Designing the Systems That Organize Science: Executable Institutions and Objective-Conditioned Optimization

Science Utopia Lab
Integrated research working paper, November 17, 2026
Not submitted; human authorship and affiliations to be finalized

Abstract

What changes when an objective selects the system that organizes science, rather than only a researcher's next action? We formulate objective-conditioned institution design around the persistent ScienceUtopia researcher world. Parameters, executable rules, mechanism composition, information routes and processes are editable; researcher-model weights and each study's measurement functions remain fixed. Black-box proposals, LLM program authors and scoped gradient estimators share a native execution boundary, without substituting a miniature numerical world for researcher activity. We distinguish executable designs, actual behavioral changes and improvements under the assigned objective. Developer and external-author designs exercise native plans and information flow and sometimes improve observed proxies. However, a matched local-designer comparison retaining 120 attempts and 90 selection-fixed evaluations loses its apparent advantage against strong simple rules, with no exercised topology change. Native gradient followups do not beat their recorded incumbents. Later source/workflow studies retain ties, adverse effects and runtime failures rather than treating validity as gain. Two same-authority hand-written topic rules yield 20 versus 36 accepted decisions at equal collected cost 200. A separate terminal hand rule lowers resource-aware loss by four units through collection avoidance, without new scientific output. A subsequent two-objective source-information pilot retains all eight own slots: seven native outcomes complete, one remains null, and all four selected systems tie their strong visible reference. Its adverse utility path recovers a known supplied rule, not a new best-known institution. A later finite method pilot exposes a document-visibility/tool-phase mismatch: its measured outcomes remain valid, but method and objective attribution is confounded. All 24 submitted native outcomes complete: 12 tie and 12 worsen their strong visible reference; no proposal improves it. The prospective repair is not evidence of useful search. A corrected four-step successor retains all 24 attempts: 23 outcomes complete, one remains an execution null, eight tie and 15 worsen their strong reference. All six primary best-known endpoints return supplied references; the separate own-only endpoints have three ties and three adverse results. Literal code edits and actual native invocation still do not suffice for useful institutional invention. Complete programs, paths, costs and input/response classes remain visible, without pooling studies as independent replications. An English pixel-art interface exposes the same systems and trajectories. The program establishes an executable research interface and its current limits, not validated science-policy objectives, universal differentiability or reliable autonomous institutional invention. An important supported main-track claim and adequate independent evidence remain open.

Preprint.

Contents

1 Introduction . . . 3
2 Problem: objectives select organizing systems . . . 5
3 An executable space of scientific systems . . . 6
4 Search methods and gradient boundaries . . . 8
5 Evidence strategy: complementary studies, not pooled replication . . . 10
6 Objective-conditioned parameter optimization . . . 10
7 Executing institutions beyond fixed parameter menus . . . 12
8 Nested free design and conditional distribution updating . . . 13
9 What native gradient experiments establish . . . 16
10 Matched local-designer comparison with stronger references . . . 17
11 A game-like observatory over the same research world . . . 23
12 Before scaling: discriminating mechanisms from interface burden . . . 24
13 Authority, authoring and measurement diagnostics . . . 30
14 Native code execution and the support of the objectives . . . 33
15 A delivered information intervention is not necessarily an effective one . . . 44
16 Capacity-limited native search preserves a strong allocation rule . . . 45
17 Concrete agendas change native planning and accepted output . . . 48
18 Free planning search preserves strong reference behavior . . . 50
19 An operative numeric parameter has a conditional, nonmonotone response . . . 53
20 A finite native parameter comparator does not improve the incumbent . . . 55
21 Changing observation and action timing in the native world . . . 57
22 Closing the compiler-feedback structural loop . . . 59
23 Actual staged construction in the original scientist world . . . 65
24 Matched Objective and Feedback-History Comparison . . . 68
25 Original Researcher Choice and Resource-Use Diagnostics . . . 71
26 Original-Topic Authority and Three-Objective System Search . . . 74
27 A reachable resource-objective gain without new scientific output . . . 80
28 Source-grounded free-system search under two objectives . . . 83
29 Restricted and free methods: a finite native pilot . . . 86
30 Corrected four-step institution search: design and scope . . . 91
31 Corrected search: complete outcomes and retained incumbents . . . 93
32 Discussion: what can be concluded across the program . . . 96
33 Related work . . . 99
34 Limitations and responsible interpretation . . . 101
35 Conclusion . . . 103
References . . . 104
A Complete parameter-study context . . . 106
B Complete structural case tables and interpretation . . . 106
C Gradient assumptions, diagnostics and complete followups . . . 107
D V2 all-candidate cross-scores and failure accounting . . . 109
E V2 costs, resources and separate longer horizon . . . 111
F Fixed-record incentives versus dynamic policy optima . . . 116
G Reproducibility, provenance and editorial scope . . . 116

Reading this manuscript. This integrated working paper develops the whole research argument without a conference page cap. The experiments are distinct studies, not one pooled benchmark. Sections 2–4 define the design and optimization interfaces; Sections 6–10 present complementary evidence; Section 11 describes the interface. Sections 12–31 retain later studies and diagnostics with their distinct inference boundaries. Appendices retain assumptions, complete case tables, cross-scores and costs. Existing short drafts and their source-bound evidence remain unchanged.

1 Introduction

Scientific organization is usually treated as the setting in which research occurs: institutions allocate resources, evaluation selects work, publication circulates results, and careers affect future participation. These arrangements can instead be design variables. Given an explicit objective, which organizing system would an optimization process construct, and what would it make researchers do differently? This question concerns the system that coordinates science, not only tuning an acceptance rate or automating an individual scientist.

ScienceUtopia supplies a persistent, literature-grounded LLM-agent ecosystem with research, submissions, peer review, publication, funding and longitudinal researcher state [Jin et al., 2026]. We extend that world rather than replace it with a smaller numerical model chosen for convenient gradients. The original organization is one representable instance, not the upper bound of the institutional language. Review is an example, not a mandatory family: new organizing programs may alter information flow, planning, resource allocation, communication or the relations among mechanisms.

The central separation is between *what can change* and *how it is searched*. Parameters, algorithms, added mechanisms and rewired processes form a spectrum of design freedom. Random proposals, loss-feedback program search, continuous optimization and score-function estimation are methods that can act on different parts of that spectrum. An institution need not be globally differentiable to admit useful optimization. Conversely, writing a tensor representation or computing a derivative is not evidence that the derivative faithfully predicts the native research world.

The research contribution and its scope. We present an executable formulation of scientific-system design, an integrated native execution/search/evidence interface, and a sequence of bounded studies that test different parts of the formulation. Parameter search demonstrates objective-dependent selection and uncertain held-out profiles. External-author cases establish that new source, protocols, project processes and feedback routes can execute in the original world. Gradient studies distinguish exact local computation, approximate response representations and conditional expected-loss gradients. A matched local-designer study tests permitted freedoms against stronger references without dropping failures. The game-like interface makes the same designs and observations inspectable.

Code-defined agent search, mechanism synthesis and scientific harness adaptation are existing research directions, not claims of priority made here [Narayanaswami et al., 2022, Hu et al., 2025, Wu and Liu, 2026]. Our empirical target is the map from a declared objective to an executable organizing system, its native behavioral consequences and its unused-condition profile. Identifying that object does not yet establish a general result about the value of free institutional design; the bounded studies below retain that gap rather than claiming novelty for a generic propose–evaluate–revise loop.

These are not claims of a new optimization identity, a universally good scientific institution or a trained self-improving inventor. The main empirical tension is that *representational freedom, exercised institutional change and effective optimization are different achievements*. Wider permission does not guarantee a submitted program; a new node need not improve loss; a lower proxy need not improve science. The experiments make these differences observable rather than enforce a favorable outcome.

Evidence at a glance: what supports which claim?

The central empirical question is not whether an institution can be encoded, or whether a chosen proxy can be made smaller than a weak starting system. It is whether objective-conditioned search finds consequential organizing rules beyond strong same-authority alternatives, and how those rules

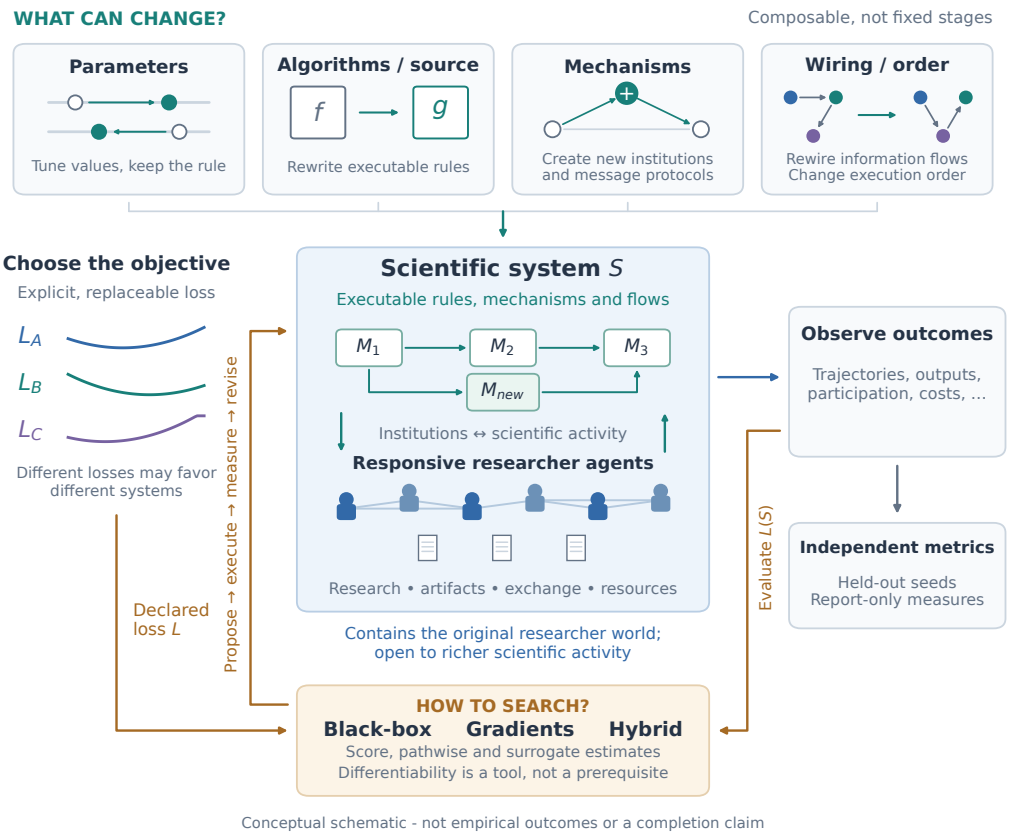


Figure 1: **Research program: optimize scientific organization.** The objective, the editable system and the search method are separate objects. Parameters, algorithms, mechanisms and wiring/processes are design freedoms, not four optimizer names. The native world is contained rather than replaced. The diagram describes the research space, not proof that every experiment uses every freedom or that any method improves scientific truth.

act in the persistent researcher world. Table 1 separates the strongest completed comparisons from claims they do not establish. The table is a reading map, not a pooled estimate or a declaration that the research program is publication-ready.

Loss symbols are study-local, not interchangeable metrics. In the matched V2 study, $T = -A$ and $C = 4A - Q + 0.1R$: A counts accepted decisions, Q sums their native peer scores and R counts reviews. In the later content/resource studies, $L_U = -U$ and $L_C = -U + 0.1K$: U sums centered opinions from a fixed external abstract-content observer, while K is collected native activity cost. Native peer scores and external opinions are different measurements; review counts and collected resource units are different costs. Neither objective is retrospectively substituted for the other. The historical parameter objectives and their unused-condition evaluation remain separately defined in Section 6.

A completed finite search is not a method advantage. At November 16, 2026, 14:05 UTC, the later priority/free-method writer exits normally with all 24 own observations and three references. Twelve own observations tie and 12 worsen their visible reference; none improves it. The 12 frozen own-only selections yield eight ties and four worse losses. The full original paired cold/cost/readout chain is still running at the 14:11 UTC verification. Section 29 includes all 24 slots, while older dated receipts remain unchanged. Its author visibility/phase defect prevents isolated attribution to objective or method; prospectively corrected software is not an observed corrected search result. Literal 24-slot values, frozen selections and original input bindings are retained in `research/results/native-priority-method-terminal-snapshot-v1.json`. The orig-

Completed comparison	Supported observation	Not established
Matched local search (Section 10)	120 slots: 64 complete, 25 generation nulls, 31 execution nulls. After frozen selection, 90 physical evaluations of 15 exact policies retain stronger rules: throughput code, technical-only and uniform selections average 6.2 acceptances, tying all-accept; LLM numeric retains the seed at 0.2. Review-cost code selections average loss 1.92, worse than reviewer-free reject at 0.	Free-design or feedback superiority; structural invention. The 90 jobs are not 90 independent optimization histories.
Same-authority topic rules (Section 26)	Two hand-written rules yield 20 versus 36 accepted decisions at equal collected cost 200. Original topic durations and scientist weights are fixed; actual topics and subsequent researcher inputs differ.	Automatic discovery, duration-only mediation or a world-level optimum.
Terminal resource rule (Section 27)	A hand rule improves declared resource-aware loss by 4: collected cost 200→160, unpaid cost 0→40, requested cost still 200. All 179 scientist inputs/literal choices and 36 accepted records match.	Less requested work, better science, an invented institution or long-run benefit. The gain is retained, not disqualified.
Source-information search (Section 28)	Eight own slots retain seven completed outcomes and one native null. All four frozen own selections tie the strong visible rule; an adverse path recovers a supplied rule rather than a new best-known program.	A general negative effect of source access, an impossibility of useful institution search or a clean isolated measure of understanding.

Table 1: Evidence and inference boundaries. Losses and budgets belong to their own studies; matching symbols do not make them a common benchmark. Hand controls establish reachable effects but are not discoveries by the local inventor. Technical nulls remain unavailable measurements, not bad numeric scores. Complete programs, all attempts, costs and separate conditions remain in the corresponding sections and appendices.

inal paired audit/cost/readout/all-eight-author chain and separate paired visibility exports subsequently close at 14:21 UTC with unchanged counts and selections. This closes the finite original study, not the corrected-inventor, independent-evidence or main-track contribution questions.

The earlier structural, gradient, nested-design and English game-interface tracks also remain in this manuscript. Their different authors, sources and conditions cannot be added to the matched local experiment as independent replications. Available design freedoms and available optimizers are not themselves evidence that those freedoms are useful or that any optimizer improves expected institutional performance. A focused main-track claim and adequate independent evidence remain open, even though substantial software and finite studies are complete.

2 Problem: objectives select organizing systems

Let $S \in \mathcal{S}$ specify an organizing system, W researcher-model weights, D the literature corpus, c an environmental condition and ϵ residual execution/model randomness. The native simulator generates a trajectory

$$\tau \sim P(\cdot \mid S, W, D, c), \qquad m = M(\tau). \tag{1}$$

It contains research directions, projects, papers, review and publication records, resources, messages and persistent state. A declared loss ℓ_λ defines the design problem

$$J_\lambda(S) = \mathbb{E}_{c,\epsilon}[\ell_\lambda(M(\tau))], \qquad \text{seek } S \text{ with small } J_\lambda(S). \tag{2}$$

Finite observations estimate only their declared conditions. A single saved trajectory is neither an exact expectation nor a global optimum. Weights W are fixed within each experiment, but behavior is not: a different information route, resource state or publication history changes future LLM inputs.

Table 2: Four separate contracts. No substantive veto is implicit.

Object	Question answered	Search role
LossSpec	Which measured system is preferred under this objective?	Proposal feedback and selection
Validity	Did literal code execute and yield the required measurements?	Numeric loss or explicit null
Evaluation	What happens under other scores or unused conditions?	Withheld when declared held out
Adoption policy	Should a measured institution be deployed?	Separate, only if explicitly chosen

2.1 Loss, validity, evaluation and adoption

Four contracts have different jobs (Table 2). The optimized loss selects candidates. Technical validity checks whether an observation can be obtained under the declared execution and budget contract. Outside metrics and unused conditions characterize selected systems. An optional adoption policy governs deployment, not the search objective.

An execution failure has a missing loss, not an invented large penalty or a zero score. We retain the consumed slot, draft, error and observed cost. An objective may select a normatively unattractive institution: we do not reject it through an undeclared fairness, novelty, participation or improvement floor. This does not mean technical completion makes the institution desirable. If evaluation informs the next proposal, it is development feedback rather than held-out evidence.

2.2 Objective design is a scientific question

Science-of-science research motivates examining careers, knowledge production, collaboration, diversity and institutions jointly [Fortunato et al., 2018]. Responsible research assessment also cautions against equating a metric with the mission it is meant to assess [Hicks et al., 2015]. We use this perspective to expose objective choices, not to declare one official national-policy index the universal target. Publication counts, topic entropy, review scores, participation and resources can each be optimized or diagnosed; their role must be declared before a campaign.

Optimization creates a further distinction: a useful measurement for describing a fixed system may become a poor objective once the system can change its measurement-generating process. A program can reduce review work by omitting review, raise throughput by changing project timing, or select historically favorable reviewers. Those are possible answers to a declared proxy, not automatic evidence of epistemic progress. We therefore preserve raw components and cross-objective profiles instead of reporting only a winner's scalar.

3 An executable space of scientific systems

3.1 Specifications, implementations and composition

A SYSTEMSPEC declares actor/artifact types, mechanism instances, input/output ports, directed connections, temporal latency, schedules, initial state and invariants. A mechanism specifies an implementation and its parameters/state contract. Its implementation can be native trusted code, isolated authored Python, an LLM operation or a supported tensor representation. A fingerprint covers literal specification bytes after canonical serialization; it is identity evidence, not a measure of algorithmic novelty.

The design freedoms are related but should not be collapsed into a fixed menu:

1. **Parameters**: change settings within a fixed implementation and graph.
2. **Algorithms**: replace executable mappings or state updates behind a compatible interface, not merely select a named policy.
3. **Mechanisms**: introduce or remove actors, storage, decision services or source-defined institutional components with explicit consumers.

4. **Wiring and processes**: change who receives what, when it arrives, the dependency order, conditional invocation and bounded repetition.

Study-specific domains can restrict these freedoms for comparison. A valid schema does not make every imaginable graph executable in a particular world. For example, an empty graph may be structurally well formed but fail the native observation contract. Omitting review requires a compatible downstream decision consumer; the host does not invent missing reviews or silently repair the publication rule.

Graph syntax and operational authority are different. An arbitrary program with the same observations, state and invocation opportunities may implement several mechanisms inside one policy. Conversely, a program invoked only after review cannot undo the reviews already performed in that period. Comparisons must therefore specify available observations, effects and execution times, not only node count or the label "free graph." A restricted parameter family can lack an operation even when another code representation could express it. Section 13 examines these boundaries without claiming that graph syntax is universally necessary.

3.2 Containment of the original researcher world

`NativeScienceKernel` executes specifications against the original Simulation lifecycle. Default nodes delegate to its phase methods rather than approximate scientist behavior numerically. A private annual frame retains native domain objects; versioned JSON ports expose bounded records to authored mechanisms. Review selection is separable from publication commit, so new decision code can still affect the actual paper tracker, retrieval archive, resource state and later researcher prompts.

Default behavior has deterministic native-lifecycle checks comparing scientific state, RNG state and full request payloads under matching scripted replies. These are engineering containment tests, not new Qwen experiments. The live studies below separately use fixed local Qwen3-8B scientists and actual literature-grounded native forwards. They are small local settings, not a reproduction of the source paper's reported Qwen3-32B configuration. The local four-year literature subset contains 103,107 CS records from 2016–2019, drawn from SciEvo [Jin et al., 2024]. Corpus and retrieval identities are bound separately from the researcher model; a local run is not a replication of the source paper's model scale or full experimental population.

The extension is conservative, not unlimited. Generated code cannot import arbitrary host objects, change researcher weights, rewrite a measurement or invent a trusted capability by naming it. Annual native phase handlers retain their own scientific semantics. A new kind of real activity needs an implemented consumer, identity/accounting transport and measurement interpretation; a record called "discovery" does not become a validated discovery.

Native effects versus a fixed historical calendar. The later workflow domain fixes native effect contracts rather than their annual schedule. It permits institutional code, nodes and typed wiring while requiring exactly-once protected effects, producers before zero-latency consumers, opening before activity and ledger closure after mutations. Descriptive ordering metadata is not execution authority. A structural-only experiment can then hold sources, parameters and nodes literal and change phases/connections alone; that experimental restriction is not the general system language's ceiling. A hand timing control changes three phases and five edges without rewriting any native effect. Its fresh scientist-world loss is an empirical question, not a consequence of graph validity. The complete new-binding software check uses seven SCRIPTED native worlds, 28 annual checkpoints, 13 bounded mock-HTTP workers and a retained generation null, with matching double audits, costs and reports. This exercises original production and cold readers, not real-model workflow invention or improvement.

3.3 New institutions and their actual consumers

Isolated Python receives declared inputs and state and returns schema-checked outputs/state. It can implement an entirely new algorithm, generate intermediate protocols or maintain institutional memory. It runs in a pinned, network-disabled container, not as model-supplied code imported by the host. Output shape, identities and required port coverage are checked; scientific merit is not.

The domain bridge supplies generic effects: persistent actor/artifact creation and revision, namespaced state, delivery of arbitrary linked records to real researchers, actual project-plan changes and conserving resource transfers. Roles and schema-declared state are data; source supplies the institution's allocation or communication rule. New actor serialization alone does not provide autonomous behavior. Custom artifacts can link native papers but do not increase accepted-paper counts merely by being stored.

Information must reach a consumer to matter. A message route records producer, port, event lineage and recipient. Researcher context delivery modifies actual LLM requests; decision outputs address real submitted papers; plan amendments alter actual native project state. Execution receipts can establish delivery and committed effects. Neither a node label nor the designer's explanation establishes that the changed code ran, was read or caused a better outcome.

3.4 Processes, time and state

Dependencies and typed data edges define deterministic within-period ordering. Delayed edges carry records to later periods. Controlled activities support source-authored Boolean guards, bounded counts, persistent state and explicit termination. Missing control can be declared a wait or an error; a skipped activity is not fabricated research. Bounds are execution budgets, not merit filters. Original single-shot annual handlers cannot simply be repeated with their old IDs and side effects. Richer repetition requires repeatable activities with genuine invocation identities and domain effects.

Checkpoints retain mechanism state, queued messages, actors, artifacts and source binding. Resume cannot silently switch to a different institution. This supports longitudinal feedback within a world. Most reported optimization trials, however, reset their worlds: repeated candidate evaluation is not the same as one society recursively redesigning itself over time.

4 Search methods and gradient boundaries

4.1 A common propose–execute–measure loop

The persistent controller accepts a loss, initial specification, design domain, information policy and finite budget. It saves a proposal before dispatch, validates its literal transaction, executes a native forward, measures canonical loss and updates eligible search history. The generic ask/tell boundary does not require simulator derivatives. A domain adapter, not the optimizer, understands coordinates, source replacements or graph transactions.

Black-box optimization includes uniform sampling, incumbent mutation, local refinement and LLM proposals. The implementation supports an objective and own-history archive, not an obligation to change or improve a candidate. Repeated designs, worse revisions and ties consume their scheduled budget. The host does not rewrite generated code, manufacture a terminal submission or replace a failed slot. Trace-informed designers may inspect more than a scalar oracle; such comparisons must disclose their richer information.

An LLM designer can emit executable definitions and structured patches. The execution engine checks the resulting program rather than interpreting prose as functionality. In the local matched study, the scientist and designer are different fixed models. External-author pilots have different provenance: they establish expressive examples and observed redesign, not local-model invention or an authenticated independent designer identity.

4.2 Pathwise derivatives of explicit representations

For a supported differentiable institution $u = f_\theta(x)$, a local backward pass computes derivatives of that represented mapping. Tensor state and explicit routes can carry derivatives through multiple represented operations and invocations. Such a derivative is exact for that computation, not for an opaque LLM, retrieval, integer parser or unrepresented native transition.

A conditional response operator $\widehat{r}_\phi(x, \theta)$ fitted to native observations can supply missing response sensitivities in an approximate path. Autograd then computes $\nabla_\theta \widehat{L}$, not automatically $\nabla_\theta J$ of Eq. (2). Hard selection may be relaxed by smooth probabilities while the native forward remains

hard. Temperature changes both fit and sensitivity. A low calibration error does not establish accurate unused-condition predictions, and a useful local fit does not make arbitrary generated source differentiable.

We keep represented derivatives, native perturbations and projected native reevaluations separate. Opaque links remain unknown, not zero-imputed. If a hybrid proposal combines estimates from several sources, its target, bias and eligible data must be named; blending estimates is not an unbiasedness proof. The native-response experiments in Section 9 examine this boundary directly.

4.3 Expected-loss gradients around an opaque world

Let a fixed generated vocabulary V map explicit discrete choices and continuous latents d to $S(V, d)$. For a declared distribution $q_\theta(d \mid V)$,

$$J_V(\theta) = \mathbb{E}_{d\sim q_\theta,\, \tau\sim P(\cdot|S(V,d))}[\ell(M(\tau))]. \tag{3}$$

Under support, integrability and differentiation conditions, and with no additional direct θ dependence of the conditional environment,

$$\nabla_\theta J_V = \mathbb{E}[(\ell - b)\nabla_\theta \log q_\theta(d \mid V)]. \tag{4}$$

This standard likelihood-ratio identity estimates a design-distribution gradient without a world Jacobian [Parmas and Sugiyama, 2021]. It covers the explicit categorical/Gaussian design law, not unknown probabilities of arbitrary LLM source synthesis or researcher text. Fixed transforms from unbounded latents to legal coordinates make the sampled density and executable value distinct objects that must both be retained.

For K IID design draws, leave-one-out baselines exclude the scored draw's own loss. For a prior-frozen surrogate $c(d)$ with tractable expectation,

$$\nabla_\theta J_V = \mathbb{E}[(\ell - c(d))\nabla_\theta \log q_\theta(d \mid V)] + \nabla_\theta \mathbb{E}_{q_\theta}[c(d)]. \tag{5}$$

The correction matters; inaccurate controls need not bias this identity but can increase variance. Fitting on the current scored batch or omitting the correction changes its justification. Composable gradient frameworks and composite estimators motivate these interfaces, not a claim to have reproduced their complete algorithms [Parmas and Seno, 2022, Parmas et al., 2023].

Validity-dependent missing losses need special care: discarding failed draws does not automatically estimate the unconditional expectation above. We keep null batches and unavailable updates explicit. A campaign may instead declare a failure-valued estimand in advance, but these studies do not invent one after observing failures. A computed update and a distribution actually evaluated after that update are also separate evidence.

4.4 Changing the design language itself

An outer author can use its own measured history to change V: new source, mechanisms, factors, coordinate meanings and wiring. Inner score updates stay conditional on the current V. If an outer revision removes a coordinate or changes its meaning, old gradients/control state are not silently transported. A point-mass design or retaining the incumbent is legal. Section 8 executes this outer-design/inner-update composition in the original world. It is an application of existing search and estimation ideas, not a new theory of differentiable program synthesis.

4.5 What finite searches and fixed-policy comparisons estimate

Let $\mathcal{C}_{\lambda mh}$ be the declared eligible candidate pool for objective λ, proposal method m and design history h. Eligibility here means the protocol's selection pool, not a hidden scientific-merit test. Write $\mathcal{V}_{\lambda mh} \subseteq \mathcal{C}_{\lambda mh}$ for the complete eligible designs. After development, retain the entire observed minimum set

$$\mathcal{B}_{\lambda mh} = \operatorname*{arg\,min}_{S\in\mathcal{V}_{\lambda mh}} \widehat{L}^{\mathrm{dev}}_\lambda(S), \tag{6}$$

and apply only the predeclared tie rule to choose an evaluation specification $S^*_{\lambda mh}$. Equal losses do not establish equivalent programs or identical trajectories. An initial/reference policy belongs to the pool only when that protocol makes it eligible; submitting an unchanged reference is legal

but is not new invention. An empty eligible complete pool leaves an unavailable selection, not an automatically substituted baseline.

For a fixed selected policy and an explicit reference S_0, the common-condition comparison is

$$\Delta_{\lambda mh}(c) = L_\lambda(S^*_{\lambda mh}; c) - L_\lambda(S_0; c). \tag{7}$$

Negative values favor the selected policy under that same objective. Different losses need not have comparable units, so comparing the magnitudes of T and U is not a method ranking. A common world seed names a matched environmental condition, not identical residual LLM randomness after two policies generate different requests. Pairing is therefore by declared condition and actual observations, not by an assumption that every stochastic event is shared.

Variation over unused c characterizes the already selected policy. Variation over independently repeated h is also needed for an expected claim about the search procedure that selects it. More evaluation seeds for one selected specification do not become more independent optimization histories. Pooling several development histories into one winner likewise yields one selected policy per arm, not one held-out result per history. Different source snapshots, design domains, horizons and observers below are not pooled into a method effect.

A technical null in either member makes that pair's loss difference unavailable. All attempted, failed and unattempted slots remain visible with their actual costs. Reporting loss only for completed pairs is conditional evidence and must accompany coverage; it is not an unconditional end-to-end risk estimate. No retrospective failure penalty is introduced to obtain a numerical ranking. Cross-objective scores reuse a trajectory's measurements when available; they are not additional world samples. Cached judgments of identical content similarly remain the same physical opinions, not independent replications.

The objective-to-system question includes both its selected outcomes and its path: literal revisions, exercised effects, native consequences and cost. Different objectives may return overlapping minimum sets. If the objective also changes a loss-selected editing parent, the contrast concerns the entire objective-conditioned search, not scalar feedback alone with the parent held fixed. None of these definitions requires different winners or a positive gain.

5 Evidence strategy: complementary studies, not pooled replication

The research questions require different tests. A fixed-topology study isolates objective-conditioned parameter selection. Structural cases test whether novel source, protocols and routes actually reach native activities. Gradient pilots test derivative meaning and subsequent native forwards. The largest completed matched study compares local-designer packages and stronger references on a common source. Later topic/resource controls and free-design pilots extend the operational question without becoming replications of that earlier study. Table 3 summarizes these roles; the opening Table 1 separates results from unsupported claims.

Across native case studies, five original researchers and fixed Qwen3-8B weights retain language decisions, literature grounding, projects and papers. Most cases use two years and development seed 42; the older parameter study uses four-year worlds and a different source/seeding protocol. A common seed does not force equal fresh LLM realizations. We distinguish reset worlds, common-prefix-coupled forwards, conditional IID design draws and genuinely unused evaluation conditions. No new inference is introduced by this synthesis.

6 Objective-conditioned parameter optimization

6.1 Design and sealed evaluation

The first study varies eight institutional settings within one fixed topology: acceptance rate, funding application cost, funding slot fraction, resubmission cost, review policy, reviewer capacity, reviewer matching and reviews per paper. The simulator, adaptive-random-search rule, five-candidate budget and development seeds 601/607/617 are held fixed while the optimized LossSpec changes. For valid run r, the losses are

$$\ell_r^{\text{part}} = 1 - a_r + 0.05 b_r, \qquad \ell_r^{\text{topic}} = -h_r, \qquad \ell_r^{\text{burden}} = b_r, \tag{8}$$

Table 3: Evidence families retained in the integrated argument. Counts are within-study budgets, not additive independent replications. Historical source/model/feedback differences prevent pooling method rankings.

Family	Design/author	Native evidence	Question addressed
Parameter study	Eight fixed-topology settings; adaptive random search	45 logical development rows; 15 sealed evaluation runs	Can objective changes select different systems and profiles?
Structural cases V8/V12	External source/protocol/graph author	Eight/six two-year worlds	Do new mechanisms, routes and plan processes execute?
Nested case V13	External vocabulary + conditional LR + redesign	Eight worlds; 592 fresh replies	Can inner updates and outer source changes share a native loop?
Gradient cases V9–V11	Response representation, local refinement, score/control estimator	Coupled paths and eight reset worlds, separately bound	Does backward mean fidelity, a native-tested update or superiority?
Matched study V2	Local 35B designer and uniform numeric control	120 attempts; 90 selection-fixed evaluations	What is achieved under matched permitted freedoms and stronger references?
Later topic/resource controls	Hand-written rules with shared authority	Separate two-world native contrasts; no designer proposals	What effects are reachable without new research laws or changed weights?
Later free-system pilots	Local code/graph authors; explicit source information	Separate six-slot and eight-slot studies, with original nulls	Do generated policies improve the strong supplied rule, not merely recover it?
Priority/free-method pilot	Uniform, adaptive and free code/graph packages	24 complete own slots; 12 frozen selections; paired original closure	Which proposals operate, and how does the interface limit method attribution?

where a_r is active-researcher fraction, b_r review burden per active researcher and h_r topic entropy. Each arm minimizes the development mean plus 0.25 population standard deviations over its three seeds. Entropy is not truth, and burden can fall by suppressing evaluation. These deliberately narrow proxies stress the objective-to-system map.

All 45 logical design–seed rows terminate: 30 complete and 15 fail at the declared reviewer-capacity boundary, retaining null loss. Twenty-nine development rows are physically executed and 16 logical rows adopt byte-identical initial trajectories under the study's provenance contract. Thus 45 logical rows are not 45 fresh native worlds. All winners are sealed before a common evaluation on unused seeds 701/709/719/727/733; all 15 arm-labelled runs complete.

6.2 Different objectives, two actual designs

Participation/review selects proposal 2 with development loss 0.321846. Topic and burden select proposal 1, with losses -1.696931 and 1.668963. In the topic arm proposals 1/2 tie, so the pre-declared earliest rule selects 1. Only participation has multiple valid adaptive proposals beyond the initial pair; later topic/burden proposals fail technical validity. The paths show an executable own-observation loop, not rich or repeatable path divergence.

The selected participation design has acceptance rate 0.4968, capacity 8, topic-based matching and three reviews per paper. Topic/burden share a design with rate 0.1132, capacity 2, random matching and four reviews per paper. Three objective labels therefore produce two designs, not three institutions. Full settings appear in Appendix A.

On unused conditions, the participation-selected design has aggregate losses 0.295 on participation, 2.076 on burden and -1.732 on topic. The duplicated topic/burden design has approximately 0.288, 1.449–1.454 and -1.712. Ordering changes with the scoring loss, but every one of the nine paired

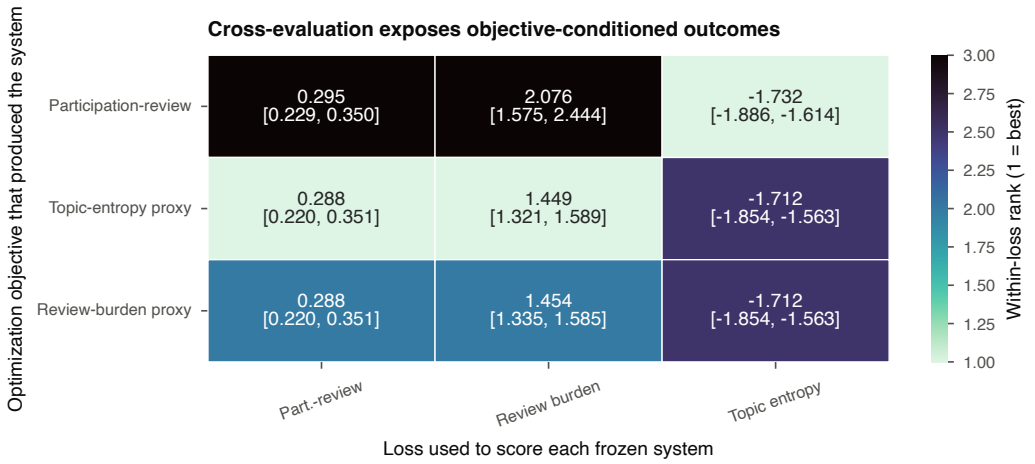


Figure 2: **Historical parameter study: sealed-system cross-evaluation.** Numbers are each complete LossSpec aggregate with the study's exact-bootstrap percentile intervals over five unused seeds; color ranks within a loss only. Topic and burden rows encode the same design but separate stochastic evaluations. All nine paired-contrast intervals include zero. Descriptive ordering is not stable superiority or evidence that every objective discovers a distinct system.

95% intervals includes zero. The apparent rank reversal is descriptive. The selected design is not necessarily best under its own unused-condition loss.

Operational differences are clearer than normative conclusions. The participation-selected design averages 11.6 accepted papers and full review coverage, versus 4.0 and approximately 0.55 for the other design. It also has higher burden, approximately 1.959 versus 1.415–1.421. Activity fractions 0.82 versus 0.80 and topic entropies 1.771 versus 1.750 remain uncertain at five seeds. This is not a welfare ranking, optimizer comparison or structural-search result.

7 Executing institutions beyond fixed parameter menus

7.1 A new planning and delayed-feedback route

The external-author V8 case permits source, intermediate protocols, node addition/removal, wiring and activity processes. Two objectives use $T = -A$ and $U = 3.25A - Q + 0.05R$, where A counts accepted native decisions, Q sums original reviewer-score means on accepted decisions, and R counts review records. Authored final scores do not create Q. Each objective has two generations and two conditional-IID draws per generation.

