

6 Economic performance and AI cost

Company productivity is an outcome measure

This is the current-level answer to the first organizing question. For suitable profiles, use gross profit per human hour with an accounting reconciliation. Value and hours must share the same scope and period. A trailing 90-day view is an option, not a universal rule; Sections 10 and 11 define its trajectory.

Use hours actually worked, including relevant support and required availability; disclose paid labor cost and reserved staffing as well. Do not substitute headcount or focused coding hours for the complete labor input. This distinction is consistent with OECD productivity measurement guidance [7].

Reconcile direct labor, AI costs, contractors, refunds, depreciation and shared overhead. Gross profit depends on accounting classification and business model. Banks, public bodies and pre-revenue organizations may need a sector-defined value or mission measure; do not force all organizations into gross profit.

Observation is different from attribution

Use this ratio only with a positive meaningful baseline. If value is zero or negative, report absolute measures and changes. Price, mix, market demand, outsourcing and accounting changes can move the ratio. A before-after increase is not itself proof AI caused it.

Cost efficiency and return are different

Gross profit divided by AI spend may be disclosed as AI spend efficiency. It is not return on investment because it credits AI with the entire business's profit. Keep it secondary and do not use it alone for performance payments.

The benefit ledger must distinguish incremental realized contribution, cash costs actually avoided, and released capacity. Unused staff time is not a realized cash saving. Do not count both imputed time savings and the margin from redeploying the same time. Assign every cost once; costs already deducted in B must not be deducted again in C.

Return ratios require positive program cost C. If C is zero or negative, publish net benefit and ledger components. Without a credible counterfactual, label attributed benefits unestablished.

AI cost includes applicable APIs, subscriptions, local hardware allocation and depreciation, energy, hosting, storage, evaluation, security and maintenance. Program costs also include implementation, integration, training and incremental human operations under the selected accounting boundary. Show marginal and full costs separately when they differ materially.

Per-task estimated dollar value is a forecast until reconciled to a real economic outcome. Several tasks contributing to one sale must not each receive the sale's full margin. Financing raised, invoice face value and program dollars spent are not gross profit or verified benefit.

7 Quality and control qualification

Publish raw performance metrics even when a quality condition fails, with an explicit qualification status. A failed result must remain visible. Never allow high throughput to average away a critical control failure.

Required view	Rule
Outcome acceptance	Publish eligible, completed, accepted, rejected, late and reopened cases, with original cohort identities.
Production escapes	Record independently discovered defects after release, including those approved by a reviewer. Report severity counts and customer impact.
Shadow reviewer tests	Use hidden known-bad and known-good cases away from customer production. Report detection, false rejection, review time and test composition.
Control effectiveness	Test permission, privacy, authorization, rollback, containment and escalation requirements applicable to the profile.
Affected segments	Predeclare relevant risk and customer groups, thresholds and sampling. Do not let aggregate success conceal material subgroup harm.
Service and capacity	Publish latency, tail latency, backlog, availability and escalation response where they affect the promised outcome.

Do not reward blind approval or suppressed escalation

A falling correction rate may indicate improved outputs, weaker inspection, easier cases or overloaded reviewers. It is not proof of learning by itself. Evaluate it with independent outcome checks, reviewer challenge performance, workload mix and escaped defects.

Correct abstention and referral can be successful outcomes when the measurement contract defines the task as making a safe disposition. In a task requiring completed delivery, an abstention remains incomplete delivery even if it is the correct safety action. Preserve both facts.

Thresholds must be chosen before evaluation

Each profile must justify quality and impact thresholds using the applicable risk, service commitment and domain evidence. A 1% error rate is not universally acceptable. A contractual threshold, a scientific estimate and a normative risk tolerance are different things and must be labeled accordingly.

Qualification states

Qualified means the stated requirements are supported for the stated scope and period. Breached means a condition failed. Insufficient evidence means data cannot establish the requirement. Provisional means the observation horizon or reconciliation is still open. These states do not certify all future behavior or every part of the company.

A localized workflow failure may invalidate only that scope. A shared identity, data-integrity or access-control failure can undermine a wider claim. Define the propagation rule before evaluation and preserve the incident record.

