AGI ALPHA FLYWHEEL

Profitable Work First. Compounding Capability When Worthwhile.

Transformational Intelligence ∞
Learning to Change the Game

FIND YOUR ALPHA. BUILD YOUR SUCCESSOR Ω.

ACTIVATE. EVOLVE. SCALE.

Independent research edition prepared for MONTREAL.AI.
Evidence cutoff: 26 October 2026. AI-assisted implementation and manuscript.
Author attribution and external submission remain subject to the project's publication owners.

Abstract

Can expertly used frontier AI deliver useful accepted work at a complete cost below a selling price referenced to equivalent human services? We study a coordination-mechanism foundry using an executable finite-cash model, documented API prices, human-team service-rate anchors and explicit workload and market assumptions. Across 64 fresh worlds per scenario, competent Beta-only delivery produces mean collected operating surplus of USD0.41, 3.70 and 6.76 million in conservative, central and favorable conditions over 104 weeks plus run-off. Initial seed capital is USD60,000 and is excluded from profit. The tested policies charge an average 29.8%, a fixed 55% and a fixed 75% of the respective human reference. No separate R&D is selected. Profitable Beta operation is therefore a positive conditional commercial result, independent of any skill improvement. Optional retention, sharing and investment are evaluated separately with three nonoverlapping cost books. Their primary enterprise increment is unestablished; the favorable network-minus-Beta mean is USD$-8,766$ with an interval spanning loss and gain. A tested Alpha-gated graph, a mixed exact mathematical witness, 6,048 simulated histories and complete ledgers make these distinctions reproducible. No actual customer sale, substantive frontier task trial, live ENS enforcement or production certification is demonstrated. Quality, scope, demand and workload assumptions bound the commercial inference; capability compounding and recursive improvement require separate evidence.

Evidence boundary. Actual local software and finite mathematical checks; modeled agents, contracts, payments, retention decisions and market outcomes. Graph fixtures test enforcement mechanics. They do not certify operational Alpha. All commercial conclusions are conditional on the supplied model.

1 Profitable useful work as the commercial core

A customer buys a useful outcome, not a graph node. Completing a difficult job may require a better representation, a new experiment, an executable protocol, or a composition of methods. If that work leaves a reusable capability, the next question is whether retaining it is worthwhile after the additional work needed to make it safe, applicable and useful to a recipient. This question is different from asking whether an independently funded research project repays its entire construction cost. Both are legitimate decisions; they need distinct ledgers.

The commercial question comes first in this edition: *Can competent frontier AI deliver accepted useful outcomes at a complete cost substantially below a selling price referenced to equivalent human services, leaving profit and available operating surplus?* Continuing profitable Beta delivery, retaining a cash reserve or declining a new skill is commercial success under that model. Retention, sharing, new capability construction and expansion are optional ways to improve that foundation. Their value must be assessed separately; they are not conditions for calling a profitable service profitable.

The frozen experiment also asks when customer-funded transformational work leaves net Alpha and when network availability adds value beyond isolated retention. On 27 October 2026 the user requested a commercial-first reporting order. We preserve the earlier protocol, its original hypothesis order, all samples and all outcomes. This is a disclosed reporting amendment, not a retroactive change to the preregistered experiment or a new policy search. Derived commercial ratios and conservative receipt-only checks are labeled postprocessing.

We interpret Transformational Intelligence as changing the problem-solving setup while preserving the authorized purpose. Boucher's supplied manuscript distinguishes purpose, formulation and evaluation, and requires an executable construction plus a falsifier [1]. The supplied Gym, Specialist ASI, Navigator and Successor manuscripts reinforce complete mission architecture, protected evaluation, retained failure and the separation of historical evidence from present authority [2–4]. Their illustrative workloads, prices, performance scores and institutional arrangements are not adopted as calibration.

The study begins from a bounded portfolio screen. It does not treat all useful inventions as commercially demanded, or all difficult jobs as sources of reusable knowledge. Retention may be rejected, deferred, superseded or retired. Separate research and expansion are optional. Waiting is a legitimate outcome.

1.1 Contributions and evidence classes

The package contains an executable finite-cash economy, a graph lifecycle with an explicit Alpha gate, a mathematically checkable coordination artifact, a source-linked calibration register, complete primary-run ledgers, paired counterfactuals and an English/French offline demonstration. Table 1 states what each component can establish. The source namespace is additive to the existing repositories. Existing owners, implementations, publications, contracts, roles and settlement routes remain intact.

Table 1: Evidence classes must not be silently upgraded.

Class	Present evidence	Practical limit
Actual local implementation	Executed simulator, exact solver, graph API and adversarial tests	Software behavior within the tested interface
Synthetic mathematical experiment	Exact finite profiles, deviations and dynamic-program certificates	Declared discrete model and mechanism family
External reported evidence	Official billing documents, service schedules, primary research	Different tasks or models do not calibrate acceptance directly
Explicit modeling assumption	Task workloads, quality, demand, trust, retention and obsolescence	Conditional scenario consequences, not observed market estimates
Simulated enforcement	Fixture membership, validators, admission and permissions	No live ENS, deployed contract or production confidentiality
Realized commercial activity	None in this edition	Zero actual study sales and collections

2 A demanding paid-job portfolio

The primary venture is a **coordination-mechanism foundry**. A consortium, procurement designer or resource-sharing organization commissions an executable allocation protocol under a written budget, incentive and delivery contract. The supported outcome is an offline R&D package: typed inputs, executable candidate mechanisms, counterexamples, finite-model checks, integration tests, documented assumptions and a deployable-design handoff. It is not an operational market, a legal opinion or a guarantee about unconstrained strategic behavior.

The recurring trigger is a change in participants, resource constraints, admissible bids, uncertainty or coordination objectives. A beneficiary may pay because a feasible, checkable design reduces trial-and-error and makes a decision possible within its development budget. That willingness to pay is modeled, not observed. A small demonstrator is not assigned the price of a complete professional engagement.

Three increasing modeled scopes share this venture. The first specifies and implements one bounded protocol family. The second adds a reusable evaluator and integration package over changed conditions. The third composes protocols, dependencies and adversarial tests into a broader design. Alternative ventures screened in the companion are distributed recovery protocols, compositional interoperability and privacy-constrained collaboration. They have executable technical precedents [5–7]; their greater deployment, environment or security obligations make them contrasting cases, not additional successes claimed here.

Mechanism design is supported by task-relevant prior research. AMD-LLM combines mechanism-code search with problem-specific repair and reports recovery of known theory as well as richer designs [8]. Program synthesis supplies another executable precedent [9]. These studies motivate the venture and its tests. They do not supply Astra success probabilities, customer demand or the study's token workloads.

2.1 From Identify to Out-Evolve

Table 2: The journey consists of falsifiable work, not stage labels.

Stage	Executable contribution	Decisive failure condition
Identify	Bind beneficiary, accepted outcome, budget, assumptions and permissions.	The mechanism solves a different objective or exceeds the authority envelope.
Out-Learn	Acquire distributional or constraint information; design counterexamples and informative probes.	New information fails to distinguish the decision-relevant cases.
Out-Think	Express complementary acquisition as a typed all-or-none allocation problem.	Independent decisions still violate the aggregate constraint.
Out-Design	Compile a budget-indexed frontier and an executable protocol.	The independent small-domain oracle finds a better feasible design in the claimed family.
Out-Strategise	Select delivery, retention, price, reserves or optional new work from remaining value.	Complete prospective benefits do not cover incremental costs and displaced capacity.
Out-Execute	Deliver, check, correct, package and record outcomes under finite resources.	A failed check, stale dependency or unaffordable obligation prevents accepted completion.
Out-Evolve	Apply a retained method to fresh work; requalify after Beta or scope changes.	A competent existing method ties or wins, or current evidence no longer supports advantage.

The model is replaceable. The retained object is mission intelligence: executable methods, representations, dependencies, evidence, cost history, failure conditions and scope. Replacing a model without reconsidering those bindings cannot preserve an unsupported claim. The earlier Successor harness explicitly retained a conventional caching tie and an expert advantage; that lesson informs our strong comparators, not our outcome assumptions [4].

3 Frontier configuration and calibrated uncertainty

The modeled demanding roles use `gpt-6-astra`, Pro mode and maximum reasoning effort. They include production, validation, correction, capability design and capital decisions. Qualified visual and computer-use support uses `gpt-6.1-terra` in standard mode with medium or high effort. Official pages were checked on the evidence date [10–14]. Only aliases were exposed in the inspected snapshot sections; no immutable dated model identifier is invented. This research session used AI assistance to build the implementation, but did not execute instrumented frontier trials over the substantive customer jobs.

Table 3: Documented token prices in USD per million; assumed quantities are separate.