All eight native worlds complete with 734 fresh Qwen replies. The developer seed has 19 nodes/27 edges; evaluated graphs have 16–18 nodes/22–24 edges. Throughput losses are $[-9, -9, -10, -10]$; utility losses are $[0, 0, 0.4833, 0.4833]$. All outcomes are retained in Appendix B; these same-seed development worlds are not an unused-condition causal comparison.

Throughput redesign replaces hash-selected plan eligibility with actual deadline inspection and annual amendments. In the first revised world, a researcher whose project would finish in year 2 has a real amended plan finishing in year 1 and an actual year 1 accepted-paper record. Five rather than four year 1 submissions/acceptances are recorded. A newly wired publication receipt route delivers earlier paper identities to the year 2 planning node, and resulting focus text reaches actual researcher requests.

Timing limits the mechanism claim. The extra year 1 acceptance occurs before any prior publication receipt exists; it cannot be attributed to that delayed feedback. Source, scope, focus and wiring change jointly, without component ablation. What is established is an exercised plan/record/information path, not the independent benefit of the added feedback loop.

Table 4: **All nested-study outcomes.** These are same-development-seed worlds, not independent scientific seeds. Q is original peer evidence. The post-update draw is not omitted when it worsens.

Optimized loss	Stage	A	Q	R	T	E
Throughput	G0 draw 1	7	0	0	-7	0
	G0 draw 2	7	0	0	-7	0
	Updated G0	6	0	0	-6	0
	Free G1	10	0	0	-10	0
Reviewed production	G0 draw 1	6	20	6	-6	-17
	G0 draw 2	6	18.5	12	-6	-12.5
	Updated G0	6	12	4	-6	-10
	Free G1	6	21	6	-6	-18

The utility revision adds review-digest feedback and restores review. It accepts one paper with mean score $11/3$ after 18 reviews, yielding $3.25 - 11/3 + 0.05(18) = 0.4833$. The initial no-review/no-acceptance observations remain better at 0. A new mechanism can be technically real and worse under the declared objective. No quality or participation veto reverses that result.

7.2 An executed new mechanism that is not retained

The V12 case gives each objective three own-history source redesigns. All six worlds complete with 565 fresh replies. Throughput installs a production office, annual scheduler and all-accept decision source, producing 10 native plan amendments and 10 acceptances per world. Later generations retain the exact first institution; the loss path $[-10, -10, -10]$ is not repeated invention.

Utility begins with an evidence office and an original-review threshold above 3.25. Its G1 introduces a genuinely executed milestone-clinic node, typed amendment transport and two wires, and changes evidence-topic matching. It shortens one project and extends another to a two-year cycle; it is not simply a universal acceleration rule. The native world completes, but utility worsens from $-1/6$ to 0 with 15 reviews. G2 returns to the G0 incumbent, giving $[-1/6, 0, -1/6]$.

The G2 parent is the best G0, not chronological G1. A zero parent diff therefore correctly records retention, while the chronology still exposes the unsuccessful clinic. This case matters because it demonstrates actual mechanism/wiring freedom without selecting only successful-looking revisions. It does not establish which simultaneous edit caused the changed loss.

8 Nested free design and conditional distribution updating

8.1 One loop, two levels of freedom

The V13 case combines an externally authored institutional vocabulary with one within-vocabulary likelihood-ratio update and a fresh native followup. Each of two losses, $T = -A$ and $E = -Q + 0.5R$, receives exactly four scheduled worlds: two IID G0 design draws, one draw from saved updated G0 weights and one free G1 source redesign. All eight complete with 592 fresh scientist replies and no response reuse. Every world resets development seed 42. The external Codex author identity is self-reported, not authenticated local-Qwen authorship.

For two draws, the implemented LOO estimate is

$$\widehat{g} = \tfrac{1}{2}\left[(L_0 - L_1)s_0 + (L_1 - L_0)s_1\right], \qquad s_i = \nabla_\theta \log q_\theta(d_i \mid V), \quad \theta^+ = \theta - 0.1\widehat{g}. \tag{9}$$

The followup uses exactly that saved vocabulary/distribution and performs no extra update. G1 then sees only its own three preceding observations, source, failures and saved update; it may change algorithms or support. Removing all coordinates explicitly creates a point mass rather than silently reusing old weights. A completed update is not itself evidence of lower expected loss.

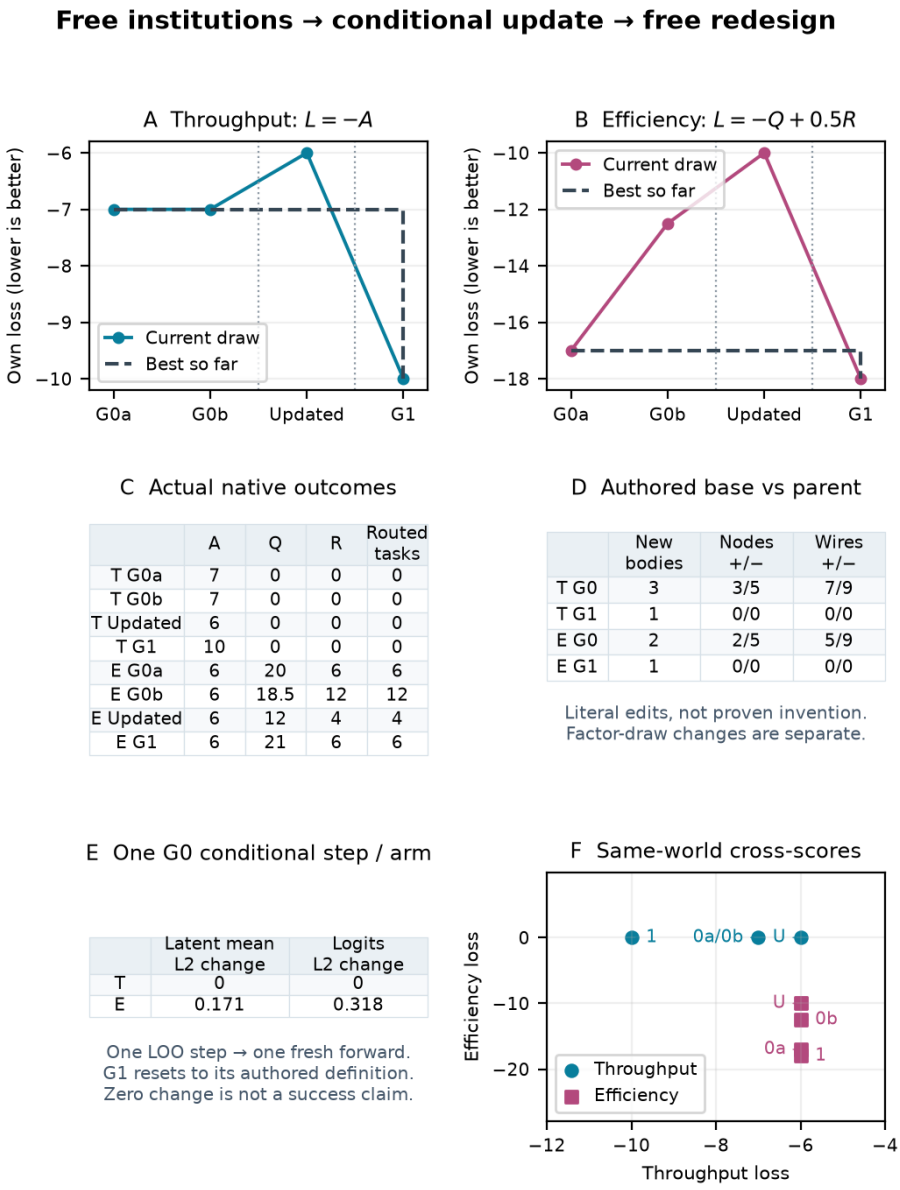
8.2 Distribution changes and native-tested redesign

Equal initial throughput losses give zero centered advantages and unchanged Gaussian mean. Its next native draw has loss -6, worse than -7; it is not reported as a successful gradient proposal.

Reviewed-production losses $[-17, -12.5]$ produce a nonzero update: the coverage latent mean changes from 1.5 to 1.671086 and reviewer-count logits from $(-1, 1, -1)$ to $(-1, 1.225, -1.225)$. The next native draw has loss -10, also worse than its prior best. Neither sample demonstrates expected improvement or estimator superiority, nor does one worsening sample refute the score identity.

The later throughput source revision replaces mixed-cycle hashing with universal annual scheduling while preserving native topic, focus and reason fields. The zero-review dispatcher and compatible all-accept rule remain. Observed loss improves from -7 to -10. Reviewed-production G1 replaces the coverage lottery with full coverage and one eligible reviewer per paper, ranked by its own historical scores. A read-only join rechecks the 22 original review records used in that ranking, including shrinkage toward a pooled mean. Task identities change, but actual Qwen reviews and the original parser still determine Q. Its observed loss improves from -17 to -18.

Initial free vocabularies already differ from their common parent; G1 changes one source body in each arm while graph sizes remain 17/25 and 16/23. Thus these are algorithmic edits without new G1 topology. Actual source/plan/wire/task receipts verify 40 source executions, 16 complete-plan effects, 16 routing events and 28 review records. These establish exercised native effects, not a causal decomposition of the gain. Favorable-reviewer selection exposes a proxy-seeking route, not improved scientific evaluation. No hidden merit veto removes it.



	A	Q	R	Routed tasks
T G0a	7	0	0	0
T G0b	7	0	0	0
T Updated	6	0	0	0
T G1	10	0	0	0
E G0a	6	20	6	6
E G0b	6	18.5	12	12
E Updated	6	12	4	4
E G1	6	21	6	6

	New bodies	Nodes +/−	Wires +/−
T G0	3	3/5	7/9
T G1	1	0/0	0/0
E G0	2	2/5	5/9
E G1	1	0/0	0/0

	Latent mean L2 change	Logits L2 change
T	0	0
E	0.171	0.318

Figure 3: **Complete native nested-design paths.** Current observations, best-so-far, actual $A/Q/R$, literal changes and conditional updates remain separate. Both gradient followups are worse than their prior best; both later external code revisions improve their observed first best. "Efficiency" here means the illustrative E proxy, not scientific efficiency or welfare. Cross-scores reuse the same checkpoints; they are not held-out evidence.

9 What native gradient experiments establish

9.1 A fitted response representation can fail unused conditions

The V9 case varies a charter's evidence weight θ, paired with originality weight $1-\theta$. The charter enters actual year 2 reviewer requests after the pre-intervention native history. Scientist weights and the original integer review parser stay fixed. The loss is $U = 3.25A - Q + 0.05R$.

Three calibration endpoints $\theta = (0.2, 0.5, 0.8)$ fit per-review raw-response operators. Calibration-loss RMSE chooses a declared relaxation temperature; the operator and proposal are frozen before unused coordinates 0.35/0.65 are read. At the center, a clipped negative-gradient step is followed by a native execution, even if it is a no-op. These are unused coordinates within one development history, not unused world seeds.

Six forwards use a verified 53-reply common prefix and then 31 fresh replies each: 504 logical responses comprise 318 replays and 186 new calls. They are not six independent societies. The fitted temperature 0.01 achieves calibration-loss RMSE1.16×10^{-5}, yet its center derivative is only 1.15×10^{-15}. The prescribed proposal is the unchanged 0.5, with native loss 0.9. The known lower calibration endpoint at 0.8 is not substituted for it.

At unused 0.65, predicted loss is approximately 0.8168 but native loss is 0.9. Across $0.35 \to 0.65$, native change is zero, the response-aware representation predicts -0.08395, and the frozen-response baseline predicts zero. The representation therefore fails the declared finite-path check and does not improve prediction here. This is a substantive boundary result: native ingress, identity joins, loss accounting and backward all execute, while fidelity and a useful native update remain unestablished.

9.2 Different update paths, the same observed incumbent

V10 compares represented gradients with local derivative-free interval refinement, charging both three shared historical calibration evaluations and three fresh attempts. The current coordinate starts at 0.5, but both already know the historical incumbent 0.8 with loss $49/60 \approx 0.8167$. The gradient method has raw responses/traces; the black-box method receives coordinates/scalars only. Equal evaluation count is not equal information.

All six fresh attempts complete, generating 186 new replies after their shared prefixes. The represented derivatives $(-.24092, -.09762, -1.24626)$ yield different coordinates, but native loss sequences coincide and neither beats the already known incumbent. This demonstrates real proposed updates and measured paths, not a gradient advantage. Appendix C retains the temperature diagnosis explaining why calibration fit alone was insufficient.

9.3 Conditional LR and corrected controls reach native execution

V11 uses explicit mean/minimum reviewer aggregation and a bounded cutoff driven by a Gaussian latent in one authored vocabulary. Eight fresh worlds compare a derivative-free feedback heuristic with score updating. Both arms observe $[0.7833333, 1.2, 1.2, 1.2]$ and retain their first incumbent; 742 fresh replies are recorded. Prior-frozen controls implement Eq. (5), but the final control update is saved after the last measured batch, not tested again within that budget.

A separately budgeted two-world V12 followup evaluates that exact saved distribution: minimum aggregation at cutoff 3.549508 gives $A = Q = 0, R = 21, U = 1.05$; mean aggregation at 3.080488 gives $A = 2, Q = 7, R = 21, U = 0.55$. These 180 fresh replies close the missing-forward question. The two outcomes do not establish lower expected loss, lower variance or a concurrent advantage over V11's historical outcomes. The distinction between a computed derivative, a saved proposal and a native-tested distribution is preserved across all these cases.

Table 5: Native-tested moving paths. Both methods know the better historical point before these proposals. A moving current iterate is not a new best.

Method	θ_1	θ_2	θ_3	Native loss sequence
Represented gradient	.6000	.6488	.7488	$(.9000, .8167, .8167)$
Black-box refinement	.6500	.7250	.7625	$(.9000, .8167, .8167)$

10 Matched local-designer comparison with stronger references

The preceding cases establish executable freedoms and bounded observations, but differ in source, author and feedback. They cannot answer whether a local code/graph designer outperforms parameter tuning. V2 therefore fixes one native source and compares proposal packages prospectively. It is complementary evidence, not a replacement for the earlier structural and gradient questions.

10.1 Objectives, domains and information boundaries

The study uses five researchers, fixed Qwen3-8B scientist weights, a four-year CS literature corpus, two-year development worlds and common world seed 42. The primary minimized losses are

$$T = -A, \qquad C = 4A - Q + 0.1R. \tag{10}$$

The earlier $E = -Q + 0.5R$ is diagnostic only, not a third search arm. The older $U = 3.25A - Q + 0.05R$ remains a different study's objective. No coefficient is retuned from V2 results. The common editable seed is not the untouched legacy institution; both are evaluated separately.

Each primary objective has five proposal policies:

1. LLM numeric: six declared coordinates, fixed source and topology.
2. LLM code/fixed graph: those coordinates and three existing Python bodies.
3. LLM code/graph: any valid specification within the native study contract.
4. Technical-only code/graph: the same broad domain without scientific-loss feedback, using the fixed initial parent and its own technical history.
5. Uniform numeric: independent samples of the same six coordinates, without a designer-model call.

The six coordinates are two existing tensor thresholds, bias, memory cost, router threshold and integer process count. Domains are checked against the initial anchor, not the previous parent, so gradual edits cannot escape a restricted arm. The code/graph domain permits structure change but cannot change scientist weights, corpus or measurement functions.

All LLM arms use the same approved local 35B designer profile: 16 charged actions, 8192 output tokens, 40960 context, temperature 0.6, thinking off, 10 s connect/900 s read allowances, without hidden transport retry or slot refill. Technical errors expose the same exception-class channel. Three fresh controller histories, each four generations per arm, give 120 candidate slots. Controller seeds 42/1031/2063 are not independent development worlds: every forward still uses world seed 42. Reference results and other arms' histories are not supplied to the designer. Fresh source/model/backend-bound interface controls precede dispatch; exact-copy equality is diagnostic, not a merit criterion.

Technical-only filtering hides losses, measurements, loss-selected parents, incumbent indicators and derived outcome metadata, not merely the scalar text. The intervention therefore includes parent selection, not only score display. Technical success remains visible and may itself be informative. This is a bounded information policy, not an adversarial isolation theorem about the entire operating environment.

10.2 References, selection and unused conditions

Nine fresh references share the same source/world: legacy, editable seed, rank/rate, all-accept, all-reject, score thresholds 3/4 and reviewer-free accept/reject. Earlier S1 controls are not reused as these fresh observations. These references prevent attributing a weak-seed improvement to successful invention when an obvious rule already achieves it.

Obj.	Proposal space/control	Meas.	G-null	X-null	M-null	Best	Ties
T	LLM numeric	0/12	12	0	0	0.000*	1
T	LLM code, fixed graph	9/12	1	2	0	-7.000	9
T	LLM code and graph	7/12	0	5	0	-7.000	6
T	Technical-only code/graph	7/12	0	5	0	-7.000	5
T	Uniform numeric	12/12	0	0	0	-7.000	8
C	LLM numeric	0/12	12	0	0	2.400*	1
C	LLM code, fixed graph	6/12	0	6	0	2.100	2
C	LLM code and graph	8/12	0	4	0	2.100	2
C	Technical-only code/graph	3/12	0	9	0	2.400*	3
C	Uniform numeric	12/12	0	0	0	2.400*	4

Table 6: **All scheduled development attempts.** Each row pools three controller histories with four proposals each, all at world seed 42. Meas.: available assigned losses; G/X-null: generation/execution failures; M-null: completed forward with unavailable assigned loss. Nulls are not penalties. Best includes the shared initial system; $*$ marks its selection, not an invention. Ties count all exact minimizers, including the initial reference when tied. Only display values are rounded; selection uses unrounded report values. Controller histories are not independent environment replications.

After all search slots, each objective/arm chooses from every completed candidate plus the initial system using only its declared loss. Every exact tie is retained; the earliest declared chronology supplies the representative. An initial selection is reported as no discovered improvement, not a generated replacement for failed proposals. The selection manifest is frozen before unused-condition evaluation and its results never return to search.

The 19 logical labels comprise 15 exact policies. Explicit sharing gives 90 physical jobs: each policy on five unused seeds 31001/31013/31019/31033/31039 for two years, plus a separate four-year diagnostic at 31001. Shared labels do not create independent samples; four years is not pooled with the five two-year observations. All 90 jobs finish normally.

10.3 All attempts, paths and exercised freedoms

Of 120 slots, 64 complete, 25 fail generation and 31 fail execution. All 24 uniform numeric proposals execute; all 24 LLM numeric proposals exhaust their action budget without a valid explicit submission. One fixed-graph code proposal also fails generation. The technical nulls are not numeric bad losses.

A bounded inspection confirms the correct six-coordinate contract in all 24 initial numeric-arm HTTP requests, with initial prompt counts 12145–12411, 40960 context and 8192 output settings without truncation. The general editor protocol and restricted session domain coexist. This refutes the narrow claim that the contract was absent; it does not establish clear salience, comprehension, later-call capacity or the cause of every failure. The end-to-end comparison therefore includes interface compliance and does not isolate intrinsic expressivity. Literal error categories appear in Appendix D.

Full-spec inspection joins all 64 complete candidates and all frozen ties. Forty candidates change selection source, and others change existing numeric settings. None changes declared node IDs/order, wiring, node ABI, initial state or schedules. Twenty change numeric process count, so unchanged graph metadata is not proof of unchanged runtime behavior. The code/graph label denotes permission, not a structural invention observed in this study.

Throughput repeatedly reaches $T = -7$, matching all-accept at development seed 42. Uniform numeric and technical-only proposals also reach it. For C, code minima 2.1 improve editable initial 2.4 but not reviewer-free reject 0. Initial policies are selected for both LLM numeric arms and for C technical-only and uniform arms. The whole path, rather than its most favorable history, provides no evidence of feedback or free-design superiority.

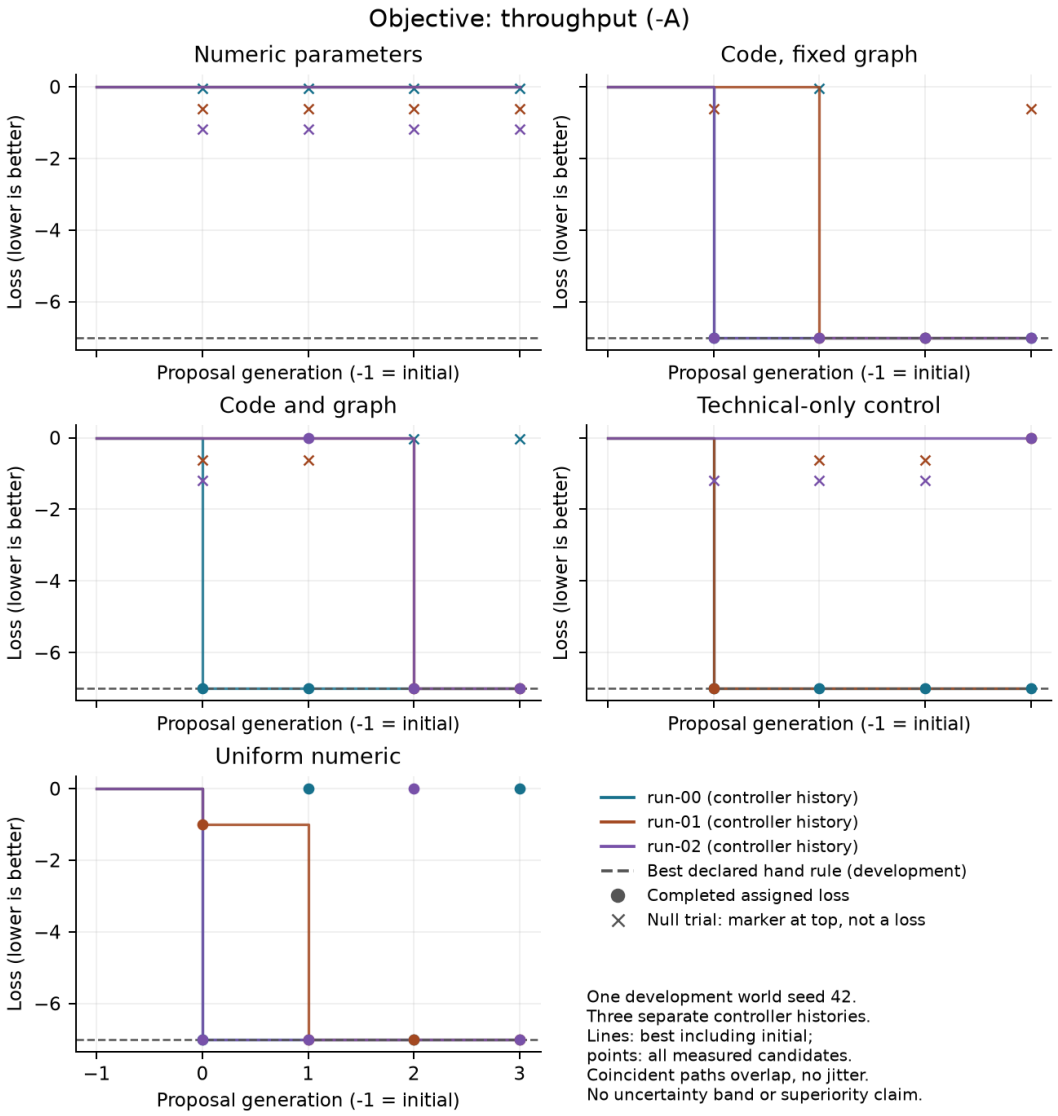


Figure 4: **All V2 throughput histories.** Every measured candidate and generation/runtime null remains. Lines show best including the shared initial observation; a line through a null does not make that attempt successful. The hand-rule line is a stronger reference, not an extra objective or veto. Three controller histories use the same development world seed.

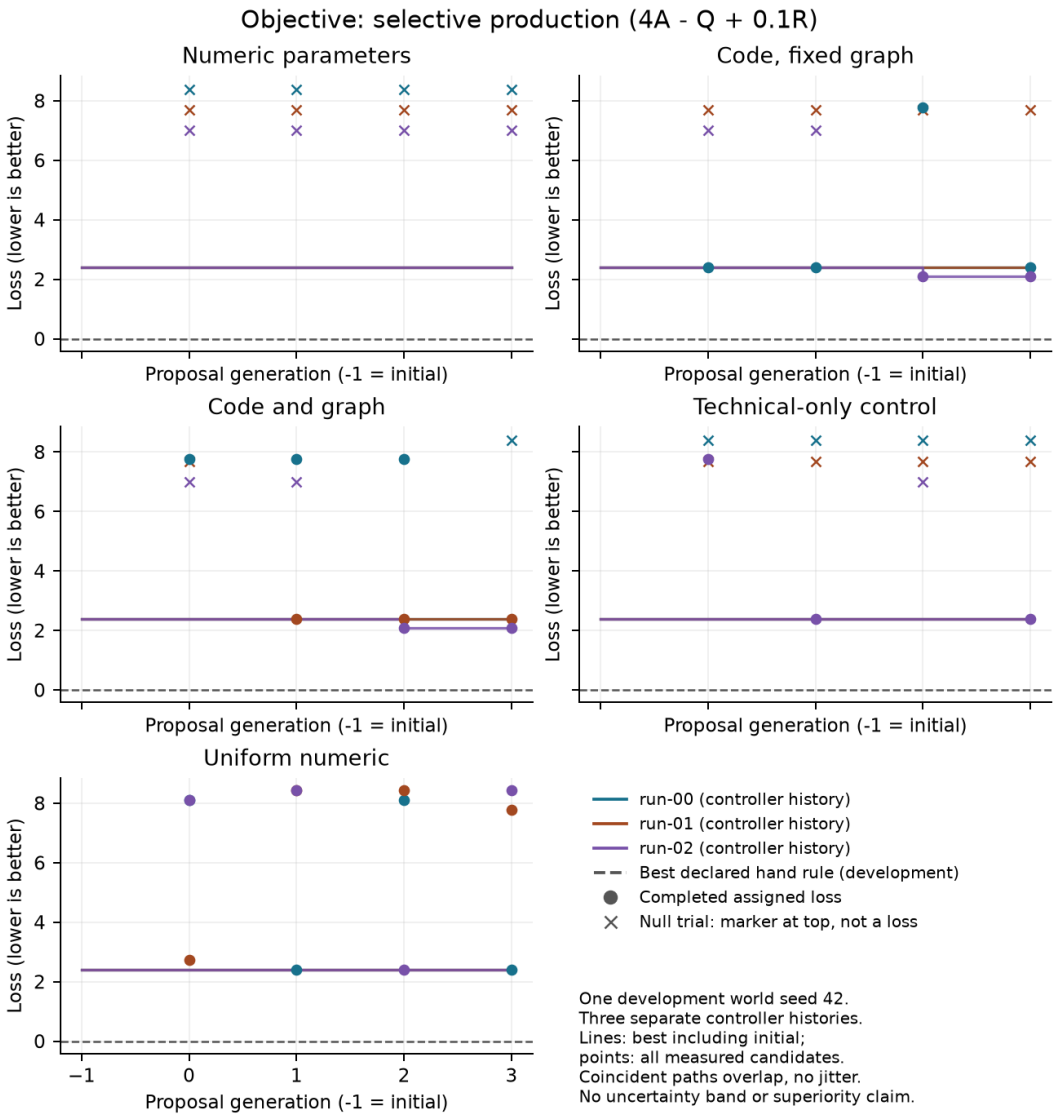


Figure 5: **All V2 review-cost-sensitive histories.** Code minima improve editable initial 2.4 to 2.1, but not reviewer-free reject 0. Uniform and technical-only controls are retained even when they tie or do not improve. Own-loss units are not comparable to the throughput panel's units.

10.4 Selected policies under common evaluation scores

Figure 6 reports every reference and selected label under common $T/C/E$. Executable throughput code/control and uniform selections average 6.2 acceptances over five unused two-year seeds, matching all-accept and score 3. Legacy averages 2.6 and the editable seed 0.2. Reviewer-free accept-all also averages 6.2, with $Q = R = 0$: equal throughput is not equal review-backed production or better science.

Both C code selections have mean $A = Q = 0, R = 19.2$, hence mean $C = 1.92$. Editable seed and review-based reject-all each average 2.58; reviewer-free reject remains 0. Diagnostic mean $E = 9.6$ is not favorable reviewed production. The different objective selects different observed acceptance/review behavior, but these data do not identify a novel optimal policy.

At the separate four-year seed, throughput executable selections and all-accept each produce 11 acceptances. Both C code selections have $A = Q = 0, R = 33$, giving $C = 3.3$, versus initial 6.9 and reviewer-free reject 0. Some equal-throughput policies differ under E. One fresh four-year realization is not another five-seed estimate or an isolated method effect.

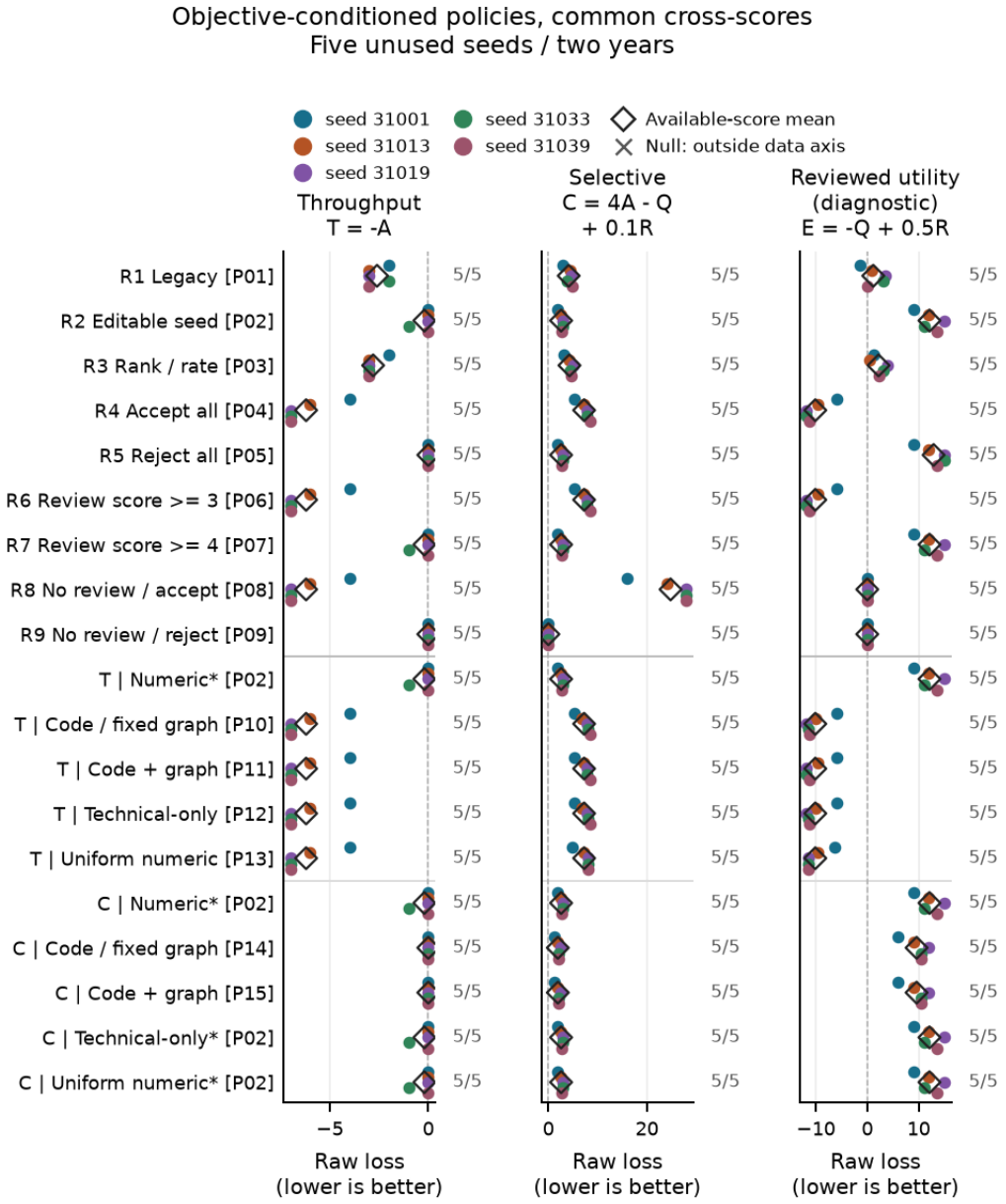


Figure 6: **All selection-fixed V2 policies on five unused two-year seeds.** Points are raw world observations; hollow diamonds are descriptive means, not confidence intervals. Separate loss scales are neither normalized nor averaged. P labels mark exact policy/job sharing, and stars mark initial selections. E is diagnostic, not an E-optimized experiment.

10.5 Mechanisms behind equal scalar losses

The development $C = 2.1$ selections have different components: code/graph gives $A = Q = 0, R = 21$, whereas fixed-graph gives $A = 1, Q = 4, R = 21$ and $4A - Q$ cancels. An authored final score 0 is not the original-review measurement Q. One fixed-graph history's G2/G3 sources have different bytes but equal Python ASTs and otherwise equal specifications. Both keep their actual costs and evaluations: AST deduplication was not a selection rule.

The selected C code/graph program rejects addressed papers and writes final score 0. The kernel applies that result to the actual paper, publication stores it in review history, and the researcher resubmission prompt prints it. Three audited year 2 requests contain the corresponding year 1 paper and score 0.00; each joins one completed response with an empty resubmission list. This traces an exercised program–record–input–action path. It does not isolate score 0 as the cause of fewer reviews: no alternate intervention world is simulated, and fresh LLM realizations are not perfectly coupled.

The distinction matters for future experiments. A generated parameter without a consumer is inert; a changed consumer can affect future behavior even if the graph is unchanged; equal losses need not imply equal institutions. Source, actual input/output receipts and downstream state are needed together. Unexercised empty-input or dummy-ID branches do not become validated because one trajectory completes.

10.6 Costs and outside metrics

The writer completes normally in approximately 21 h 56 min. Across the 194 physical response journals, 16,043 recorded scientist responses contain 17,021,423 total tokens. Proposal authoring records 1,387 requests and 1,385 final-response signals, not 1,385 successful candidate submissions. Its measured host model-call time is approximately 19,682 seconds, a component rather than GPU time or an additive end-to-end compute total. The 293 role loads have unknown native usage rather than zero usage. Appendix E separates all scopes.

Search uses 95 world directories: 64 complete and 31 partial runtime-null worlds; 25 generation-null slots dispatch no native world. All 64 completed development worlds and 90 evaluation jobs end with five active original researchers. That final snapshot gives no observed separation, not equal lifetime careers. Resources vary descriptively: two-year summed means are 616 for initial, 627 for C code and 632 for throughput uniform, with overlapping ranges. These endogenous resources are not money, welfare or a new optimization target.

11 A game-like observatory over the same research world

The interface is part of the research program, not a separate fictional simulator. An English pixel/cyberpunk city provides a spatial view of researchers and institutions; the Institution Foundry exposes source, mechanisms, wiring, conditional distributions, losses and all attempts. Canonical event/replay exports and verified checkpoints supply the numerical state. The visual language makes the ecosystem inspectable without making artistic choreography a measured research trajectory.

The useful interaction is to move between levels: inspect an objective's history, select an observation, read the literal institution, follow its graph and examine which effects reached native scientists. A local editable branch can start from an observed system; a draft is not executed merely because it is visible. Runtime validation, an explicit queued job and actual compute authorization remain separate. Likewise, installing a new institution into a continuing society is different from evaluating it in a reset world.

This establishes an evidence-reading and design interface, not a completed management game. No user study, learning benefit, enjoyable gameplay or human policy-design advantage is demonstrated. Those require separate evaluation. The longer-term game can make information circulation and institutional decisions interactive, but it must retain the same identity/state boundary rather than manufacture a more flattering world for display.