8 Confidence observation windows and missing data

Every important estimate needs its evidence context

Report counts, exposure, sampling design, coverage, measurement uncertainty and observation dates. An error rate of 0/12 is not equivalent to 0/12,000. Separate production observations from synthetic challenge tests. Correlated cases or one repeated template do not constitute independent evidence.

Illustrative zero failure calculation

For independent, representative Bernoulli trials with zero observed failures, the exact one-sided 95% upper bound on the failure probability is 1 minus 0.05 raised to the power 1/n. The examples below follow that model; they are not universal audit sample requirements. Exact binomial methods are described by NIST [8].

Zero failure observations	Upper bound	What the example means
59	4.9508%	Just enough under these assumptions to place the upper bound below 5%.
299	0.9969%	Just enough under these assumptions to place the upper bound below 1%.
2,995	0.099974%	Just enough under these assumptions to place the upper bound below 0.1%.

Clustered workflows, distribution shift, selective inspection and rare catastrophic failure modes require a different design. Use clustered or hierarchical inference where appropriate. Report measurement error in human time and reference weights separately from sampling uncertainty; a narrow statistical interval cannot repair biased inputs.

Live and mature views

Operational telemetry may update continuously. Financial measures follow a reconciled accounting cadence. Quality outcomes mature only after their profile-specific exposure and discovery horizon. A 28-day dashboard cannot finalize its newest cases under a 60-day defect rule.

Publish provisional and matured cohorts with clear dates. Late defects, refunds and data corrections restate the original cohort, retain its original publication and explain the change. The profile may require longer follow-up than 60 days; choose it for the failure process, not for convenience.

Missingness and supplier estimates

Unknown time, costs or quality must not become zero. Prefer measured or attested supplier hours. If estimation is necessary, separate labor from materials, equipment, overhead and margin, then publish a plausible rate and labor-share range. Dividing every vendor invoice by one blended hourly rate is not a defensible universal method.

Repeated qualification decisions require prespecified evaluation dates and rules, or an appropriate sequential method. Joint statistical requirements need a disclosed approach to multiplicity. Fixed-sample intervals do not automatically support continuously monitored or joint claims.

Undefined ratios

Publish ratios only when denominators are established with the required sign. No due workload means no observed coverage; no AI spend means no finite spend-efficiency ratio; absent reference weights prevent weighted totals. Publish underlying quantities and the reason, never zero, infinity or an invented default.

9 Comparisons and causal claims

Four levels of inference

Claim	Evidence required	Permitted conclusion
Measured level	Defined scope, verified inputs and reproducible calculation.	This is the observed performance within this boundary.
Observed improvement	Comparable predeclared baseline and current cohorts, with mix and accounting differences shown.	Performance changed; causation is not established.
Attributed improvement	A justified controlled, randomized, phased or credible quasi-experimental comparison with sensitivity analysis.	An estimated AI-related effect under stated assumptions.
Generalizable benchmark	Independent replication across a defined population, held-out organizations and stable profile definitions.	Comparison applies to the supported population, not every business.

Fair peer groups

Before publishing a percentile, establish a reference population with comparable outcome definitions, reference effort, task complexity, quality floors, industry, business model, company scale, outsourcing boundary and accounting conventions. Report recruitment, sample size, period and uncertainty. Country and currency adjustments require their own explicit method.

Self-selected early adopters do not represent all businesses. Do not manufacture a percentile from a handful of aBOS customers or rank a company against a peer set chosen after seeing its result. No cross-industry league table is proposed in version 0.51.

Prospective measurement and blinding

Prospective means evaluation cases arise after the protocol is frozen. Blinded means specified parties cannot access test labels, assignments or hidden challenges in ways that bias the result. Live work is not automatically blinded. Record model, prompt, workflow, policy and evaluator versions.

Independently maintained holdouts must remain separated from optimization. Publish the evaluation specification and representative examples, while rotating protected challenge content. Releasing every live test answer would defeat the test.

Comparative trial design

Where feasible, assign comparable cases or periods to the new and reference processes before outcomes are known. Retain failed and redirected assignments in the analysis. Evaluate displacement and spillovers: an apparent improvement in one department may move work to another.