Model	Standard input	Cached read	30-minute write	Output
GPT-6 Astra	10.00	1.00	12.50	50.00
GPT-6.1 Terra	1.00	0.05	1.25	5.00

Pro bills aggregate underlying model work at the selected token rates; reasoning tokens are billed output. There is no fixed "expert use" price or capability multiplier [13, 15]. The primary cost model conservatively charges all fresh input as 30-minute writes and counts cache reads separately. A no-cache

sensitivity bills every input at standard uncached rates. Workloads aggregate multiple adequately contextualized calls; per-request planning stays below 128,000 input and 32,000 output tokens. The documented long-context surcharge above 272,000 input tokens therefore does not apply to the planned request envelope. Both model pages list 1,050,000 context and 128,000 output limits [10, 11]. Actual account entitlements are unverified.

The register separates observed documentation, external reported findings, derived references and explicit assumptions. Tool budgets include search, executable checks, computer use and job-specific compute. Fixed infrastructure, storage and access are additional. API billing is separate from ChatGPT/Work subscriptions [16]. No subscription or token issuance subsidizes the modeled service.

External evidence supports uncertainty rather than a universal transfer gain. SkillsBench reports benefits from curated skills, while SWE-Skills-Bench reports many redundant skills and sometimes substantial token overhead [17, 18]. SkillLens and skill-conflict research motivate heterogeneous and harmful transfer [19, 20]. Correlated judge errors motivate a shared-fault component rather than independent-vote certainty [21]. No benchmark score is converted into a customer-acceptance probability. Our quality, demand and workload values remain scenario assumptions with broad stress tests.

4 Automated enterprise and finite-cash simulation

The primary horizon is 104 weeks plus 16 weeks of run-off, with six producer agents, two validator slots and USD60,000 seed. An assumed USD9,500 bootstrap cost is expensed initially. All workflow costs are automated in the primary configuration. No routine human provider review or manual onboarding is hidden in the simulation. Governance over real deployment and authority remains outside routine job execution.

Agents begin with heterogeneous ordinary experience and specialization. Every competent arm can reconstruct a missing method, retain ordinary experience and benefit from outside diffusion. A successful observable delivery can yield a rights-cleared candidate with a declared probability; it can also leave no reusable learning. A promising but unbuilt direction may enter optional C. A later job-learned artifact can supersede that direction without being trapped behind its old research requirement. Full source history remains.

Candidate extraction, generalization and validation consume B resources and can fail. Admission compares matched recipient-by-depth cost probes with the cheapest competent reconstruction or currently eligible retained equivalent. The proposed bundle is optional. These deterministic response surfaces are model assumptions, not frontier reasoning over actual jobs. Fresh simulated jobs exercise the admitted asset's modeled consequences; they do not empirically validate the assumed response surface.

Production has specification errors, implementation errors, component failures and shared weekly faults. Validation combines assumed executable-error detection with semantic detection; common faults weaken the latter. False passes, false rejections, one correction attempt, correction failure, abstention, lateness, defaults and later refund/recovery are charged. Hidden correctness is used by the simulated world and the analyst, not as an immediate public trust signal. Trust responds to observed delivery and subsequent refunds.

Transfer includes integration, failed adaptation, fallback reconstruction, harmful adaptation and current-status checks. Failed transfer does not retain the skill's quality or validation shortcut. Redundant or overly expensive capabilities can be declined. Each work tick checks the applicable skill and bundle dependencies, so stale or revoked history cannot authorize continued selection. Revalidation has cost, capacity demand and its own errors.

4.1 Cash, profit and capacity are different quantities

Prices are offers, not receipts. Orders reserve worst-case permitted delivery and correction costs before work starts. Collections arrive after the modeled delay and can default. Warranty funds and commitments reduce available cash. No future receipt funds current retention or research. The ledger uses integer cents and never permits negative cash; unpaid obligations remain visible and reduce spendable funds.

Let R be collected revenue, A, B, C, F accrued cost books, K closing unspent commitments and Q future refund obligations not yet expensed. The reported *collected operating surplus* is

$$\Pi = R - (A + B + C + F) - K - Q. \tag{1}$$

Unpaid expenses already in the books are not deducted again. Receivables are reported but excluded from this conservative terminal measure. Seed is not earnings. These are constant-price, pretax operating scenarios, not audited financial statements. The actual historical research-session bill is unknown; it is not reported as zero or recovered through simulated collections.

Five subweekly work ticks share finite token budgets, producer availability and validator capacity. Retention and optional construction occupy those same resources. Maintenance and evaluation consume quota and staff-slot time. Adding agents alone does not add validators or quota. The scale menu charges additional worker setup and recurring support. Minute-level API scheduling is not emulated; a paced production broker and verified account quota would still be needed.

5 Human reference, buyer game and economic policies

The commercial comparator is a human-delivered service with the same supported outcome and responsibilities. Table 4 combines observed US contract service rates with explicitly assumed role-hours. RAND's schedule gives research rates; Z Federal's 2026–27 contractor-facility schedule gives engineering and project-management rates [22, 23]. Neither document is a matched project quotation or evidence that workers use no AI.

Table 4: Derived human-service references, USD, US domestic.

Role	Hourly rate	Scope 1 hours	Scope 2 hours	Scope 3 hours
Economics policy researcher	334.18	24	48	96
Senior software engineer	213.72	40	80	160
Quantitative research analyst	264.95	24	48	96
Project manager	200.09	12	20	32
Total team-hours		100	196	384
Derived service reference		25,329.00	49,857.64	98,114.56

These scopes include bespoke specification, development, experiments, evidence, integration and handoff. Team-hours do not establish elapsed duration. Familiar instances could take much less effort; sensitivities reduce the reference to 10%, 25% and 50%. No price is assigned to an unexecuted speculative idea or to the tiny witness as if it were a complete engagement.

Buyers choose the offering, matched human service, postponement or no purchase. The random-utility model includes value, expected quality, trust, integration, price and urgency-weighted delay. For our offer, one normalized form is

$$U_{\text{ours}} = vq\tau - p/H - \iota - \eta L + \epsilon, \tag{2}$$

where H is the human reference, τ trust, ι integration burden and L delay. The other options have their own utility terms. The seller sees declared task information and history, using population priors rather than the buyer's private realized value. Priors are supplied to the model; actual demand is not estimated in this edition.

Prices are decision variables. The adaptive rule chooses a one-period profit-maximizing grid response to the modeled buyer distribution and current capacity. Thirty-two policy candidates combine fixed or adaptive pricing with delivery-only, optional retention at two hurdles, or retention plus separate research. Selection maximizes training mean surplus subject to the declared sample cash-shortfall constraint. It is optimal only within this finite family, not a universal market equilibrium or globally optimal dynamic strategy.

Retention and research decisions use remaining value, construction/validation uncertainty, reserves, available cash and estimated displaced delivery margin. The enterprise can continue delivery, wait, retain a bundle or abandon a weak candidate. No mandatory investment count or reinvestment percentage exists. A separate capacity menu uses one set of simulated worlds to choose and another to test an initial capacity configuration. This bounded expansion comparison does not claim optimal hiring or expansion timing over every possible future.

6 Protocol, counterfactuals and confirmation

The initial separate-build protocol was frozen and partially trained before the user clarified that paid-job learning must be primary. Its 18-file source freeze and 909 completed training records remain archived. No economic confirmation was run under that protocol. Version 2 is a disclosed design amendment with new training, confirmation and sensitivity seed ranges. It does not relabel earlier results as fresh confirmation.

The final protocol and code are hashed before policy selection and confirmation. Training suffered a storage interruption after 1,576 complete ordered records. The exact prefix and incomplete tail were preserved; an infrastructure-only wrapper resumed the remaining tasks through the unchanged frozen functions. The recovery receipt binds the source, prefix, wrapper and final training hashes. No seed, model, selection rule or confirmation result was changed or reused. The three scenarios are conservative, central and favorable joint assumptions. Twenty-four training worlds per scenario choose among the 32 policies; 64 fresh worlds per scenario and arm then evaluate the selected policy. Primary arms are competent Beta, isolated retention, network sharing and network sharing without separate R&D. All begin with the same seed capital, agents, quotas, experience and coupled exogenous market draws. Differences in real modeled sharing or retention overhead remain charged.

The unit of uncertainty is the independent world. Jobs descended from an artifact are not independent invention campaigns. Mean and paired-difference intervals describe Monte Carlo error conditional on the model; Wilson intervals describe simulated frequencies. They do not quantify uncertainty about real acceptance, service effort or market demand. Fifty-seven predeclared parameter cases test retention cost and success, demand, prices, copying, ordinary learning, moving Beta, transfer, obsolescence, capacity and cash.

Adaptive enterprise comparisons include downstream changes in pricing, queues, investment and learning. A separate frozen-recipient experiment removes only the candidate, retaining competent reconstruction and existing equivalents under the same resource envelope. It charges failed and harmful transfer, adaptation, recovery, checking and lifecycle upkeep, with separate historical job memos. This deterministic panel was smoke-tested in development; its replay is an arithmetic reproducibility check, not an unseen stochastic trial.

7 Commercial core: confirmed modeled operation

7.1 Useful accepted work, prices and complete costs

The answer is positive within the declared model: competent Beta delivers useful accepted coordination-design work, collects customer payments, covers complete delivery and business costs, and leaves operating surplus. It needs no graph advantage or separate R&D to meet that commercial criterion. The outcome remains conditional on task performance, human-scope equivalence, buyer demand and workload assumptions; it is not realized revenue.