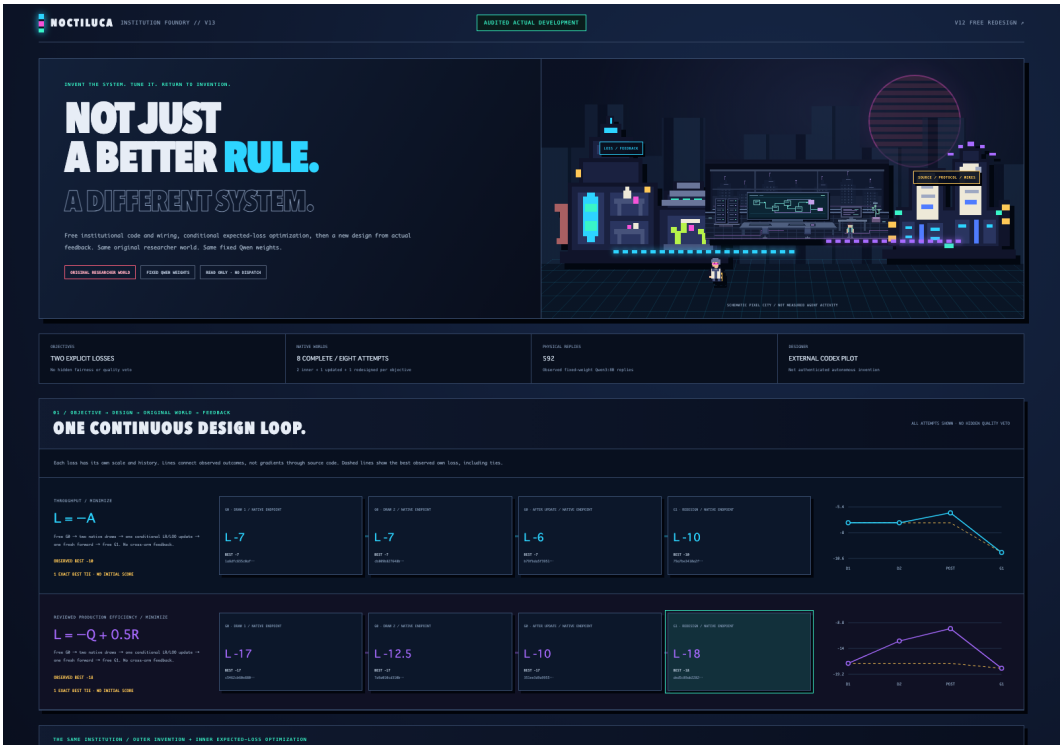


Figure 7: **Actual English Institution Foundry view of the V13 evidence.** Cards show the complete eight-world nested case, including worse post-update draws and external-designer provenance. Pixel city movement is schematic, not measured researcher motion. This historical capture is bound to V13, not a claim that the screen displays the later V2 study or that gameplay improves research understanding. Full-size captures remain in the artifact repository.

12 Before scaling: discriminating mechanisms from interface burden

The next stage is not simply more seeds. We first separate three tractable questions: whether a proposer can submit its intended design, whether an institution has an actual native consumer, and whether objective-conditioned revision changes native outcomes. Only the last is new optimization evidence. The following additions have distinct statuses; they are not folded into the closed V2 benchmark or its unchanged selection.

12.1 What the closed record already explains

A read-only analysis of all 120 V2 statuses retains 64 measured profiles and nine reference profiles reconstructed from their three affine loss scores. Those nine reconstructions are not new raw measurements. In the observed coefficient family $L_{\alpha,\beta} = \alpha A - Q + \beta R$, the original C coefficients $(4, 0.1)$ favor reviewer-free rejection over the observed high-Q profile. This follows from proxy geometry, not necessarily a weak optimizer. The map is only over observed profiles; it is not a bound or optimum for untried dynamic policies (Figure 8).

Likewise, the numeric failures are not an absence of submission attempts. Their 384 editor steps contain 47 finish rejections: 43 outside the declared numeric-only domain and four out-of-range. Of 228 reads, 166 target existing selection source. This describes a presentation/compliance confound and does not identify a single proven cognitive or transport cause. A new interface must be tested prospectively rather than repairing these failed slots.

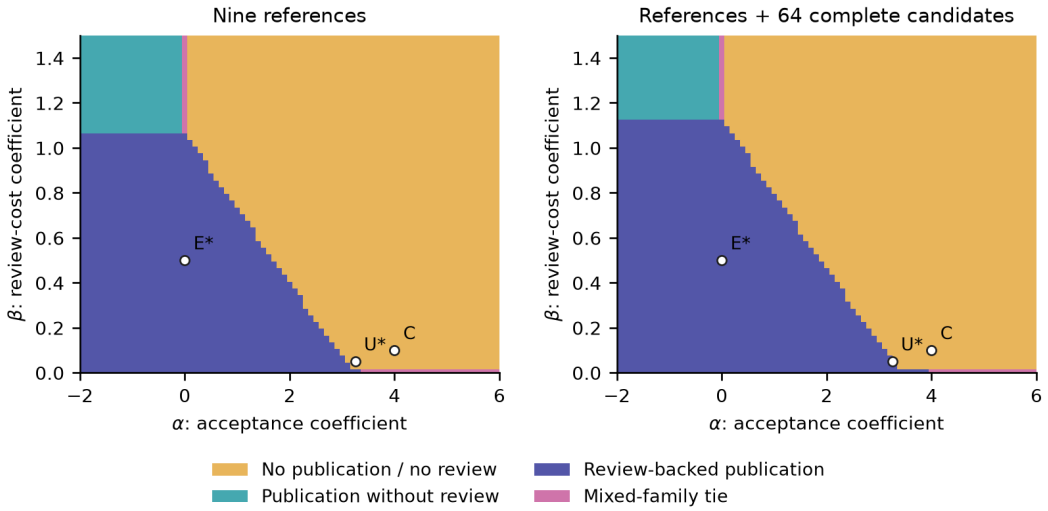


Figure 8: Post-hoc objective geometry over observed development profiles, not new objective-conditioned search. Left: nine reference-score inversions; right: those references plus 64 measured completed candidates. Colors identify observed outcome families, not general policy equivalence. The coefficient grid is a plotting choice, not additional samples or global optimality. C is the original V2 loss; E*/U* mark diagnostic formulas, not new searches.

Case	Prompt tokens	Output tokens	Literal action	Native loss
T	3196	285	Explicit parent retention	Not observed
E	3340	993	Different valid coordinates	Not observed

Table 7: Actual local 35B numerical readiness, two generation POSTs. Recorded HTTP call times are 6.972 and 14.056 seconds, excluding separate role loading and other overhead. Both submissions cold-replay. This is not a search comparison.

12.2 Literal numerical submission: a limited real-backend result

The new numerical interface asks for exactly six typed coordinate values and an uninterpreted rationale. Values compile literally to a patch under the same declared domain; there is no clipping, code repair or implicit baseline. A complete explicit parent assignment is valid retention. Finite semantic-call and whole-session wall budgets, durable exact wire bytes, raw responses and cold replay distinguish known invalid replies from unknown completion.

At pinned source `215311c`, two predeclared local 35B readiness cases both complete on their first generation call (Table 7). No native world is evaluated. In particular, the E rationale incorrectly conjectures that limiting delivered history records could directly reduce the review-count metric; preserving a valid literal submission does not validate that explanation. The result establishes interface execution for these two requests, not an optimized loss, general editor readiness or designer competence. Later runtime changes require a new source-specific certificate.

12.3 A native information-path intervention, not another toy world

A developer-authored history office now reads actual native papers, statuses and review summaries, with bounded records per researcher. M0 runs the producer without delivery; M1 delivers own history; M2 delivers society history; M3 delays own history an additional year. All retain the original scientific activities and remove the old example's direct project reassignment. M2 is not an exactly token-matched inert placebo. Original-world tests with scripted HTTP/corpus and real isolated code execution verify the information route and full pre-delivery payload equality under different host clocks. Actual first consumption depends on the next eligible research-direction call, not just message delivery; long projects can defer that call.

Seed	Condition	A	Q	R	$-A$	$-Q + 0.5R$	First input year
41001	M0	0	0	78	0	39	–
41001	M1	0	0	87	0	43.5	2
41001	M2	0	0	87	0	43.5	2
41001	M3	0	0	87	0	43.5	3
41011	M0	0	0	48	0	24	–
41011	M1	0	0	48	0	24	3
41011	M2	0	0	48	0	24	3
41011	M3	1	4	45	-1	18.5	3
41017	M0	0	0	42	0	21	–
41017	M1	0	0	42	0	21	3
41017	M2	0	0	42	0	21	3
41017	M3	0	0	42	0	21	3

Table 8: Completed native information comparison, three development seeds and four years per world. A is accepted decisions, Q their archival mean-review score sum and R review records. First input year is the first completed direction request containing the delivered history, not its delivery date. No-dossier M0 entries are absent, not missing observations. All five researchers remain active in each world. These are engineering proxies, not measured scientific truth or social value.

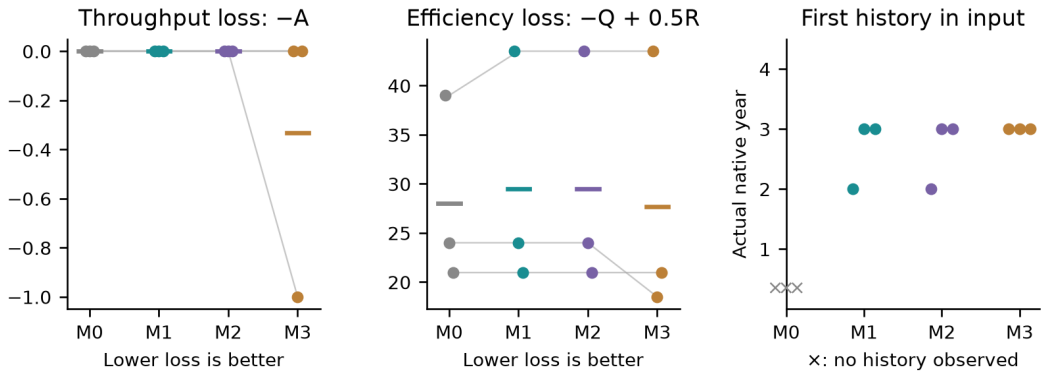


Figure 9: Actual closed information study. Each condition has three world observations; thin lines join the same declared seed across conditions, and horizontal bars show condition means, not confidence intervals. Right: actual first history-bearing direction inputs. Delayed information does not imply an extra year before the next eligible call in every seed. The small study is descriptive and does not isolate content from prompt length or future noise.

The actual finite study now closes all 12 four-year worlds: four conditions at seeds 41001, 41011 and 41017, with five fixed Qwen3-8B researchers. Two independent full audits using the unchanged original source `215311c` produce identical complete audit bytes. They rejoin raw requests, responses, code executions, checkpoints and independently recomputed measurements. All five first-direction payloads match across the four conditions within each seed. This does not couple all subsequent fresh response noise. Table 8 retains every world, rather than only the condition means.

M1 and M2 deliver 20 researcher histories per world, versus 15 for M3; the histories reach actual subsequent model inputs in all connected conditions. Nonetheless, 11 worlds have no accepted decision. A read-only replay of all 48 annual hard-selector traces reproduces their recorded outputs and memory using the original tensor compiler. In those 11 worlds, every observed review mean is below the selected effective threshold 4.0. The remaining M3 world at seed 41011 includes a score exactly 4.0, accepts once, and retains the resulting memory penalty (later effective threshold 4.01). Thus flat throughput is consistent with the fixed selector's exercised rule; it is not evidence that the information route is inert or that information cannot matter under another institution.

The mean T losses for M0–M3 are respectively $0, 0, 0, -1/3$; mean E losses are $28, 29.5, 29.5, 27\frac{2}{3}$. In M1/M2 the paired E changes from M0 are $+4.5, 0, 0$; in M3 they are $+4.5, -5.5, 0$. This variation does not establish a reliable delayed-information advantage. The study records 2,248 completed scientist responses, 2,185,747 reported input tokens and 202,556 output tokens; 6,938.423 seconds is the sum of recorded scientist-call intervals, not the end-to-end experiment duration. Cached prompt counts are a subset and are not added to input tokens. Corpus indexing, model loading, code execution and host overhead are outside that call-time component.

12.4 Closed objective-conditioned native search

The next study completes an objective-conditioned loop in the same original researcher world, rather than a simplified numerical society. It uses an unchanged executable source, `62388a7`, five fixed Qwen3-8B researchers, four annual cycles and the existing local 35B designer. Before scientific dispatch, its own two numerical and 28 staged-editor controls pass, including cold replay. This is interface evidence, not a guarantee that free proposals will execute. No post-hoc repair is applied to a received scientific proposal.

Finite comparison and fixed selection. There are two assigned losses, T and E, and four proposal methods: six literal numerical coordinates, generated source with fixed wiring, generated source and graph, and a uniform numerical comparator. Each arm consumes two proposal slots on development seed 41001. Generated proposals can use their own prior recorded result; the open-seed loss is included in final comparison but is not exposed before the first proposal. Six separately executed references include the open seed, legacy system, reviewed accept-all, no-review accept-all, no-review reject-all and an information-disconnected system. The production, calendar and accounting boundary is fixed for this pilot; institutional algorithms, entities, memory and routes remain editable. This is an explicit comparison domain, not an outcome veto or the global design ceiling. C is evaluated only as a diagnostic and does not select a proposal.

All 16 slots are consumed: 11 complete and five fail during execution, with no generation failures, replacement slots or imputed losses. Table 9 retains every slot; Figure 10 shows best-so-far paths including the common initial comparison. Selection freezes at 07:47:45 UTC on November 9, before evaluation starts at 07:47:57. Within each arm, the lowest available own loss wins, including the initial system; exact ties retain the earliest declared row. Three previously unused seeds (51001, 51011, 51017) then evaluate each selected policy and four references. Twelve labels times three seeds give 36 logical rows but only 27 physical jobs, because exactly identical policies share execution. All 27 complete at 11:29:22 UTC. Two full original-source audits subsequently match byte for byte, covering 1,859 raw files and all 49 physical worlds: six development references, 16 submitted candidates and 27 evaluations. Five of the 49 are the retained partial runtime failures, not completed worlds.

Table 9: All 16 development slots. Entries are assigned proxy losses, lower is better; F denotes a runtime failure with unavailable loss. Selection also includes the common initial (T: 0, E: 43.5), hence an initial can win.

Objective	Method	G0	G1	Fixed selection
T	Numerical	-14	-14	G0 (earliest tie)
T	Code / fixed graph	-14	F	G0
T	Code and graph	F	F	Initial
T	Uniform	0	-2	G1
E	Numerical	29.333	-22.667	G1
E	Code / fixed graph	F	F	Initial
E	Code and graph	27	43.5	G0
E	Uniform	43.5	34.667	G1

What the optimizer actually changes. E/numerical G0 has A=2, Q=$6\frac{2}{3}$ and R=72, hence E=$29\frac{1}{3}$. The next request contains that observed loss. Its submitted G1 changes the memory-router threshold from 1 to 0 while retaining the other five literal coordinates; A=14, Q=$43\frac{2}{3}$ and R=42 give E=$-22\frac{2}{3}$. The router's actual native consumer is operative: the G0 rule can reject after its accumulated-memory penalty, whereas G1 accepts at the changed threshold. This establishes an executed feedback-history

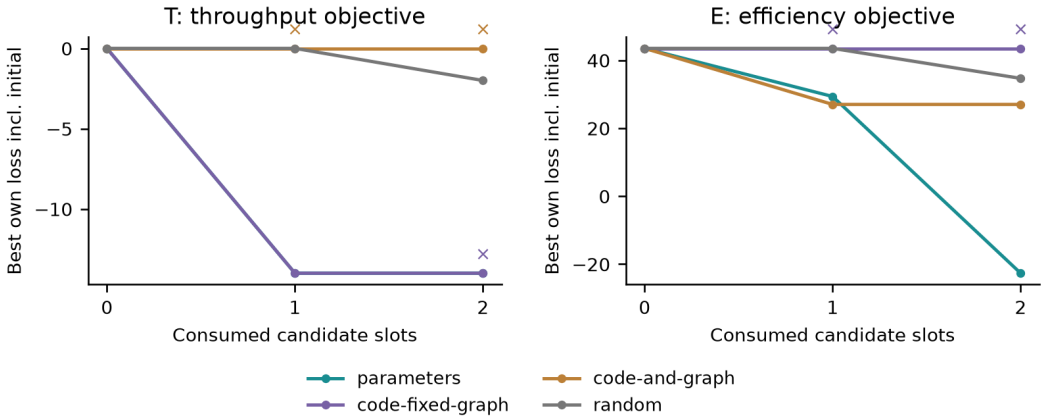


Figure 10: Actual two-slot optimization paths, not additional simulations or fitted response surfaces. Slot 0 is the common initial included for final comparison; subsequent points are the best eligible loss including it. Crosses mark consumed failed slots and do not assign those proposals a loss. The T numerical and fixed-graph code paths coincide exactly. The E numerical G1 decrease is an observed development result, not established optimizer superiority.

update, not that the author's explanation is correct or that feedback caused the improvement relative to another search method.

Free-source proposals do not simply select from built-in rules. The completed T/fixed-graph G0 writes an always-accept decision rule; E/code-and-graph G0 writes always-reject with a literal zero final score; its G1 uses the original review-score average with a threshold above 5 and changes addressed thresholds and bias. Those free-code values need not lie in the six-coordinate numerical box. Nevertheless, all 16 submissions retain the initial 21 mechanisms, 29 connections and schedules: eight change source, none changes literal wiring or scheduling. Allowed structural freedom therefore does not establish realized topology invention. Rejecting after review does not itself avoid the earlier review operation, contrary to one generated comment; source rationale is not a validated mechanism explanation.

The five failures are substantive execution records: an unwrapped world payload, missing decision fields, `score` instead of `final_score`, a float treated as a sequence, and incomplete submitted-paper identity coverage. Readiness did not prevent these free-design mistakes. They remain unavailable-loss outcomes and consume their declared slots. A separate later output-envelope advisory is tested but is not backported into this experiment, and it does not automatically repair generated source.

Selection-fixed common evaluation. Table 10 and Figure 11 report all selected labels and four references under the same three losses. The E/numerical policy averages -17.333 on E versus -16.889 for reviewed accept-all, a paired mean difference of $-4/9$. The three differences are $-1, -2/3, +1/3$: the sign varies, and this small study does not establish superiority. Their accepted counts and review counts match; the difference is in endogenous review-score sums, not verified scientific value or an isolated causal information effect. T/numerical, T/fixed-graph and reviewed accept-all have identical observed score vectors but different literal policies and separately executed jobs. By contrast, the two selected uniform labels really do share jobs; the initial reference is also shared with the selected E/fixed-graph and T/code-and-graph labels. Neither equal outcomes nor duplicate labels create independent evidence.

Measured components, costs and reproducibility. Across the 49 physical forwards, scientist journals contain 8,025 completed responses, 8,453,441 reported input and 723,376 output tokens; 1,416,327 cached prompt tokens are a subset of input, not additional work. All five partial failures retain their charged replies. Designer proposal records include six numerical calls (25,651 input, 8,461 output tokens) and 104 staged-editor calls (2,076,602 input, 39,372 output tokens). Readiness is separate: three numerical calls and 28 editor controls. Numeric caller intervals total 201.790 seconds; editor transport intervals total 1,533.432 seconds, and role loads are kept separate. These

Table 10: All selection-fixed labels: three-seed means of native components and common losses (all 3/3 available). A counts accepted decisions; Q sums endogenous accepted peer scores; R counts review records. Q=0 without review is a measurement convention, not evidence of zero scientific quality.

Label	A	Q	R	T	E	C
Ref: open seed	0	0	67	0	33.500	6.700
Ref: reviewed accept-all	10.667	32.889	32	−10.667	−16.889	12.978
Ref: no-review accept-all	10.667	0	0	−10.667	0	42.667
Ref: no-review reject-all	0	0	0	0	0	0
T: numerical	10.667	32.889	32	−10.667	−16.889	12.978
T: code / fixed graph	10.667	32.889	32	−10.667	−16.889	12.978
T: code and graph	0	0	67	0	33.500	6.700
T: uniform	1.667	6.111	61	−1.667	24.389	6.656
E: numerical	10.667	33.333	32	−10.667	−17.333	12.533
E: code / fixed graph	0	0	67	0	33.500	6.700
E: code and graph	0	0	35	0	17.500	3.500
E: uniform	1.667	6.111	61	−1.667	24.389	6.656

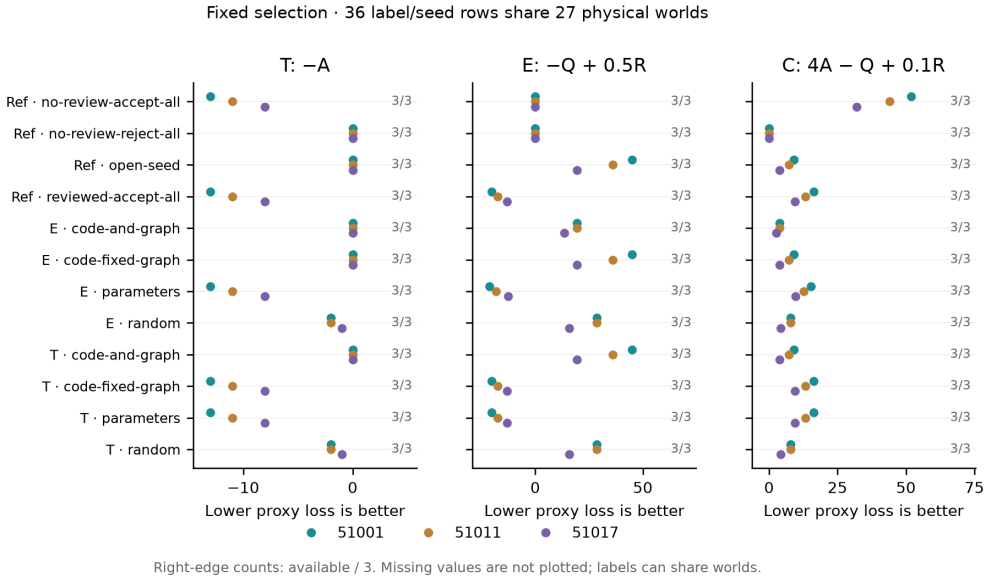


Figure 11: Every selection-fixed seed score under T, E and diagnostic C. Colors identify seeds, not independent repeats of shared policies. The right-edge counts retain score availability. Strong reference rules expose proxy geometry: unreviewed acceptance has no recorded Q or R, whereas reviewed acceptance can make E negative. C was not searched in this pilot; its apparent reference ordering is not an optimized-C result.

component counters and overlapping timing aliases are not added into project FLOPs, energy, money or total compute. The scientific search and evaluation stages occupy about 7 hours 18 minutes end to end on this local binding, excluding earlier readiness, audits and all software-development work.

Retained data contain every executable submitted system, physical-world join, raw component status, exact selection tie, shared-job identity and 96 paired reference comparisons. Independent numerical rebuilding recomputes their cross-scores and regenerates both figures byte-identically; a separate literal change projection records actual source/interface/wiring differences. Full raw/source auditing, retained-data numerical rebuilding and fresh inference remain different reproducibility levels. The English Foundry now inspects this closed study's actual development code, ports, connections, year snapshots and failures; it neither turns development records into unused-seed evidence nor submits another simulation.

12.5 Verifiable activities and continuity at the same boundary

An opt-in claim-verification consumer checks an institution's literal assertion about counts of actual native accept/reject decisions or review records over a completed interval. It reads the simulator's archive independently of the candidate's observation or proposed verdict. A correction has a new identity, retains the false predecessor and binds its verification hash. Retained state is checked against actual evidence, rather than accepting a designer-seeded `matches=true` flag. Scientific text remains literal; a source-owned operational review-time field is excluded only from the count-evidence identity.

Original-world/Docker software tests connect authored false assertion, independent verification, correction and actual next-year researcher input. This is internally checkable empirical work about the simulated society, *not* verification of the truth of retrieved scientific papers or a new scientific-value objective. No default phase or hidden merit gate is added.

Additional scripted native tests place the existing antithetic score estimator on the current information institution, execute its updated parameter in a fresh native forward and retain a no-gradient control. A same-society test changes the history mechanism while preserving researchers, papers, resources, mechanism memory, RNG state and pending delayed deliveries. These establish current-boundary connectivity, not new live-model gradients, practical method advantage, autonomous invention or recursive improvement. The English pre-scale observatory presents only timestamped journal observations, explicitly distinguishing a start without a terminal record from a verified completion.

13 Authority, authoring and measurement diagnostics

The closed comparisons leave several explanations for limited structural search: the objective may favor a simple rule, an operation may be unavailable in the declared domain, the author may misunderstand its execution point, or the submitted function may fail on native inputs. The following diagnostics separate these possibilities without adding candidate losses to the earlier studies or pooling different sources as method replications.

13.1 Available operations and execution time

In the information institution, costs execute at phase 80, review at 90, feature extraction at 100, decision code at 130 and publication at 140. A late decision function can affect subsequent resubmission and future work, but cannot remove the current year's completed review operation.

An existing typed review-routing capability binds institution-authored assignments to the actual submitted papers, eligible reviewers and exact input-batch hash. However, the earlier fixed-production domain rejects that implementation as part of its protected production boundary. A prospective `fixed-native-production-review-routing-v2` scope makes this additional authority explicit while retaining the original planning, production and accounting semantics. It does not redefine the old domain retrospectively. Its pre-review snapshot/producer must also precede the actual review stage.

These facts identify what an edit *could* change, not why a previous designer omitted an unattempted edit. A later source/domain change cannot by itself establish that earlier non-rewiring was caused

by a missing capability. Nor does extra authority establish that a graph representation is necessary: an equally authorized policy program with the same callbacks may express the same computation. Native connection checks use scripted researcher replies and real isolated execution; they are not autonomous invention results.

13.2 Eight closed authoring sessions and fixed-input probes

At source `a2fb13c`, a finite authoring assay crosses T/E, initial versus one recorded own-history request, and access versus no access to an optional execution-order reference file. These historical requests are new authoring inputs, not fresh world observations. Each local 35B session has 16 charged actions, 8192 output tokens, 40960 context and 1200 seconds. Both condition-specific readiness sessions pass 29 controls at their own binding.

All eight sessions finish with static-valid explicit submissions, reproduced by two byte-identical original-source cold audits. Seven change selection source; one changes four upstream router coordinates. None changes node, connection or schedule metadata. Five intermediate actions are rejected within the fixed budget; no slot is refilled and no received code is repaired.

The explanatory file is not read in any of the four access-enabled sessions; capability-detail files are also unread. Thus this is a comparison of offered file access, not evidence that understood execution-order information was ineffective. Its request-bound decode seeds also differ between conditions, so it is not a common-random-number treatment estimate.

Separate network-disabled, read-only Docker probes use four saved native selection inputs containing 2, 7, 9 and 11 papers. The original selector first replays exactly on all four. Each candidate selector is then tested on those same frames, resetting to the recorded input/state rather than carrying the candidate's state through years. Table 11 retains all eight candidates. No probe commits a native decision, runs an upstream mechanism or evaluates a new scientific world.

Table 11: All authoring submissions and partial selector probes. H0 is the initial request; H1 contains a recorded history. File means optional access, not actual reading. Rows are not native-world validity or loss results.

Loss	History	File	Edit	Fixed-input result
T	H0	No	Source	Accepts all 29 addressed papers
T	H0	Yes	Source	Accepts all 29 addressed papers
E	H0	Yes	Source	`KeyError: world` on all 4 inputs
E	H0	No	Numeric	Unchanged selector rejects 29; edited upstream is not executed
T	H1	Yes	Source	Empty-cycle output on all 4 nonempty inputs; identity coverage fails
T	H1	No	Source	Accepts all 29 addressed papers
E	H1	No	Source	Accepts one per input, rejects the remaining 25
E	H1	Yes	Source	Accepts all 29 addressed papers

Of 32 candidate function calls, 28 return and 24 cover every addressed paper. These are component checks, not six valid whole institutions. In particular, the numerical candidate's modified upstream is not exercised by a selector-only probe. Runtime exceptions and empty outputs show why successful static submission cannot certify arbitrary code's behavior on actual inputs.

Generated explanations also differ from operative effects. One plan proposes stricter late selection to avoid earlier review work. Another source raises `final_score` to at least 10, but the E loss reads original reviewer scores, not that authored field. A later source comment correctly acknowledges the late selector's lack of direct review-count control. These are literal observations about particular programs, not a universal diagnosis of the model's reasoning or evidence of measured loss improvement.

Authoring uses 110 chat POSTs with 2,225,666 reported input and 23,017 output tokens; readiness uses 58 POSTs with 618,249 input and 3,574 output tokens. The interval from first start to last chat response is 1861.22 seconds. Ten separate empty-prompt loads complete with unreported token counts, not observed zero usage. The assay and probes make no researcher-model calls and produce

no new candidate losses. Inline order context and more detailed ABI cards are later prospective interfaces, not retroactive changes to this assay.

13.3 A fixed measurement function can have endogenous observations

The measurement code is immutable within a study, but the observations it receives need not be institution-independent. For the original-review proxy,

$$Q(\tau) = \sum_{\substack{p \in \mathcal{A}(\tau) \\ |\mathcal{R}_p(\tau)|>0}} \frac{1}{|\mathcal{R}_p(\tau)|} \sum_{j \in \mathcal{R}_p(\tau)} y_{pj}, \tag{11}$$

with zero contribution for an accepted decision without reviews, as defined by the adapter. Both the accepted set $\mathcal{A}$ and reviewer set $\mathcal{R}_p$ can depend on the institution. Excluding an authored final score from Q therefore does not turn Q into independent scientific quality.

A read-only analysis revisits the 49 physical worlds of the closed `62388a7` T/E pilot. It reconstructs A/Q/R from the 44 completed four-year checkpoints, retains five partial failures without imputing complete outcomes, and checks the full 1859-file inventory against the original-source full audit. These records contain 217 reviewed and 46 unreviewed accepted decisions, not 263 independent papers or new world replications.

For each already observed review panel, we enumerate mean-score readouts of all subsets of size k. Accepted papers and actual review work stay fixed. A uniformly selected subset has the same expected mean as the original panel; the minimum and maximum are sensitivity endpoints and can require hindsight. The analysis neither simulates choosing those reviewers in advance nor subtracts the cost of reviews that were actually performed.

Twenty-four completed worlds have a nonzero one-score Q-readout envelope. Figure 12 shows all three old evaluation seeds for the E/numerical selection and reviewed accept-all. Within each row, the recorded accepted set is fixed; artifacts and responses need not match between rows. At seed 51001, for example, original Q values are 39.667 versus 40.667, with one-score readout ranges 39–41 versus 39–43. These are not confidence intervals, independent quality estimates or an explanation of the observed between-policy difference. They do not invalidate that measured difference.

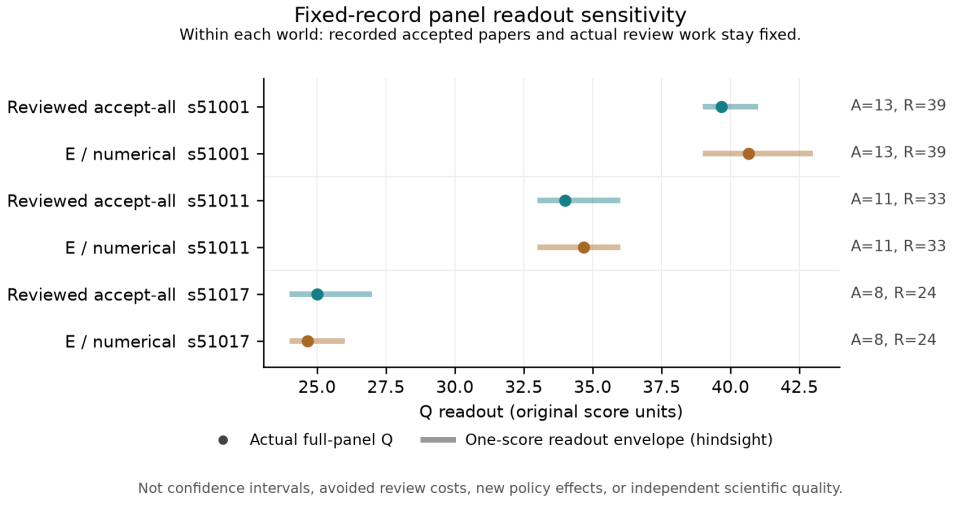


Figure 12: **Recorded panel readout sensitivity.** Dots are actual full-panel Q; bars reuse one score per already reviewed accepted decision, with hindsight endpoints. Actual A and R remain unchanged. This is not a reviewer-routing intervention, reduced review cost, new optimization or scientific quality. Both original-source analyses and both PNG/SVG builds reproduce identically.

The practical implication is to name the optimization target precisely. T/E remain legitimate declared proxies; this diagnostic adds no new loss, fairness condition or adoption veto. A separate blind evaluator could measure another fixed-panel opinion, but would not itself verify scientific truth. Claims

about knowledge gain would require an independently justified measurement or actual verification activity within an extension of the native researcher world. The later routing-authority campaign and its stopped outcome are treated separately in Section 14, rather than pooled with this fixed-record diagnostic.

14 Native code execution and the support of the objectives

Permitting arbitrary code is not the same as obtaining a working institution, and obtaining a working institution is not the same as answering an important optimization question. This section separates these issues in a later native campaign and a read-only analysis of the earlier completed T/E cohort. The loss definitions remain unchanged; no participation, novelty or fairness condition is added to the optimizer's acceptance criteria.

14.1 A stopped routing-authority campaign

The new campaign retains the original five author-capable agents, fixed Qwen3-8B researcher weights, the other native agents, the SciEvo literature and four-year Simulation. It crosses T/E with numerical, fixed-graph source and code-and-graph design. Two proposals per arm and eight references are declared in advance. The references include original reviewed accept-all and typed one- and two-reviewer accept-all institutions, giving simple controls access to the same new review-routing authority as generated institutions. These controls change actual review work, not a retrospective subsampling of already paid reviews.

All eight reference worlds complete. Of twelve declared proposal slots, eleven are submitted: five complete native execution and six fail technically. The final numerical T proposal is never submitted because its required fresh interface calibration fails on the E control. That response is a final HTTP 200 envelope with `done=true`, a length termination and 8192 reported output tokens, but no complete literal submission. The campaign stops at that point; its unsubmitted slot is neither refilled nor assigned a loss. Table 12 retains all twelve slots. Two cold stopped-prefix audits agree on the nineteen physical worlds and their original responses, checkpoints and measurements. This audit scope verifies a stopped experiment, not normal completion of its planned search and evaluation.

Table 12: All declared slots in the stopped native campaign. Numbers are development losses, not unused-seed results. Technical nulls have no imputed loss; generation labels index each arm, not the global dispatch sequence.

Objective	Allowed design	Generation 0	Generation 1
T	Numerical	-7	Not submitted
E	Numerical	-11.167	-11.167
T	Fixed-graph code	-2	Float used as index
E	Fixed-graph code	Missing world envelope	-11.167
T	Code and graph	Missing feature key	Missing world envelope
E	Code and graph	Float used as index	Float used as index

Seven of the eight code-permitted submissions change source, and the remaining one changes numerical router coordinates. None changes connections, node membership or scheduling. The completed fixed-graph T submission is the numerical-only edit; the completed fixed-graph E submission changes source. Therefore neither the method label nor static acceptance establishes that structural freedom was exercised successfully. The six execution failures are three float-index errors, two missing output envelopes and one missing feature-key error. Their literal programs remain unmodified.

The nineteen worlds produce 1796 completed scientist responses, with 1,615,207 reported input and 181,308 output tokens. Separate auxiliary records contain 256 editor-readiness, 112 staged-proposal, 13 numerical-readiness and three numerical-proposal intents, plus 25 designer and nineteen scientist loads. Unreported usage on load calls is unknown, not zero. Readiness and proposal costs are not silently excluded because the campaign stopped.

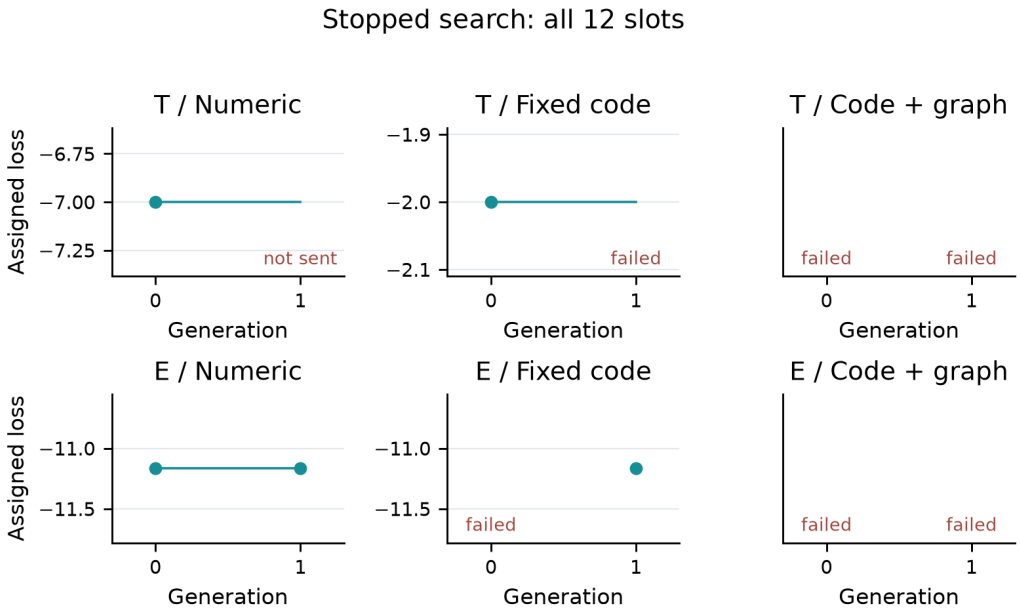


Figure 13: **Every original proposal slot.** Dots are completed candidates' assigned development loss; lines retain the best completed submitted value. Failed calls and the final unsubmitted slot have no numerical loss. The initial policy is not a fallback. These are two-step histories, not repeated independent searches or unused-seed estimates.