Current level and improvement trajectory are required separate views. The trajectory protocol follows in Sections 10 and 11. A faster-growing early-stage business is not automatically more capable than a stable efficient business. Test sensitivity to baseline, case mix, time estimates and cost allocation.

10 Performance trajectory and measurement rules

The second organizing question asks how reliably value per human involvement is maintained or improved. Publish current level and trajectory together. Neither is a weighted maturity component, and reporting a trajectory does not require a company to show improvement.

Value relationship and observed change

V is the profile-defined realized value, H is all attributable human hours, b is the declared baseline and t is the current period. Publish V, H, outcome volume and both ratio levels. Relative change requires a positive meaningful baseline; otherwise show absolute change. A tiny baseline requires explicit sensitivity, not a large growth headline.

State the elapsed interval and cohort reference dates. If a rate is reported, define its time unit and estimation method; a two-point descriptive rate is ΔP divided by elapsed time. A percentage change over one interval does not establish a continuing growth rate. Prespecify series, cadence, seasonality and dependence handling before making a trend claim.

Track the same performance dimensions

Dimension	Longitudinal disclosure
Value and human effort	Value per all human hours; raw value and hours; accepted reference workload and labor intensity. Reference workload must not be relabeled as money.
AI execution and oversight	Coverage changes in percentage points; qualified AI oversight leverage; required staffing, failures and manual rescues.
Economics and resources	Full cost per accepted reference workload, declared currency basis and a bridge for relevant costs outside the value numerator. Never deduct a cost twice.
Quality and service	Acceptance, escaped defects, severity, latency, backlog and affected segments, with exposure and mature observation horizons.
Evidence	Completeness, uncertainty, method and profile versions, comparison adjustments, verification and trajectory status.

Comparison contract

Freeze baseline selection, outcome definitions, boundary, complexity rules, time collection, cost allocation, acceptance criteria and observation horizon. Record model and workflow changes; do not freeze away the improvement being studied. Compare mature cohorts and aligned accounting periods. Show observed results and any justified adjustment for price, mix or other material changes.

Retain failed, redirected and abandoned assignments. Restate original periods for late defects or financial corrections. Treat overlapping rolling windows as dependent. A material method or boundary change needs a defensible series bridge; otherwise mark a break and start a new series. A net economic improvement claim must reconcile relevant costs outside its numerator.

11 Interpreting change and a worked trajectory

Trajectory status is separate from quality status

Profiles must predeclare the primary metric, guardrails, each metric's favorable direction or acceptable range, and consistent, nonoverlapping decision regions. Coverage and autonomy changes are descriptive unless the profile establishes a justified performance objective.

Trajectory state	Evidence required
Supported improvement	A comparable change in the favorable direction beyond the profile's predeclared materiality margin is supported by its uncertainty analysis.
Maintained within bounds	Evidence supports equivalence within a predeclared acceptable change range. Failure to detect a change is insufficient.
Supported deterioration	Evidence supports deterioration beyond the predeclared materiality margin.
Inconclusive	The comparison is interpretable, but uncertainty does not establish improvement, maintenance or deterioration.
Not comparable	A material scope, method, mix or accounting break has no defensible bridge. No usable baseline is reported as trajectory unavailable.

Profiles must justify margins and inference methods for the decision and measurement precision. Quality breaches, immature outcomes or unsupported cost tradeoffs prevent an unqualified claim of dependable improvement even when the raw ratio rises. Publish the ratio and every status. A claim of sustained improvement also needs the profile's prespecified confirmation horizon and evidence; repeated overlapping windows are not independent confirmations.

Example C Current level and trajectory

Illustrative arithmetic only. Three equal-length periods have the same declared boundary and accounting basis. The selected value numerator is contribution before the listed program costs; those costs have not already been deducted. Quality, uncertainty and attribution evidence would still be needed for qualified claims.

Measure	Baseline	Period 2	Period 3
Realized value before listed costs	$10,000	$11,000	$11,500
All attributable human hours	100	100	100
Value per human hour	$100	$110	$115
Program costs outside numerator	$1,000	$1,500	$1,600
Contribution after listed costs per human hour	$90	$95	$99

Value per human hour rises 10.0%, then 4.55%, for 15.0% overall. The level improves while the percentage pace slows. After the listed costs, contribution per human hour rises 10.0% overall. These are observed changes, not ROI, proof of an AI-caused effect or a forecast of continuing growth.