Customers receive an executable bounded coordination mechanism, task-specific checks and counterexamples, declared assumptions, integration artifacts and a handoff. They do not buy an untested idea or an operational market guarantee. The three human-service references are USD25,329, USD49,857.64 and USD98,114.56 for assumed teams of 100, 196 and 384 hours, using the documented US rates in Table 4. They are derived matched-scope scenarios, not actual customer quotations or verified human-only labor.

Table 5: Commercial core: Beta-only means over 64 fresh worlds per scenario. USD unless noted; 104 intake weeks plus 16 run-off weeks.

Metric	Conservative	Central	Favorable
Useful accepted outcomes	48.39	163.00	208.36
Price / human reference	29.8% adaptive	55% fixed	75% fixed
Mean completed-job quote	13,960	25,844	35,075
Collected receipts	677,921	4,192,972	7,377,908
A: full delivery expenses	209,756	432,391	556,141
Common setup and business costs	56,744	57,715	58,288
Closing obligations	0	471	0
Collected operating profit	411,421	3,702,395	6,763,479
Available closing cash	470,952	3,761,488	6,821,122
Buffered surplus above seed	408,312	3,698,570	6,758,482

B and C expenses are zero in these Beta arms. A includes learning during delivery, failed work, tools, verification, correction, packaging, refunds and recovery. Common costs include the USD9,500 bootstrap and shared business infrastructure. Initial USD60,000 capital is not revenue or profit. The central closing-obligation mean is deducted conservatively; already-accrued unpaid expense is not deducted twice.

The policy search used buyer choices among our offer, a human service, postponement and no purchase. These prices were selected in the declared 32-policy network family before confirmation and then applied unchanged to Beta. They are evaluated, economically selected policies, not an independent proof of the globally best Beta price. Conservative realized offer ratios range from 25% to 45%; central and favorable fixed ratios are 55% and 75%. A lower price alone does not establish real demand.

7.2 Surplus, usable cash and correct-outcome receipts

Available cash removes the ledger's protected commitments, warranty reserves and unpaid expenses. It is different from profit and includes seed. The final table row is a reporting-only planning measure: $\max(0, \min(\Pi, \text{available cash} - 60{,}000) - 2{,}640)$. It preserves initial seed, caps the figure by reported profit, and leaves a six-week Beta operating buffer. The USD2,640 buffer is not another expense or a claim that all future liabilities are known.

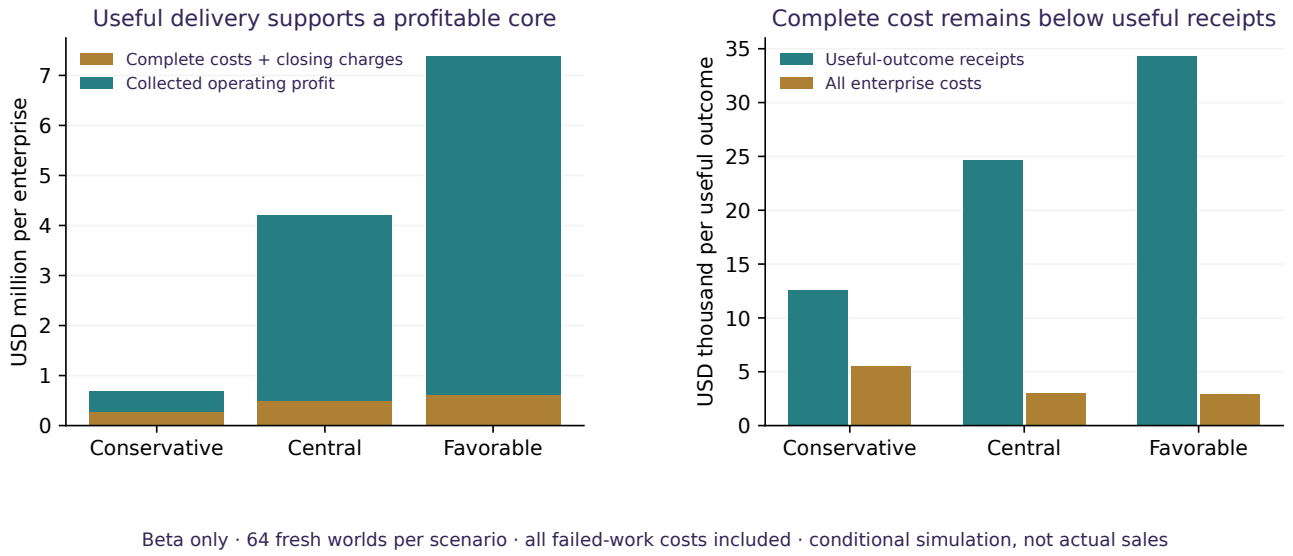


Figure 1: The commercial core before optional skills. All quantities are modeled; every cost of failed work remains in the accounting.

To avoid crediting the commercial conclusion to erroneous deliveries, a retrospective accounting check keeps only receipts linked to on-time, truly correct outcomes and still deducts every enterprise expense and unexpensed closing obligation, including refunds and recovery of incorrect work. This is a deliberately conservative cash-revenue sensitivity, not a newly tuned policy or another sample.

Table 6: Commercial checks using only useful-outcome receipts. Pooled per-outcome ratios use the same useful-outcome denominator.

Metric	Conservative	Central	Favorable
Useful receipts per useful outcome	12,604	24,622	34,279
All costs per useful outcome	5,507	3,010	2,949
Conservative surplus per useful outcome	7,097	21,612	31,330
Useful-receipt-only total surplus	343,435	3,522,731	6,527,982
Worst world in this check	-30,129	1,517,350	2,826,354

Mean surplus remains positive in every scenario under that stricter check: USD343,435, USD3,522,731 and USD6,527,982. The conservative minimum is USD$-30,128.78$, so the result is not a guarantee that every world remains profitable under every accounting sensitivity. Pooled ratios summarize a heterogeneous portfolio; they are not a matched-job quote or an empirical productivity multiple.

The modeled customer savings on useful outcomes average USD1.67, USD3.43 and USD2.43 million across the three scenarios relative to their declared human references. These are price differences for supported useful outcomes, not measured consumer welfare, willingness to pay or evidence of actual sales. Quality is a simulated latent property. Positive economics cannot certify a real professional service or a real minimum quality floor.

7.3 Throughput, capital and commercial downside

The six-producer, two-validator configuration supports mean useful throughput of 0.47, 1.57, 2.00 outcomes per intake week, respectively. Dividing by the full 120-week horizon gives lower rates; neither

measure is observed production or a steady-state capacity guarantee. The primary scenarios assume 1.8, 4 and 7 initial qualified leads per week, with repeat demand modeled separately.

The primary runs begin with USD60,000 and collect after seven, four and two weeks in the three scenarios. They record no Beta cash shortfall in 64 worlds each; the Wilson 95% upper bound is nevertheless 5.66%. A dedicated central-prior stress fails in all 16 worlds at USD10,000 seed and has zero observed shortfalls at USD25,000. This brackets tested outcomes, not an identified minimum required capital or a guarantee at USD25,000.

Demand sensitivity brackets the mean-profit crossing between 0.05 and 0.25 initial qualified leads per week at the frozen central price and other assumptions. Reducing the human reference to 10% leaves positive mean Beta surplus of about USD80,204, while the combined low-demand, low-reference, high-cost, low-trust and slow-collection case loses about USD16,661. Thus the commercial core has explicit failure conditions even though the three primary scenario means are positive.

A split-sample capacity menu selects doubling producers, validators and quotas together. Its eight fresh test worlds improve Beta mean surplus by USD1,331,416, with conditional Monte Carlo interval [USD685,519, USD1,977,314], after the modeled capacity costs. That resource decision does not depend on skill sharing. It tests an initial capacity choice, not global optimal expansion timing. Spending still requires currently available cash.

These findings support continued modeled Beta delivery and retained reserves as economically meaningful choices. The next sections ask whether a candidate skill, sharing policy or extra investment adds enough value to improve that already-profitable foundation. Optional enhancements are allowed to fail their own tests without discarding the commercial core.

8 Beta, Alpha and three nonoverlapping cost books

Let $B_{a,t}$ denote agent a's competent frontier configuration at time t, including ordinary tools, context, experience and available equivalent methods. Let $K_{a,t}$ be its additional eligible retained capabilities. Beta is not defined by the absence of a skill file. It can reconstruct methods, use ordinary dynamic programming, retain experience and possess a functional equivalent.

A candidate k has scoped prospective Alpha when the same mission and resources produce a positive net increment relative to $B_{a,t} + K_{a,t}$ after all remaining incremental costs. For a planned reuse count R, paired per-mission gross gain g_i, declared nonnegative uncertainty allowance u and remaining cost C, the research gate computes

$$N = R\bar{g} - C, \qquad N_{\text{lower}} = R(\bar{g} - u) - C. \tag{3}$$

Admission requires $N_{\text{lower}} > 0$ and a nonredundant, scoped comparison, followed by technical validation and the separate admission decision. Explicit weights represent recipient and task mixtures. The allowance is a declared modeling margin, not a statistical confidence interval. A numeric gate does not establish that its inputs are true.