14.2 Separate evaluation of the available submissions

The stopped campaign has four arms with at least one completed submission; the two code-and-graph arms have none. A separately frozen study selects the minimum assigned development loss within each available arm, preserves all ties, and uses the earliest tied submission. It then evaluates those fixed policies and six references at seeds 62001, 62011 and 62017, whose prior use is checked before dispatch. No new designer proposal is generated. Thirty physical jobs are declared without semantic deduplication. An unavailable arm has no policy and no score; it does not inherit the seed or a hand reference.

This study evaluates policies available after a stopped prefix. It cannot complete the missing proposal, establish an equal-budget method comparison or turn three evaluation seeds into independent design histories. All thirty physical worlds now complete; two original-source cold audits and two cost projections agree exactly. Table 13 and Figure 14 retain all selected-policy labels, including both unavailable arms. The new development loop described below does not receive these unused-seed losses.

The T numerical selection ties every accept-all reference's throughput: A is 6, 8 and 17 on the three seeds. The E numerical selection's paired E difference from reviewed accept-all is $+0.667$, -0.667 and -2.000. Its mean advantage is therefore 0.667 loss units, not a stable gain established by three seeds. Against one-reviewer accept-all, the same differences are $+8.667$, $+7.000$ and $+14.333$, all worse. The E fixed-code selection matches reviewed accept-all's measured A/Q/R on each seed, whereas the T fixed-code selection matches the initial seed. Equal summary measurements do not prove source equivalence or identical complete trajectories. Thirty actual jobs remain thirty jobs, not additional replications from matching labels.

The one-reviewer reference changes actual work: its mean R is 10.333 versus 31 for reviewed accept-all, while mean Q is 33.333 versus 33. This is not removing already performed reviews or establishing better scientific truth. Routing controls use the full domain's authority; their advantage over a restricted numerical arm is not an isolated effect of graph syntax or search. The selected E numerical policy's different Q at the same A/R does not identify the cause of its small score difference. The

Table 13: Separate fixed-submission evaluation: three-seed means. A counts accepted decision records, Q sums their original mean review scores and R counts actual reviews. Lower T/E is better. No policy exists for the two unavailable arms; their entries are not zero or baseline values.

Policy	A	Q	R	T	E
Initial seed	.667	2.667	65	−.667	29.833
Legacy	5.333	18.778	53	−5.333	7.722
Reviewed / all	10.333	33	31	−10.333	−17.500
No review / all	10.333	0	0	−10.333	0
One reviewer / all	10.333	33.333	10.333	−10.333	−28.167
Two reviewers / all	10.333	32.833	20.667	−10.333	−22.500
T / numeric	10.333	33	31	−10.333	−17.500
E / numeric	10.333	33.667	31	−10.333	−18.167
T / fixed code	.667	2.667	65	−.667	29.833
E / fixed code	10.333	33	31	−10.333	−17.500
T / code and graph	–	–	–	–	–
E / code and graph	–	–	–	–	–

unoptimized selective cross-score remains available in the retained report, not another optimized arm.

This separate evaluation takes 16,496.75 seconds (about 4 h 35 min), including 4590 completed scientist responses with 4,923,237 reported input and 426,066 output tokens. The 784,809 cached input tokens are a subset, not additional input; thirty explicit scientist loads have unknown token usage. No new designer proposal is generated. Two portable reconstructions also reproduce the report, CSV and original four PNG/SVG files byte for byte, without another raw audit, model call or authored-code execution. A reporting-only join error is preserved as an earlier stopped analysis: arm-grouped candidate records must be joined by identity to dispatch-ordered worlds, not zipped by position.

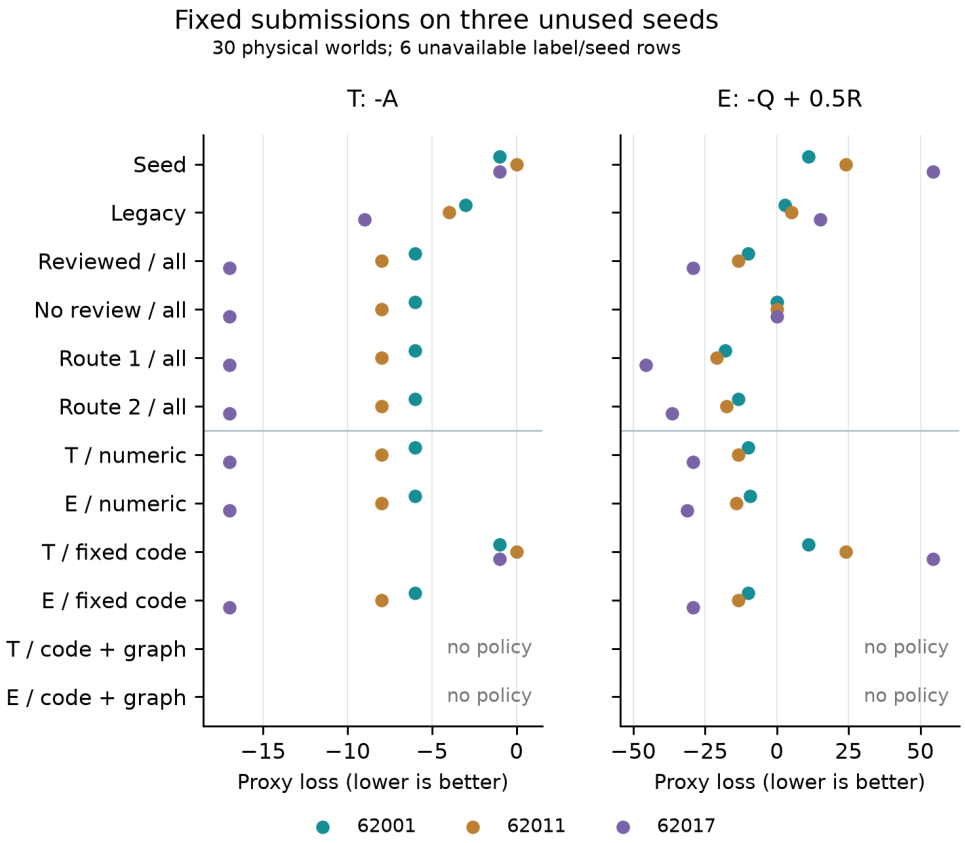


Figure 14: **Fixed submissions on unused seeds.** Each colored dot is an actual world at seed 62001, 62011 or 62017, not a confidence interval. The upper six rows are references, the lower six selected-arm labels. "All" means accept all; Route 1/2 assigns one/two actual reviewers. Both graph arms remain unavailable. No selected policy beats the newly authorized simple controls on its assigned objective in these observations.

14.3 Execution feedback without host repair

The execution failures motivate an optional component-testing interface, not an automatic repair of received programs. A developer bank binds 24 literal native inputs to three earlier development references, two components and four saved years. A proposed code artifact can be tested in the existing network-disabled container; the input, received code, returned value or exact exception, timeout and charged attempt are retained. Component tests do not replace native evaluation or define another loss.

Three real-container diagnostics illustrate both utility and limited support. On the same saved year-one input, the baseline reproduces its recorded output and a retained missing-envelope program fails. A retained float-index program returns because that input's paper batch is empty: a passing component test does not show that the program handles a later nonempty batch. Multi-year coverage is therefore useful, but still cannot certify a whole trajectory whose inputs change under a new institution.

A prospective post-edit mode connects literal edits to a fixed suite of four saved yearly inputs, with at most eight component attempts per proposal. Errors and compact results enter the next designer prompt; the complete records remain archived. A failed component test neither vetoes submission nor causes the host to wrap, revert or replace the generated source. Edits after the component allowance remain permitted and explicitly untested. Scripted full-path tests connect calibration, the public certificate API, the production role factory, the next feedback-bearing request and cold replay.

They show software connectivity, not that the actual designer can repair its code or improve a native objective.

A separate finite development campaign now closes all four own-history proposal slots and seven physical native worlds: three complete references, two complete E candidates and two partial failed T candidates. Its fresh real-mode calibration passes all 28 controls. Both E proposals obtain -11.1667, tying reviewed accept-all and losing to one-reviewer accept-all (-18.5). T has no completed submitted policy and receives no fallback. The reused development seed 61001 is not held out. Two original-source audits, cost projections and seven-file report exports agree exactly.

All four submissions edit only selection source, with no node, connection or schedule changes. Each session completes 14 authoring steps. Of 32 actual component attempts, 28 return and four raise. Same-fixture later-code comparisons show four returned-to-raised and twelve returned-to-returned transitions, with no raised-to-returned transition. Three final submitted artifacts are untested after the fixed allowance is exhausted; eight earlier returned probes do not certify a later code version.

The first T submission lacks an output payload envelope. Its next request contains the observed errors, but a later read removes them from the final request. A separate prospective persistent view keeps observations attached to their exact code versions; this experiment is not hot-patched, and the availability gap is not a demonstrated causal explanation of the failure. The other T submission fails on an empty native cycle without the required explicit empty-decision batch. The host neither repairs source nor adds a component-test pass requirement to scientific acceptance.

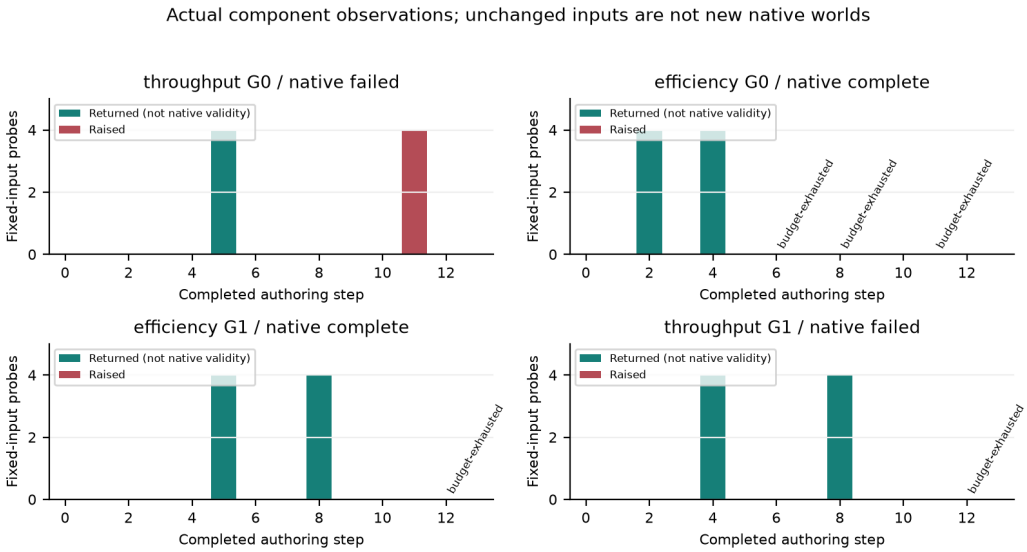


Figure 15: **Actual component observations in all four sessions.** Each bar counts four executions on saved developer inputs, not four new worlds. Returned values do not establish whole-world validity; later edits may be untested after the budget is consumed. The T G0 session introduces four observed errors after four returned probes, and no session demonstrates an error-to-returned repair.

The seven worlds report 663 completed scientist responses, 600,856 input and 66,114 output tokens. Cached input (110,645) is a subset, not additional input. Separate records contain 114 readiness and 56 proposal requests, eight designer loads and seven scientist loads with unknown load-token usage. The loop connects actual code tests, feedback, native execution and retained nulls. It does not show repair success, structural invention or an advantage over the strong reference.

14.4 Why different proxies can select the same rule

Fix the submitted records, their observed review scores and actual review work R. Write $z_i \in \{0,1\}$ for accepting record i and $q_i \geq 0$ for its original mean review score, with $q_i = 0$ for an unreviewed record. Then

$$L_T(z) = -\sum_i z_i, \qquad L_E(z) = -\sum_i z_i q_i + \tfrac{1}{2}R. \tag{12}$$

Both conditional losses are minimized by accepting every record; an acceptance with $q_i = 0$ ties under E. Different coefficients therefore need not imply different acceptance rules. An improvement over the initial threshold rule alone is weak evidence for a need to invent a complex institution, making accept-all with the same actual authority an essential control.

This is not a global optimum of the persistent world. An institution can change reviewer assignments, actual work, future submissions, information and resource trajectories, which are fixed in the conditional statement. Moreover, changing the reviewer panel changes the observed q_i itself. Neither the algebra nor a reduction in E establishes improved scientific truth. Adding an acceptance penalty, a blind observer or a participation term would define a new prospective objective, not correct this experiment's loss.

14.5 Observed annual resource and activity support

An original-source readout of the earlier closed T/E cohort checks all 176 saved annual checkpoints from its 44 complete physical worlds against the existing full audit. Five partial failures are retained without complete outcomes. All 880 observed actor–year boundaries are active; all 44 final activity stocks are five. Table 14 shows the individual year-end resource range, not only pooled final totals.

Table 14: Saved annual support in the earlier 44 complete worlds. Each year contains 220 actor observations. Extrema can come from different actors and worlds; rows are not a single actor's trajectory.

Year	Active / observed	Minimum resources	Maximum resources
1	220 / 220	90	150
2	220 / 220	75	200
3	220 / 220	55	250
4	220 / 220	30	300

Two source-bound readouts and two portable numeric exports agree exactly. All recorded initial-funding fields are 100; repeated stock fields are not new funding allocations. The historical within-year resource ledger was disabled, so saved boundaries cannot exclude temporary inactivity or identify within-year minima. Nor do 44 completed worlds constitute 44 independent optimizer histories or show that other institutions could not reduce activity. The final-activity component alone cannot distinguish these observed completed worlds. Increasing their seed count is not by itself a remedy for an unobserved or structurally flat target. A more informative main comparison must first identify a consequential information, action or timing opportunity in an extension that retains the persistent researcher world.

14.6 A fixed content instrument separate from institutional review

A separately frozen retrospective readout rates all 76 distinct contents in the earlier 44 complete worlds' accepted sets. The fixed local Qwen3-8B receives only title, abstract and topics under one anchored 1–5 rubric, without author, paper identifier, institution, policy, review score or acceptance label. Three seeded positions give 228 actual complete replies. One cached panel is reused for identical content; repeated accepted decisions are not independent ratings. The five original partial worlds stay excluded.

Panel means are 2 for two contents, 3 for twenty, $3\frac{1}{3}$ for five, $3\frac{2}{3}$ for one, and 4 for forty-eight. The predeclared contribution $\bar{b}_i - 3$ is therefore negative for two contents, zero for twenty and positive for fifty-four. Unlike the nonnegative conditional contribution in Equation 12, accepting another record need not increase this utility. This does not identify a dynamic optimum or establish that the instrument has already been used for objective-conditioned search. All 263 historical accepted-decision occurrences receive their content's fixed panel; the 44 world utility sums range from zero to $9\frac{1}{3}$.

Two cold exports reconstruct every literal rating and produce identical data and figures. Actual usage is 101,340 input and 32,493 output tokens, with no unknown counts. The summed HTTP duration is 859.568 seconds, not total project compute. These are retrospective abstract-based opinions, not

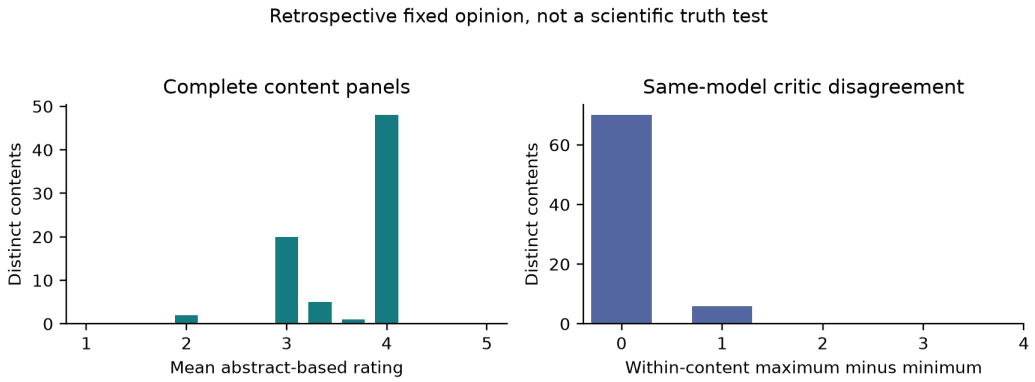


Figure 16: **Observed support of the fixed abstract-based instrument.** All 76 content panels are complete. Seventy have identical scores across the three positions, and six have a range of one. Agreement within the same model does not establish independent errors or scientific truth.

rejected-content measurements, truth judgments, held-out evidence or policy reselection. Neither historical T/E losses nor their selected institutions are changed.

14.7 Resource allocation with the original production laws

A prospective four-world developer assay varies original initial endowment and one conserving-transfer edge, retaining five native Qwen3-8B researchers, the original four-year corpus, projects, costs, grants and activation rules. The original two-band blueprint assigns initial resources $[15, 15, 40, 40, 40]$ at base30 and $[30, 30, 80, 80, 80]$ at base60, in seed-determined founder order. An explicit developer program redistributes whole units toward the active authors' floored mean just before annual costs. Its source, state, nodes and schedules are identical within each pair; only delivery to the native conserving resource consumer is connected or disconnected.

All four worlds complete, and all 80 founder-year ledger reconciliations have zero residual. At base30, connection changes final active founders from four to five and accepted decisions from five to seven; final total resources change from 360 to 350. At base60, both conditions finish with five active founders, seven accepted decisions and 500 resource units. The connected worlds each execute 14 transfers, totaling 154 debit units at base30 and 184 at base60. Each debit equals its actual credit; transferred units are not new funding. Later total stocks can differ because native costs and grants continue after the conserving intervention.

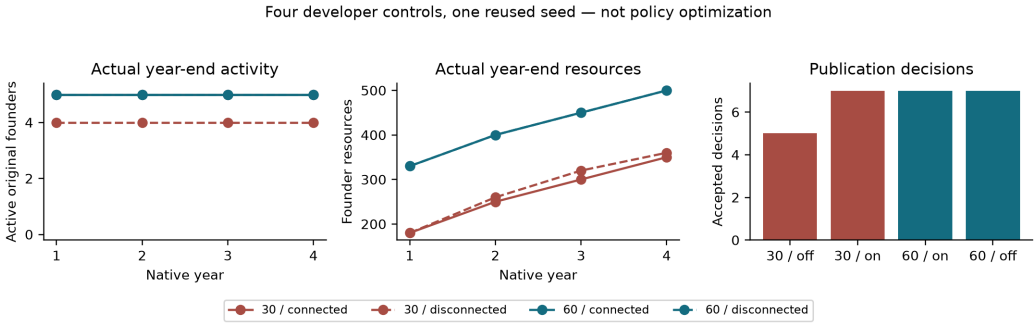


Figure 17: **A consequential action under one initial resource regime.** All four developer controls are retained. The activity curves for both connected worlds and base60/disconnected coincide at five. Resource curves for the two base60 worlds coincide. At base30 the delivery edge changes activity and output, not just command logs. One reused development seed is not independent optimization replication or a held-out policy comparison.

The first five full researcher request payloads match within each endowment; subsequent sequences differ. Two cold audits, cost projections and seven-file exports agree. Actual usage is 499 completed scientist replies, 444,035 input and 49,595 output tokens, plus four loads with unknown token usage and no designer-generation calls. This assay shows action support in the unchanged research production world, not an automatically invented policy. Output and participation align in the observed low-endowment contrast; different objective-optimal institutions have not thereby been established.

14.8 An objective tie can conceal different research records

Five separately audited developer references retain the same low-endowment world and fixed abstract-based instrument, with count loss $T = -A$ and content loss $U = -\sum_i (\bar{b}_i - 3)$. The three connected accept-all rules (original reviews, one reviewer, or no reviews) each obtain $T = -7$, $U = -4$. Reviewed accept-all with the conserving-transfer delivery disconnected obtains $T = -5$, $U = -4$; reject-all obtains zero under both criteria. The observed T minimum set therefore has three worlds and the U set four, sharing three. Choosing a shared tied point would not establish identical preference sets. These finite observed minima are neither global optima nor newly optimized institutions. The five complete references belong to a study that subsequently stops technically before any candidate native world; two matching original-source stopped-prefix audits preserve that distinction.

The connected and disconnected accepted sets are not a common five-record set plus two additional papers. Matching year, paper and exact content gives three shared occurrences, four connected-only and two disconnected-only occurrences. Repeated decisions retain their multiplicities. Figure 18 decomposes both observed outcomes. The common three contribute three utility units. The four connected-only records contribute one unit, as do the two disconnected-only records: each side has one score 4 record, with three versus one neutral score 3 records. The total score sums are 25 and 19, so

$$\Delta T = -(7 - 5) = -2, \qquad \Delta U = -\big[(25 - 3 \cdot 7) - (19 - 3 \cdot 5)\big] = 0. \tag{13}$$

Score 3 is neutral under this declared preference, not evidence of bad science. The same cached abstract opinion is not an independent rating replicate.

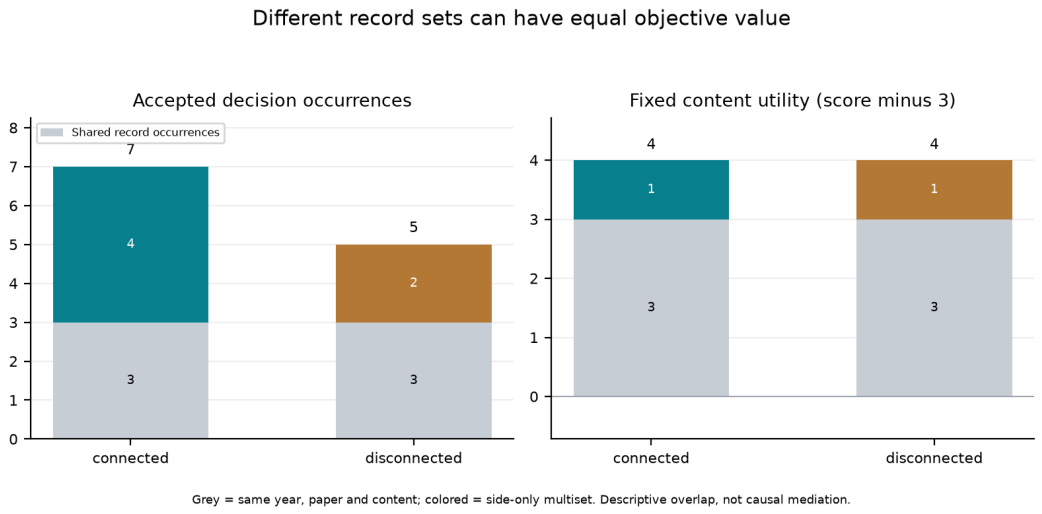


Figure 18: **Different observed records, equal content utility.** These are actual completed developer references at one reused seed, not automatically invented policies or unused-seed evaluations. Grey represents the same year/paper/content occurrences; colors represent each world's side-only multiset. Seven versus five accepted decisions yield utility four in both worlds. Side-only occurrences are not identified as papers causally created or prevented by transfers.

For these fixed realized records, replacing the center by c changes their utility difference to $6 - 2c$; the tie occurs at $c = 3$. This conditional sensitivity neither changes the loss nor predicts adaptation after optimizing a different loss. It is not causal mediation or truth validation. Two read-only exports agree in all data and figure bytes, with no new inference or generated code execution. The contrast explains an aggregate tie, not autonomous invention or a selection-fixed optimized-policy advantage.

14.9 Connecting additional objectives to adaptive search

An optional operator-bound measurement hook now connects such additional metrics to adaptive native loss evaluation, the next institution proposal and cold campaign reopening. Missing or changed contracts reject before dispatch; unknown critic completion stops later calls without a fabricated loss. A scripted original-Simulation integration check verifies this connection and an inert adapter preserves the tested researcher request payloads. The contract stamp identifies orchestration, not observer accuracy. Actual-model search under a new objective is a separate experiment, described next rather than inferred from this connection test.

14.10 Closed count-versus-content institutional search

A prospective study uses $T = -A$ and $U = -\sum_{i \in \text{accepted decisions}} (\bar{b}_i - 3)$, the latter measured by the fixed blind content instrument rather than institution-dependent review scores. Five native Qwen3-8B researchers, the original four-year corpus, project durations, costs, grants and activation laws remain unchanged. The declared fixed-production review-routing domain permits arbitrary institutional code, mechanisms, entities, wires and schedules, but not direct project overwrites or changes to the observer. A base-30 two-band endowment, the same history-information institution and the conserving liquidity program provide a consequential starting world. Neither loss contains a fairness, participation, novelty or review-cost term.

Five fresh references precede two objective-conditioned searches with four proposals each, using one reused development seed 61001. The actual editing parent minimizes the assigned loss over the common measured references and that objective's own completed history; it need not equal the design anchor. This is end-to-end loss-conditioned search, not an intervention holding the editing parent fixed. The existing local 35B designer has fresh interface calibration per proposal, a finite sixteen-action authoring allowance and at most 64 original-container component probes. No host repairs a received scientific program or replaces a failed slot.

The first attempt stops before any candidate world because its outer cold replay incorrectly requires the historical eight-probe allowance instead of the declared 64. Its five reference worlds, 28 readiness and thirteen proposal replies remain a distinct stopped record. A separately frozen corrected-source study imports none of those observations. That study closes all eight submissions: six complete native worlds and two have technical nulls (Table 15). Both original-source audits and both cost projections agree exactly. This normal closure is not improvement.

Table 15: All eight count/content proposals, in actual dispatch order. The assigned loss is development-only. Probes execute saved component inputs, not full counterfactual worlds; every edited final artifact is probed on its exact version. All 52 attempts return, including those in failed native runs.

Slot	Native status	Own loss	Edit from parent	Probes
T/g0	Complete	-7	Accept all, final score 1	8
U/g0	Failed	–	Add third feature column	8
U/g1	Complete	-4	Unchanged reference	0
T/g1	Complete	-7	Accept all, mean review score	8
T/g2	Complete	-7	Accept all, final score 0	8
U/g2	Complete	-4	Rescale two feature columns	4
U/g3	Complete	-4	Accept all, mean review score	12
T/g3	Failed	–	Invalid empty-cycle fallback	4

U/g0's three-column output violates the receiving tensor's two-column input declaration. T/g3 adds a decision field to the empty-cycle sentinel and fails the native decision-batch contract. Returned component results do not certify downstream or whole-world contracts. All failures retain null losses. No candidate changes nodes, wires or schedules; U/g1 is a literal identity submission, not a repaired program or invented institution.

All six completed proposals obtain the same cross-scores: $T = -7$, $U = -4$. The three connected accept-all references—original reviews, one reviewer and no reviews—also obtain these values. Reviewed accept-all with liquidity disconnected has $T = -5$, $U = -4$; no-review reject-all has zero under both losses. Each objective has three tied own submissions, with the predeclared earliest tie

selecting T/g0 and U/g1. The latter is literally its submitted reference, not a host fallback. Figure 19 retains every slot, including the first unavailable U value and final T failure.

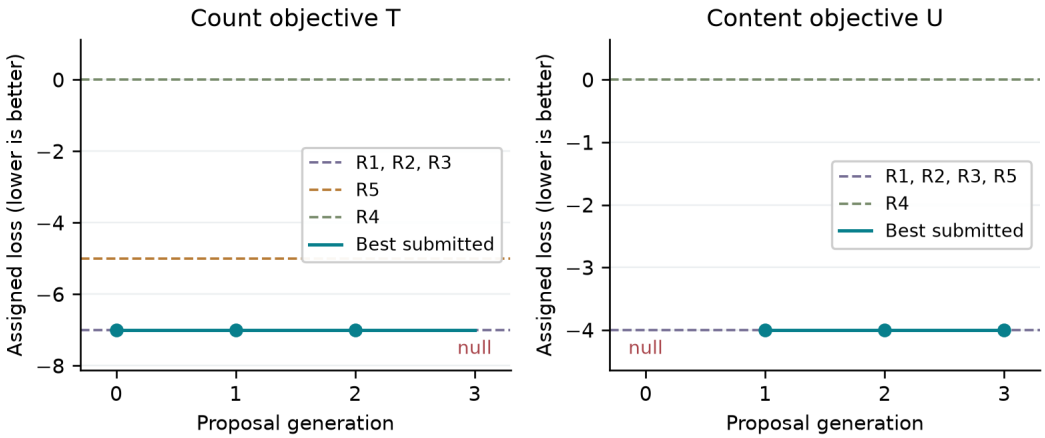


Figure 19: **A completed loss-conditioned search can still tie strong rules.** Both own-history paths and all five common references use the declared T/U losses. Coincident reference lines retain all labels: R1 is reviewed accept-all, R2 one-reviewer accept-all, R3 no-review accept-all, R4 no-review reject-all (all connected); R5 is reviewed accept-all with liquidity disconnected. Null proposals receive no invented score. These are one history per objective at a reused seed, not independent search replications or unused-condition performance.

Across the eleven complete physical worlds, the observed T minimum set has nine members and the U set ten. Of 55 within-seed pairs, 36 jointly tie, ten have the same strict preference and nine tie only under U. There is no observed strict preference reversal. These counts characterize the finite common-completion pool, not 55 independent trials or a global-optimum proof. The six successful submissions do not improve the strong references under their assigned objective.

Same loss is weaker than same observed behavior. A separate read-only comparison retains full original model payloads, including messages and generation settings, without removing score, identity or timestamp text. U/g1, T/g1, U/g2 and U/g3 have the reviewed reference's identical ordered sequence of 129 requests and identical literal response choices. Their complete provider envelopes do not all match. T/g0 and T/g2 change input and response sequences but still tie aggregate losses. Thus an unchanged loss, unchanged observed inputs and unchanged scientific state are distinct statements. Seven completed observed input classes are not program-equivalence classes, independent replications or a reason to skip future worlds. The failed prefixes remain listed with 34 and five requests. Actual source/consumer analysis also retains 236 code calls, 45 annual checkpoints and 225 founder-boundary states, including partial evidence; seven uncommitted bundles are not completed events.

Cost and evaluation boundary. The thirteen native attempts produce 1383 completed scientist responses, 1,258,863 reported input and 140,155 output tokens; 214,438 cached inputs are a subset. Separate accounting retains 230 readiness and 93 proposal POSTs, sixteen designer loads and thirteen scientist loads, with unknown load-token usage. The blind observer makes 33 physical POSTs for eleven cached panels, with 14,940 input and 4,671 output tokens. Shared content is not independently rerated. These components do not measure total project compute or money.

A separately frozen evaluation compares both fixed selections and all five controls at previously unused world seeds 64101, 64107, 64113. Literal identity between U/g1 and its reference yields eighteen physical worlds for 21 logical label–seed comparisons. No new designer, loss change, reselection or failed-slot replacement occurs. All eighteen complete normally; two original-source cold audits and cost projections agree before the two independent score/effect exports. Table 16 and Figure 20 retain every label and seed. Earlier studies elsewhere in the manuscript are not pooled with this evaluation.

Table 16: **All fixed T/U policies and unused seeds.** Entries are ordered 64101, 64107, 64113; both losses are minimized. Transfer denotes conserving equalization. U-selected is the literal reviewed reference, not three more physical executions. Fractions summarize recorded floating values.

Fixed policy/control	T	U
Reviewed all + transfer; U-selected	$-11, -7, -9$	$-25/3, -16/3, -16/3$
One-reviewer all + transfer	$-11, -7, -9$	$-25/3, -16/3, -6$
No-review all + transfer	$-11, -7, -9$	$-31/3, -16/3, -16/3$
No-review reject-all + transfer	$0, 0, 0$	$0, 0, 0$
Reviewed all; transfer cut	$-10, -5, -8$	$-26/3, -10/3, -16/3$
T-selected literal program	$-11, -7, -9$	$-28/3, -16/3, -6$

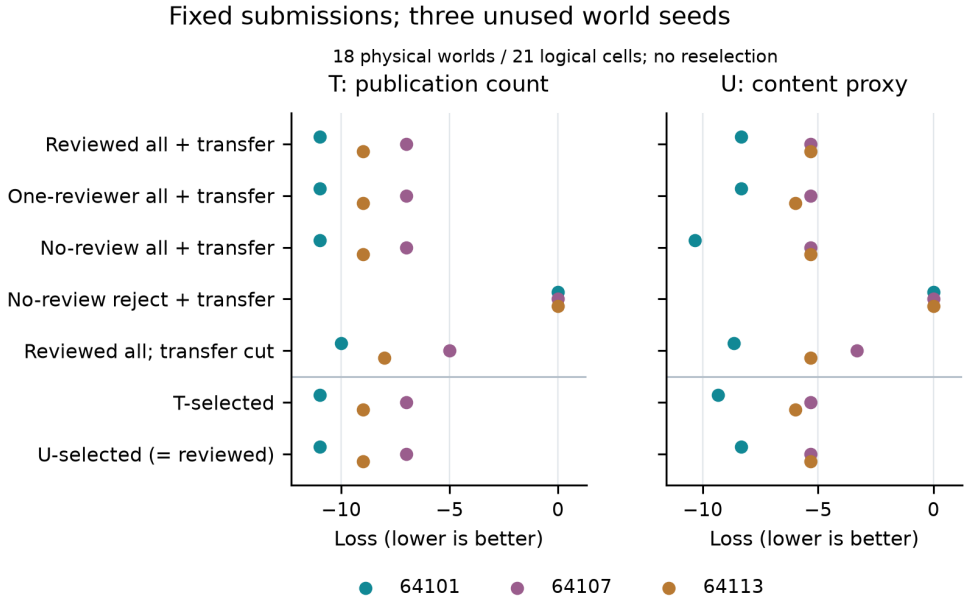


Figure 20: **Unused world seeds do not establish optimizer superiority.** T-selected ties all three connected accept-all controls on publication count. Its content differences against no-review accept-all are $+1, 0, -2/3$, so advantage depends on the comparator. U-selected repeats its literal reference. The horizontal separation is actual loss; slight vertical offsets identify seeds only. These are fixed-policy worlds conditional on one history per objective, not independent optimizer replications.

T-selected improves U against reviewed-all on two seeds and ties one, but U was not its optimization objective. Against no-review-all its differences are mixed. All connected all-accept controls and both selections have identical T at each seed. The transfer-cut reference produces fewer accepted decisions on each seed, without identifying the optimizer's unique contribution or a complete causal mediation mechanism. No uniform strong-reference advantage or general objective-to-institution ranking follows. Native traces preserve 372 recorded code calls and 72 annual boundaries; observational execution traces are not a replacement for separately controlled mechanism tests.

The evaluation lasts about 2h42m for eighteen serial native worlds, including content scoring; final cold closure adds about 2m21s. Original costs record 2377 scientist replies, 2,431,115 input and 225,508 output tokens. The 441,737 cached input tokens are a subset, not extra input. The separate observer makes 144 physical POSTs (67,044 input and 20,611 output tokens). Eighteen load records retain unknown token usage. Three world seeds do not supply independent search histories, independent critic errors or validated scientific truth.

15 A delivered information intervention is not necessarily an effective one

The count/content study changes some inputs without exceeding strong rules. Before attributing this to absent information or launching a larger search, we test whether a particular institutional message reaches the relevant native decision at the intended time. This is a developer intervention, not a newly invented institution or an independent optimization history.

Fixed contrast and native world. At source `5925d5b`, three original four-year worlds use the same five Qwen3-8B researchers, explicitly reused development seed 61001, base-30 resource profile, M1 history, conserving equalization and unchanged T/U measurement. The literal message recommends choosing a permitted short-duration topic; it does not change the displayed catalog, choose a topic for a researcher or overwrite project dates. The three conditions cut delivery, deliver at phase 2 before the first direction choices, or deliver at phase 235 after those choices. The message and all three conditions are fixed before dispatch, not chosen from favorable responses. A permission-only preparation stop precedes the separately recorded successful launch; no scientific slot was consumed there or replaced. All three worlds subsequently close normally.