A business maintaining $200 per human hour is not automatically less capable than one rising from $20 to $30. Show both levels and trajectories. A ratio can also rise while total value falls; always retain the underlying value, hours, volume and service results.

12 Capability evidence profile

The original drafts assign points to five dimensions and label score bands as organizational maturity. Version 0.4 replaces that total with an evidence profile until construct validity and weighting have been tested. Composite indicators can be sensitive to their inputs, normalization and weights; the OECD and European Commission guidance makes those choices explicit [6].

Domain	Observable evidence	Status to publish
Operational AI coverage	Material AI role; complete process inventory; accepted outcomes and labor boundary.	Observed coverage with scope, uncertainty and operating modes.
Data integration	Authoritative source joins, freshness, lineage, permissions and handling of unavailable data.	Demonstrated, partial, insufficient evidence or breached.
Control and accountability	Named accountability; policy enforcement; escalation, recovery, change approval and applicable authority constraints.	Evidence of tested controls and outstanding failures.
Evaluation and measurement	Independent acceptance, calibrated evaluators, hidden challenges, time and cost reconciliation, reproducible reports.	Coverage and independence of evidence; unmeasured risks.
Controlled improvement	Versioned changes, held-out tests, regression detection, rollback and evidence-based retention of effective versions.	Tested change-control capability and the separate measured trajectory; justified decisions not to change.

What does not earn maturity points

- Buying more tools, spending more tokens or launching more agents.
- Hosting locally merely because the founder prefers local compute. Test actual data control and applicable constraints instead.
- Collecting feedback without validating the resulting change.
- Improving quickly from a weak baseline or hiding humans behind outsourcing.

Observed gains and AI-led improvement

Observed improvement, controlled change capability and AI-led improvement are distinct claims. The last requires records showing AI's role in proposing or executing changes. Include the human effort and resource cost of improvement itself. Model upgrades, feedback volume and deployment counts do not prove gains. AI causation requires a justified comparison; no past trajectory guarantees future improvement.

If a composite is proposed later

Require a theory of the construct, stakeholder agreement on its purpose, evidence that dimensions measure distinct relevant properties, reliable scoring by independent assessors, sensitivity analysis, validation on held-out organizations and transparent handling of missingness. Publish the dimension vector as well as the total.

A weighting rule is a design judgment, not a law of nature. A validated composite may serve a specific decision and population; it should not be marketed as a precise universal percent AI native. Score bands and badges require separate calibration and an appeals process.

13 Evidence model and reproducibility

Portable records with stable identities

Accept interoperable exports, APIs and auditable manual records. Continuous event telemetry is one implementation. Small firms must be able to participate through proportionate collection using the same definitions and explicit verification limitations.

Record family	Minimum evidence
Business case	Entity, stable outcome ID, process and complexity stratum, assignment and original due date, scope and acceptance contract.
Execution	Run and attempt IDs, parent case, material AI role, system and policy versions, artifact references and substantive human handoffs.
Human effort	Pseudonymous actor or role, time category, duration, collection method, allocation rule and uncertainty.
Resource and cost	Usage, rate source, currency, full or marginal costing, local resource allocation and duplicate-charge controls.
Acceptance and incident	Independent evaluator or acceptance source, tests and versions, result, severity, release, discovery and reopening dates.
Economic realization	Accounting source, value type, adjustments, refunds, allocation and benefit attribution method.
Period and method	Window, cohort status, calculation version, missingness, report digest and amendment history.

Four linked ledgers

Maintain outcome, human-effort, resource-cost and incident ledgers. Financial realization can join them later using a stable business outcome ID. Retries and subagents never create extra outcomes. Financial estimates and recognized results must remain separate.

Every record needs an event identity, occurrence time, recording time, source and correction history. Evidence may be confidential while the method is public. Hashes can protect integrity after capture; they do not prove that the original event was truthful or the source population complete.