Table 7: Each actual modeled expense is charged once.

Record	Included activities	Prospective treatment
A: paid job	Intake, source work, reasoning, learning, experiments, production, in-job integration, correction, validation, packaging and recovery	Historical delivery remains in enterprise profit. It is not recharged to retention.
B: additional retention	Additional extraction, sanitization, fresh validation, packaging, admission evaluation, distribution, revalidation and upkeep	Only remaining incremental amounts enter the retention decision.
C: optional new R&D	Deliberate additional search or construction beyond the paid outcome, including failed attempts	A separate cash-funded decision; not required for every skill.
Common	Setup, shared orchestration, execution support, storage and run-off overhead	Charged once to the enterprise and disclosed separately.

Integration performed as part of a receiving paid job is charged in A. The counterfactual Alpha calculation includes its incremental effect on that job's cost; it does not issue another ledger expense in B. Fixed graph service overhead is charged once to the enterprise's retention book. A marginal admission decision does not require every new skill to repay the already operating service again. The full network comparison still has to recover that overhead.

Historical source-job cost is a non-additive memo in each retention record. Changing that memo from zero to a million dollars leaves the prospective Alpha decision unchanged; the job's actual ledger remains unchanged and fully charged. The model distinguishes a priced, delivered source job from a useful source job whose customer payment was collected. Defaults cannot be turned into funding.

Complementary components can qualify jointly. The bundle bears its complete incremental costs and pins the components required for the benefit. Components that lack standalone advantage remain candidate-only dependencies, rather than being relabeled as independently positive skills. Failed approaches belong in separate failure memory and never participate as positive graph entries.

9 An actual executable transformational witness

Consider buying a complementary bundle of inputs from suppliers with independent private integer costs. The useful purpose is to obtain the whole bundle within a hard payment budget. Choosing individually attractive procurement decisions can violate the shared budget. Truthful incentive labels alone do not repair it.

The executable representation uses posted thresholds t_i. Acquire the complete bundle and pay t_i to each supplier only when every reported cost satisfies its threshold. Otherwise acquire nothing and pay nothing. With public value V, budget B and known independent cumulative distributions F_i, optimize within this declared family:

$$\max_{\mathbf{t}:\,\sum_i t_i \le B} \left(V - \sum_i t_i \right) \prod_i F_i(t_i), \tag{4}$$

including the no-purchase option. These are mathematical utility units, not commercial USD. The finite model excludes collusion, externalities, strategic distribution reports and continuous unbounded type spaces.

The transformation replaces independent threshold selection with a budget-indexed dynamic-program frontier and a typed all-or-none protocol. The construction must preserve budget, allocation and payment semantics, and the certifier must check individual rationality and finite unilateral deviations.

A Bellman certificate supplies checkable upper-bound conditions. An exhaustive small-domain oracle is independently informative about the optimizer; a signed claim about it is not a substitute.

The strongest cold baseline already knows dynamic programming and the relevant theory. No claim that frontier Beta lacks this algorithm is made. The potentially retained asset is a compiled frontier for an applicable input family, together with its protocol and evidence. A recipient with the same artifact gets no additional Alpha from another copy. This known-method prototype tests a change in representation; it is neither a novel mechanism theorem nor autonomous frontier invention.

9.1 Frozen local results and the mixed transfer outcome

The original witness protocol was frozen before its confirmation: 48 fresh small cases match the exact optimizer, all 48 pass the finite property checks, and five larger domains pass their certificates. The finite checker covers 22,992 profiles and 440,256 unilateral deviations. The largest declared threshold-vector space has 129,746,337,890,625 possibilities, while its dynamic program uses 14,040 transitions. That is not exhaustive enumeration of the large space or universal optimality across all mechanisms.

The twelve fresh same-family reuse queries produce identical solutions in the strong cold and retained branches. Construction transitions fall from 13,926 to 1,716. Aggregate certificate inequalities rise from 13,926 to 22,308; witness multiplications rise from 7,022 to 12,103. Selection states remain 360, and 74,496 payload bytes are hashed. The shared-route certificate counts include the creator and all recipients. The result establishes output-preserving transfer relative to competent reconstruction, not newly feasible frontier work. It does not establish a net time, token or dollar saving. Reducing repeated construction can move work into verification. Trusted-attestation reuse is a separate operational hypothesis requiring its own cost and trust evidence.

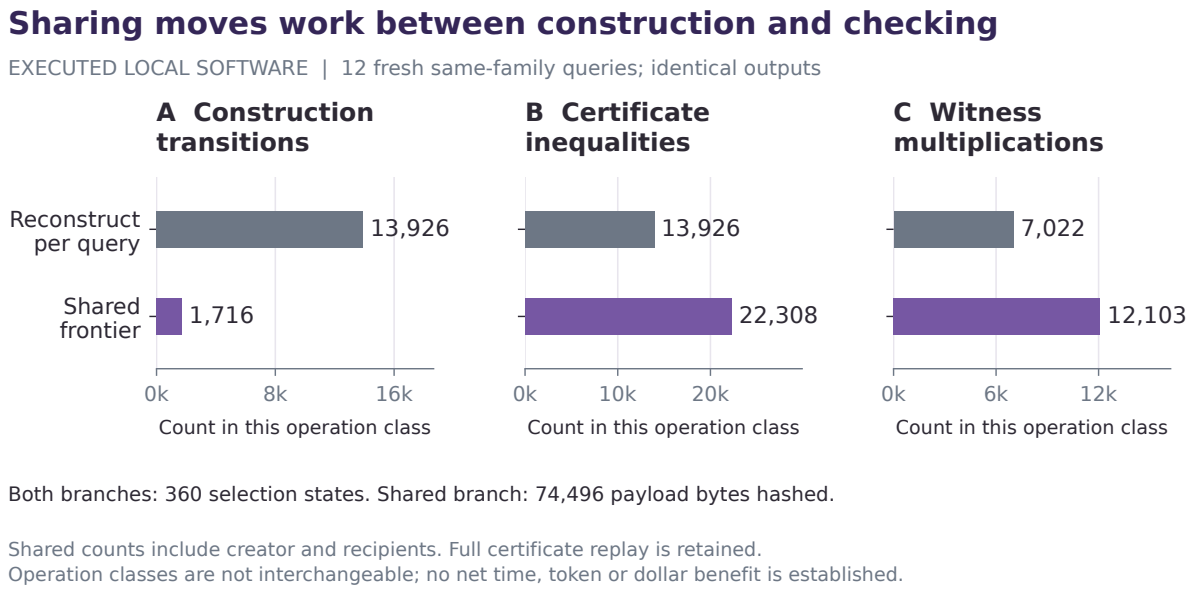


Figure 2: Actual finite software counts. Different operation classes remain separate; they are not summed as interchangeable dollars or frontier tokens.

The retained evidence includes negative budget examples, duplicate recipients and changed-distribution refusals. A post-confirmation JSON tuple/list wrapper comparison defect was repaired with a regression

test; the original frozen computation and result bytes are retained. No numerical algorithm, seed or scientific result was changed to make the witness pass.

10 Graph lifecycle, availability and authority

The research API implements candidate, evidence, Alpha assessment, validation, admission, discovery and optional authorized reuse. Revalidation repeats the relevant current checks. Retired, revoked and superseded versions are terminal. Canonical promotion requires current admission and cannot bypass it. Manifest, evidence, policy, Beta configuration and dependency identities are bound to the records. The full current capability-set comparator is checked before validation and admission. Reuse checks the bound assessment, current Beta, policy, evidence, scope and dependency state. A later unrelated admission does not automatically recompute every earlier comparison: detecting new overlap and requesting an updated assessment remains a controller responsibility.

Every eligible shareable capability is discoverable and retrievable by all authenticated direct `*.agent.agi.eth` fixture members. A member may inspect and export the capsule, including the inputs needed for a composition. Creating a graph record and executing an action require their separate permissions. There is no preferred recipient subgroup. Candidate-only bundle components are retrievable through the admitted bundle, not falsely advertised as independent Alpha. Receiver checks include scope, dependencies, freshness and revocation.

Identity, record writing, eligible validation, admission and execution are separate authorities. Nested names and ENS-looking strings do not acquire direct-label authority. The fixture requires eligible distinct validators and prevents a producer or attestor from admitting its own record even when that principal holds multiple roles. Existing AGI Jobs `*.club.agi.eth` roles, governance, contracts and settlement are preserved.

Local AES-GCM, Ed25519 and SHA-256 operations execute, but their keys are publicly derivable research fixtures. Consequently the package provides no production confidentiality or identity enforcement. It has no live ENS, on-chain registry or zero-knowledge deployment. Inclusion and signatures bind bytes; they cannot prove useful capability, truthful measurements or economic Alpha. Authentication cannot prevent an authorized exporter from copying material.