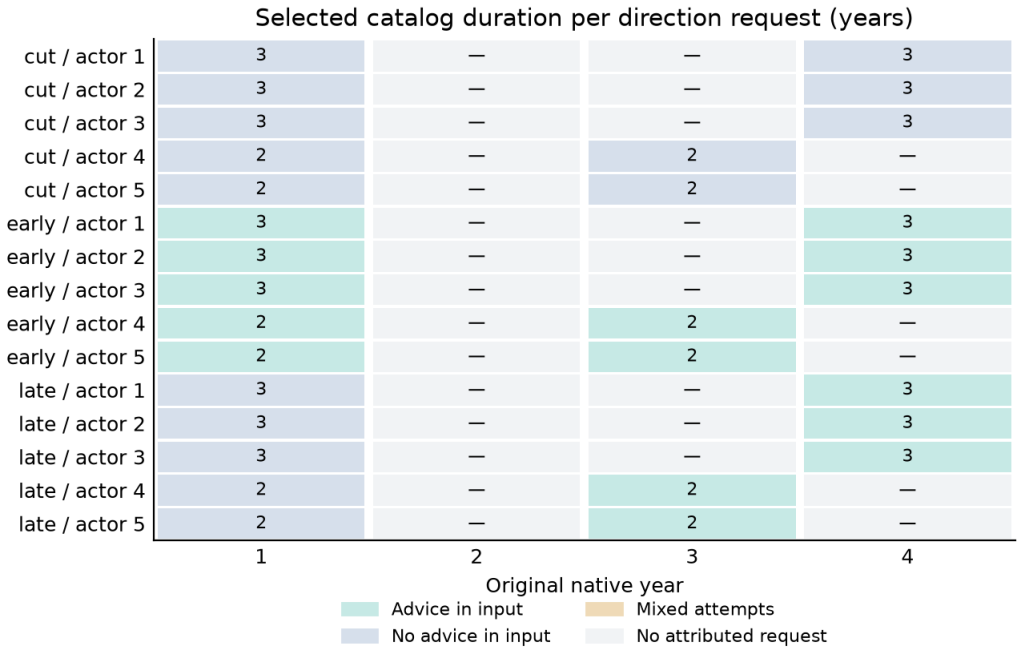
Condition	Advice in inputs	Shortest choice	Mean duration	T	U
Cut	0/10	4/10	2.6	-7	-4
Early	10/10	4/10	2.6	-7	-4
Late	5/10	4/10	2.6	-7	-4

Table 17: All three literal advice-timing controls. Advice presence is checked in actual direction-request inputs, not inferred from node execution. Duration is the selected topic's displayed catalog duration in years, averaged over ten requests per world. Shortest is relative to that request's permitted catalog; following the advice is not a validity requirement. The late message is absent from first-year direction inputs, as declared. These are one-seed developer contrasts, not independent search replications.

Delivery established; the proposed choice effect absent here. All 30 direction requests are uniquely joined to their original stored memories and displayed catalog. Early and late delivery each store five addressed instruction records; cut stores none. The exact literal text occurs in 0/10, 10/10 and 5/10 direction inputs, respectively. Nevertheless, every corresponding actor/year chooses the same topic and catalog duration across all three worlds (Table 17). Each chooses a shortest displayed option on four of ten requests, with mean selected duration 2.6 years, and all yield T=-7/U=-4. Non-shortest choices are not schema violations: the message is advice, not a mandatory native law. We retain those choices and their losses rather than rewriting them to satisfy the instruction.

Equal selected choices do not imply equivalent worlds. The complete ordered model-input and literal-response sequences differ among all three worlds. Cut has 129 completed scientist replies, early and late 130 each. Total author resources at year 4 are 350, 310 and 350, respectively. Thus unchanged topics, durations and T/U are weaker than behavioral equivalence or identical resource trajectories. Resource diagnostics do not add an adoption veto or another term to either optimization loss.

Audit, costs and inference boundary. Two original-source cold audits and two cost projections agree; the choice, effect and complete-input exports and repeated summary/figures also agree. Actual native execution takes about 24 min, followed by about 1 min 15 s of cold/report closure. There are 389 scientist replies with 360734 input and 38542 output tokens; 63873 cached input tokens are a subset. The unchanged content observer uses 39 separate requests. Tokenless model-load costs remain unknown; there are no designer proposals. The result establishes that this literal message was delivered and did not improve these particular choices or T/U at this seed. It does not establish general instruction ineffectiveness, equal full behavior, scientific quality, a policy ranking or a benefit of graph search. All exact data, message identity, source and closure receipts are retained in `research/data/native-instruction-timing-v1/`.



? = unknown duration; comma list = multiple attempts; actor labels are local to each control.

Figure 21: Actual annual direction-input delivery and catalog-duration choices in the three closed native worlds. The original displayed catalog and literal received action determine each point. Delivery varies at the declared times; the observed selected durations coincide. This supports a scoped delivery/choice distinction, not a general inability of information institutions to affect researchers or a causal account of why this message did not change topics.

16 Capacity-limited native search preserves a strong allocation rule

16.1 A separate activity objective under original production laws

The preceding count/content search does not identify a useful objective-to-system distinction. We therefore conduct a separately declared native pilot with existing capacity limits, not a retrospective change of its losses. Five original Qwen3-8B university researchers operate for four years over the same 103,107-record corpus, with base-30 balanced resources. The original `papers_per_project=2` cap permits partial submissions. Fixed funding at fraction 0.2 supplies one total annual award slot across programs. Native research cost remains 10 when an active researcher has a working project; the original per-paper flow has no direct submission debit. Neither a paper fee nor a new resource law is activated to force a trade-off. This campaign also fixes native research planning: it does not grant the `native.project` action. Broader planning-authority cases elsewhere in the program are not pooled with this restricted-domain search.

Alongside $T = -A$, define $I_{iy} = 1$ when initial founder i is author-capable and active at the end of year y, and zero otherwise. Activity here means the saved native `is_active` flag. The separate loss is

$$P = -\sum_{y=1}^{4} \sum_{i \in F} I_{iy},$$

where F is the initial five-researcher cohort. New actors do not enlarge the cohort. The formal minimum is -20; the indicator does not certify resource sufficiency, useful work or continuous survival. In particular, native zero-resource activity semantics are preserved. P is an objective in its

System or own slot	T	P	Annual active founders
Reviewed all / equalization	−10	−12	5,5,1,1
Reviewed all / transfer cut	−10	−12	4,4,2,2
No-review all / equalization	−10	−12	5,5,1,1
Reviewed all / reserve-proximity	−12	−15	5,5,3,2
Reviewed all / completion-priority	−12	−15	5,5,3,2
T/g0; selected T	−12	−15	5,5,3,2
P/g0; selected P	−12	−15	5,5,3,2
P/g1	—	—	Technical null after year 2
T/g1; ties T/g0	−12	−15	5,5,3,2

Table 18: All nine physical attempts in the capacity pilot. Lower losses are better. The failed world has two saved annual boundaries and recorded work in year 3, not a completed four-year outcome. Equal sums for the first three rules conceal different activity paths. All complete own submissions tie both strong allocation rules; there is no generated-policy advantage.

own search, not an additional constraint on T or an adoption veto. Every complete world receives both scores from its original checkpoints.

Five common developer rules precede two own slots per objective at explicitly reused development seed 61001. All rules use their declared native authority: reviewed accept-all with equalization; its transfer cut; no-review accept-all with equalization; and reviewed accept-all with reserve-proximity or completion-priority allocation. The latter rules conserve resources and seek a pre-cost reserve of 21. They are strong hand controls, not generated inventions. Each actual editing parent minimizes the assigned loss over common references and its own history; all four slots start from reserve-proximity. Final selection is instead restricted to complete own literal submissions, including retention, with the earliest tie and no reference fallback. The local 35B designer has 16 proposal steps per slot and 64 optional historical component probes. Initial and four history-bearing technical calibrations precede dispatch.

16.2 All submissions, strong ties and a real execution failure

All four slots submit literal programs. Three complete native execution; P/g1 fails and receives neither a numerical loss nor a replacement. Both original cold audits and cost projections agree, as do two exports of each score, effect, input and preference analysis. Table 18 retains the full comparison.

No own submission changes nodes, connections or schedules. T/g0 replaces the already all-accept selection source with unconditional acceptance and constant final score 1. T/g1 retains the mean native review score while rewriting unconditional acceptance. P/g0 likewise retains that mean and replaces direct field access in the allocation code with existing-value defaults. These literal edits do not establish a new allocation algorithm. The full scientist input sequences of P/g0, T/g1 and both strong allocation references are identical at this observed history, with identical literal response choices; full provider envelopes are not identical. T/g0 has a different full input sequence despite the same T/P and activity path. Observed input equality is not program equivalence on all possible states.

P/g1 attempts a different operation. Its final allocation code includes initial-founder support and a virtual-source fallback. In year 3 it emits two transfers of 10 from the nonexistent actor `system`. The native conserving-transfer receiver rejects the batch because endpoints must be existing author-capable actors. The original code result and two committed annual boundaries are retained; partial work is not promoted to a complete loss. This is an authored invalid command, not evidence that the framework silently repaired the institution or that the policy performed poorly on P. Four component probes of that final artifact had returned on historical fixtures. Their success did not certify the later native consumer or its trajectory: local code return and valid downstream scientific execution are distinct tests.

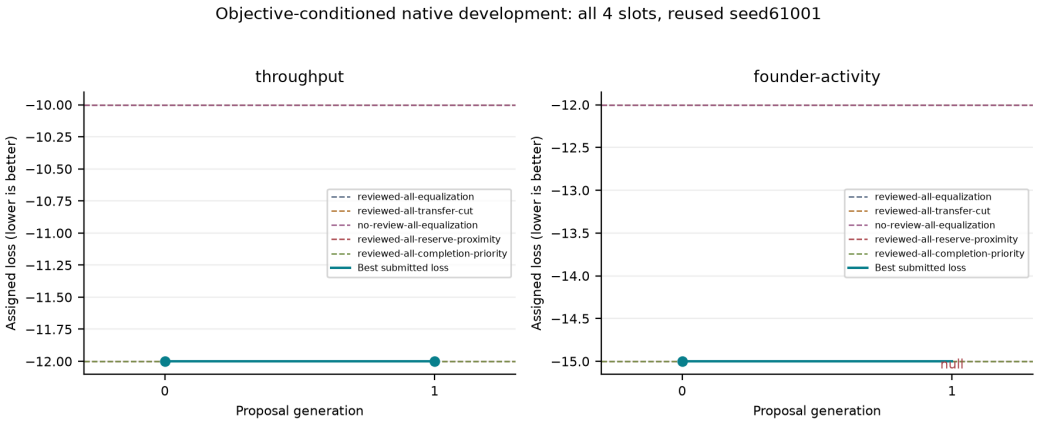


Figure 22: Both actual objective-conditioned search paths, all five coincident reference labels and the P/g1 null. The first own proposal already ties the strong editing parent; the flat path is not an improvement from that parent. There is one design history per objective, not independent optimizer replication.

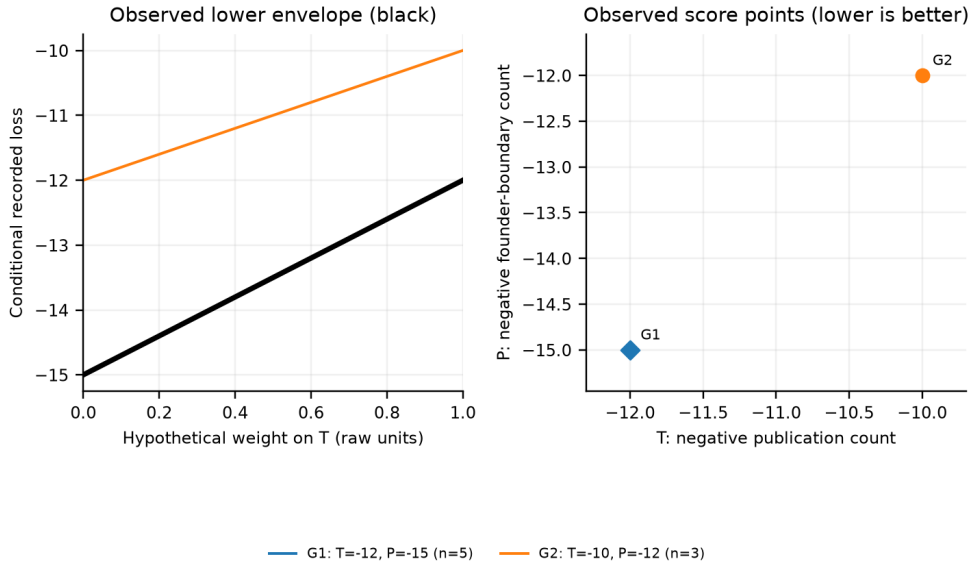


Figure 23: Conditional geometry of the complete capacity archive, including all tied programs and excluding the one null from numerical coordinates. The lower envelope selects the same score group across raw-unit weights. Five worlds at one point are neither equivalent programs nor independent optimizer histories. No new objective, candidate selection or inference is introduced by this readout.

16.3 What the common score point does and does not explain

The eight complete worlds have only two observed score pairs: $(-12, -15)$ for the two strong controls and three own submissions, and $(-10, -12)$ for the other three controls. The first strictly dominates the second in both recorded coordinates. Consequently, every hypothetical raw-unit linear mixture $wT+(1-w)P$, $0 \leq w \leq 1$, has the same minimizing score group in this archive (Figure 23). This is an exact finite-record calculation, not optimization at an interior weight, a global Pareto frontier or evidence that T and P are intrinsically aligned. The search did not explore materially different complete organizing systems.

The lower-loss hand rules and own programs have 12 rather than 10 accepted decisions and 15 rather than 12 saved founder boundaries in this development condition. That comparison concerns allocation rules; it is not evidence of autonomous invention or a causal mediation estimate for a particular transfer. The original traces include 172 code calls and 34 committed annual boundaries, including the failed prefix. Recorded scientist cost is 1,171 completed replies, 1,117,703 input and 111,959 output tokens, with 213,168 cached input tokens as a subset. Separate designer transport records contain 58 proposal POSTs and 142 readiness POSTs;18 role loads have unknown token usage. No critic is called. Execution takes about 2h12min, with cold/reader closure about 4min17s later.

We retain this finite negative search result rather than expanding it until a desired institution wins. The prepared unused-seed evaluator is not dispatched for this flat pilot: no unused seed has been selected or scored. The remaining research question is how to obtain and explain a nontrivial objective-to-system relationship, distinguishing the explored behavior, designer interface, world dynamics and objective incentives. More fixed-policy seeds alone would not answer that question or establish publication readiness.

17 Concrete agendas change native planning and accepted output

17.1 Matched authority without changing production laws

The capacity pilot protects original project selection; it does not test whether institutions can improve that selection. We therefore conduct a separate three-arm developer intervention, retaining the same five initial Qwen3-8B university researchers, four native years, corpus, reserve-proximity allocation and reused development seed 61001. All three specifications have the same initial graph, available grants and agenda-generation rule. The institution chooses the shortest topic in each eligible researcher's actual displayed expertise catalog, breaking ties lexically. Only its actuation differs: no effect, concrete advice before original planning, or direct legal assignment. These are explicit hand-written controls, not generated discoveries.

The new catalog-assignment action must preserve the exact displayed topic, native catalog content and duration, and can apply only to an active idle author. It cannot shorten ongoing work or create an artificial output. This is an explicitly named extension of institutional authority, not a weakening of the earlier fixed-production scope or unrestricted permission to replace native production. Original native costs, funding, submission capacity, calendar and accounting remain operative. The shortest-topic rule is a particular control, not a requirement imposed on generated institutions.

Every arm receives both unchanged losses T and P. Neither is a constraint on the other. No criterion requires advice to be obeyed, a topic to change, or either loss to improve. All three worlds complete without replacement. Two original-source audits and two cost projections agree; the choice, effect and complete-input readers each export identically twice, as does the final figure and table export. Table 19 retains all conditions.

17.2 Observed choices and an actual institutional consumer

Without agenda actuation, initial projects have catalog durations 3,3,3,2,2 years. Both effectful arms start with 2,2,2,2,2-year projects and start another two-year project for each eligible founder in year 3. Saved native start and end dates retain these original catalog durations (Figure 24); no existing project is retroactively shortened. A project visible in an inactive founder's saved state is not thereby completed, and an assignment is not itself a publication.

In the advice arm, all ten original direction requests include their actual actor/year agenda, and every uniquely attributed literal selected topic matches that agenda. There are zero native assignment effects. This is observed uptake by the fixed researcher model on this history, not an assumption inferred from delivering a prompt. Direct assignment instead produces ten validated native effects and zero original direction calls: it replaces the choice rather than demonstrating researcher obedience.

The effectful arms each have 18 rather than 12 accepted decision records and 16 rather than 15 active founder boundaries. Their annual active counts are 5,5,4,2 versus 5,5,3,2 for no effect (Figure 25). All three end with two active founders. Final total founder resources are 50,40,40 in the table's order,

Agenda actuation	T	P	Scientist replies	Planning calls	Assignments
No effect	−12	−15	148	9	0
Concrete advice	−18	−16	180	10	0
Direct assignment	−18	−16	168	0	10

Table 19: All three native planning controls at one reused development seed. Planning calls are original research-direction requests; assignments are catalog-validated native effects, not completed projects. Both effectful arms improve accepted-count loss and saved-founder activity together. This is not objective-conditioned search or independent replication.

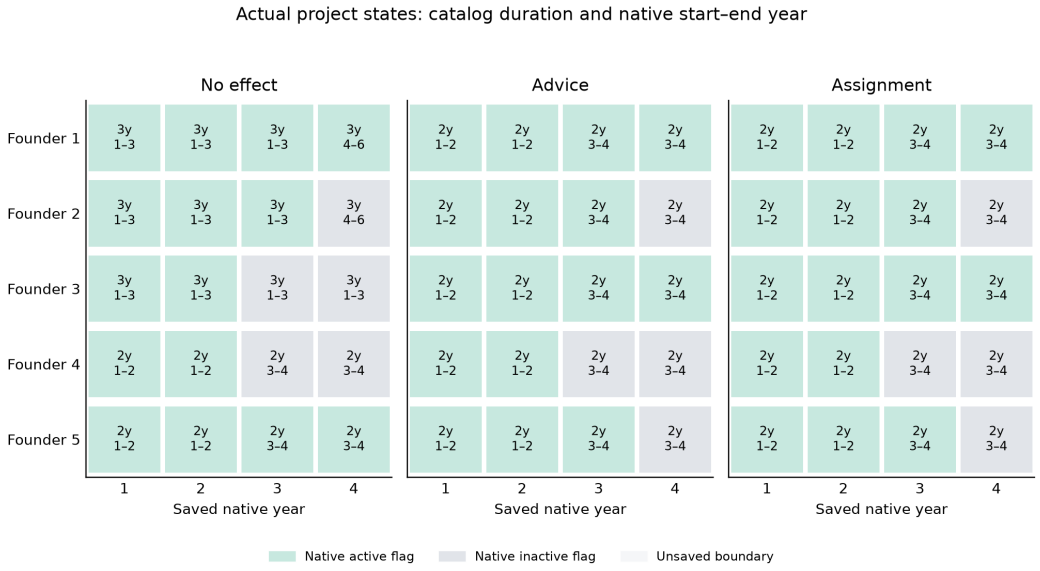


Figure 24: Actual saved project states for all five founders and four years in each arm. Cells show the catalog duration and native start–end years; colors encode original active flags. Project dates and flags are different quantities. The two-year effectful agenda contrasts with the initial mixed two/three-year path, without changing original production laws.

so this is not an improvement in every reported quantity. Accepted decisions remain a proxy, not verified discoveries; P remains a native flag sum, not meaningful work or continuous survival.

17.3 Inference boundary and measured cost

The comparison establishes an effective planning path available to an institution, not autonomous invention or a demonstrated objective-to-system distinction. Identical initial rules and permitted authority do not imply identical later inputs. All three full ordered scientist-input sequences differ, including the two equal-loss arms. Their later resource and model response histories can evolve differently. Changed plans and accepted output therefore do not isolate project duration as the sole mediator. The earlier generic-advice null uses a different message and world scope; it is not a matched message-granularity ablation of this study.

All 496 scientist requests have committed replies, with 518,946 reported input tokens and 45,741 output tokens. The 90,562 cached prompt tokens are a subset of input, not an additional charge. Three scientist-load POSTs have unknown token usage. There are no designer-generation or critic POSTs. Native execution takes 31min15s, followed by about 46s to original double audit and cost closure. These are local observed wall times, not energy or fresh-machine benchmarks.

This result supplies useful same-authority reference policies for subsequent free planning/allocation search. A count/content study must still evaluate its own actual submitted designs, both objectives and every strong reference, then fix selections before unused-condition evaluation. A new graph is not required; neither is a positive result. The later planning study is separate: its first initial-

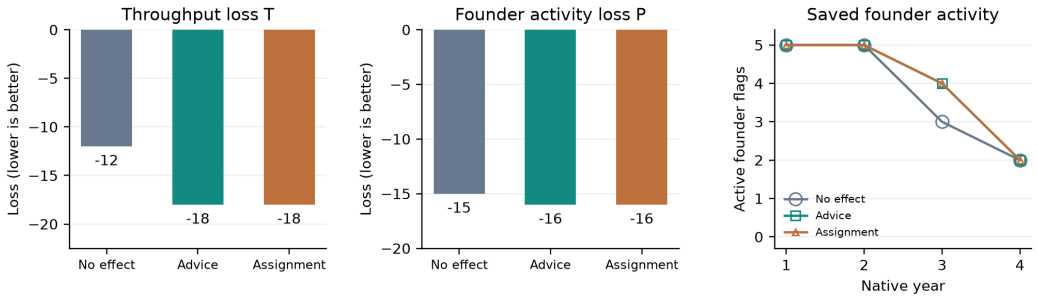


Figure 25: Actual T/P losses and annual founder flags in all three completed worlds. Advice and direct assignment tie both recorded losses despite distinct request counts and full scientist-input histories. There is no observed T/P trade-off in this comparison and no seed uncertainty estimate.

readiness attempt stops when the local service receives a termination signal, before any reference or own candidate world. A separately predeclared fresh attempt cannot erase that stop or its unknown response costs. The completed three-arm evidence alone does not close the important empirical question or justify a main-conference-ready judgment.

18 Free planning search preserves strong reference behavior

18.1 A separate finite objective-conditioned study

The preceding three-arm comparison establishes an operative planning path, not what a search invents. We next permit original catalog assignment in a separate count/content search. Five initial fixed Qwen3-8B university researchers and other native actors run for four years against the same 103,107-record corpus. Base resources remain 30, the project-output cap remains two, and the annual funding fraction remains 0.2. Native production, costs, calendar and accounting are unchanged. Institutions may edit source, nodes, protocols, wiring and schedules; catalog assignments must preserve the actual displayed topic, content and duration. Existing projects cannot be shortened. This named design scope is not the limit of the broader system framework.

The losses remain $T = -N$, where N counts accepted decisions, and $U = -\sum_{j \in \mathcal{A}} (\overline{q}_j - 3)$, where the three-position fixed-content panel supplies an abstract-based opinion. These opinions are not scientific truth. P and resource histories are diagnostics, not additional objectives or constraints. Eight fresh developer references share the permitted authority (Table 20). Three proposal slots per objective use the installed local35B designer at optimizer seed 68011. The interleaved order is T0, U0, U1, T1, T2, U2. Each editing parent is selected by assigned loss from the common references and own completed history; all 14 worlds explicitly reuse development seed 61001. These are two histories, not independent optimizer replications or unused-condition evaluations.

All six programs submit and complete. Final selection is the earliest tied completed own submission, T0 or U0, with all three own ties retained per objective. No reference fallback, novelty requirement, participation floor or improvement veto is added. Two original-source audits and cost projections agree; five pairs of score, choice, effect, complete-input and signed-content exports agree. The stopped initial calibration remains separate from this normally closed second attempt, with its unknown response costs retained.

18.2 Content preferences refine observed count ties

Every own count loss is -18 and every own content loss is approximately -13.3333 (Figure 26). The selected U institution has observed U lower by 3 than selected T, at equal T. Parent selection belongs to the declared objective-conditioned procedure: all T parents are R2, and all U parents are R7. Their difference is a valid observation of the end-to-end procedure, not an isolated gain caused by the generated edit or a general expected objective-label effect from one history per branch.

ID	Institution or own submission	T	U	Point
R1	No agenda effect, reserve 21, reviewed all	−12	−8	G5
R2	Advisory agenda, reserve 21, reviewed all	−18	−11.3333	G3
R3	Assignment, reserve 21, reviewed all	−18	−13	G2
R4	Assignment, reserve 10, reviewed all	−12	−8	G5
R5	Assignment, equalization, reviewed all	−12	−8	G5
R6	Assignment, reserve 21, native threshold 3	−18	−13	G2
R7	Assignment, reserve 21, no-review all	−18	−13.3333	G1
R8	Assignment, reserve 21, no-review reject	0	0	G6
T0	First count proposal, selected	−18	−10.3333	G4
U0	First content proposal, selected	−18	−13.3333	G1
U1	Content redesign	−18	−13.3333	G1
T1	Count redesign	−18	−11.3333	G3
T2	Last count proposal	−18	−10.3333	G4
U2	Last content proposal	−18	−13.3333	G1

Table 20: All eight fresh references and six own proposals, in execution order. Every own assigned loss ties its best fresh reference. Display rounding does not determine ties: selection and analysis retain original floating-point losses. Point labels map every exact tie in Figure 27.



Figure 26: All three consumed proposals per objective and the unchanged fresh reference losses. Overlapping gray lines retain the eight references listed in Table 20. Own losses and incumbent paths are flat; this is not independent-seed uncertainty or evidence of generated advantage.

R7 minimizes U and is also one of four references minimizing T. The complete 14-world archive therefore has overlapping, but unequal, minimum sets. Of 91 pairs, 43 have the same strict preference, 36 tie only in T and 12 tie in both losses; none has a strict preference conflict. U refines part of a count-neutral plateau rather than revealing an observed unavoidable trade-off (Figure 27). This finite archive is not a global frontier or an upper bound on attainable institutional performance.

18.3 Literal code edits versus actual researcher behavior

All six submitted programs change only `world.selection` source, explicitly accepting all papers from parents already configured accept-all. No program changes planning, allocation, nodes, wiring or schedules. T0 and T2 also set `final_score=1.0`; T1 retains mean native review scores. The U programs retain mean scores, with a default-valued field lookup in U0. A literal source change or equal assigned loss does not prove behavioral equivalence.

The complete-input reader instead finds eight exact observed-input classes. U0/U1/U2 and R7 have identical ordered full scientist payloads and literal response choices. T1 likewise matches R2. T0 and T2 match each other but differ from R2, with outside U worse by 1. Full provider envelopes

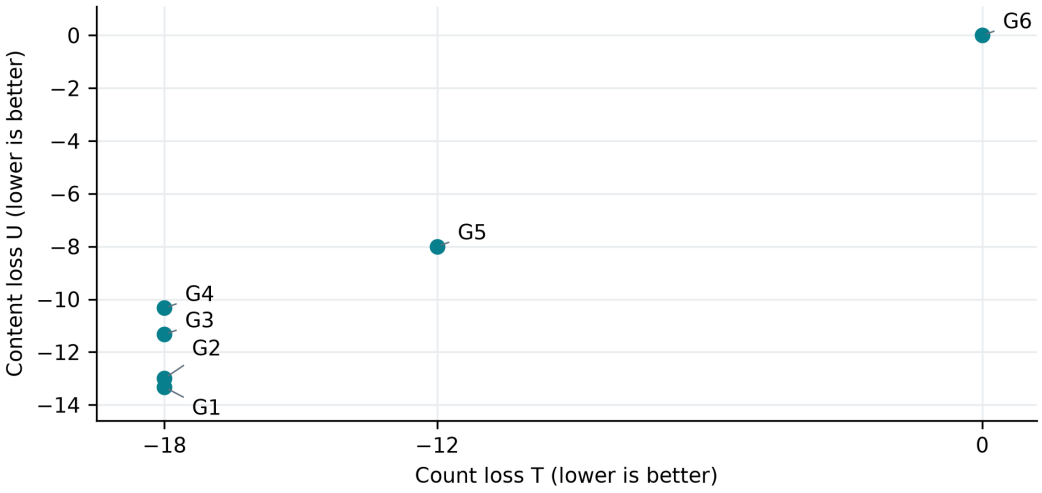


Figure 27: Both losses on all 14 completed native worlds, grouped only at exact shared score coordinates. Table 20 retains every label. Annotations move for legibility; data coordinates are not jittered. G1 belongs to both observed minimum sets. Different selected institutions do not imply that the objectives must trade off.

differ within these multi-world classes. Equality on these observed paths is not equivalence on unobserved histories or permission to merge different policies in future runs.

Each T program retains ten original direction calls and no catalog assignment; each U program retains ten validated assignments and no direction call. All own annual active-founder paths are 5, 5, 4, 2. The effect trace retains 336 code calls and 56 committed annual boundaries across all 14 worlds. Literal first plans predominantly target selection and adjacent scoring code. U2 also proposes editing the catalog agenda, but its final program does not. The plans include mistaken hypotheses about already all-accept parents. This documents narrow actual exploration; it does not identify parent presentation, file ordering, incentives or designer capability as its cause.

18.4 Signed content support and measured cost

The 216 accepted occurrences represent 54 unique contents. Every observed centered contribution is positive or zero, never negative (Figure 28). R7 and each U proposal have 13 positive and 5 zero occurrences; T0/T2 have 10 positive and 8 zero occurrences. Rational utilities $40/3$ and $31/3$ explain those latter sums but never replace operational floating-point losses. For a fixed accepted multiset, discarding a positive contribution cannot increase total utility. A selective institution may still change future production and the multiset; rejected or unproduced content is not rerated. Occurrences, unique contents and critic errors are not independent replicates of the same estimand.

Execution takes 13,014.325 seconds (3 hours, 36 minutes, 54 seconds), with about 5 minutes 15 seconds to double audit/cost closure and another 66 seconds to the last reader. There are 2,019 completed scientist responses, 2,196,668 input and 194,184 output tokens; 368,423 cached prompt tokens are a subset of input. Fixed-content measurement uses 162 critic POSTs for 54 three-position panels, with 77,940 input and 24,002 output tokens. Designer authoring records 86 POSTs with 3,598,980 input and 8,820 output tokens; readiness records 200 POSTs with 3,843,374 input and 9,844 output tokens. Thirteen designer and fourteen scientist loads have unknown token usage. Categories stay separate; missing costs are not zero-filled, and no energy or money estimate is inferred.

This closes a functioning feedback-bearing native search, not useful automatic invention. The prepared selection-fixed evaluator has not been dispatched for these programs; no unused seed is chosen. More flat evaluation alone would not distinguish objective support from narrow proposal coverage. A substantive follow-up must separate action choices and proposal information before making method-general claims, and independently evaluate whatever fixed systems and design procedures that question concerns.

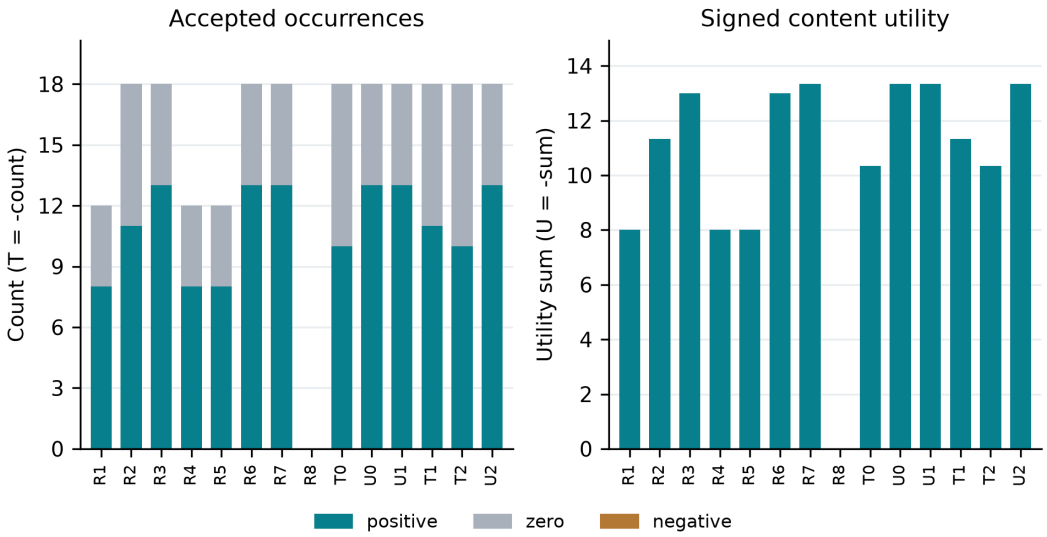


Figure 28: Every accepted occurrence and its signed fixed-content contribution. All 14 worlds remain in the original order; R8 has zero accepted records, not a missing observation. No negative contribution appears in these realized accepted sets. The panel is an opinion instrument, not a truth oracle or a rerating of rejected and unproduced contents.

19 An operative numeric parameter has a conditional, nonmonotone response

19.1 Measure the component without inventing counterfactual worlds

The planning search leaves the strong reference's allocation program unchanged. Before interpreting a numerical search over its reserve, we ask a narrower question: does this existing parameter change the commands on the actual inputs that its consumer received? We retain R7's complete original inputs for years 1–4 and execute its exact archived allocation artifact in the same pinned sandbox image for every integer reserve $r \in \{1, \dots, 60\}$. Only the already declared input parameter changes. All 240 calls complete; at $r = 21$, all four complete component returns match their original records. This adds no scientist-model call or native world and does not alter any optimizer's feedback or the original study's inputs.

Let $x_y^{(21)}$ denote R7's recorded year-y component input under its original reserve21 trajectory, and let C_r be the unchanged program with reserve r. The measured object is

$$\left(C_r(x_1^{(21)}), \dots, C_r(x_4^{(21)})\right), \qquad r \in \{1, \dots, 60\}.$$

It is *not* a rollout on the future states $x_y^{(r)}$ induced by each alternative system. Outputs are requested transfer commands, not newly committed transfers, participation outcomes or loss evaluations. Consequently, neither adding commands across recorded years nor maximizing transferred funds would estimate the optimization target.

The 60 reserve values produce respectively 19, 17, 15 and 8 distinct exact transfer-command outputs on the four historical inputs. The joint four-frame signature has 23 distinct values. The empty command output recurs in disjoint low and high reserve intervals; we do not merge those intervals into an unobserved continuous range. Figure 29 exposes why increasing a parameter need not increase even its immediate output. The program requests complete rescues up to the target reserve, keeps that same reserve at donors, and skips a recipient if the remaining surplus cannot cover its whole need. Increasing the target can reduce feasible transfers. For example, the year3 input gives the following exact command projections.

This establishes an actual parameter consumer and an irregular exercised command surface. It does not show that any reserve optimizes T or U. Equal outputs on these four inputs also do not establish

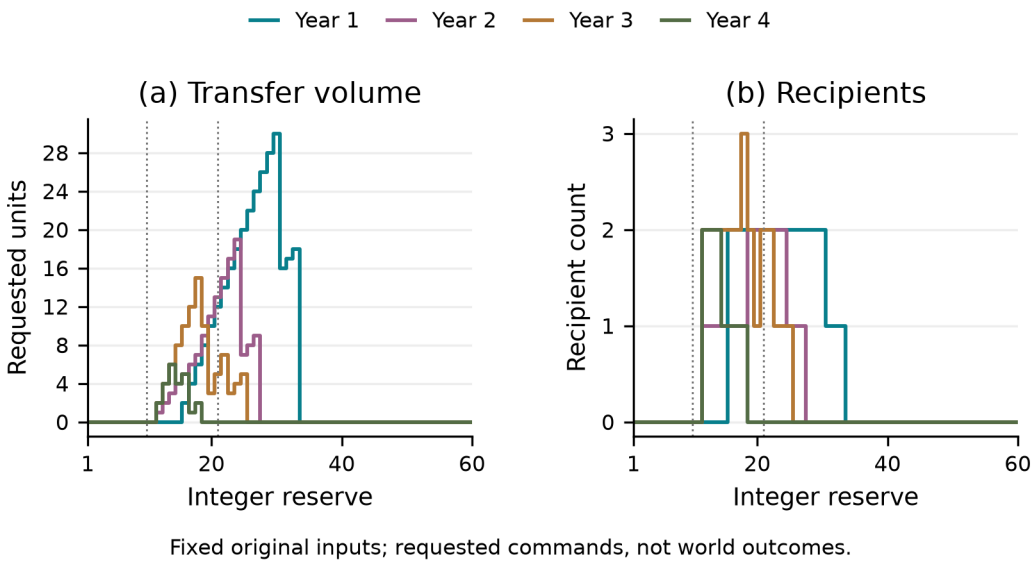


Figure 29: The exact archived allocation component on four fixed historical native input frames, varying only its reserve parameter. Left: requested transfer units. Right: recipients in those commands. Dotted lines locate the existing reserve10/21 controls, not selected optima. Curves are conditional command responses, not alternative scientist-world outcomes or loss paths.

Reserve	18	19	20	21	26
Requested transfer units	15	10	3	5	0
Recipients	3	2	1	2	0

Table 21: A nonmonotone conditional response on the same original year3 input. No reserve is ranked by scientific performance in this diagnostic.

global policy equivalence, authorize sharing future native jobs, or provide a gradient of full-world loss.

19.2 Join commands to the one trajectory that actually occurred

The original reserve21 year3 commands can separately be joined to its real native resource ledger. The five founders open with balances $(31, 17, 11, 11, 20)$. The component requests 5 units from founder0, with 4 to founder1 and 1 to founder4; these commands match the realized transfers. Every founder pays the original annual cost of10, and founder2 receives the original20-unit grant. Table 22 retains all five records, including the deactivated researcher. Every accounting residual is zero and total resources reconcile as $90 - 50 + 20 = 60$. Transfers conserve money within the group; they do not increase the grant budget or remove costs.

Native attrition occurs after the award: the inherited implementation removes active authors whose balance is truthy and at most10. Thus a different founder with pre-award balance1 survives after receiving20, while founder3 at1 exits. The truthiness check also excludes exactly zero from this particular branch; we report that simulator convention rather than silently replacing it with a cleaner threshold. This is an existing world rule, not a fairness floor, participation objective or promotion veto added by our optimizer.