Reproduction package

Export the measurement contract, versions, data dictionary, authorized evidence extracts, input digests, calculation code, allocation rules, expected outputs, exceptions, quality cohorts and restatements. Independent implementations should reproduce deterministic values to the specified rounding and documented numeric tolerance.

Use data minimization, access controls and consented sharing. Public rankings must not require publication of customer records, employee-level surveillance or trade secrets. Auditors can verify private evidence under controlled access; public aggregate disclosure is a separate permission.

Assurance disclosure

Name the preparer and verifier, independence and conflicts, procedures, evidence examined, sample and scope, findings, limitations and report date. Calculation reproduction is distinct from checking source truth or population completeness. Do not imply a standardized assurance level before such levels are defined and validated.

14 Independent validation program

Validity must be demonstrated through a preregistered research and verification program. The following are proposed work packages, not claims that validation has already occurred.

Test	Design and acceptance decision
Content validity	Businesses, workers, domain experts and buyers identify the decisions the benchmark must support. Remove irrelevant vendor preferences.
Construct validity	Verify that coverage, leverage, economics, trajectory and control are distinct; test meaningful operating differences without rewarding weak baselines.
Reliability	Two independent teams map and assess the same cases. Report disagreement and its causes; set tolerances before evaluating the holdout sample.
Reproducibility	At least two independently built calculators process reference fixtures. Require consistent results, explicit rounding and identical treatment of missing cases.
Causal sensitivity	Compare observations with controlled or phased pilots. Test price, mix, outsourcing, seasonality, timing and allocation; validate trajectory and maintenance claims.
Robustness and gaming	Try task splitting, agent multiplication, hidden labor, baseline inflation, trivial AI inserts, selective intake, suppressed complaints and duplicated value.
External validity	Replicate in organizations not involved in design, including smaller businesses, low adoption, public or mission-driven organizations and non-aBOS systems.
Operational burden	Measure time and cost of complying. Simplify collection without weakening definitions or disguising uncertainty.

Pilot recruitment proposal

A useful initial planning target is 12 to 20 organizations across three or four contrasting sectors, with multiple technology providers and independent assessors. This is a feasibility sample, not a statistically representative population or a sufficient basis for industry percentiles. Include organizations outside the sponsor's customers and at least some conventional baselines.

Choose confirmatory sample sizes after measuring variance, clustering, missingness and the minimum decision-relevant effect. Freeze the analysis and acceptance criteria before collecting the held-out validation sample. Do not declare success using thresholds selected after seeing the data.

Release gates

A version can advance only when its scope and terms are resolved, reference fixtures reproduce, material gaming failures are addressed, pilot limitations are published, and independent reviewers accept the remaining claims. Unresolved issues should narrow the claim rather than disappear into an average score.

15 Governance and path to adoption

Independent stewardship

Alien Cortex may initiate and fund the work. For an industry standard, the governing process should represent businesses, competing providers, researchers, accountants or assurance specialists, workers and domain experts. No single vendor should hold a controlling vote or veto over changes that affect its product score.

Publish a charter, conflict disclosures, meeting decisions, method versions, change proposals, public comment dispositions and an appeals process. Separate methodology decisions, commercial implementation and independent assurance. An auditor's compensation must not depend on whether a client scores highly.

Open implementation rights

The common definitions, schemas, reference fixtures and calculation methods should be available on terms that allow competing implementations. Before launch, determine intellectual-property and trademark terms, essential-IP disclosures and implementation rights. A standards claim cannot conceal a requirement to buy a proprietary product. This draft does not commit or license any existing patent rights.

Existing work to build on

Source	Relationship to this proposal
ISO IEC 42001	Existing AI management system requirements. Map governance evidence where appropriate; do not claim this benchmark is certified or endorsed by ISO [4].
NIST AI Risk Management Framework	Useful risk and evaluation vocabulary, including independent assessment and ongoing monitoring. The benchmark adds defined business-output and human-input measurements [5].
OECD and European Commission methods	Foundations for careful labor-input and composite-indicator design. Use them to test comparability and uncertainty [6,7].
Open standards development	Consensus, stakeholder participation and public review are established standards-development practices. Formal recognition is a separate process [9].