The implementation is an in-memory Python object with a trusted fixture directory and harness clock. It checks public-API operations, not arbitrary access to the Python process. A reuse permit is rechecked against current state under a lock and consumed once; its receipt explicitly records `code_executed=false`. No capsule program or external action executes through that API. The documented compact Python-JSON serializer and SHA-256 tree are not an Ethereum, EIP-712 or production protocol compatibility claim.

Separate failure memory preserves sanitized unsuccessful approaches, their context and evidence. It has a distinct namespace and commitment root, with no admission, canonical-selection or reuse-permit path. Declarations reject protected evaluation cases and scorer internals from shared capsules and memory; the research API does not claim automatic semantic sanitization. Confidential customer information and restricted third-party material are outside this study.

New model, tool, routing, verifier, data or environment configurations require versioned Beta and affected requalification. An external controller must detect and submit that change; the graph does not observe a provider runtime or infer changed competence automatically. Old assessments remain inspectable without restoring influence. Admission, acceptance, operating authority and capital allocation remain distinct decisions. The supplied GoalOS drafts are design references, not evidence that their dashboards or protocols were already implemented [24, 25].

The evidence-only adapter `integration/modeled_admission.py` connects the deterministic recipient assessment to this graph. It preserves the numeric Alpha decision while binding its declared Beta path, recipient equivalents, exact scope, evidence and component identities. The predetermined retention and optional new-R&D cases run in separate fixture graphs; positive inputs still require validation and distinct admission, while rejection remains in failure memory. All-member retrieval and separate execution permissions are exercised. The capsules contain public modeled-work descriptors, and their evidence checks schema and accounting consistency, not task performance. The Beta binding is a declared lifecycle projection with assumed future reassessment, not a presently qualified production configuration. The Monte Carlo simulator uses its own reduced skill-state records rather than calling this cryptographic graph API for every simulated job. These linked components do not constitute an operating invention service or establish empirical Alpha.

11 Optional retention, sharing and investment

11.1 Incremental value beyond the profitable commercial core

The positive Beta commercial result does not depend on the following enhancements. The same 768 confirmation histories separately test whether the shared graph increases enterprise surplus. No primary scenario establishes that additional benefit. Declining retention or continuing with existing frontier capabilities preserves the profitable operating option; it is not a failed enterprise. All values remain conditional model results, not realized margins or empirical demand estimates.

Table 8: Same-policy enterprise comparison, USD. The interval is for the paired network-minus-Beta increment.

Scenario	Beta	Network	Increment	95% MC interval
Conservative	411,421	397,474	-13,947	[-14,104, -13,790]
Central	3,702,395	3,690,332	-12,063	[-18,516, -5,610]
Favorable	6,763,479	6,754,714	-8,766	[-118,842, 101,310]

The conservative policy uses the adaptive price rule with zero offset and chooses delivery only; its mean offer is 29.8% of the scope-matched human reference. Central selection chooses a fixed 55% price and delivery only. Favorable selection chooses a fixed 75% price, retention hurdle 1.5, and no separate R&D. These discounts emerge from the finite search. They were not imposed. Fixed-price selection does not establish that a fixed price is universally best.

No primary history makes a separate C investment. Network and network-without-C results are therefore identical. Conservative and central policies do not retain graph capabilities; operating the available network still incurs overhead and can shift cash timing. Their small outcome differences cannot be called Alpha. A manager could avoid that unused service by operating the Beta-only configuration; the experiment deliberately reports its cost.

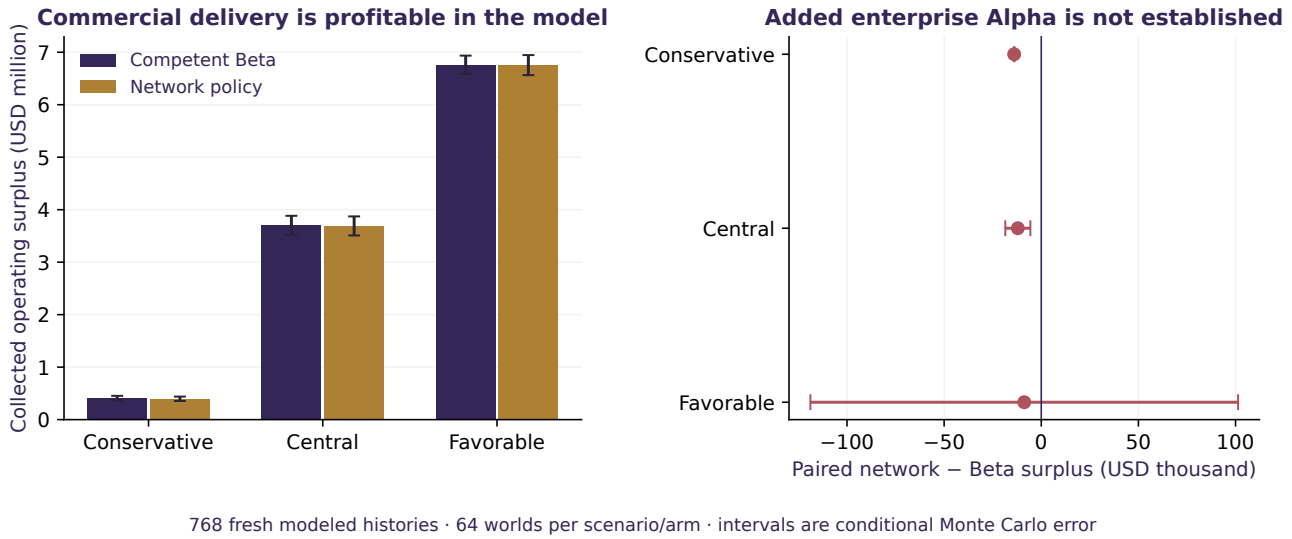


Figure 3: Primary modeled commercial results and their internal increment. Large absolute surplus does not establish a shared-capability advantage.

In the favorable condition, the network averages 7.84 retention campaigns, 7.38 admissions and 44.84 useful recipient jobs per world. Those are adoption and activity counts, not added outcomes. Its accepted-useful outcome difference from Beta is −0.875 jobs (95% conditional MC interval $[-4.034, 2.284]$). The surplus difference is unresolved and has a negative point estimate. The network-minus-isolated comparison is also unresolved: USD$-51,118$ with interval $[-180,073, 77,838]$. Selection on training data does not guarantee a confirmation gain.

11.2 Complete costs, cash and the source of retained learning

Table 9: Mean network cost books and collections, USD. Common costs are charged once; B includes graph overhead even without retained skills.

Scenario	Collections	A jobs	B retention	C new R&D	Common
Conservative	677,726	209,639	13,870	0	56,744
Central	4,194,516	431,819	14,197	0	57,696
Favorable	7,368,084	523,584	31,711	0	58,076

Closing obligations average USD471 in central network runs and zero in the other two; they are deducted in surplus. Seed capital is additional cash, not revenue. All 64 network worlds in each scenario have positive surplus and no recorded cash shortfall. Zero shortfalls in 64 worlds still has a 95% Wilson upper bound of 5.66%; it is not proof of a sub-5% real-world risk.

The independent retained-ledger audit reconciles every one of the 768 confirmation worlds, including expense-book partition, commitments, warranties, collection delays, source-job memos and finite cash. It found no invariant failure. These repeated accounting checks verify the implementation, not the truth of the workload or demand assumptions.

Across favorable network worlds, 502 B campaigns occur. Of these, 396 are labeled funded from available operating surplus and 106 from seed or mixed funds. Source-job collection often comes after retention:

381 recorded source links collect later, 132 before or at retention, and 9 remain uncollected at closing. A customer order therefore does not imply that its receipt financed the next extraction.

An illustrative trace uses the prespecified first favorable network world, seed 770001. Source job 9 incurs USD974.32 of direct A cost, including a USD536.53 learning memo that is already inside that cost. Retention reserves USD183.94 at week 5 and spends it at week 6: USD58.94 extraction and USD125 creator bounty. A separate USD51.30 portfolio appraisal occurs in week 5; ongoing maintenance and network service also remain in B. The reserved campaign subtotal is not its full lifecycle cost. Its USD18,996.75 customer collection arrives at week 7. The first useful receiving job in this lineage completes at week 13. This early retention is seed/mixed-funded. No source A amount is charged again. The sequence illustrates the modeled lifecycle; it does not prove that this receiver could not have succeeded with competent reconstruction, and this entire network world earns USD1,062,359 less than its paired Beta world.

Validation remains imperfect. Conservative, central and favorable network worlds average 5.23, 7.00 and 5.70 false passes respectively. Remediation costs are charged, but these errors and the finite liability assumptions prevent treating modeled profitability as operational service qualification.

11.3 A conditional recipient-level Alpha result

The separate deterministic recipient panel holds the task and resources fixed. Its five competent recipient profiles yield expected initial delivery savings of USD750.08, USD148.43, zero, USD505.78 and zero. The specialist and already-equivalent recipient decline acquisition; missing a skill file does not imply missing capability. Every declared delivery route remains feasible in this panel, so it demonstrates no newly feasible job.