The next process-design question. The timing of observations, transfers, costs and awards can affect eligibility, actual charged costs and subsequent researcher inputs. Changing that timing is not the same intervention as changing a reserve within a fixed calendar. The conditional analysis motivates a separately declared workflow comparison with unchanged native effect implementations, rather than a larger search over an allegedly smooth scalar response. It does not predict which

Founder	Opening	Net transfer	Cost	Grant	Closing	Active
0	31	−5	10	0	16	Yes
1	17	4	10	0	11	Yes
2	11	0	10	20	21	Yes
3	11	0	10	0	1	No
4	20	1	10	0	11	Yes
Total	90	0	50	20	60	4/5

Table 22: All original R7 founder resource records in year3. Founder indices abbreviate the fixed original institution0000–0004 researcher identities; the retained ledger preserves their full identifiers. This is the realized reserve21 trajectory, not outcomes for the other 59 parameter values.

Order	Proposal	Reserve	T	U	Assigned loss
1	T0	3	−12	−7	−12
2	U0	13	−12	−8	−8
3	U1	27	−14	−9.3333	−9.3333
4	T1	10	−12	−7	−12
5	T2	12	−12	−7	−12
6	U2	25	−14	−10	−10

Table 23: All six new parameter proposals, in execution order. All complete. Both losses are negative sums, so lower is better. T's warm-start loss is -18 and U's is -13.3333; no new proposal improves its assigned incumbent. Display rounding does not determine comparisons or ties.

workflow will win, demonstrate autonomous structural invention, or validate an attrition convention as a real-world science-policy goal.

20 A finite native parameter comparator does not improve the incumbent

20.1 An unchanged world, a deliberately narrower search

We next run a parameter-only comparator against the closed planning archive. This is a separately declared post-hoc comparator, not another execution of its 14 worlds or a controlled causal intervention on design freedom. Original native scientist, production and effect code is checked against 265 unchanged source files. Fixed Qwen3-8B scientists, the original corpus, four years, world seed 61001 and all original T/U measurements remain literal. Only the already consumed integer reserve may vary from 1 to 60; code, nodes, wiring, schedules and other parameters cannot change in this family.

The earliest tied best reference supplies each warm start. T starts from advisory R2 with review; U starts from assignment-based R7 without review. Both use reserve 21. The existing adaptive random search at seed 68011 draws one initial uniform proposal per objective, then bounded mutations around its best measured system, with mutation scale 0.2 and the existing decay. Three own slots per objective run in order T0, U0, U1, T1, T2, U2. No designer LLM or readiness call generates these numerical proposals. Eight reference outcomes and the six earlier free-program slots are reused transparently, not counted as fresh worlds or independent optimizer histories. Unknown or failed observations would remain null, with no repair, refill or invented penalty.

The original presentation-dependent resource coordinator stopped without starting this experiment. A separate prospective dispatcher starts the same previously unused plan only after checking its exact original source/bank, the other study's stop, retired processes and two fresh empty-backend observations 30 seconds apart. This is not resumption of the stopped study: its unknown response remains unknown and its partial feedback is not used.

20.2 Improving an own-output path is not improving the known solution

All six new worlds complete and both original-source cold audits, cost projections and complete report exports agree. The best-own T path is flat at -12, and its earliest selected submission is T0,

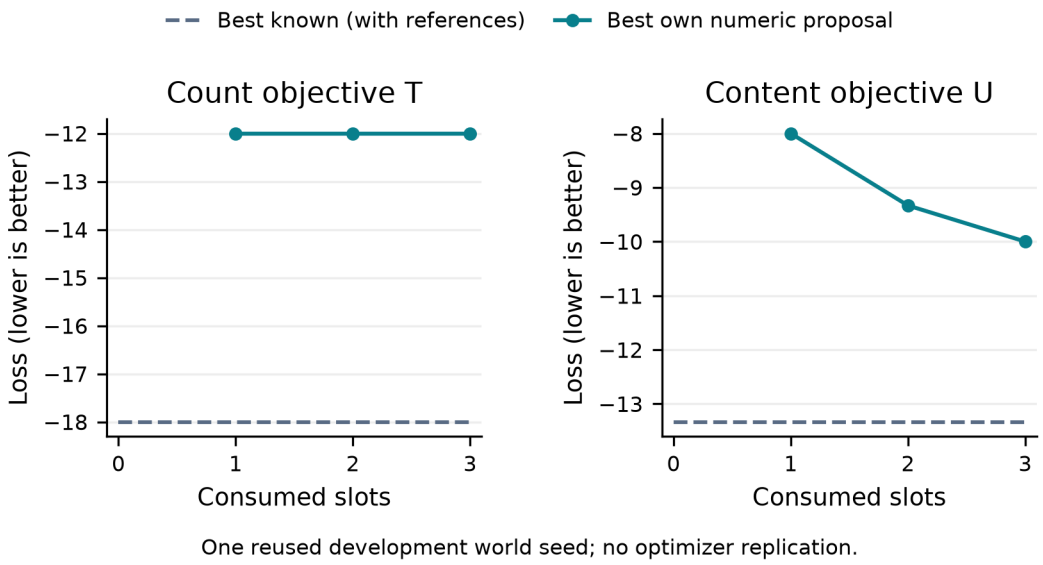


Figure 30: Reference-inclusive and best-own numeric paths on the same complete development observations. Zero consumed slots contains the measured warm start, not an invented own proposal. The improving own U curve never improves the known incumbent. These curves are not independent-history uncertainty or an unused-condition evaluation.

with all three own ties. The best-own U path falls from -8 to -9.3333 to -10, selecting U2. Neither reaches the corresponding original warm start. The best-known path, which retains completed references, therefore remains T=-18/U=-13.3333. Own-only selection and the optimization incumbent are different readouts; we do not replace a worse generated output with a reference and call that reference an invented improvement. Nor do these six trials establish that reserve 21 is globally optimal over the declared 60 values.

The earlier free-program own submissions attain those strong reference losses. That is an observed package-level comparison, not an intrinsic advantage of program search: information, interfaces, generation expense and design families differ, and this comparator was designed after the pilot. One history per objective and one reused world seed cannot rank optimization methods in expectation. A larger run of the same comparison would not by itself isolate those confounds.

Literal paths and observed experience. The six new worlds form four exact observed full-input classes. T0/reserve3, T1/reserve10 and T2/reserve12 each contain 133 complete scientist requests. Their ordered full model-input payloads and literal response choices match; all four actual allocation returns in each run request no transfers. The provider envelopes remain different and are retained separately. This checks these observed histories, not global equivalence of the policies or permission to share future world jobs. The complete all-world input-class readout is a separate double-exported audit-bound analysis, not an assumption inferred from matching scalar losses.

Measured expense and boundary. The six-world scientific interval spans about 79 minutes 7 seconds, including its native scientist and observer phases. Recorded scientist usage comprises 664 responses, 704,991 prompt tokens and 67,549 completion tokens, with no wire input lacking a committed reply. The common observer adds 93 physical critic requests, 44,481 input tokens and 13,766 output tokens; content cache hits are not extra independent ratings. Six ordinary scientist role loads are recorded separately. This is end-to-end local elapsed time, not a dedicated model-speed benchmark, GPU FLOP estimate, energy measurement or dollar cost. The later audits/reports are outside that scientific interval. No unused evaluation condition is selected or run by this comparator.

Observed quantity	Original R7	Hand timing/routing
Throughput loss T	−18	−20
Content-utility loss U	−13.3333	−15.3333
Completed scientist request/reply pairs	115	127
Actual four-year grant income	80	80
Actual four-year research fees	190	200
End-year-4 active original authors	2	5
End-year-4 original-author resource stock	40	30

Table 24: All completed reference cells in the stopped workflow experiment. Both declared losses prefer the hand control in this one condition. Counts and stocks explain behavior; they are not additional objectives or promotion vetoes. Display rounding does not decide ties. Neither control is an automatically invented improvement.

21 Changing observation and action timing in the native world

21.1 An operated structural slice, not a replacement researcher world

The preceding parameter study leaves an important axis untested: when existing institutions observe the world and act on it. We therefore construct a structural-only search domain over the same native scientist world. Existing mechanism source, parameters, actor population, corpus, Qwen3-8B scientists, loss definitions and four-year horizon remain fixed. Phase assignments and connections among existing typed ports may change. This controlled slice does not exhaust the framework's code-defined mechanism and composition freedom.

The original assignment/reserve-21/reviewer-free reference, R7, supplies the initial graph. A strong developer control moves liquidity allocation, its effect consumer and research-cost collection after funding. Three phases and five connections change, without replacing existing mechanism implementations. The proposed experiment allows six own submissions, three under each of T and U, with both controls measured in the same reused development condition, seed 61001. It does not require novel wires, participation preservation or improvement for a candidate to be valid. Such requirements would change the declared optimization problem.

This experiment does *not* normally close. Both reference cells complete; two model-authored proposals fail before a native world, and the next history-bearing readiness check returns an invalid graph. Two consumed own slots, zero own native worlds and no final selection are retained. Remaining slots are not refilled. The cause is not a missing scientist response in this attempt: the relevant designer replies have known final completion signals.

21.2 A completed-reference contrast within a stopped experiment

The completed cells remain analyzable without fabricating whole-study closure. A separately scoped reader runs against the exact original scientific source, checks checkpoints, scientist replies, code-execution receipts, fixed-content critic outputs, declared losses and reconciled resource ledgers. Two executions produce byte-identical reports, retaining all 333 original raw files and an exact closure of 304 source files. This is repeatable analysis of the same two observations, not independent experimental replication. The missing original whole-study observer-audit file is not manufactured, and these reports cannot supply a selection-fixed evaluation prerequisite.

The original has 5, 5, 4 and 2 active original authors at the four annual boundaries; the hand control has 5 at every boundary. Aggregate resource stocks are 120, 90, 60 and 40 versus 120, 90, 60 and 30. Each receives 20 grant units per year. Actual conserving internal transfers total 12, 13, 5 and 0 versus 6, 10, 17 and 8. The observed output difference is therefore not financed by additional aggregate grant income. Equal aggregate stocks through year 3 also do not imply equal distribution, eligibility or activity.

Why this is not a pure cost-order effect. Moving a stage changes the available observation snapshot. The legal hand graph consequently reroutes inputs as well as changing relative timing. Original funding-stage summaries occur before later mutations in the hand workflow and cannot be read as common end-of-year stocks. The native funding cycle's existing deactivation rule and its truthy-zero

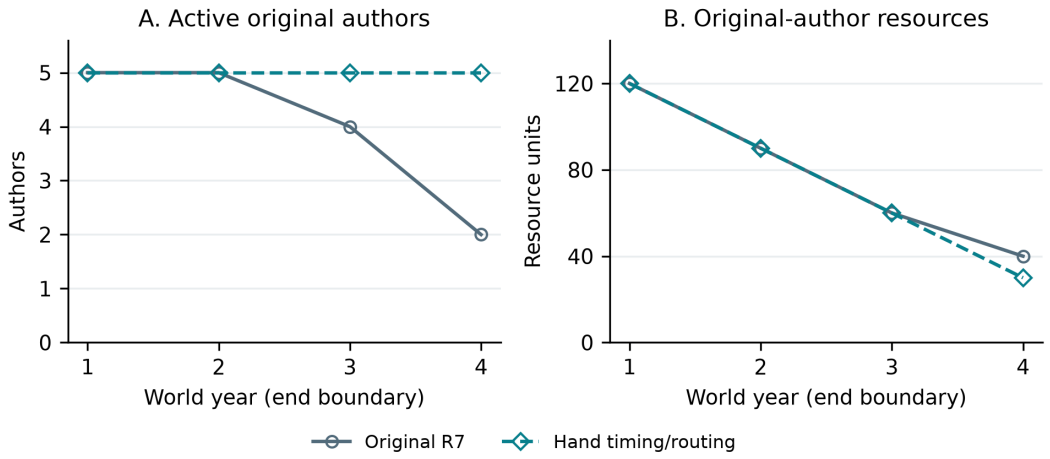


Figure 31: Actual annual boundaries for both completed developer controls. Each receives 20 grant units per year. Equal aggregate resource stocks through year 3 coexist with different activity; the hand control finishes with more active authors but less total resource stock. Both timing and observation routes change. These are neither within-year minima, independent replications nor additional participation/resource objectives.

exception remain literal; they are not repaired to make the comparison cleaner. End-year ledgers do not recover within-year minima or isolate the moment of attrition. Thus Table 24 is an operated joint timing/information-routing contrast, not an isolated causal estimate of funding-before-fees. Four legal static timing corners have been constructed, but no factorial worlds are executed by this analysis.

Both T and U prefer the same hand control here. The contrast therefore demonstrates a consequential additional design axis, not objective-dependent autonomous discovery, an optimum, stable superiority or validated real-world science policy. Existing institutional-intervention precedents already motivate studying such effects; our main scientific question still requires the objective-to-generated-system comparison, rather than claiming timing interventions themselves as a priority contribution.

21.3 Designer execution remains a separate empirical boundary

The two rejected own proposals respectively duplicate a graph connection and use an unexpected output property. The next readiness reply duplicates edges and leaves required inputs unconnected. Original public replay reconstructs those exact failures. A copy diagnostic is not a dispatch veto: the graphs are invalid under the unchanged executable domain. No scientific losses are assigned to unexecuted proposals.

A separately declared four-call local diagnostic varies only JSON Schema formatting and the model's thinking switch on the retained failed-history frame. Schema/nonthinking still returns an invalid graph; both thinking settings return valid directed calibration controls. The predeclared technical priority chooses schema/thinking, not an objective score or a novelty test. Its one free-design call reaches the 8,192-output-token limit with 29,444 thinking characters but zero final-content characters. The finish reason is `length`, not an HTTP timeout or completion unknown. We do not parse a graph from thinking text or count directed calibration as free design. The diagnostic consumes 103,439 input and 13,110 output tokens including thinking, without any new scientist world or loss observation.

These observations separate readable calibration, a complete free submission and an effective institution after native execution. A one-call followup changes only the output allowance to 16,384 on the same free frame. It returns final JSON using 23,862 input and 12,563 output tokens; the earlier thinking text is an exact prefix. Its six phase edits nevertheless put selection at phase 40 before four connected producers at phases 70, 100, 110 and 120. The unchanged compiler rejects it. This probe ends without repair, further allowance increases, certification or scientific execution. Finalization does not establish graph validity or institutional effectiveness.

22 Closing the compiler-feedback structural loop

22.1 A fresh finite study, not a repaired stopped experiment

The preceding stopped study and technical probes do not supply complete own institutions or a final selection. We therefore conduct a separate prospective campaign at source `7f1fc61f`, preserving the original scientist world, two controls, T/U objectives, four-year horizon and reused development seed 61001. The local 35B designer proposes phases and connections among existing typed ports; mechanism source, parameters and nodes are fixed in this slice. This authoring interface is deliberately narrower than the general code-and- graph framework, not evidence that broader freedom has been exercised here.

The campaign uses schema-formatted thinking output, a 57,344-token context, 16,384-output-token allowance and finite transport limits. Known invalid proposals consume slots. The unchanged compiler's exact error, offending causal-edge witnesses and original-input hashes can enter a later proposal's history; no submitted graph is repaired. The actual history projection carries these diagnostics from the original result boundary, rather than assuming a recursive prior-request object. Six proposals are interleaved T0, U0, U1, T1, T2, U2, with separate objective histories. Seven directed readiness calls and six free proposals are recorded. Directed controls are not free-design evidence.

The original process normally closes after all six slots, with four complete own worlds, two generation nulls and both complete controls. Two original-source audits, cost exports and objective reports agree byte-for-byte, as do separate choice, effect and full-model-input readouts. These are repeatable analyses of one development history, not independent optimizer replications.

System/slot	Outcome	T	U	Literal own edit
Original R7	Complete	-18	-13.3333	Reference
Hand timing/routing	Complete	-20	-15.3333	Reference
T0	Complete	-20	-15.3333	Three phase values
U0	Complete	-20	-15.3333	Explicit parent retention
U1	Generation null	—	—	Invalid causal order
T1	Complete	-20	-15.3333	Explicit parent retention
T2	Generation null	—	—	Required inputs removed
U2	Complete	-20	-15.3333	One phase value

Table 25: Every control and consumed proposal in the normally closed compiler-feedback study. Both losses are evaluated on each complete physical world. Nulls have no scientific loss. Display rounding is not used for selection. T0/T1 tie under T; U0/U2 tie under U; chronological first-tie selection retains T0 and U0. All complete own outputs tie the hand control.

22.2 Phase labels are not necessarily behavioral changes

T0 changes the director, effect consumer and cost phases from 161/162/163 to 163/164/165, but leaves connections and the actual total annual order unchanged. U2 changes only the cost phase from 163 to 165. U0 and T1 explicitly retain their parent. All four complete own systems and the hand control issue the same ordered 127 full scientist request payloads and receive the same literal response choices, including recorded finish metadata. Provider envelopes differ and remain unmodified. R7 instead has 115 requests and forms the second observed input class. Thus six physical completions yield only two observed full-input classes, not six behaviorally independent institutions. Equality on these histories is not global program equivalence or permission to skip future worlds.

The two invalid proposals are also informative about the actual search path. U1 places cost collection at 175 while retaining its outgoing zero-latency edge to communication at 170. T2 removes required selection inputs and puts consumers before connected producers. These are known final replies, not unknown transport completions. The actual U2 request contains U1's exact failure witness. U2 then compiles, but this uncontrolled sequence does not identify a feedback benefit. Some designer rationales assert parent violations or optimality unsupported by the received graph and loss definition; those assertions are not host-implemented rules or proofs of effectiveness.

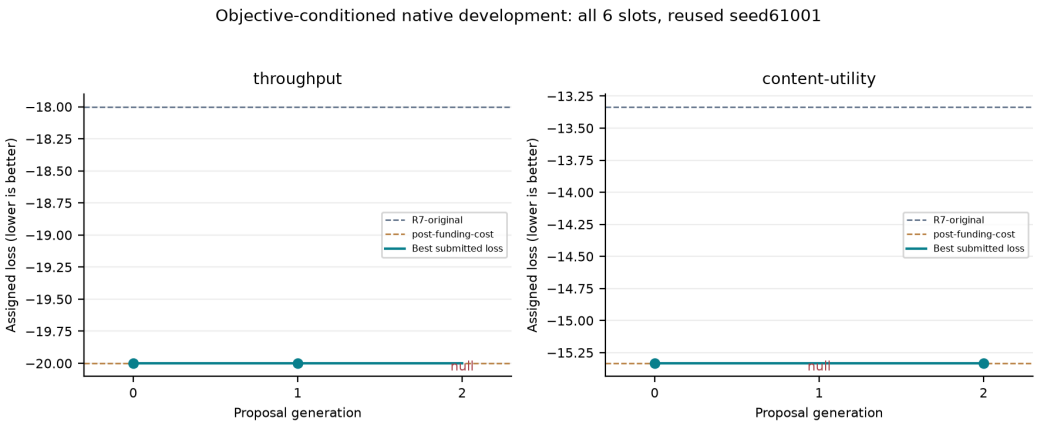


Figure 32: The complete objective-conditioned development paths. Dashed lines show both references; markers show complete own evaluations and the solid line their best submitted loss. Invalid proposals remain labeled null, not assigned a penalty or replaced. No own gain over the hand incumbent occurs. This is one reused development seed, without independent error bars.

Across this campaign there are 750 complete scientist responses, 75 physical content-critic calls and no scientist wire without a committed reply. Recorded scientist usage totals 937,108 prompt and 77,254 completion tokens. Free-design calls use 147,740 input and 46,206 output tokens including thinking; readiness uses another 169,711 and 5,490. Science wall time is 6,082.936 seconds on shared local hardware. Token accounting and wall time do not establish GPU FLOPs, energy, monetary cost or dedicated-device throughput.

22.3 What this closure establishes, and what remains

This campaign establishes an operated proposal–failure-feedback–native- evaluation–selection loop, not an automatically invented improvement. Its chosen parent is already the hand timing/routing control. Retaining that parent, or changing inert phase labels, cannot be presented as discovering the control's advantage over R7. No successful own wiring change occurs.

There is also a concrete headroom question before simply adding repetitions. At the inspected original initialization, each of five researchers has a shortest expertise duration of two years. Under serial uninterrupted catalog projects, two outputs per project, no extra producers or authors, and at most one acceptance occurrence per output, the conditional four-year envelope is $2\sum_{i=1}^{5}\lfloor 4/2 \rfloor = 20$. The hand reaches this conditional value. This is not a theorem over all allowed graphs, resubmission paths, decision occurrences or other initializations. U also depends on content, so neither its optimum nor a necessary T/U trade-off follows from the count.

22.4 A normally closed, selection-fixed unused condition

Before seeing own outcomes, a separate note fixed one hash-derived evaluation seed, 1858637637. Both own labels and both controls are retained regardless of development improvement. Literal identity alone shares physical jobs: the U-selected submitted system is exactly the hand control; T0's phase alias is a separate physical job. Thus four labels require three worlds, not four independent evaluations. No evaluation outcome is fed back to the designer or used to reselect a system.

The first preparation stops before inference when the usage inventory encounters non-production unit-test stores. That failed preparation supplies no evaluation scores and remains failed. A separately declared correction binds 72 inspected test-only files by exact path and hash; changed, missing or unlisted malformed stores still stop preparation. Metadata-bearing scientific journals cannot be exempted. After software checks and repeated inventories find no use of the fixed seed, the new evaluation at source `2133ffc0` normally closes all three jobs. Both original-source cold audits and double score, choice, effect and full-request readouts agree.

Frozen label	T	U	Physical job
Original R7	-32	-24.6667	000
Hand timing/routing	-36	-26	001
T-selected (T0)	-36	-26	002
U-selected (U0)	-36	-26	001 (literal identity)

Table 26: Selection-fixed evaluation on the single predeclared unused seed. Lower losses are better. Both own selections tie the hand on both objectives; their differences from R7 are -4 under T and -1.3333 under U. U0 is the hand literal, not an independently invented success. No confidence interval, independent optimizer replication or stable superiority follows from one seed.

The separately executed hand and T-selected systems issue identical ordered full scientist inputs and receive identical literal choices. Raw provider envelopes differ. Across the three worlds there are two observed input classes, so the equal scores are accompanied by observed behavioral retention, not merely a scalar tie. Neither observed equality nor physical identity is used retrospectively to skip a required world.

The evaluation records 503 complete scientist responses, 761,779 prompt and 46,319 completion tokens, and no scientist input without a committed reply. The separate common observer makes 123 physical calls, with 55,617 prompt and 18,046 output tokens. Wall time is 2,867.393 seconds on shared local hardware. An omission in the cost reader's study-name allowlist initially excludes this observer component from its export, although the original scientific audit already records it. A new role-presence-based reader and matching double supplement retain those calls; old exports, scientific losses, frozen choices and raw records remain unchanged. Cache subsets and nested durations are not added to provider token counts or treated as dedicated-device performance.

22.5 Connecting the broader institution space to the native loop

The structural interface above cannot change institution source or create a new mechanism. A prospective optional `open-institutions-ir-v1` adapter therefore connects the existing arbitrary Stable-ID compiler to the same bounded author/certificate/replay boundary. It admits literal code, parameters, new mechanisms and protocols, wiring and schedules within the explicitly declared native effect laws, without replacing the original scientist world. The host does not implement a designer's prose or supply missing consumers.

In a committed-source software integration check, a prescribed authored planning office replaces the catalogue director, changes its executable source from shortest- to longest-duration assignment, and explicitly reconnects both its input and its downstream native actuator. The full native pipeline verifies changed scientist inputs, seven development worlds and three fixed-evaluation worlds, including a known null, actual sandbox/source execution and agreeing double cold readouts. HTTP, corpus retrieval and upstream origin are scripted; there are zero real-model calls. This proves an active end-to-end software path, not an LLM-invented institution, a beneficial intervention or a new scientific comparison.

The prospective study keeps the strong hand control as a measured comparison but exposes only R7 as initial optimizer feedback. The complete hand configuration and outcomes are withheld, not its common building blocks: the hand and R7 have identical mechanism sources and parameters, differing in timing and routing. This is not discovery of previously unseen component code. The changed information access also means a comparison with the earlier interface cannot isolate expressivity. That implementation check alone is not real backend calibration or free invention. The actual subsequent pilot below keeps those distinctions explicit.

22.6 First real open-code proposal: an exercised path without improvement

At source `fe4281f4`, both fresh controls and the first free open-code proposal complete in the original researcher world. R7 alone is supplied as the optimizer's parent/reference; the hand remains comparison-only. The designer submits two literal operations: retaining accept-all and replacing selection Python with unconditional acceptance and `final_score=1.0`. The parent already accepts all, but stores the mean review score, zero when reviews are absent. There are no new nodes, connections or schedules.

Completed system	T	U	Scientist replies
R7 reference	−18	−13.3333	115
Strong hand, comparison-only	−20	−15.3333	127
First own T0	−18	−12.3333	114

Table 27: Completed cells inside a subsequently stopped open-institution pilot, not a normally closed search. Both losses are minimized. T0 ties its parent on T and is worse on U, without reaching the strong hand. These are one-seed observations, not independent optimization replications. The remaining five own slots have no result; no selection or unused evaluation is inferred.

The first 70 full scientist request payloads match R7. At zero-based request 70 in year four, the same earlier accepted paper's history contains score 1.0 instead of 0.0. Later full inputs and literal choices differ. Both worlds have 18 accepted records and 18 distinct accepted papers, but four paper IDs differ on each side. The common panel scores those R7-only papers 4,4,4,4 and T0-only papers 3,4,4,4; the 14 common papers have identical panel means. This content difference accounts for the one-unit utility decrease. Writing the institutional score does not directly overwrite the fixed common panel or its loss calculation. Equal throughput therefore does not make this edit globally inert; the observed information path is not an isolated causal effect estimate for the changed score or evidence of improved science.

The following U0 directed readiness call returns a final HTTP 200 response, but omits the schedule operation for its added control node. The unchanged domain rejects it and the writer stops before a U0 proposal. All four actual author calls recompile identically in two original-source cold reads; the 374-file raw inventory remains unchanged. Separate original-source checks also verify the role bindings, four annual checkpoints, native code receipts and common-observer loss reconstruction for each of the three completed worlds. These completed-world and author readouts do not supply full optimizer-history replay, a whole-study audit or normal optimization closure. The unfavorable T0 result and stopped attempt remain separate, without repair, refill, a fabricated selection or a scientific loss for missing slots.

The calibration frame contained both a directed task and the shared instruction to optimize; the received response expresses confusion between them. A prospective calibration-only correction separates that task, preserves the complete context and repeats the full directed target last. Copy equality remains diagnostic, while actual invalid IR still stops. This supports an interface correction, not a uniquely identified behavioral cause or guaranteed future compliance. The separate corrected study is reported next; the earlier stopped attempt is not repaired, refilled or pooled with that new history.

22.7 Closed open-code search: construction failures and one unfavorable policy

At source `25be72cd`, the separate task-corrected campaign closes all six prospectively ordered slots on development seed 61001. Arbitrary literal code, new mechanisms and legal wiring remain available; only R7 is optimizer-visible. All seven directed readiness calls pass, but free construction remains unreliable. The original writer exits normally, and matching pairs of original-source audits, cost exports, objective, choice, effect and full-request readouts cover the complete history. This is one optimizer history per objective, not an independent replication of earlier searches.

T0's generated source attempts 500 synthetic accept decisions but fails first because its world output lacks the required payload envelope. It does not produce 500 native papers, and the observed failure is not a downstream test of origin protection. U0 drops required code-artifact schema fields; U1 edits a nonexistent parameter path; T1 omits the explicit empty-cycle decision batch; T2 names a runtime image different from the pinned sandbox. The host does not repair these submitted values. Directed calibration success alone does not establish free-design reliability, and these failures do not all have a single demonstrated cause.

U2 replaces selection source with unconditional acceptance of actual submitted papers and `final_score=100.0`, handling the required output and empty cycle formats. Its 116 scientist responses differ in count from R7's 115 and the hand's 127; all three completed worlds form different full-input classes. The earlier score1 proposal has the same rounded scalar losses but 114 scientist re-

System/slot	Outcome	T	U
R7 reference	Complete	-18	-13.3333
Strong hand	Complete	-20	-15.3333
T0	Runtime null	—	—
U0	Generation null	—	—
U1	Generation null	—	—
T1	Runtime null	—	—
T2	Generation null	—	—
U2	Complete	-18	-12.3333

Table 28: All slots of the normally closed open-institution v3 campaign. Losses are minimized. Five physical native worlds include two failed own worlds; three generation nulls never enter native execution. Nulls receive no scientific loss. U2 is worse than both references under its assigned U loss. No new node, wire or schedule occurs in the sole completed proposal.

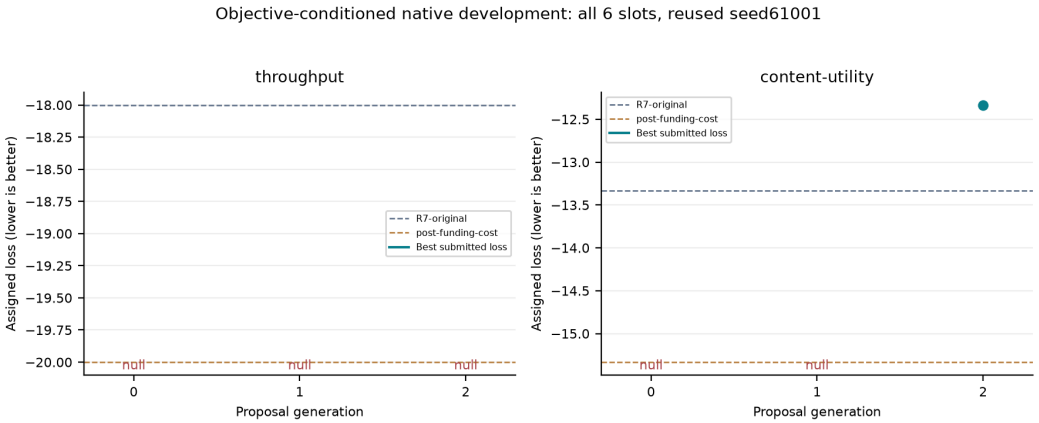


Figure 33: Complete own-loss paths in the open-institution v3 study. The T arm has no complete point or selected program; the U arm has one complete point, worse than both dashed reference losses. Red null labels near the lower plot boundary are annotations, not numerical loss penalties or points on the hand reference. No failures are replaced, and no independent error bars are inferred from the three generations of one objective history.

sponses. Equal scores therefore do not establish equal trajectories, and this cross-study comparison does not isolate a numerical score's effect.

An additional fixed-record decomposition uses every accepted occurrence in the three completed development worlds. R7 and U2 share 13 same-year/paper/content occurrences, with centered utility $28/3$. Each has five side-only occurrences: their utilities are four for R7 and three for U2. That difference accounts for U2's one-unit disadvantage without changing publication count. The hand and U2 share ten occurrences; their ten and eight side-only occurrences contribute eight and five utility units respectively. These are descriptive multiset differences, not causal extra or missing papers. Positive/zero/negative counts are 13/5/0 for R7, 15/5/0 for the hand and 12/6/0 for U2. No accepted record in these completed cells has negative centered opinion, and the hand dominates both other observed points. This archive therefore supplies no strict T/U tradeoff; it does not establish global objective alignment, an attainable subset optimum or a reason to alter the declared loss after observing results.

The U2 rationale claims its institutional score directly raises common-panel utility. The actual measurement instead presents title, abstract and topics to a separate fixed observer and sums panel means minus three over accepted decisions. Written institutional scores can affect subsequent scientist behavior but do not directly overwrite those panel inputs or ratings. The actual exact loss specification contains that formula. Its separate arithmetic view lacks a registered text definition for the newer common metric: a possible presentation gap, not a wrong loss or an identified cause of these failures.

Designer prose is neither a measurement nor proof of optimality. No prompt, loss or running policy is retrospectively amended.

The campaign records 358 complete scientist responses, 434,380 prompt and 37,362 completion tokens, with zero scientist wires without a committed reply. The common observer makes 90 physical calls, using 44,331 prompt and 14,146 output tokens. Six free proposals use 145,384 prompt and 35,802 output tokens; seven readiness calls use 165,705 and 16,316 respectively. Role-load token fields that are unavailable remain unknown. Shared-host science wall time is 3,590.225 seconds; cache subsets and nested durations are not additional provider tokens or dedicated-device performance.

Own-only selection leaves the T policy absent because every T slot is null. U2 is frozen as the only complete own U policy despite its disadvantage; no hidden improvement veto or replacement with a control is applied. After repeated pre-dispatch unused inventories, a separate evaluation starts on the prospectively declared seed 1516320444 with exactly R7, the hand and that frozen U policy. It makes no new designer call and does not reselect. The original evaluation normally closes all three jobs on November 14 at 11:19 UTC. There is no synthetic T-selected row.

Frozen system	T	U
R7 reference	-30	-25
Strong hand	-32	-26
U-selected (U2)	-28	-23

Table 29: All three physical policies in the open-code study's selection-fixed evaluation on unused seed 1516320444. Both losses are minimized; the generated U policy is worse than both references on both criteria. No T policy exists because all its own development slots failed. A held-out world condition does not create a new independently repeated optimizer history, confidence interval or reliable institutional advantage.

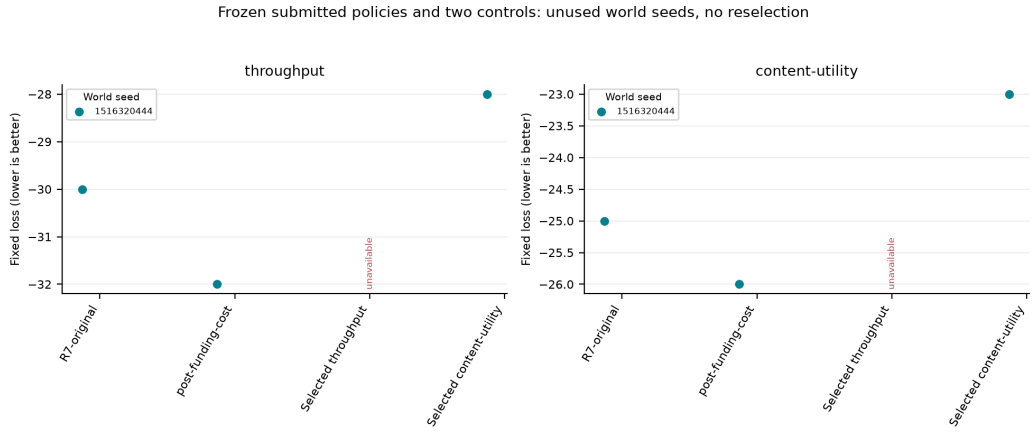


Figure 34: The complete frozen-policy comparison on the predeclared unused world condition. The missing T selection remains visibly unavailable, not assigned a baseline score. The U selection loses under both fixed criteria. One marker per physical policy/criterion is not an optimizer replication or an uncertainty interval. Both criteria are measured on the same native world.

The frozen U policy's differences from R7 are $+2$ in T loss and $+2$ in U loss; relative to the hand they are $+4$ and $+3$. Its written institutional score100 therefore does not produce the claimed common utility improvement on either observed condition. This is not an isolated causal effect of that score, a universal failure of code search, or permission to tune the objective and select again using the unused result. The missing T policy, all development nulls and the unfavorable completed policy remain part of the result. All five original read-only continuations exit normally; paired original-source audits, cost exports and result, choice, effect and full-request reports agree. The three completed worlds issue 156, 163 and 148 scientist requests respectively, forming three different full-input classes. In total, the evaluation records 467 complete scientist responses, 668,024 prompt and 40,433 comple-

tion tokens, with no scientist wire lacking a committed reply. The separate common observer makes 138 physical calls, with 62,169 prompt and 19,547 output tokens. Shared-host science wall time is 2,441.556 seconds; there are no new designer calls. These are bounded-study costs, not project-total compute or a dedicated-device benchmark. Neither normally closing this search nor completing software tests supplies an important independently supported objective-to-institution answer.

Staged construction is not yet an optimization result. Subsequent software work connects an optional draft, compile, component-test, revise and explicit-finish profile to the same native objective campaign and selection-fixed evaluation. Its ten SCRIPTED native worlds and 48 real container probes verify the public worker, cold audits, costs and handoff, with zero actual model calls. Tests use 20 uniquely attributed historical R7 inputs, not hand-control or unused-condition data. They report failures without repairing source or requiring a fixture pass before submission; new interfaces remain legal and may be untested. The declared six-slot profile permits at most 16 designer calls and 64 component probes per proposal, plus request-bound readiness sessions, rather than the old one-call interface. These are budget ceilings, not observed model usage. At this software-proof snapshot, its real-model plan had not passed the shared-resource gate or begun scientific dispatch. The subsequent actual study is reported separately in Section 23. No model-authored repair, new institutional advantage or matched causal effect of component feedback follows from the software proof.