Two adoption routes

Begin with an open industry consortium and working draft that multiple vendors can implement. Use published pilots to establish demand and resolve defects. In parallel, explore a relevant standards-development channel through participating experts or a national standards body. Do not imply that an informal consortium has ISO authority.

OpenExO already offers an AI-native organizational model and an eight-dimension readiness self-assessment [14]. Explore collaboration and evidence crosswalks. Position this proposal around reproducible operating measurements; do not claim to have invented AI-native assessment.

A practical proposed sequence is: first month, charter and measurement contracts; months two and three, pilot implementation and independent replication; following months, public review, additional sectors and held-out validation. These are planning stages, not promised completion dates. Formal standardization and widespread adoption depend on external participation and demonstrated utility.

Make adoption useful: procurement comparisons, operating improvement, implementation acceptance and voluntary assurance. Free basic measurement plus optional implementation and independent review can reduce entry barriers. Do not make paying a fee equivalent to earning a good result.

16 Open release and community process

The project is intended to become an open specification with open-source calculation tools. Any company must be able to run a released version without an aBOS account, proprietary endpoint or permission from a vendor. This working draft has not been publicly released or assigned final licenses.

Public release component	Purpose
Specification and glossary	Definitions, formulas, claim limits and examples in editable source form.
Schemas and report templates	Portable inputs, completeness disclosures and reproducible output fields.
Profiles	Versioned workflow and sector definitions, with their supported comparisons.
Reference calculator	Local execution with documented dependencies; one implementation of the rules.
Fixtures and conformance tests	Synthetic examples, edge cases, expected outputs and adversarial cases.
Contribution and governance rules	Public issues, proposals, review decisions, conflicts, release rules and appeals.
Licenses and implementation rights	Explicit terms for use, modification, redistribution and competing implementations.

Independent execution is a requirement

No normative calculation, required profile or conformance test may depend on an exclusive commercial service. Each integration, including aBOS, must map its evidence into the public schema and apply the same published method and profile versions. Extensions must be identified and must not silently alter a standard result. Passing tests does not verify a company's records.

Changes must be open to challenge

Anyone may propose a defect, improvement or profile without being a customer. A substantive proposal must state its rationale, evidence, changed wording or code, worked cases, affected tests, comparability impact and conflicts. Maintainers publish decisions and reasons; disputed methodology has an independent appeal path. Community voting alone does not establish measurement validity.

Stable releases preserve meaning

Each report names immutable method and profile releases. Changes to definitions, weights, scope, qualification or missingness require a compatibility review. Stable releases require public fixtures to pass, independent calculation reproduction, disclosed limitations and public review of material changes. Historical reports remain available; bridges or restatements are explicit.

Separate four claims

Open source establishes rights and practical access. Conformance establishes implementation consistency. Report assurance describes checks on specific evidence. Validation supports an interpretation for a defined population and purpose. None of these claims automatically supplies the others.

Public methods do not require public customer records. Keep confidential company evidence local or under authorized review. Rotate protected evaluation cases while publishing the methodology and representative examples. The founding sponsor may provide the best implementation without controlling competitors' access to the rules.

17 Worked examples and conformance challenges

Example A Human leverage without agent counting

Illustrative only. An AI-supervised cohort produces 150 accepted outcomes. The common reference is one human hour per outcome. Total attributable human work is 75 hours, of which 15 are active oversight. Of 150 due cases, 120 meet the governed-execution rule and 30 require substantive human execution. All 150 meet the acceptance and timing policy.

Measure	Calculation	Result
Accepted reference workload	150 outcomes × 1 reference hour	150 reference hours
Total human leverage	150 / 75	2.0×
Qualified AI oversight leverage	120 qualified AI reference hours / 15	8.0×
Total human minutes per outcome	75 × 60 / 150	30 minutes
Oversight minutes per qualified AI outcome	15 × 60 / 120	7.5 minutes
Qualified AI operating coverage	120 / 150 with equal reference weights	80%

These results do not prove profitability, company-wide AI nativeness or sustainability at greater load. Disclose reserved supervision staffing, setup allocation and quality maturity. Replacing five agents with fifty changes none of these figures if the accepted work and human effort are unchanged.