At the declared 80 uses, the paid-job-retention panel projects USD34,740.79 less complete delivery cost and USD9,495.70 additional lifecycle cost, leaving USD25,245.09 modeled net Alpha. Remaining upfront B is USD849.84; upkeep, revalidation and reassessment dominate the rest. Historical A is unspecified in this panel, retained as a memo, and never charged again. Its uncertainty allowance is zero; this arithmetic is not a measured or statistically confirmed return.

The illustrative optional new-R&D campaign instead projects USD$-5,295.69$. It assumes one failed construction attempt (USD11,617.10 within USD23,234.20 total C construction) and campaign success 0.648, versus 0.90 for retention. This is not a cost-only matched contrast or evidence that new R&D generally fails. The strict reuse thresholds at the declared lifetime are 22 uses for retention and 97 for new R&D. These thresholds keep the declared age/task mixture; they are not interchangeable with the enterprise's component-level screening. A first positive projected lifetime of two weeks in the retention panel is a prospective utility threshold, not an observed two-week cash repayment.

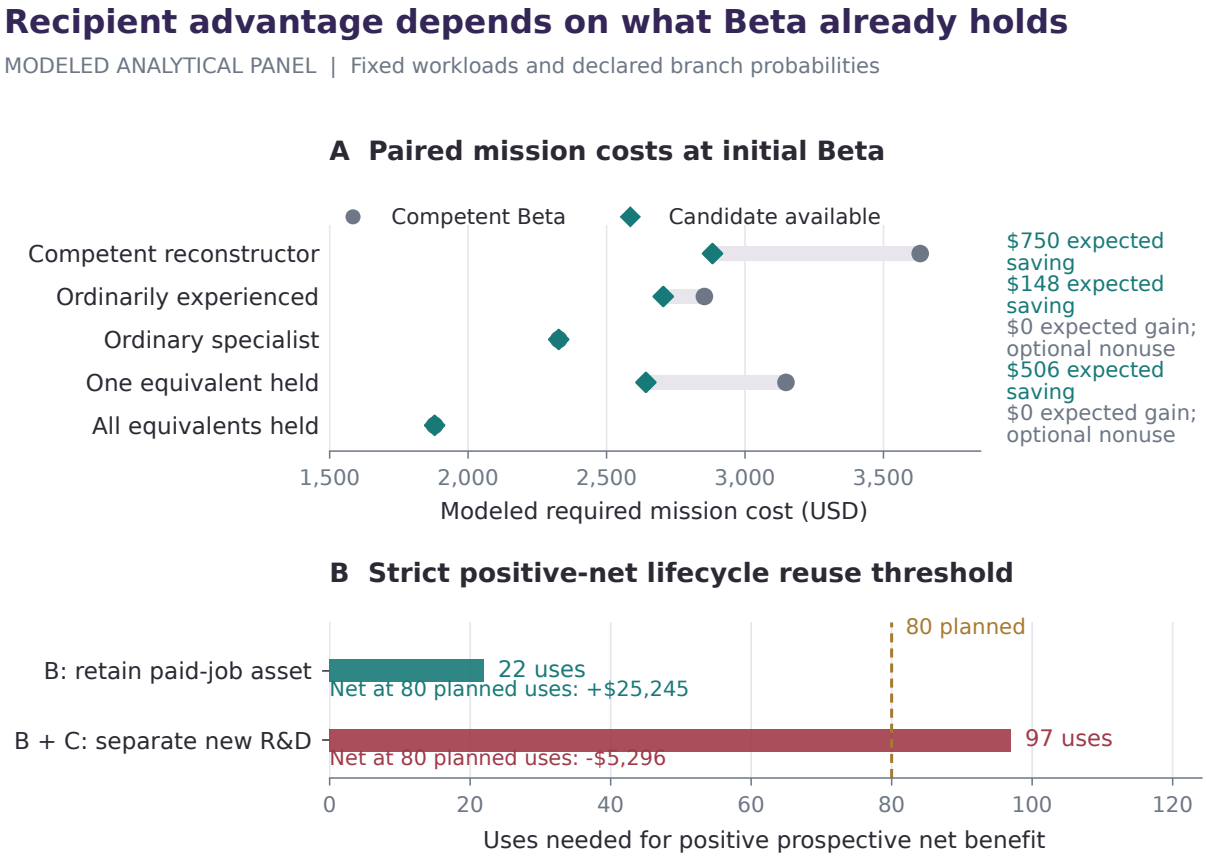


Figure 4: Separate deterministic modeled recipient probes. Existing equivalents and optional nonuse are preserved. These values do not repair the unresolved enterprise comparison.

The actual graph adapter admits the positive modeled bundle and rejects optional new R&D into separate failure memory, using distinct fixture producer, validators and admitter. Six authenticated direct-label fixture members retrieve it; lacking an action permission blocks an execution permit while preserving access. A declared Beta change invalidates its old authority. This establishes a source-bound software lifecycle over modeled evidence, not operational Alpha.

11.4 Sensitivity, demand and capacity

All 2,736 predeclared sensitivity histories and 240 ablation histories completed. Sensitivities explicitly enable optional B retention under the frozen central pricing rule; they are not the central delivery-only policy relabeled as a success. Separate-C-enabled responses are retained alongside no-C responses. Their

finite grids describe conditional mechanisms and parameter sensitivity, not market confidence or a universal optimum. Intervals are pointwise normal Monte Carlo intervals without multiplicity correction.

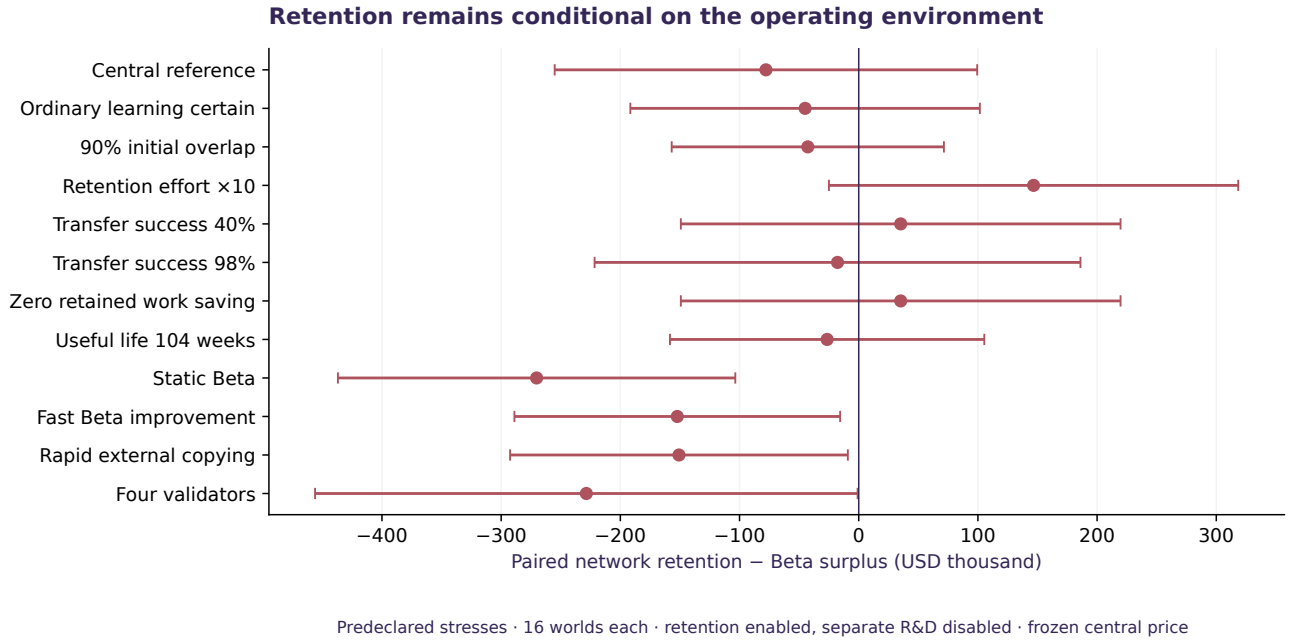


Figure 5: Selected predeclared stresses. Bars are paired conditional Monte Carlo intervals across 16 independent modeled worlds; all 57 cases remain in the companion.

Table 10: Selected conditional stresses, USD mean surplus and paired increment.