23 Actual staged construction in the original scientist world

23.1 Free design, component observations and whole-world evaluation

We next execute the staged interface with the local 35B designer and original Qwen3-8B scientists, rather than attributing the preceding scripted software behavior to an autonomous author. Source `f47270b9` preserves the original corpus, persistent four-year world, fixed T/U losses, R7-only initial optimizer reference and comparison-only hand timing control. Three proposals per objective are interleaved T0,U0,U1,T1,T2,U2. Arbitrary mechanism source, nodes, ports, protocols, connections and schedules remain expressible in the general native interface. There is no novelty, fairness or improvement admission test.

The designer can inspect and edit its literal draft, observe no-network component execution on 20 historically recorded R7 inputs, revise and explicitly finish. Eligible edited code is tested against its four original yearly inputs; new interfaces remain legal but may be untested. Failed component returns do not veto submission. All component tests and HTTP calls share a finite worker allowance; the host neither repairs source nor supplies a missing consumer. Each proposal permits at most 16 model steps and 64 probes. Initial and per- proposal directed readiness tests remain separate from free-design evidence.

System/slot	T	U	Probes	Actual own edit
R7 original	−18	−13.3333	—	Reference
Hand timing	−20	−15.3333	—	Reference
T0	−18	−13.3333	8	Selection source
U0	−18	−13.3333	0	Literal retention
U1	−18	−12	16	Agenda source
T1	−18	−13.3333	4	Selection source
T2	−18	−13.3333	20	Selection source
U2	−18	−13.3333	16	Selection source

Table 30: All eight completed physical worlds in the actual staged study. Both losses are minimized; the hand is comparison-only. All 64 component attempts return, without an error-repair demonstration. No own graph adds nodes or connections. Display rounding does not enter selection. Five own outcomes tie R7 and U1 worsens content utility; none beats the hand.

23.2 Execution is not improvement or an invented graph

All six proposals and eight native worlds complete. Each proposal consumes 14 actual model steps, for 84 total, including rejected actions and explicit finish. The 64 probe returns do not constitute 64

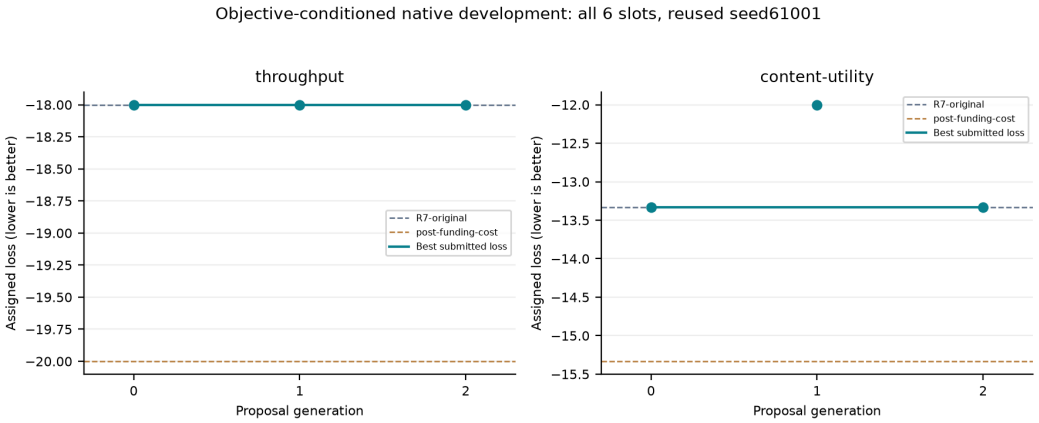


Figure 35: The actual objective-conditioned staged development paths on reused seed 61001. Markers are all three completed own proposals per objective; solid lines retain best submitted loss, and dashed lines show both controls. The U1 marker is worse while the best-own line stays flat. Neither objective improves over its optimizer-visible R7 reference, and the comparison-only hand remains better. This is one development history without independent error bars.

independent scientific evaluations. T0/T2/U2 change selection source, T1's selection AST equals R7, and U0 explicitly retains the literal parent. No new mechanism or connection is generated. This is a completed construction-and-evaluation path, not an automatic discovery of the hand timing rule or an isolated feedback benefit.

U1 changes agenda code to prefer smaller declared project duration with a keyword-count tie heuristic. The 23rd scientist request changes its research topic from distributed systems to software engineering. The two worlds accept 18 occurrences each: 12 shared, full-rating-identical occurrences contribute utility 8; the six R7-only occurrences contribute $16/3$, while the six U1-only occurrences contribute 4. Thus its U loss worsens by $4/3$ despite equal T. This diagnoses observed content replacement, not a causal advantage of a proxy for topic quality. The heuristic is authored code, not a changed loss.

The actual U1 planning frame contains the exact U arithmetic but lacks its inline semantic definition; the inspected T2 frame includes a count definition. The full measurement contract remains retrievable in the request archive, and U1's recorded reads do not request that archive. This is a bounded information-presentation limitation, not proof that a target was absent, an API was broken or that the omission caused the unfavorable result. The study does not isolate an equal-information objective intervention. Prospective interface corrections do not retrospectively repair its inputs or failures.

23.3 Complete costs, frozen selection and remaining evidence

Two original-source audits and cost projections agree byte for byte; all files in the paired result, choice, effect, request-class and construction exports also agree. These are reconstructions of one development history, not independent optimizer replications. Scientist caches contain 937 complete responses, 1,117,688 prompt and 98,457 completion tokens, with no input lacking a committed reply. The 190,032 cached prompt tokens are a subset. Separately recorded calls include 200 readiness, 84 proposal steps, 13 designer loads, eight scientist loads and 93 physical content-critic requests. Missing load usage stays unknown. Campaign wall time is 13,394.180 seconds on the shared host, including construction and readiness, not dedicated inference speed.

The unchanged own-only minimum/chronological-first-tie rule selects T0 for T and U0 for U. U0 is literally R7, not a fallback. Before own outcomes, a prospective note declares unused seed 1891355522; it is retained regardless of the absence of development improvement. Broad production-journal inventories and an immediate pre-dispatch repeat find no prior use, with only exact hash-bound nonproduction fixtures excluded. Four logical labels require three literal-distinct physical jobs, with no new author, repair or reselection. The original fixed evaluator normally closes

all three worlds; its two cold audits/cost projections and all paired result/choice/effect/request exports agree. No unused outcome is fed back to the designer or used to reselect.

Frozen label	T	U	Physical job
R7 original	-32	-28.6667	000
Hand timing	-36	-31	001
T-selected (T0)	-32	-28.6667	002
U-selected (U0)	-32	-28.6667	000 (literal identity)

Table 31: All frozen labels on the single predeclared unused seed 1891355522. Both losses are minimized. Each own selection ties R7 and is worse than the hand by 4 T-loss units and $7/3$ U-loss units. U0 is the same physical policy as R7, not another independently invented success. Display rounding is not used for selection; one condition is not an independent optimizer replication.

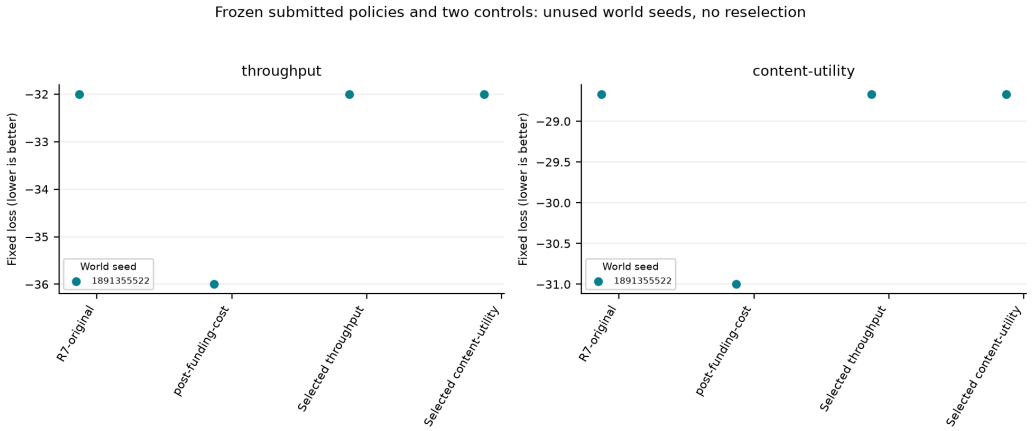


Figure 36: The complete selection-fixed comparison for actual staged construction. Four logical labels represent three physical worlds. Both selected systems tie R7 on both objectives and lose to the hand. The shared R7/U0 marker is not a repeated measurement or an uncertainty estimate. No successful own invention or objective-specific trade-off follows from these equalities.

The separate R7 and T0 worlds issue the same ordered 161 full scientist payloads and receive identical literal choices; provider bodies differ and remain unmodified. The hand issues 176 requests and forms a second observed input class. This is observed-history equality, not global program equivalence or a reason to skip required future evaluations. The three worlds record 498 completed responses, 757,983 prompt and 44,296 completion tokens, without a scientist input lacking a committed reply. The separate common observer makes 132 physical calls, using 59,418 prompt and 19,188 output tokens. Shared- host evaluation wall time is 3,381.233 seconds. No new design call occurs.

The apparent catalog saturation also requires careful scope. On the inspected original initialization, core-expertise-only assignment gives the conditional duration-only output envelope 20, whereas the original legacy direction parser admits direct related topics and gives the corresponding relaxation 36. The latter ignores resources, actual choices, retrieval and content; it is not an attained outcome or global optimum. A graph can stop preempting idle authors with catalog assignments and restore native direction choice. Thus 20 cannot explain failure over the whole permitted institutional space as an impossibility.

Completing all construction slots removes an important technical missingness in this study, but does not supply an important general objective-to-institution answer. Both selected programs still lose to the hand on the unused condition. The negative own result, absent topology exploration and unequal inline information coverage remain explicit. Simply repeating this question at larger scale would not by itself establish main-track significance.

Feedback	Assigned loss	G0	G1	G2
Scalar	L_T	-18	-18	null
Parent trajectory	L_T	-18	-18	-18
Scalar	L_U	-13.33	-13.33	-13.33
Parent trajectory	L_U	-13.33	null	-13.33

Table 32: All twelve scheduled own slots; smaller loss is better. Utility is rounded here from -13.333333333333332. Ten measured assigned losses tie the original reference. The hand comparison achieves $L_T = -20$, $L_U = -15.33$. Nulls are unknown losses after native runtime failures, not assigned penalties or silently replaced samples. One reused development condition supplies no independent error bars or isolated information-condition effect.

24 Matched Objective and Feedback-History Comparison

Question, authority and finite design. The next study compares scalar feedback with access to the parent's native trajectory under throughput loss L_T and common-content utility loss L_U. The searched object remains the organizing institution: both information arms can edit executable rules, mechanisms, wiring and schedules. They retain the original four-year persistent researcher world, corpus, fixed Qwen3-8B researchers and approved local 35B designer. Each of four own histories receives three proposal slots, for twelve total. Seed 61001 is reused across this development comparison; neither candidates nor feedback steps are independent optimizer replications. The original reference is optimizer-visible; the developer-authored post-funding-cost comparison is withheld from the designer. Here R7 is the developer-authored seed institution, not the published simulator's unmodified default institutional configuration; the original native researcher engine is retained while the organizing institution is editable. Losses, technical validity and held-out evaluation remain separate concepts.

Closed paths, without improvement. All twelve proposals are submitted; ten complete native measurements and two fail at the publication consumer. There is no generation error, host source repair, failed-slot refill, novelty requirement, fairness veto or loss change. Every measured assigned loss ties the original reference and is worse than the hand comparison by 2. Nine own candidates share the original cross-score vector, $L_T = -18$, $L_U = -13.333333333333332$. Final parent-trajectory T2 instead has $L_T = -18$, $L_U = -12.333333333333332$: matching its assigned throughput target does not mean unchanged cross-objective behavior or improved utility.

The finite measured matrix contains only three distinct T/U vectors: the reference plus nine own worlds, the hand, and final parent-trajectory T2. The hand strictly dominates every other measured row in both recorded losses. Any nonzero nonnegative linear weighting of these two losses therefore keeps it best within this measured set. This does not identify a global or expected Pareto frontier. Nor does the absence of a preference reversal establish that objectives do not matter: the search has not generated a measured tradeoff on which to test objective-conditioned institutional choice.

Source diversity is not behavioral diversity. The complete journals contain 168 charged author steps and 108 component probes; every probe returns, but those producer probes do not execute the complete downstream publication contract. All 27 document-changing operations address `world.selection`. Ten submitted systems change selection source relative to the reference, while two retain it literally. None adds or removes a node, connection or schedule or changes another system field. The existing file API supports complete authored nodes and root-level composition edits, so these results do not establish that structural invention is unavailable in the representation. They show that this finite local search did not produce it.

Final-checkpoint inspection finds zero rejected records in every completed world: 18 accepted in the reference and each valid own world, 20 in the hand comparison. Accepting every paper already saturates immediate throughput on each actually recorded decision batch. Concentration in selection edits thus targets a stage with no immediate fixed-batch acceptance headroom. In contrast, the hand changes timing/routing and the eventual submission volume. This is not a global throughput bound or proof of the author's search bias: another decision could alter later information, resources or submission batches.

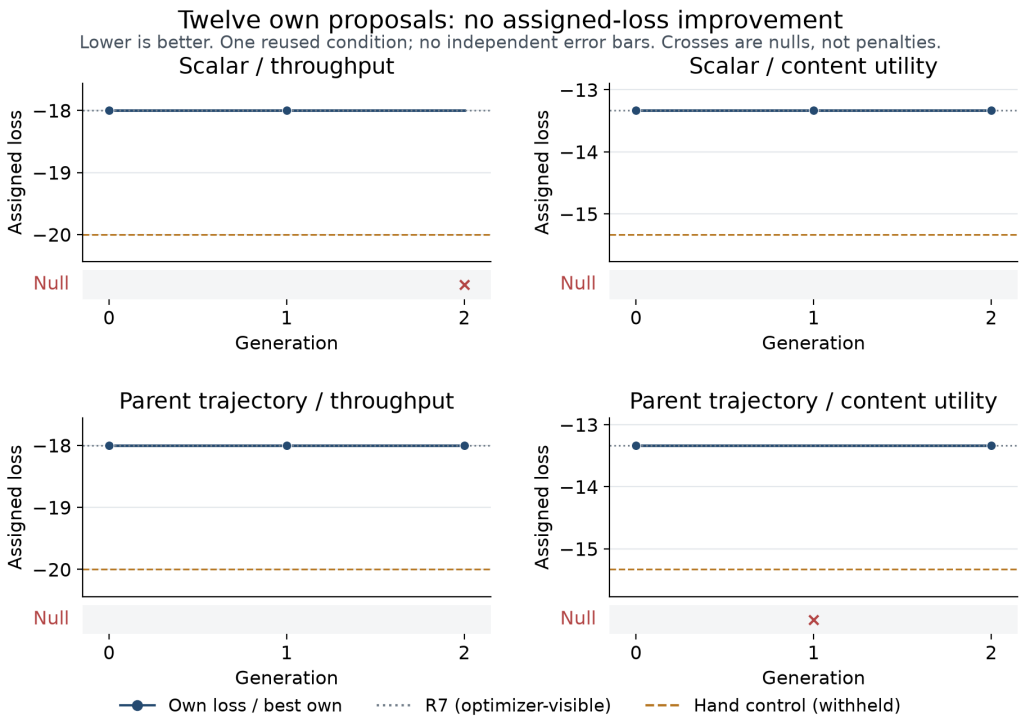


Figure 37: Complete assigned-loss paths in the matched comparison. Dots are completed proposals; lines carry forward the best own measured loss. Crosses occupy a separate nonnumeric null strip, not the loss axis. The original R7 reference is optimizer-visible; the hand comparison is withheld. Cross-objective worsening of final parent-trajectory T2 is reported in the text, not plotted as its assigned throughput loss. One reused condition gives no independent optimizer-replication error bars.

Cold comparison of full scientist request payloads, rather than source hashes or scalar losses alone, places the reference and nine valid own worlds in one exact observed-input class. Their literal response choices also agree; full provider envelopes do not all agree. The last valid proposal and hand comparison each have a different full input sequence. Failed worlds do not become complete path-equivalence classes. This is observed-path equivalence under a shared condition, not global policy equivalence, independent replication or permission to deduplicate nonidentical systems in subsequent evaluation.

An operative score route without an assigned-loss gain. The final valid proposal accepts real submitted papers with constant `final_score`=1. Its first 70 scientist request payloads match the reference. Request 71 (index 70), a year 4 round-intention prompt, first differs in delivered own-history dossier review scores and native past-paper scores, 0.0 versus 1.0. The complete text difference consists of six changed digits: four dossier fields and two *Your Past Papers* scores. The actual consumer copies the authored final score into the paper; publication archives it in review history, and both the information office and native past-paper prompt renderer read that archive into researcher context. Thus a score absent from the direct critic-loss formula can still be operative through future information. Declaring it inert from the loss formula alone would miss this path.

Both worlds accept 18 content occurrences. Fourteen common year/content pairs have identical retained content and critic values. Four reference-only pairs contribute 4 centered utility units, whereas four last-proposal-only pairs contribute 3, accounting exactly for the utility worsening by 1. Annual author accounting, final-active count 2 and final resources 40 match. This is a decomposition of one coupled trajectory, not a score-only average causal effect, scientific truth measurement or policy-general welfare conclusion.

Frozen label	L_T	L_U	Physical job
R7 reference	-34	-25.67	000
Hand reference	-36	-30	001
Scalar / T selection	-34	-25.67	002
Scalar / U selection	-34	-25.67	000
Parent trajectory / T selection	-34	-25.67	003
Parent trajectory / U selection	-34	-25.67	004

Table 33: All six frozen logical labels on the single reserved unused condition. Job 000 is shared only because the complete policies are literally identical. The other jobs execute separately despite their observed input equality. Utility is rounded from 25.666666666666668. These are not six independent replications, and the condition never enters the historical optimizer bank.

The two nulls expose a producer/consumer boundary. Parent-trajectory U1 returns a scored empty-cycle marker instead of the exact explicit empty decision batch. Scalar T2 double-wraps already wrapped decisions. Both reach and fail the unchanged native empty-cycle consumer contract. Their losses remain unknown. Returning successfully from a component probe is not whole-world validity. Regression cases cover the observed shapes without repairing these historical candidates or unwrapping their decisions on the host.

Costs and frozen evaluation. The original runner closes normally after 5h56m39.98s on the shared host. This is not a dedicated inference benchmark or the project's total cost. The cost reader records 1,391 completed scientist responses and 84 separate physical common-content observer requests, all 84 with known terminal attempts. Auxiliary groups retain 372 editor-readiness attempts, 168 staged proposal attempts, 25 designer-load attempts and 14 scientist-load attempts. Readiness and proposal transport intervals total 6,401.32s and 5,708.03s respectively; they are operations inside, not additions to, overall elapsed time. One of 168 proposal records lacks the final-response flag despite a retained HTTP record; HTTP completion is not conflated with explicit finished designs or native worlds. Missing usage is not imputed, overlapping durations are not added and response records are not treated as additional independent scientific observations.

Two original-source cold audits and cost reports match byte for byte. Each of five paired readout file sets and the paired allocation reports also matches. These are reconstructions of one experiment, not repeated inference. The original own-only minimum and earliest-tie rule selects generation 0 in every history, with no hand-policy fallback. Reserved seed 676047151 supplies the predeclared unused condition after broad source-bound usage inventories. Four frozen own selections and two references share five physical jobs solely through literal system identity, with zero new designer-generation POSTs.

Closed unused-condition outcomes. All five physical jobs complete normally on reserved seed 676047151 (Table 33; Figure 38). The original reference and every frozen own selection yield 34 accepted records and centered content-opinion utility 25.666666666666668; the hand yields 36 and 30. Thus each selected institution ties R7 but has loss differences of $+2$ on T and $+13/3$ on U against the hand. No candidate is selected again, repaired or substituted using these results. All five cold initial physical-society identities agree. The four distinct own/R7 physical worlds also share an exact full scientist request sequence and literal response choices; complete provider envelopes are not all equal. Sharing the one literal-identical R7/scalar-U job is a separate fact from this observed-path equality.

The hand/reference content difference is not merely two additional records. Their 28 common year/paper occurrences have identical frozen critic values and contribute $68/3$ utility units. Six R7-only occurrences contribute 3; eight hand-only occurrences contribute $22/3$. The net utility difference is therefore $22/3 - 3 = 13/3$. This exact fixed-record decomposition uses shared content ratings, not new critiques or scientific truth. The hand changes timing and information routing jointly, so this is not an isolated timing-effect estimate.

The fixed runner takes 1h09m35.57s on the shared host, separately from development. It records 825 completed scientist responses, 126 physical common-observer POSTs and five empty scientist-load attempts; there are zero designer-generation POSTs. Scientist provider usage reports 1,305,668 prompt and 78,658 completion tokens (1,384,326 total); 191,272 cached-prompt tokens are a re-

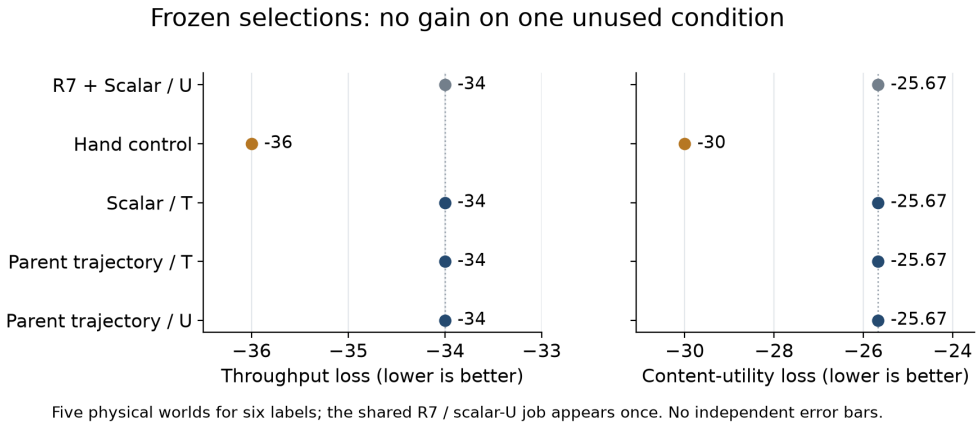


Figure 38: Frozen evaluation shown once per physical job. R7 and the scalar-U selection share the top row; the remaining selected programs execute separately. Lower loss is better. The additional trajectory-access condition does not yield an observed gain here. This single unused condition is neither an independent optimizer replication nor a statement of expected superiority or equivalence.

ported subset, not extra tokens added to that total. Two original-source audits and cost reports match, as do paired score, full-input, project-choice and mechanism-effect readouts. No successful reconstruction creates another independent scientific sample.

Unused evaluation does not revise the earlier selections or establish a causal benefit of trajectory access. These tied and adverse outcomes locate the limits of this finite search without showing that objectives never matter, that broader institution search is impossible, or that the project is main-track publication-ready.

25 Original Researcher Choice and Resource-Use Diagnostics

25.1 A completed three-control native-choice pilot

Question and authority. The preceding searches did not find a useful structural change. A separate developer-control study asks whether removing direct agenda assignment and supplying executable institutional advice exercises the original researcher choice path. It is not another optimizer run. All three institutions retain the original four-year world, five founders, 103,107-paper offline corpus, fixed Qwen3-8B scientists, original T/U measurements and development seed 61001. Their independently audited cold initial physical-society identities agree. The known seed is reused; this is neither unused evaluation nor independent optimizer replication. There are zero designer-generation calls.

The first control is the existing strong post-funding-cost institution with core-topic agenda assignment. The second cuts that existing project-assignment actuation and leaves topic choice to the original native researcher routine, which permits core and related topics. The third adds a code-defined information office and an addressed context receiver: two nodes and two edges. Its actual sandboxed code recommends the shortest original eligible duration, with lexical topic tie-breaking. The host does not force that recommendation, change original durations or introduce a new topic-assignment API. A larger nominal choice set is not evidence that a researcher uses it effectively.

Advice delivery, topic choice and calendar are different observations. All ten advised direction inputs contain current-year advice; ten addressed messages are stored. Exact recommended-topic matches are 0/10, whereas matches to the recommended minimum duration are 2/10. The text asks researchers to consider a topic, not obey a mandatory command, so exact mismatch does not establish general disobedience. Both unforced conditions make ten native direction requests with uniquely matched saved direction memories and no host project-assignment effects. Their full scientist input sequences differ from each other and from direct assignment.

Developer control	L_T	L_U	Project effects	Direction requests
Direct core assignment	-20	-15.33	10	0
Original native choice	-14	-10	0	10
Native choice with advice	-16	-12	0	10

Table 34: Every scheduled control completes. Smaller losses are better; the first utility is rounded from -15.333333333333332. Advice improves both losses by 2 relative to unadvised choice, but loses to direct assignment. These hand controls do not establish a newly invented policy, an expected advantage or a global optimum. Both T and U prefer the same control in this finite set.

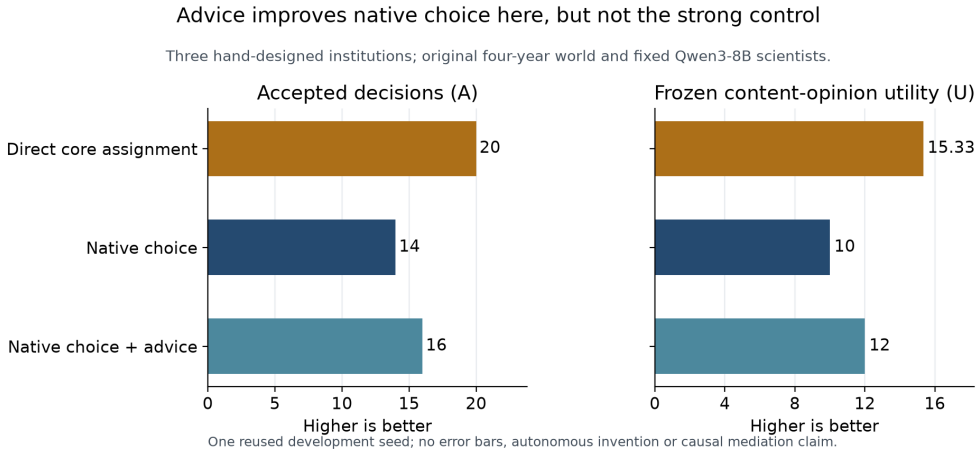


Figure 39: Actual output and fixed content-opinion utility for the three completed controls; both are plotted positively, unlike their negative losses in Table 34. One coupled condition supplies no independent error bars. Neither information delivery nor improvement over an unadvised control establishes superiority over direct assignment.

No one-year topic is selected. All twenty observed selections occur in the displayed candidate table, but their submitted topic schemas are unrestricted strings without a topic enum. An enum-forced explanation is therefore not supported for these actual requests; this is not proof of absence of every restriction or an identified cause of the choices. The initial actor ending 0002 changes from three-year `formal_methods` to two-year `performance_analysis`; subsequent calendars also differ. Other altered information and choices prevent attributing the output difference solely to this actor or to duration.

Native choice and advised choice share six accepted year/paper/content occurrences. Eight are native-only and ten are advice-only. Fixed retained critic opinions give utility totals 10 and 12: the difference of 2 is not simply the same fourteen records plus two more. Relative to direct assignment, native choice shares three occurrences and advised choice shares four, with the remaining content different. This is exact record decomposition, not a scientific-truth measure or isolated causal estimate of the advice mechanism.

Completion and inference costs. The real pilot closes normally in 34m37.04s on the shared host. It records 386 completed scientist responses (127/127/132 by control), 120 separate physical common-observer POSTs with known terminal attempts, and three empty scientist load attempts. Scientist usage is 467,505 prompt plus 40,304 completion tokens, 507,809 total; 97,796 cached prompt tokens are a subset, not extra tokens. There are no wire-only scientist reply gaps. Observer durations are nested inside elapsed time and not added again. These are provider counts, not FLOPs, energy, monetary cost or a dedicated machine benchmark. Two original-source audits and cost exports match, as do all nine paired choice/request/effect readout files. Readout agreement is reconstruction, not repeated inference.

Original-choice condition	One-year choices	Two-year choices	Three-year choices
Native choice	0	4	6
Native choice with advice	0	6	4

Table 35: Original catalog durations of the ten observed direction requests per condition. These are events along two evolving trajectories, not twenty independent randomized researcher trials. A topic change can change later request timing, so the two rows do not identify duration-only mediation.

Closed development study	Observed system(s)	A	K
Longitudinal	R7 and all ten valid own policies	18	190
Longitudinal	Strong hand control	20	200
Choice pilot	Direct core assignment	20	200
Choice pilot	Original native choice	14	200
Choice pilot	Native choice with advice	16	200

Table 36: New retrospective resource measurements, not a new objective-driven search. The two failed longitudinal proposals keep unknown K. All measured activity-uncollected costs are zero. The studies remain separate; coincident points do not represent independent optimizer replications.

25.2 Collected activity resources in the original world

A passive, prospective-capable instrument. Inference costs above are distinct from resources inside the researcher world. An additive native measurement adapter now reconstructs collected scientific activity costs from year 0 and every declared annual checkpoint:

$$K = \sum_{y=1}^{Y} \sum_{a \in \mathcal{A}_y} \left(c_{ay}^{\text{research}} + c_{ay}^{\text{production}} + c_{ay}^{\text{resubmission}} + c_{ay}^{\text{application}} \right). \tag{14}$$

Here $\mathcal{A}_y$ contains every accounted native author, including inactive actors and later entrants. The terms are actual collected ledger costs, not requested charges or an active-count approximation. Grant/industry income and both sides of institutional transfers are excluded: reallocating resources does not consume them. Requested but uncollected activity costs are reported separately and are never imputed as collected consumption or made a veto. Units are model-internal resources, not national expenditure, compute or welfare.

The instrument checks the complete declared horizon, fixed system/world contract, literal annual actors, cumulative ledger prefixes and between-year balances, then invokes the native closed-year ledger reconciler. A complete known empty ledger can measure zero. Missing or inconsistent ledgers and partial runs have unknown cost, not zero or a numerical penalty. Live and cold paths append metrics without altering existing measurements, researchers, RNG or raw files. The existing public native loss/observation boundary accepts the additive adapter; software checks use scripted models/corpus, not new scientific runs.

What the new measurement does and does not change. Two byte-identical cold exports per study use the original scientific checkpoint decoder and resource-ledger reconciler with matching original audits, fixed scientific sources and unchanged complete raw inventories. Twelve of fourteen longitudinal worlds measure; two remain unknown. All three choice controls measure. No new model call, native scientific world or institutional code execution is performed, and unused evaluation is not newly rescored.

Unlike the observed T/U matrix, the fixed longitudinal R7/hand pair trades output against resource consumption. A hypothetical new loss $L_{T,K} = -A + \lambda K$ has equal values for that pair at $\lambda = (20 - 18)/(200 - 190) = 1/5$, in accepted decisions per native resource unit. This algebra does not adopt that weight, validate a policy objective, find a global frontier or show optimization under a new loss. Neither T/U nor their historical selections are changed. A new objective-conditioned search would need its own declared meaning, instrument, same-authority strong comparisons, finite budget and selection-fixed fresh evaluation. Resource-saving attrition and other undesirable consequences must remain visible rather than being removed with an undeclared fairness or improvement filter.

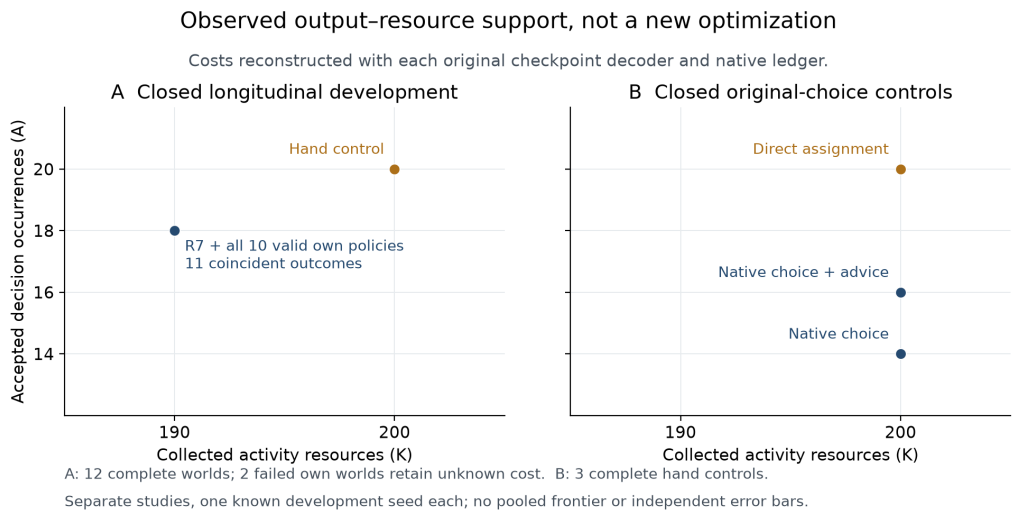


Figure 40: Separate observed output–resource supports. The longitudinal reference and ten valid own worlds coincide at $(K, A) = (190, 18)$; its hand comparison is at $(200, 20)$. Every choice-pilot world consumes 200. No pooled or expected frontier, independent error bars or new optimized design is inferred.

Developer candidate rule	L_T	L_U	K	Projects	Outside core
Core only	-20	-15.33	200	10	0
Core plus directly related	-36	-28.67	200	18	18

Table 37: Both same-authority controls complete on one reused condition. Utility losses round the retained values -15.333333333333332 and -28.666666666666668. Smaller losses are better. The broader hand rule improves both observed objectives at equal collected native activity cost, but this is not an automatically invented institution, an expected advantage or a new objective-conditioned search. Projects are actual assignment effects along two evolving trajectories, not independent trials.

26 Original-Topic Authority and Three-Objective System Search

26.1 A same-authority developer contrast in the original world

An interface distinction, not a new research law. The original unforced researcher can choose core or directly related topics, whereas the earlier typed direct-assignment action permits only core expertise. This mismatch is local to the historical fixed-effect design domain: the general engine already has a broader project API, excluded there because it also permits arbitrary durations and ongoing-project changes. It would be incorrect to conclude that the platform cannot express related-topic projects. A new explicit action exposes the original eligible union while preserving the original catalog contents, durations, idle-researcher preconditions and actual native project consumer. Historical actions and domains stay unchanged.

In a separate completed pilot, both developer systems receive the same new action permission and the same source, graph, schedule and resource authority. Both use the hand-written shortest-original-duration rule with lexical topic tie-breaking; only candidate membership differs: core-only versus core plus directly related topics. This rule belongs to each submitted institution, not to a host-enforced search restriction. The four-year, five-founder society, 103,107-paper corpus, fixed Qwen3-8B scientists and original T/U observer remain. The reused seed 61001 is development data, not an independent replication. Two original-source cold audits verify equal initial physical societies and 48 original native-effect witnesses per world. No designer generates a policy.

Actual calendar, content and resource paths. Core-only assigns ten two-year projects. The broader rule assigns sixteen one-year and two two-year projects, all eighteen outside core expertise;

no original duration is shortened. Accepted year/paper/content multisets share zero records, with twenty core-only and thirty-six broader-only occurrences. The exact utility difference is $40/3$, not the same twenty records plus sixteen more. The first full scientist request already differs at index 0 in year 1, so this contrast does not isolate duration-only mediation.

Each original annual ledger collects 50 research-resource units, for total $K = 200$ in each world. Collected production, resubmission and application costs, and separate uncollected activity costs, are all zero here. Equal K is not equal inference compute, monetary expenditure, research quality or welfare. Both original objectives prefer the same hand rule; the experiment supplies no preference reversal or global optimum.

Conditional calendar headroom. A read-only relaxation of the completed trajectories asks how many new paper slots the original calendar permits, not how a new scientist would behave. For fixed founder i, let d_i be the minimum original duration in that institution's permitted candidate set, $H = 4$ the horizon and $P = 2$ the native paper slots per completed project. Under sequential projects, no initial project and no repeated acceptance, the calendar relaxation is

$$B_{\text{calendar}} = P \sum_{i=1}^{5} \lfloor H/d_i \rfloor . \tag{15}$$

Core-only has durations $(2, 2, 2, 2, 2)$ and bound 20; directly related topics give $(1, 1, 2, 1, 1)$ and bound 36. Both observed acceptance counts reach their respective relaxation. This is conditional saturation of a fixed calendar, not a global bound for arbitrary institution-workflow designs: changing actors, effects or support invalidates these assumptions. The relaxation also ignores funding, retrieval failures and content generation and supplies no utility or resource-cost bound. Core-only is a submitted rule's restriction, not the broader shared authority of the two institutions.