Example B A large spend efficiency ratio can hide a loss

Illustrative only. A company reports $100,000 gross profit and $10,000 AI spend. The ratio is 10. That does not establish positive return if an equivalent counterfactual would produce more profit without the AI program. Calculate the incremental benefit and cost separately.

In a separate illustrative benefit ledger, verified benefits before incremental program cost are $30,000 and total incremental program cost is $20,000. Net benefit is $10,000, the benefit-cost ratio is 1.5, and ROI is 50%. These figures are valid only if the attribution and accounting boundary support them and no costs or benefits are counted twice.

Required adversarial fixtures

The reference test suite should show that splitting a workflow, multiplying agents, retrying a case, moving labor to a contractor, delaying a due date, rejecting every shadow case, omitting a defect, inflating a baseline or booking one sale's margin many times cannot produce an unqualified improvement. A policy-required approval must count as human time without being mislabeled as a defect.

18 Review of earlier drafts

Nonnormative background. This review explains design decisions; conformance does not require these earlier drafts.

Draft 0.1 and draft 0.2 are each one page and dated 5 November 2026. Source references are [1] and [2].

Draft element	Decision in version 0.4
Value per human hour	Retain as an economic measure with aligned periods and accounting scope. Separate it from AI nativeness.
Draft 0.2 baseline multiple	Retain for matched within-company trends. Do not treat it as causal proof or a universal peer rank.
Draft 0.2 outsourced hours	Retain the boundary; replace spend divided by one rate with measured or disclosed estimated labor effort.
Draft 0.2 total AI spend	Retain full-cost scope. Rename whole-company profit per AI dollar as spend efficiency, not ROI.
Draft 0.2 removal of oversight	Reverse. Restore qualified AI workload per oversight hour, alongside total human leverage.
Draft 0.2 growth and token fluency points	Retain trajectory as a core reported result. Remove growth points from the maturity total; report tokens as resource usage.
Draft 0.2 vendor-neutral data control	Retain. Local compute is an implementation option; actual control obligations matter.
Draft 0.2 shadow tests and harm floors	Retain with false-rejection rates, cohort maturity, independent sampling and risk-specific thresholds.
Proposed span of 1 to 50 or 100	Do not adopt as a universal target. Agent topology is easy to change without improving work.
Proposed continuous measurement	Retain as an implementation option already present in draft 0.1. Reconcile economic outcomes rather than invent task-level value.
Automatic payment and maturity bands	Commercial settlement is outside measurement conformance. Withhold universal maturity bands pending validation.

Attribution corrections

Solve Everything, by Alexander D. Wissner-Gross and Peter H. Diamandis, supports evaluation before automation and outcome-oriented benchmarks. The sentence about a benchmark living in an operating system is wording in draft 0.1, not a verified quotation from the essay. Neither a universal 1:50 nor 1:100 ratio was established in the reviewed sources [3].

Official YouTube auto-captions confirm the value-to-human-involvement proposal at 37:05–37:22, the related person-years and AI-use-efficiency discussion at 37:40–37:58, and process-improvement rates at 56:03–56:27. Captions lack reliable speaker labels and were not independently checked against audio. This specification's formulas and thresholds remain proposed design decisions [10].

19 aBOS as one reference implementation

Nonnormative implementation notes. This appendix is not required for conformance. aBOS (Alien Business Operating System) is the commercial product of Alien Cortex, the founding sponsor: it installs a private team of AI agents for a business, with people approving consequential actions and finance work kept on local hardware. Because the sponsor sells aBOS, this section is a disclosed interest, not an endorsement. The public standard is independent; aBOS will include a built-in adapter and calculator using the same published releases available to every other company.

Nothing in the protocol requires aBOS, its agent structure, hardware, local-first policy or commercial model. Other systems must be able to conform and receive equivalent results under the same evidence rules.

aBOS component	Proposed measurement responsibility
AI Audit	Establish process inventory, business boundary, baseline, outcome contracts and evidence sufficiency.
Governance and review queue	Link decisions, authorizations, substantive corrections and active human time to the business case.
Loop closer and source systems	Verify acceptance, unresolved cases, reopening and downstream outcome status.
Model routing and local Finance	Capture resource costs and reconcile actual economics inside the authorized data boundary.
Console and export	Display scope, metrics, uncertainty and qualification; export reproducible evidence and calculations.
Factory and workflow packs	Version the measurement service and profile adapters without changing control authority.