Stress	Network	Network − Beta	95% MC interval
Central, B enabled	3,322,575	-77,839	[-255,085, 99,406]
Human reference at 10%	55,483	-24,721	[-38,251, -11,191]
B effort 10×	3,547,114	146,700	[-25,016, 318,415]
90% overlap	3,655,114	-42,716	[-156,919, 71,488]
Faster Beta	3,496,427	-152,226	[-288,873, -15,579]
104-week useful life	3,373,915	-26,499	[-158,373, 105,375]
Joint commercial downside	-18,757	-2,096	[-2,096, -2,096]

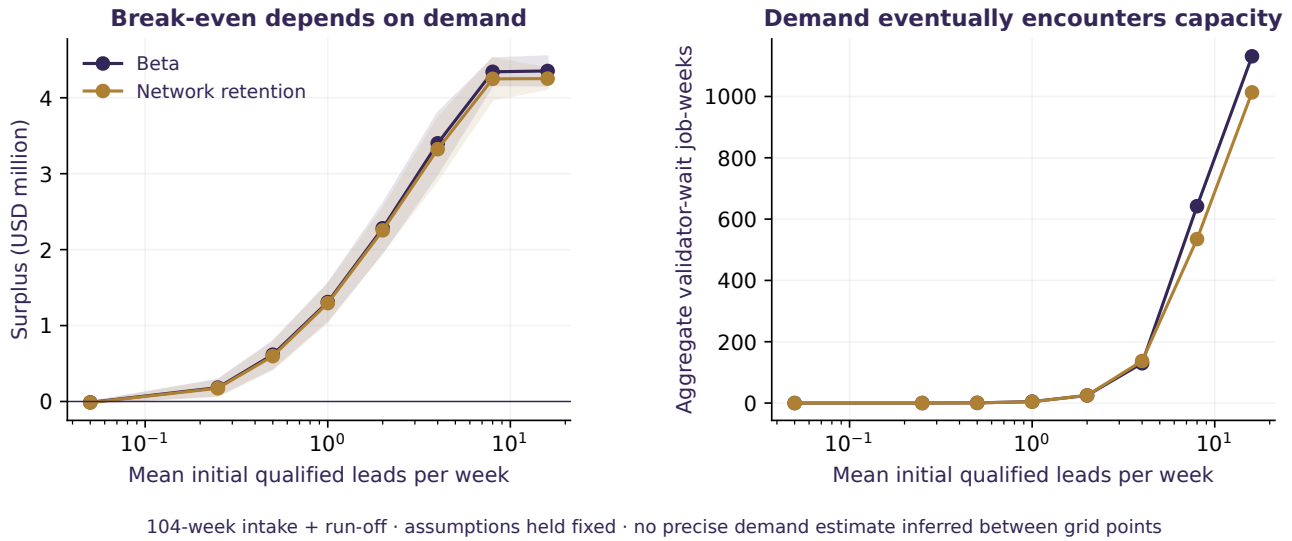


Figure 6: Demand and queueing under unchanged prices and other scenario assumptions. Aggregate waiting sums time across jobs; it can exceed the calendar horizon.

For beta under the demand sweep, the modeled mean-profit sign changes between 0.05 and 0.25 initial qualified leads per week. No precise break-even arrival rate is identified between these grid points.

For network under the demand sweep, the modeled mean-profit sign changes between 0.05 and 0.25 initial qualified leads per week. No precise break-even arrival rate is identified between these grid points.

The split-sample initial-capacity menu selects scale complete double for beta. Its eight fresh test worlds give mean gain USD1,331,416 relative to the central capacity, with conditional interval [685,519, 1,977,314]. This tests the chosen starting configuration, not globally optimal expansion timing.

The split-sample initial-capacity menu selects scale complete double for network. Its eight fresh test worlds give mean gain USD1,448,477 relative to the central capacity, with conditional interval [890,348, 2,006,607]. This tests the chosen starting configuration, not globally optimal expansion timing.

11.5 Counterfactuals and the moving frontier

The component-removal and duplicate controls use the same initial resources and seeds. Central and conservative policies retain no graph capabilities, so component-removal effects there cannot establish a skill mechanism. The deliberately forgetful no-retention diagnostic is not the fair Beta comparator; ordinary competent learning remains in all primary comparisons.

Table 11: Favorable component controls: network minus altered network, first 16 confirmation seeds.

Control	Mean surplus difference (USD)	95% MC interval
remove representation	1,591,250	[1,215,509, 1,966,990]
remove falsifier	607,215	[354,033, 860,397]
remove composition	13,805	[-17,871, 45,480]
duplicate	94,431	[-84,834, 273,697]
no retention	407,496	[155,367, 659,625]

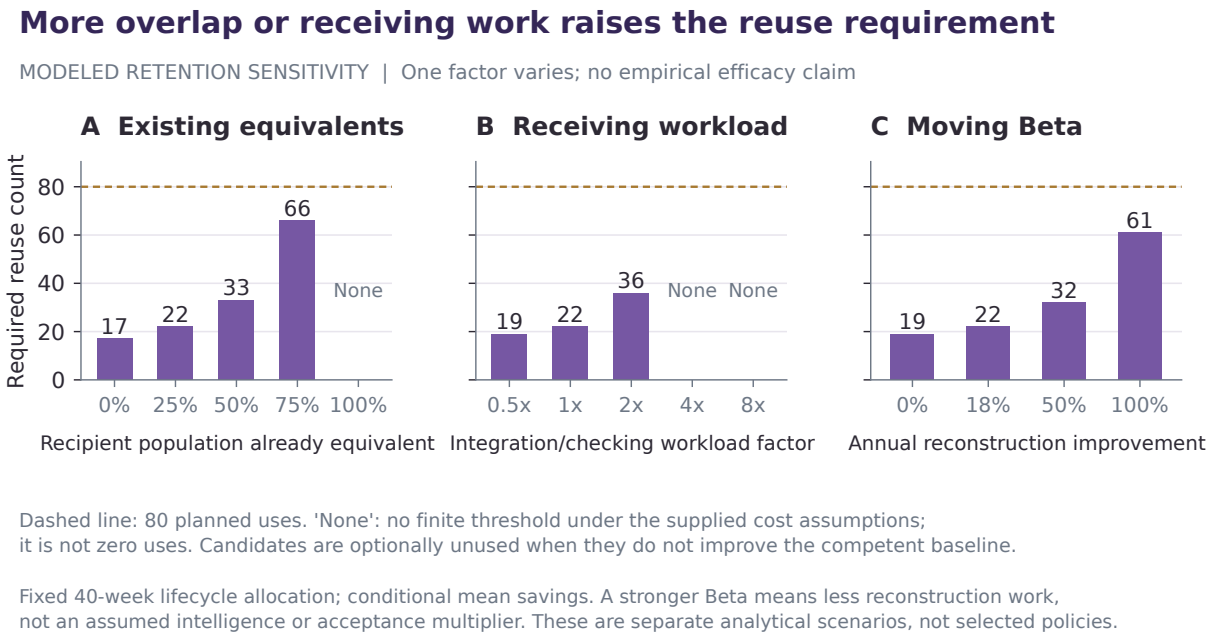


Figure 7: The separate recipient panel exposes how overlap, checking and improved Beta raise the strict reuse threshold. No finite threshold remains a distinct negative result.

Large representation and falsifier removal losses are total intervention effects. The removal masks runtime eligibility while prospective appraisal can still value the removed component, allowing futile retention attempts and changed routing or queues. These controls diagnose dependence and policy brittleness; they are not clean estimates of a component's isolated net Alpha and do not overturn the fair Beta comparison. The duplicate control explicitly marks the proposed asset redundant, testing rejection mechanics rather than independently discovering the frequency of redundancy.

Among the 57 retention-versus-Beta stresses, the only positive paired interval occurs at USD10,000 seed. Both arms suffer cash shortfall in all 16 worlds and retain no capability; the USD334 difference is a smaller failure loss, not Alpha. Fast Beta and rapid copying instead produce negative retention increments. Some optional-C enablement contrasts are positive despite zero C campaigns in those conditions: evaluator overhead and timing can change outcomes without producing a new asset. The complete corpus records three C campaign events in reduced-construction-cost conditions, not three independent invention replications: one seed appears in two alternative conditions. It establishes no separate-R&D capability feedback loop.

The commercial core is a positive modeled finding within its stated assumptions. The optional capability result is narrower: paid-job accounting can make remaining retention much less expensive than a new invention campaign, and the executable system can test and enforce a conditional Alpha claim. Full overhead, capacity, adaptation and demand still determine whether that enhancement pays. This edition does not establish a durable moat, a new commercially validated offering unlocked by sharing, an extra-R&D feedback loop or self-acceleration. None is required for profitable Beta delivery.

12 What the mechanism would need to demonstrate next

The companion `evidence/OPERATIONAL_VALIDATION_PLAN.md` specifies a proposed qualification sequence, instrumentation and decision gates. No operational pilot is claimed by that plan.

A practical next step is a bounded, buyer-defined coordination-design engagement with a written acceptance contract, matched human quotation and task-specific frontier baseline. Log the entire underlying model usage, tool calls, retries, elapsed time, errors and delivery cost. Preserve the paid job's learning cost in A. Propose an eligible sanitized capability and meter only its remaining extraction, fresh validation and distribution in B. Hold back fresh recipient tasks and protected scorer details. Test an already capable recipient, including a recipient with an equivalent method. Let rejection or optional nonuse stand.

Before operation, the existing owners would need to qualify the exact model, tool, verifier and runtime configuration, commission actual identity and key management, and authorize the particular deployment and external effects. This research creates no new credentials, chain transactions, financial authority or permission to bypass existing gates. Public deterministic fixture keys must not be treated as a production secret. Live contracts and native settlement require their own current configuration and eligibility checks.