Preserve aBOS commitments

The supplied 18 October 2026 architecture requires named-human approval for irreversible actions and makes money movement recommend-only. An accepted benchmark milestone can create an invoice-eligibility record; it does not authorize aBOS to move money. Required approvals count as human effort without disqualifying governed execution [11,12].

Financial calculations remain local under that architecture. The vendor fleet is content-free. A cross-company research dataset requires a separate authorized disclosure process; the benchmark must not silently expand fleet telemetry or repurpose client data as a shared learning pattern.

Readiness boundary

The architecture reports override capture, operator-minutes measurement and evidence surfaces as incomplete or in build on 18 October. It also states that no cell yet ran on the new core and that delivery proof required real customers. These are dated document statements, not a current production assessment. Do not assign a live benchmark score from architecture documents alone [11,12].

First implementation pilot

Instrument the AlienCortex audit-to-proposal process or a tightly scoped sales-funnel stage. Define the accepted outcome, count all founder and operator effort, retain correct STOP decisions, capture downstream defects and reconcile costs. A draft or meeting is not recognized margin. Use an independently reproduced report as the first gate.

The sponsor's pilot proves instrument feasibility only. Industry validation must include competing implementations and organizations outside the sponsor's network. Workflows whose output is financing rather than gross profit, such as investor relations, need their own outcome profile.

20 Reference sources and evidence notes

References document the starting drafts, existing frameworks and factual examples. The proposed formulas, operating modes, validation program and governance choices are this working draft's design decisions. They have not been approved by the cited organizations.

[1] The AI Leverage Benchmark, draft 0.1. One page, 5 November 2026.

[2] The AI Leverage Benchmark, draft 0.2. One page, 5 November 2026.

[3] Wissner Gross and Diamandis. Solve Everything, especially Chapters 2 and 6. https://solveeverything.org/

[4] ISO. ISO/IEC 42001:2023 public description. AI management system requirements; no certification equivalence is claimed. https://www.iso.org/standard/42001

[5] NIST. AI RMF Core, particularly Measure. https://airc.nist.gov/airmf-resources/airmf/5-sec-core/

[6] OECD and European Commission. Handbook on Constructing Composite Indicators, 2008; European Commission Step 8 Sensitivity Analysis. https://doi.org/10.1787/9789264043466-en ; https://knowledge-for-policy.ec.europa.eu/composite-indicators/toolkit_en/navigation-page/10-step-guide_en/step-8-sensitivity-analysis_en

[7] OECD. Measuring hours worked, Compendium of Productivity Indicators 2019. https://www.oecd.org/en/publications/oecd-compendium-of-productivity-indicators-2019_b2774f97-en/full-report/component-40.html

[8] NIST. Exact Binomial Confidence Limits, Dataplot reference. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/exacbici.htm

[9] ISO. Developing standards. https://www.iso.org/developing-standards.html

[10] Moonshots. Episode 308, official YouTube video. English auto-generated captions reviewed 5 November 2026, especially 35:09–39:25 and 54:13–56:27. https://www.youtube.com/watch?v=x09f_BuJsoY

[11] aBOS. Full System Description v3, 18 October 2026, especially sections 1, 3, 4, 7, 8 and 11. Supplied HTML attachment.

[12] aBOS. System Explorer, 18 October 2026, review queue, autonomy, evaluation, operator and internal status components. Supplied HTML attachment.

[13] Anthropic. Early progress on accelerating drug development with Claude, 27 October 2026. https://www.anthropic.com/research/early-progress-drug-development

[14] OpenExO. ExO Model 3.0 and Your Readiness Score. https://openexo.com/exo-model-3 ; https://openexo.com/exo-ready

Mithril example

Anthropic confirms acquiring Mithril Medicine and describes one founder coordinating 11 outside laboratories [13]. This illustrates why internal headcount is an incomplete labor boundary. The public case does not supply the gross profit, complete labor hours, AI cost and counterfactual needed to validate this benchmark numerically. It should not be described as proof of the metric.