The claim of an evidenced reinvestment feedback loop would additionally require that collected surplus finances extra work and that the resulting capability improves later accepted outcomes after its complete costs. A job-born retention loop may exist without separate research. Neither increasing job volume nor more graph nodes establishes self-acceleration. That stronger claim concerns improving the effectiveness of discovering or validating worthwhile capabilities across independent invention campaigns.

13 Limits and interpretation

This study provides calibrated billing and service-rate anchors, not a fully empirically calibrated venture. The largest uncertainties concern matched scope, actual frontier task performance, willingness to pay, repeatability and transfer. The high-cost human references and modest assumed machine workloads can produce large modeled margins; their ratio is an assumption-sensitive result, not a measured productivity multiple. The sparse finite venture family cannot establish economy-wide demand or unrestricted invention.

Exact local correctness is conditional on the modeled types and distributions. The witness's increased aggregate checking counts are retained. Graph arithmetic and cryptographic integrity do not make an assumed response surface true. Model improvement can absorb a former advantage; copying can remove rarity; continual maintenance can consume the remaining value. Access restriction is a policy boundary, not a demonstrated economic moat.

The simulator uses bounded policy families and supplied population priors. Capacity expansion is tested as an initial menu, not solved as a full dynamic control problem. Tax, external financing contracts and jurisdiction-specific professional liabilities are outside the supported operating scope. The finite refund model is not a universal liability estimate. Historical actual research cost is unknown. All of these limits remain visible rather than being filled with token appreciation or hypothetical sales.

The English/French demonstration has deterministic JavaScript interaction tests and independently cross-checked small solvers. Browser rendering could not be qualified because the available runtime lacked Chromium and automatic approval review rejected installing it under the resource restrictions. The study does not claim a rendered-browser test. The PDF is separately rendered and inspected.

Reproducibility, contribution and publication status

The companion contains source, tests, calibration, protocols, retained failures, complete primary cash ledgers and exact result identities. From its root, `python3 -m unittest discover -s tests` executes the tests; `python3 -m src.run_study all -out results/replay -workers 4` replays the declared study in a distinct directory. The witness has a separate verification command and retained source identity. The demonstration rebuilds from recorded results without an account, wallet or network dependency.

The simulator and arithmetic probes require Python; the graph additionally uses the recorded cryptography package. Native accounting modules are reused with their license and provenance. Their adapter preserves modeled receipts as assumptions, never observed collections. Node 24.22.1 local checks are not a substitute for the repositories' required installed/release qualification.

Research design, implementation, source review, experiments and manuscript were AI-assisted. No independent customer trial, external institutional review or peer-reviewed publication is claimed. The source contribution is a draft for the existing owners under `studies/network-invented-economy-v1/`, coordinated through issue 4850. Earlier owners, studies and evidence remain unchanged.

References

[1] Vincent Boucher. Transformational Intelligence: Learning to change the game. MONTREAL.AI research manuscript, October 2026. Supplied manuscript, October 2026; inspected 26 October 2026; not redistributed in this research package.

[2] Vincent Boucher. Gym, Specialist ASI & Successor Ω: The sovereign mission environment, the proven intelligence institution, and the successor to the successor in singularity time. MONTREAL.AI Research, August 2026. Version 2.1.0, 10 August 2026. Supplied manuscript; inspected 26 October 2026; not redistributed in this research package.

[3] Vincent Boucher. Mission-Sovereign Successors: GoalOS Singularity Navigator Ω + SEIZE—Alpha, Beta, Spread, and the verified succession institution in singularity time. MONTREAL.AI and QUEBEC.AI, August 2026. Version 5.0.0, Authoritative Publication Edition, 9 August 2026. Supplied manuscript; inspected 26 October 2026; not redistributed in this research package.

[4] MONTREAL.AI Research and Engineering. SUCCESSOR Ω HARNESS: Make every mission build what comes next. Experimental research edition, October 2026. 24 October 2026. Supplied manuscript; inspected 26 October 2026; not redistributed in this research package.

[5] James Bornholt, Rajeev Joshi, Vytautas Astrauskas, Brendan Cully, Bernhard Kragl, Seth Markle, Kyle Sauri, Drew Schleit, Grant Slatton, Serdar Tasiran, Jacob Van Geffen, and Andrew Warfield. Using lightweight formal methods to validate a key-value storage node in amazon s3. In *SOSP*, 2021. URL https://www.amazon.science/publications/using-lightweight-formal-methods-to-validate-a-key-value-storage-node-in-amazon-s3. Accessed 26 October 2026; deferred-venture technical precedent only.

[6] Bernd Finkbeiner and Noemi Passing. Compositional synthesis of modular systems (full version), 2022. URL https://arxiv.org/abs/2106.14783v2. Accessed 26 October 2026; version 2 dated 12 August 2022.

[7] I. Damgard, V. Pastro, N. P. Smart, and S. Zakarias. Multiparty computation from somewhat homomorphic encryption. Cryptology ePrint Archive, Paper 2011/535, 2011. URL https://eprint.iacr.org/2011/535. Accessed 26 October 2026; last revised 6 November 2012.

[8] Jiayuan Liu, Mingyu Guo, and Vincent Conitzer. An interpretable automated mechanism design framework with large language models, 2025. URL https://arxiv.org/html/2502.12203v1. Accessed 26 October 2026.

[9] Sai Kiran Narayanaswami, Swarat Chaudhuri, Moshe Vardi, and Peter Stone. Automating mechanism design with program synthesis, 2022. URL https://ala2022.github.io/papers/ALA2022_paper_30.pdf. Accessed 26 October 2026.

[10] OpenAI. Gpt-6 astra model, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed 26 October 2026.

[11] OpenAI. Gpt-6.1 terra model, 2026. URL https://developers.openai.com/api/docs/models/gpt-6.1-terra. Accessed 26 October 2026.

[12] OpenAI. Using gpt-6, 2026. URL https://developers.openai.com/api/docs/guides/latest-model. Accessed 26 October 2026.

[13] OpenAI. Reasoning models, 2026. URL https://developers.openai.com/api/docs/guides/reasoning. Accessed 26 October 2026.

[14] OpenAI. Introducing gpt-6.1 terra, 2026. URL https://openai.com/index/introducing-gpt-6-1-terra/. Accessed 26 October 2026.

[15] OpenAI. Openai api pricing, 2026. URL https://developers.openai.com/api/docs/pricing. Accessed 26 October 2026.

[16] OpenAI. Managing usage with gpt-6 astra in work and codex, 2026. URL https://help.openai.com/en/articles/20001516-managing-usage-with-gpt-6-astra-in-work-and-codex. Accessed 26 October 2026.

[17] Xiangyi Li et al. Skillsbench: Benchmarking how well agent skills work across diverse tasks, 2026. URL https://arxiv.org/abs/2602.12670v4. Accessed 26 October 2026.

[18] Tingxu Han et al. Swe-skills-bench: Do agent skills actually help in real-world software engineering?, 2026. URL https://arxiv.org/abs/2603.15401v1. Accessed 26 October 2026.

[19] Zisu Huang et al. From raw experience to skill consumption: A systematic study of model-generated agent skills, 2026. URL https://arxiv.org/abs/2605.23899v1. Accessed 26 October 2026.

[20] Chaoliang Yan et al. One skill too many: How co-installed skills conflict in coding agents, 2026. URL https://arxiv.org/abs/2610.11647v2. Accessed 26 October 2026.

[21] Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine llm evaluation panels, 2026. URL https://arxiv.org/abs/2605.29800v1. Accessed 26 October 2026.

[22] RAND Corporation. Rand gsa federal supply schedule price list, 2025. URL https://www.gsaadvantage.gov/ref_text/47QRAA24D008Z/110BBS.3WQO6G_47QRAA24D008Z_47QRAA24D008ZCPA19672FNL.PDF. Accessed 26 October 2026.

[23] Inc. Z. Z, inc. gsa mas price list modification po-0018, 2024. URL https://www.zfederal.com/wp-content/uploads/2024/10/Z-INC-GSA-MAS-Pricelist-Mod-18.pdf. Accessed 26 October 2026.

[24] Vincent Boucher. Merkle roots for the Validated Skill Graph. MONTREAL.AI Research, GoalOS Protocol Note Series, . URL https://github.com/MontrealAI/goalos-agialpha-sovereign-machine-economy/blob/1e9a5675b1662e98074fb1c71cf2c5b94c7d259f/docs/paper/Merkle_Roots_for_the_Validated_Skill_Graph_v114.pdf. Research Edition v4.0 / GoalOS V114; publication date not stated in the inspected edition. Public primary-source snapshot accessed 26 October 2026; source manuscript not redistributed in this research package.

[25] Vincent Boucher. PROOF_LOOPS.md: Field notes on machines that turn work into capability. MONTREAL.AI Research, GoalOS Protocol Note Series, . URL https://github.com/MontrealAI/goalos-agialpha-sovereign-machine-economy/blob/1e9a5675b1662e98074fb1c71cf2c5b94c7d259f/docs/paper/PROOF_LOOPS_research_paper_v2_author_edition.pdf. Research Edition v2.0; publication date not stated in the inspected edition. Public primary-source snapshot accessed 26 October 2026; source manuscript not redistributed in this research package.