Fixed content support and tied objectives. All 56 accepted occurrences in these two recorded worlds have nonnegative centered observer utility: the core pool has 15 positive and 5 zero records, and the related pool 30 positive and 6 zero records. Their exact rating sums are $46/3$ and $86/3$; original floating loss values are retained rather than replaced by a differently rounded rational conversion. If one only filters each already generated fixed pool, accepting all records maximizes count and utility; zero contributions yield respectively 2^5 and 2^6 utility-optimal subsets. These are combinatorial subsets, not 32 or 64 native worlds. Removing an acceptance in a real world can change later funding, activity and content, so this diagnostic is neither a deployable oracle nor a world-level optimum. At fixed K, the new resource-aware objective below orders these systems just as utility does; a new objective is not required to produce a different winner.

The original pilot closes in 31m13.34s on the shared host, with 306 completed scientist responses (127/179),168 separate common-observer POSTs and two scientist role-load attempts. Scientist usage totals 458,048 tokens: 428,849 prompt and 29,199 completion; cached 68,876 prompt tokens are a subset. Critic usage is separately 78,177 prompt and 24,728 output tokens. Load usage is unknown, not zero. There are zero designer-generation POSTs. Matching paired cold audits/costs and all three readout files reconstruct the original records; they do not replicate inference or establish an exclusive machine.

26.2 Free-system optimization under three declared losses

Purpose and prospective boundary. The hand-rule gain motivates a stronger visible starting bank, not another weak-comparator claim. A separately frozen study evaluates two fresh references and schedules six own proposals: two each for original throughput, original content-opinion utility and a prospective resource-aware content loss,

$$L_T = -A, \qquad L_U = -U, \qquad L_C = -U + 0.1K = L_U + 0.1K. \tag{16}$$

Here U is the original centered fixed-panel opinion, not truth. The coefficient is a developer preference exchanging ten native resource units against one opinion unit; it is not an estimated social value. Earlier resource diagnostics were observed when choosing this new objective, which is disclosed. The new loss is fixed before these free proposals and never substituted into historical selection. Missing cost or outcome remains unknown. No fairness, activity, novelty, improvement or promotion veto is added.

Assigned objective	Generation	L_T	L_U	L_C	Native status
Throughput	0	-36	-28.67	-8.67	Complete
Content utility	0	-20	-15.33	4.67	Complete
Resource-aware content	0	—	—	—	Runtime null
Resource-aware content	1	-36	-28.67	-8.67	Complete
Content utility	1	-36	-28.67	-8.67	Complete
Throughput	1	-36	-28.67	-8.67	Complete

Table 38: All six consumed own slots in the original order. All five complete outcomes have $K = 200$ and no additional institutional work calls. Rounded utility/resource losses correspond to the literal retained floating values. The strong visible related-topic reference has the same losses as rows 1/4/5/6; row 2 matches the weaker core-only reference. The null has no completed loss or resource measurement; missing cost is not zero. No failed slot is refilled or replaced by a reference.

A free organizing system, within an explicit native-effect scope. The new institution-workflow domain permits authored Python, new mechanisms, ports, protocols, wiring, schedules and bounded internal processes. It also permits institutions to assign typed work to existing fixed-model original researchers and consume their response artifacts. Exact tasks, responses and canonical receipts are recorded. Work does not itself publish papers, assign review scores, overwrite ongoing projects, mint resources or train weights. The original native research, production, funding and resource effects remain fixed; historical domains are not silently widened. This finite study's once-per-year outer invocation is not a universal engine limitation.

Both developer references are visible in full to every objective's author. The current minimum-loss measured reference or own system is the parent. Only explicit own completed submissions enter the final own selection; a missing proposal is not replaced by a reference. Unchanged-parent submissions are legal and labelled. The order is T,U,C,C,U,T, with optimizer seed 68011; these histories are not independent optimizer replications.

The approved local 35B designer uses sessions bounded to 16 model calls each, with 8192 output tokens and 64k context. Fresh source/model/backend/request-bound readiness precedes initial references and each proposal. Directed-control success is not free-design capacity or invention evidence. Optional component probes return actual errors to later turns but are not a finish veto and do not trigger host code repair. Twenty original developer inputs are development fixtures, not fresh worlds or held-out evidence. Scientist models remain Qwen3-8B, with all additional work counted against the same 1024 HTTP-attempt bound per world. The new activity loss does not charge LLM inference compute; that separate per-role cost must remain visible.

The maximum is eight physical worlds, six consumed own slots, 376 designer POSTs and 384 component probes, with finite declared transport/session/world allowances and no failed-slot refill. Transport completion-unknown or failed, stale or unknown readiness stops inference. Every technical null and adverse outcome remains recorded. A prior full public software construction exercises all eight new native worlds, six prescribed programs, twenty-four institutional work calls and matching paired cold/cost reconstruction, but HTTP, corpus, critics and child supervision are SCRIPTED. It establishes neither real-model capacity nor autonomous discovery.

All six slots close without a gain over the visible strong reference. The real study starts 00:11 JST November 16 from frozen `48e1d268`. Initial directed readiness passes all 28 controls; exact-copy equality remains diagnostic. It closes 03:30:59 JST with both references complete and all six own slots consumed: five own native worlds complete, and one fails before a completed measurement. The finite protocol, loss specifications and scientific source remain unchanged. Two source/physical audits and two cost exports match, followed by two identical seven-file readouts. This closes the declared development comparison, not an independent optimizer replication or new unused-condition evaluation.

Throughput generation 0 makes five writes to existing tensor parameters, but no source, nodes, wires or schedules. Its complete 179 scientist request payloads and literal response choices equal the strong parent on this trajectory. The author's claimed score-to-selection actuation is contradicted by the actual consumer: that original selection branch ignores the delivered score input. This is an observed path and consumer finding, not global program equivalence. Content-utility generation

0 instead edits the agenda source to remove related topics. Its 127 complete requests and choices equal the weaker core-only path, and all three losses worsen relative to its strong parent. Metadata still labels the rule core-plus-related, but the new code does not read that label. Actual consumers and full inputs, not names, determine the exercised policy.

Resource-aware generation 0 expands two-hop topics. Its code returns canonical catalog directions and original durations, but four of five first-year assignments lie outside the original actor's core-plus-directly-related support. The existing native validator correctly rejects the first invalid assignment. All 16 component probes return because their output contract does not execute the native command consumer; returned is not accepted. Read-only application of the same original pure validator to the final four fixture results finds 4/0/4/0 invalid assignments. This is additional diagnosis, not feedback that the active author had received. The broad runtime error text mentions content and duration but does not establish either was modified.

Resource-aware generation 1 is the predeclared second slot, not a refill. Fourteen completed editor steps include one rejected tool-call envelope and later explicit finish; eight component probes return. Its final source differs only in continuation indentation and has the strong parent's AST. The native world completes and ties that parent on all three losses, with unpaid cost 0, five active authors and final resources 30. An AST match is not full input equality, and fourteen editor steps are not inferred charged model-call totals.

Content-utility generation 1 explicitly retains the complete strong parent. Its fourteen steps contain a plan, three successful reads, nine rejected ordinary reads at the existing implementation checkpoint, and finish, with no edits or component probes. Every actual proposal wire supplies that exact complete parent; all fourteen provider responses end normally, not at length. The maximum reported prompt is 35,960 tokens, below the declared 57,344 limit. This rules out missing parent source and observed length termination for this session; it does not identify the effect of checkpoint design, context size or model ability. The native outcome completes and ties the strong parent.

Throughput generation 1 edits only the selection source to hardcode accept-all, already the parent's configured rule. Eleven completed steps include a rejected plan and edit, two file writes and explicit finish; eight probes return. Its new empty-cycle sentinel additionally contains decision=accept. Two actual recorded developer fixtures have no known papers, and the trusted original consumer rejects those saved output values, whereas the original parent sentinel returns. This is a conditional incompatibility, not a failed observed world: the actual related-topic trajectory completes and ties the parent. Submitted code is neither repaired nor declared globally equivalent.

All four strong-tied own worlds and their reference have identical 179 full scientist request payloads and literal choices on this condition; U0 and the core-only reference have identical 127-input paths. Full provider envelopes differ. No own system adds nodes, connections, schedules or additional institutional scientist work. The observed source/input/objective map is in Figure 42. None of the six slots demonstrates a useful new local institution or a gain over the strong known rule.

Frozen selection, complete construction and separate costs. Own-only selection fixes T0 (also tied with T1), U1 and C1, with no reference fallback. All three selected losses equal the strong reference. U's own path improves from -15.333333333333332 to -28.666666666666668, while its reference-inclusive incumbent is flat (Figure 41). The raw selection and all eight exact SystemSpecs, including failed C0, are retained so collaborators need not regenerate the author's proposals. No new unused condition has been evaluated for these frozen systems.

There are 81 complete editor steps and separately 81 actual charged proposal POSTs, 199 readiness POSTs, 19 applied file writes and 56 component probes, all of which return. Counts reconcile against complete chronological journals, terminals and physical wire records, not each other by assumption. The 81 steps include 48 ordinary read attempts (thirteen rejected), five applied plans, one rejected plan, nineteen applied writes, one rejected edit, one invalid response and six explicit finishes. Applied writes need not change the document, and returned probes need not be accepted native commands.

The writer takes 3h19m28.15s on the shared host, including references, per-request readiness and free proposals, not just native evaluation. This is not a dedicated model-speed, GPU-FLOPs, energy or monetary benchmark. Paired original-source audits cover 2,124 raw files. Paired graph/input/loss

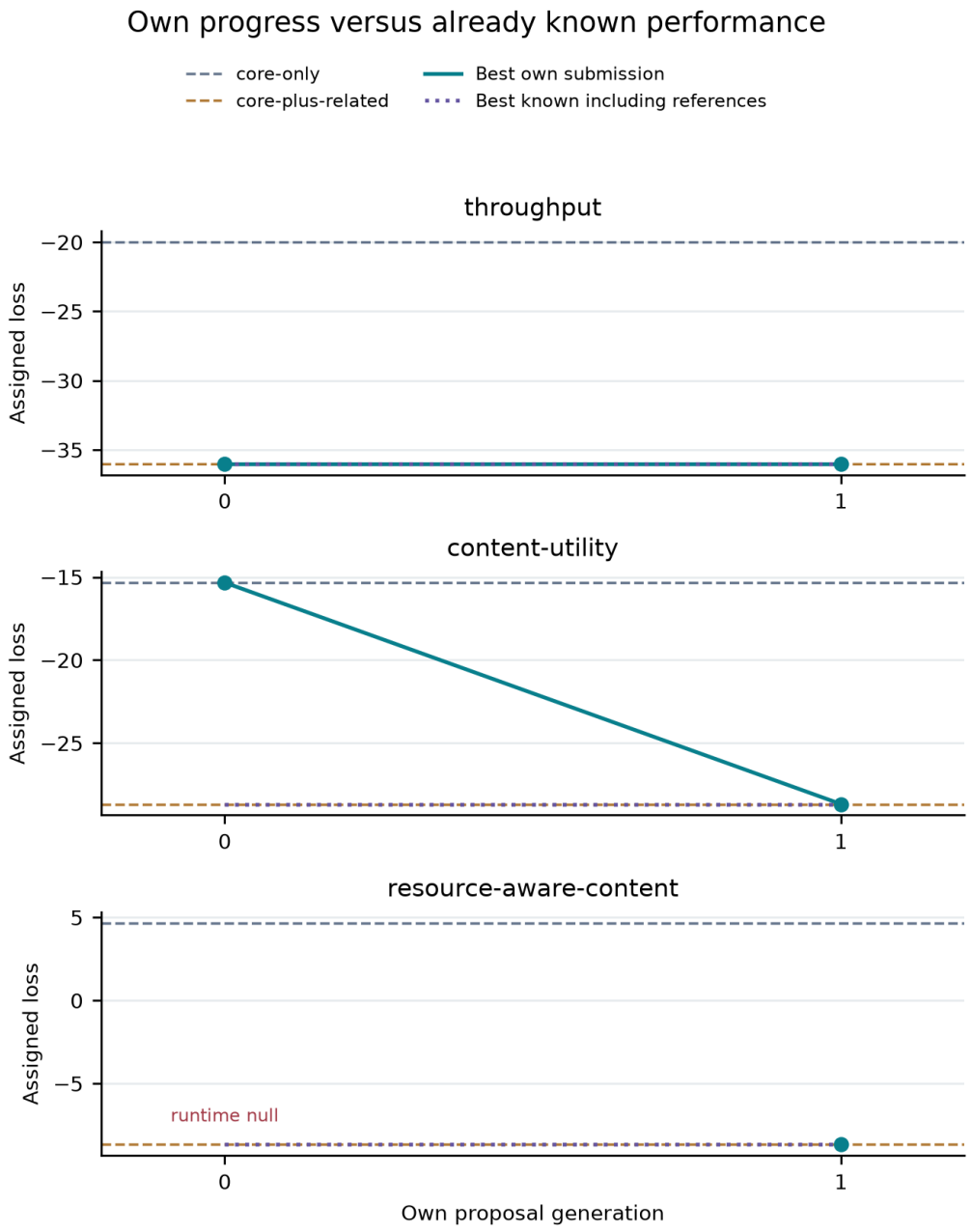


Figure 41: Best own submissions versus reference-inclusive incumbents under all three declared losses. U improves its own result by returning from a weaker reconstructed rule to explicit strong-parent retention, but best known is already at that value. C0 is a retained runtime null, not a zero loss or an extra trial. The selected T/U/C systems all tie the strong visible reference. One reused development condition; no independent optimizer replication.

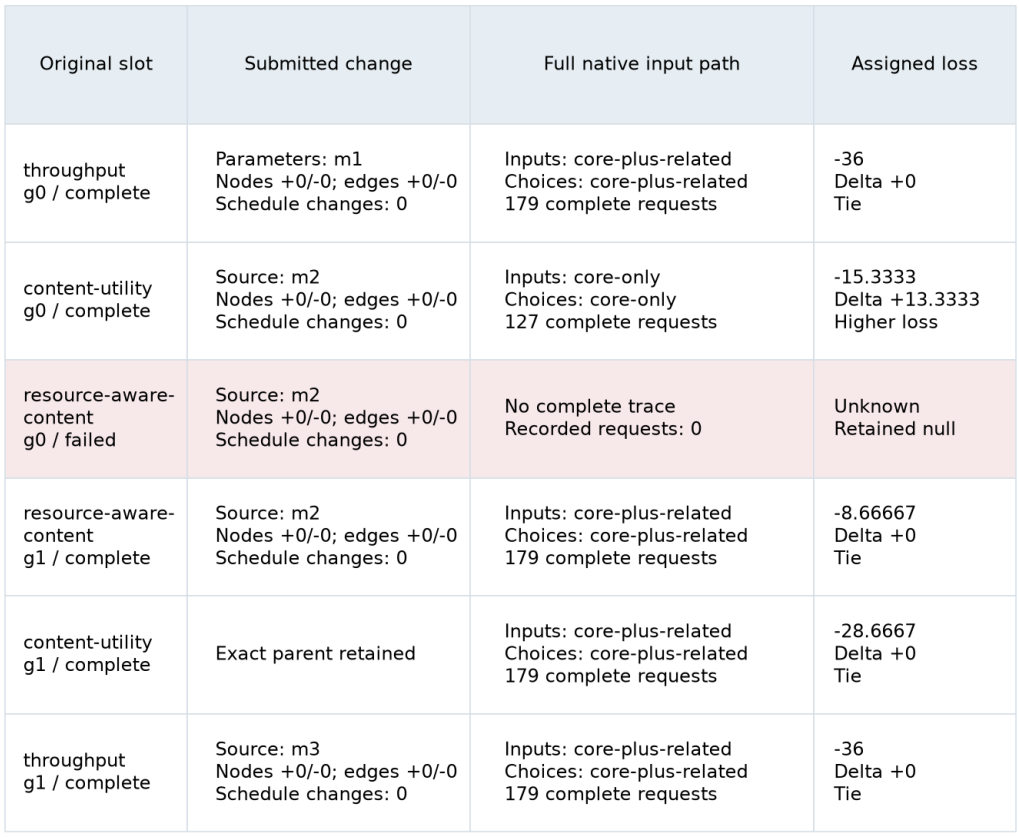


Figure 42: Literal edits, complete observed model-input/choice matches and assigned losses are different evidence. C1's changed source is only continuation indentation; U1 retains the exact parent. The C0 null remains visible. Matching reference inputs are not global program equivalence or causal mediation; in particular T1 has a recorded empty-case incompatibility outside its completed observed trajectory. All mechanism IDs remain in the retained executable specifications.

Role/component	Actual requests	Prompt tokens	Output tokens
Fixed scientists	1,149	1,663,354	106,971
Common observer	168	78,177	24,728
Designer readiness	199	4,056,056	9,541
Designer proposals	81	2,857,825	20,966
Role loads (designer/scientist)	13/8	Unknown	Unknown

Table 39: Separate provider components for the whole normally closed study. Cached prompt counts 256,855/53,542/179,640 for scientists/readiness/proposals are subsets, not additional prompt tokens. Load usage is unknown, not zero. These costs are distinct from collected native activity resources K; nested duration fields and aliases are not summed.

exports and readable native-width print layouts are independently generated from those records; this document work adds no model calls or worlds.

A separate prospective, opt-in diagnostic now compares returned project fields against the original pure assignment validators on the exact saved input. It distinguishes illegal topic support from canonical content/duration, binds the current document and fixture, and persists as non-veto author feedback. It never executes generated code, applies commands, repairs source or widens permissions.55 light SCRIPTED/pure-validator checks include the next staged prompt and cold replay. A separate 30-case generic public-worker check includes actual calibration/certificate/role-factory/execute/replay machinery, with HTTP, executor and child supervision SCRIPTED. This covers configuration transfer, not full native usefulness or every downstream API; it does not cover T1's publication sentinel. It is not installed into this closed study or retroactively presented as feedback to its author.

The outcome distinguishes exercised actions and observed paths, but does not establish global optimality, a general objective conflict, why the designer failed to improve, or superiority of a search representation/method. Additional independent histories, same-authority method controls, mechanism interventions and selection-frozen conditions must answer an important identified question, not be automatic repetitions until a win. Successful reconstruction and plotting alone are not main-track readiness.

Initial prospective cost-opportunity hypothesis. Before the separate execution in Section 27, the research agent prepares one developer source-only hypothesis after observing the T/U prefix and old pilot, before inspecting C0's candidate or outcome, but not before C0 actually runs. The parent already runs publication and funding before an institution-authored liquidity transfer and original annual costs. A final-year source branch would consolidate active researchers' balances into an existing author. Original transfers conserve resources; an exact-zero donor remains active until an unpaid annual charge deactivates it. On the measured final handoff with total resources 80, isolated original transfer/update/ledger primitives collect 10 instead of 50, with 40 unpaid, one active author and 70 resources remaining. These are SCRIPTED handoff commands on a new component ledger, not execution of the proposed code or a complete scientific world.

The conditional forecast is total $K = 160$ rather than 200 and a four-unit L_C gain if prior content and the trajectory stay unchanged. This would be collection avoidance and an end-horizon incentive, not a reduction in requested activity expense or better science. Unpaid charges and attrition must be reported, not used as an undeclared veto or hidden loss replacement. Seven initial static/isolated primitive checks complete at that preparation stage, before sandbox/public-native execution, real losses and paired reconstruction. Section 27 reports the subsequent completed study separately. A positive hand control demonstrates opportunity, not automatic local 35B invention or a global optimum.

27 A reachable resource-objective gain without new scientific output

The closed free-system search establishes no advantage over its strong visible parent, but does not establish that parent as a world-level optimum. We next execute the terminal-cost hypothesis described in Section 26, rather than increasing the same search budget. This separate, development-informed control study tests a reachable difference under the already declared $L_C = L_U + 0.1K$. It is not auto-

Table 40: Actual two-control outcomes. Lower loss is better. K is collected native activity cost, Q unpaid activity cost and $R = K + Q$ requested expense. The last two columns are measured terminal snapshots, not eligibility gates.

Hand control	L_T	L_U	L_C	K	Q	R	Active	Balance
Reference	-36	-28.67	-8.67	200	0	200	5	30
Consolidation	-36	-28.67	-12.67	160	40	200	1	70

matic institution invention, a comparison of optimizers or unused evaluation. No designer proposes either program, and no objective is replaced after its outcome is observed.

Exactly two source-defined controls. Both fresh worlds use the same four-year original simulation, five initial researchers, openly reused development seed 61001, fixed Qwen3-8B scientists, offline 103,107-paper corpus and original production, funding and annual costs. Both inherit the core-plus-directly-related topic rule, original topic content and durations, workflow-v5 authority, graph, ports, schedules and other source. The reference retains its existing liquidity rule. The control changes only that producer's Python body: after final-year publication and funding, but before native annual costs, it transfers each other active researcher's positive balance into the richest existing active researcher (lexical ID breaks a balance tie). Transfers conserve resources; neither new money, researchers, papers, graph edges nor a cost exemption is introduced.

The final-year intervention exploits a distinction in the unchanged native resource consumer. A transfer leaving exactly zero does not deactivate its donor. The subsequent negative annual charge clips the balance at zero and deactivates that researcher. Requested negative charges and collected debits therefore differ. This is an exercised original-world path, not a numerical replacement of researchers or a new activity-cost law.

Normal completion and measured outcomes. The original writer exits normally with both scheduled worlds complete in 1,952.56 seconds (32m32.56s) on the shared local host. This elapsed interval is not a dedicated GPU-speed benchmark. Two original-source/physical audits and two cost projections match; two independently invoked readout exports match in all three files. Both controls' full 179 scientist request payloads and literal response choices are identical in order, with no unmatched requests. All 36 accepted year/paper/content records are common, and their exact centered opinion sum is $86/3$. Equal observed inputs/choices are not equivalence of the two programs or an independent replication of researcher behavior.

Annual collected expense is 50 in both worlds for years 1–3. At the final handoff each world has 80 resource units. The control's four actual transfers, 11 units each, leave 80 in one existing researcher and zero in the other four. Final-year collection is 10 rather than 50; requested annual expense remains 50, and the other 40 is unpaid. Closing active researchers are one versus five and total remaining resources are 70 versus 30. All transfer receipts and annual actor/ledger reconciliation are retained, without an attrition or unpaid-expense veto.

A real declared-loss gain, not real efficiency. The measured L_C improvement is exactly four units:

$$\Delta L_C = \Delta L_U + 0.1\Delta K = 0 + 0.1(-40) = -4.$$

Thus this study supplies a feasible lower L_C than the prior strong parent, without improving the recorded output count or content opinions. For interpretation only, define the requested-cost accounting diagnostic $D_R = L_U + 0.1(K + Q)$. Then

$$\Delta D_R = \Delta L_C + 0.1\Delta Q = -4 + 4 = 0.$$

D_R is not optimized, substituted for the loss, used to select a program or used to disqualify this gain. Requested native charges are themselves an accounting instrument, not independently validated real-world resource use. The observed gain is collection avoidance at the horizon, not reduced requested expense, better science or demonstrated long-run welfare. No fifth-year continuation is evaluated here.

Separate computational costs and research consequence. The physical caches contain 358 completed scientist responses, 537,104 input and 32,382 output tokens, with a reported cached-input

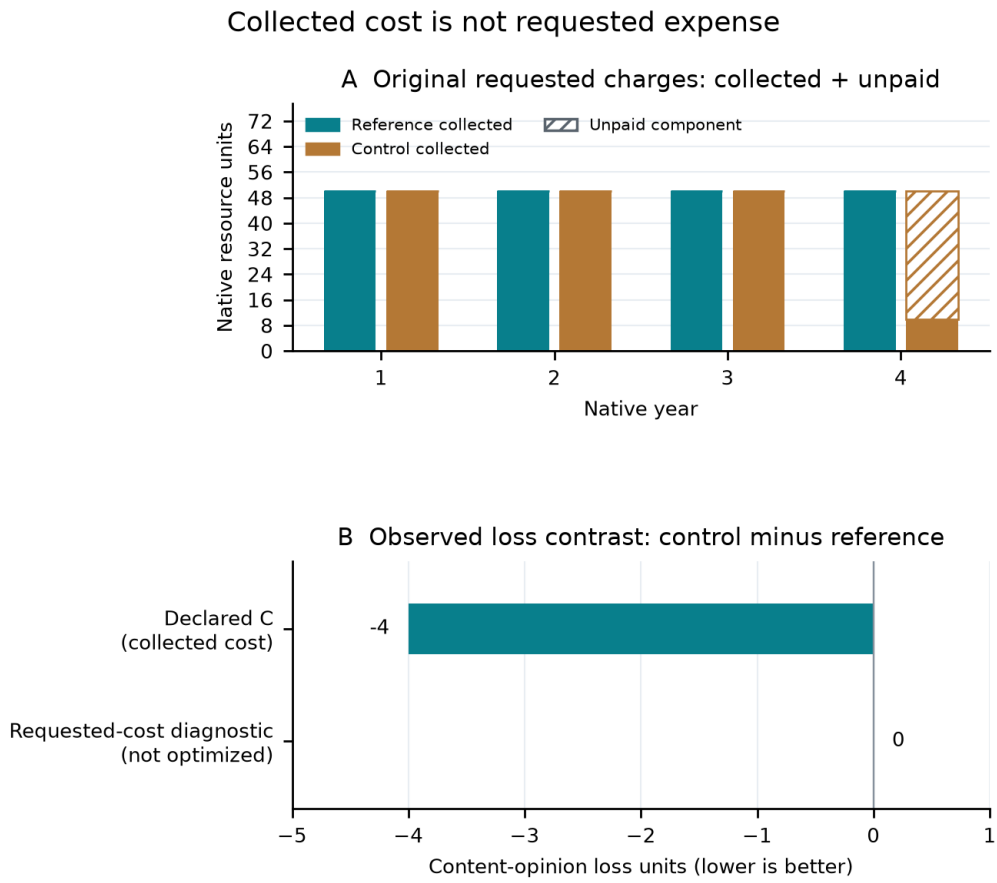


Figure 43: Original annual expense and the loss contrast. Hatched segments are unpaid charges, not saved requested work. The four-unit declared L_C gain vanishes in the separately labelled requested-cost diagnostic; the declared objective and original outcomes remain unchanged. One reused development condition and hand-authored controls provide no optimizer confidence interval.

subset of 79,402 (not added to the input total). The fixed content observer makes 108 physical POSTs, using 48,375 input and 15,337 output tokens. Cache reuse across the identical accepted records is not another physical rating. Two scientist-load attempts have missing token usage, which remains unknown. There are no proposal, designer-readiness or institutional extra-scientist calls in this hand study. These inference measurements are separate from native resource units K; they are not measured FLOPs, energy or monetary expense.

This contrast rejects a global-L_C-optimality interpretation of the previous parent on this condition. It does not show that the local designer could discover the mechanism, that a new topology is necessary or that code-defined search is superior to a strong simple rule. It also does not create a strict T/U preference reversal: these two objectives tie. The next free-system study must distinguish available design authority, actual native consumer effects, objective opportunity and discovery under matched information and budgets. Elementary end-horizon accounting and reconstruction alone are not an important main-track contribution.

28 Source-grounded free-system search under two objectives

28.1 A finite information-package comparison

The preceding studies separate available authority from exercised effects. They do not identify why a local author retains a known rule or edits an unconsumed parameter. We therefore compare two information packages in a separate, development-informed pilot, rather than claim that another fixed seed evaluation can establish effective invention. The question is whether exposing the native world's actual consumers changes the systems generated and improved under a declared objective. It is not comprehension in isolation.

Both conditions retain the original four-year world, five founders, openly reused development seed 61001, 103,107-paper offline corpus, fixed Qwen3-8B researchers and the approved local 35B author. Original production, project durations, funding and resource consumers remain unchanged. Both authors may edit institutional Python, parameters, mechanisms, ports, protocols, wires, schedules and bounded internal processes under workflow-v6 authority. They may ask the same fixed researchers typed institutional work, but cannot invent extra research production, resources or measurements. Thus the search is not restricted to the two hand-rule algorithms or an acceptance-rate menu.

The *interface-only* condition receives the existing editor, complete current system, visible measured parent, archive and component probes. The *literal-consumers* condition additionally receives six sealed native source cards on project eligibility/canonical content and duration, resource clipping/deactivation, transfer/publication consumption and cost measurement. Their literal lines and source hashes remain user data, not trusted instructions or an installed institutional policy. The package supplies no terminal transfer rule or recommended edit. Additional tokens, attention allocation and changed stochastic outputs belong to the treatment: equal proposal budgets are not equal measured computation, and a selected-loss difference would not isolate source understanding.

The already declared losses remain

$$L_U = -U, \qquad L_C = -U + 0.1K.$$

U is the three-position centered content opinion, not scientific truth; K is collected native activity expense, not inference cost. Requested/unpaid charges, participation and L_T are separate readouts, not added constraints. Three fresh hand references precede own evaluation: core-only and core plus directly related topics are visible to both optimizers; the known terminal- consolidation rule and its outcome are withheld from both. The latter is a disclosed comparison control for us, not an unknown human discovery or a selectable parent.

There are four isolated information/objective histories, with two own proposals each. Generation 0 orders interface/U, consumers/C, interface/C, consumers/U; generation 1 reverses that order. Both conditions use outer seed 78011, not independent search replications; actual per-step decoding seeds differ. Each 16-step editor session has a 65,536-token context and 8,192-token output reservation, with the same implementation checkpoint and bounded probes. The minimum-loss completed visible/own parent is retained, with first ties. After all slots, selection is minimum-loss *own completed submission*, earliest tie, without reference fallback. Nulls and adverse outcomes consume their slots; no hidden novelty, fairness, activity or improvement veto applies. No new unused-condition evaluation is attached to this pilot.

28.2 All outcomes, not only successful submissions

All eight scheduled own slots are consumed. Seven native worlds complete; one generated/compiled proposal fails original topic eligibility and retains unknown loss. Table 41 includes every slot and all three references. No candidate adds a node, connection or schedule, despite the offered authority. Every selected own system ties the strong visible reference under its assigned loss. The consumers-minus-interface selected-own contrast is zero for both L_U and L_C on this condition. There are no independent optimizer samples from which to construct an uncertainty interval.

A real adverse change followed by recovery to a supplied rule. Consumers/U generation 0 replaces shortest-duration-then-name choice with lexical-topic-first choice, retaining each original topic duration. All five initial founder assignments change: end years 1/1/2/1/1 become 3/3/3/3/3

Slot	Info	Target	g	L_T	L_U	L_C	K
Visible core-only hand rule				-20	-15.33	4.67	200
Visible related-topic hand rule				-36	-28.67	-8.67	200
Known terminal hand rule (withheld)				-36	-28.67	-12.67	160
$1^\star$	I	U	0	-36	-28.67	-8.67	200
$2^\star$	S	C	0	-36	-28.67	-8.67	200
$3^\star$	I	C	0	-36	-28.67	-8.67	200
4	S	U	0	-10	-7.00	13.00	200
$5^\star$	S	U	1	-36	-28.67	-8.67	200
6	I	C	1	—	—	—	—
7	S	C	1	-36	-28.67	-8.67	200
8	I	U	1	-36	-28.67	-8.67	200

Table 41: All actual own slots in chronological order, with cross-scores under the unchanged losses. I is interface-only; S adds literal consumer source. Stars mark the four frozen own-only selections, using full stored losses before rounding, not the withheld hand rule. Slot 6 is a runtime null, not zero or a numeric failure penalty. All complete own worlds have collected $K = 200$ and unpaid cost 0. The withheld hand has unpaid 40 and unchanged requested expense 200; it is neither automatically invented here nor admitted/rejected by an unpaid veto.

and agree with native validation events and saved researcher state. The world accepts 10 rather than 36 decisions, with $L_U = -7$ rather than -28.67 and unchanged collected $K = 200$. Its 91 complete scientist inputs/responses differ from the strong reference's 179-input path. This establishes native actuation, not a source-information benefit or duration-only causal explanation.

The following actual author wires contain that adverse own outcome and the stronger parent. The generated system instead rewrites an already present selector empty-cycle fallback. Its completed loss returns to the strong reference, so own best improves from -7 to -28.67 while best including visible parents remains flat. This is recovery to a known supplied system, not a new best-known institution or identified causal learning from feedback.

A technical null is not a scored bad institution. Interface/C generation 1 gathers related topics from the whole catalog instead of each actor's core-plus-directly-related set. Its one retained native code call returns five assignments to canonical `3d_vision`, all outside their actual actors' allowed sets. Contents, original durations and calendars are otherwise preserved. The original compound error mentions content/duration, but the failing predicate here is eligibility, not duration shortening. Later entries in that rejected batch are inspected data, not committed effects or five extra world trials. Loss remains null; neither no retained scientist-task wire nor missing measurement implies zero role-load or physical compute cost. The failure is neither repaired nor refilled. A later diagnostic-only source change names the failed predicate but is not installed into this closed run.

Observed equivalence classes and interaction restrictions. The other six completed own worlds share the strong reference's full 179 scientist input payloads and literal response choices, with equal measured metrics. Provider envelopes differ. The known terminal hand also belongs to that input/choice class but differs in collected/unpaid resources. The core reference has 127 inputs and the adverse fourth has 91. These are complete observed paths, not semantic equivalence of arbitrary programs, independent researcher replications or proof of world-level optimality.

All 101 editor steps remain, including plan-length/tool-shape rejections and implementation-checkpoint read refusals. Some requests for valid source files are blocked by that declared workflow even though complete source is already inline. Offered executable authority therefore does not mean unrestricted dialogue. Directed readiness success does not establish useful free-design capacity, and these observations identify no causal reason for the final proposals. The active source, interaction policy and loss are not repaired after seeing their outcomes. The later retrospective check in Section 29.3 finds no instance of its particular serialization/phase inconsistency across this pilot's complete eight author sessions and 101 requests. That scoped negative diagnostic neither certifies every interaction behavior nor rewrites the original measurements or selections.

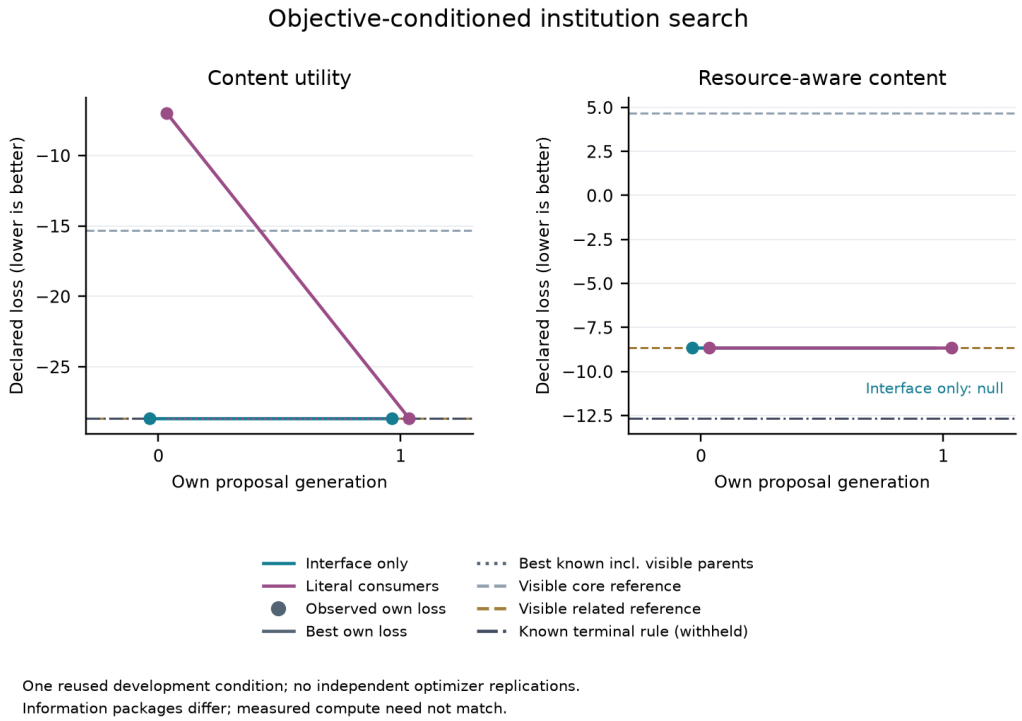


Figure 44: Four actual search histories under two objectives. Points are observed own losses; solid paths track best own submission and dotted paths include visible parents. The adverse source/U path recovers only the known reference. A null annotation has no numeric vertical meaning, and the incumbent may remain after failure. The withheld terminal line stays outside both optimizers' parent pools. One reused condition and unequal information packages do not supply independent optimizer error bars, isolated comprehension effects or equal measured compute.

Role	Responses/POSTs	Input tokens	Output tokens	Cached-input subset
Scientists	1,650	2,395,104	153,603	358,440
Content critics	198	94,806	29,111	—
Designer readiness	284	5,825,760	14,027	77,500
Designer proposals	101	4,018,739	26,804	212,450
Designer loads	18	—	—	—
Scientist loads	11	—	—	—

Table 42: Separate complete role journals, not a sum of nested timers or inferred counts from displayed author steps. Missing load usage and unreported critic cache usage remain unknown. Cached-input counts are subsets, not added to input totals. The 92 returned component probes make no extra model calls or native worlds. These computational quantities are separate from native K; FLOPs, energy, monetary cost and dedicated-host speed are not measured.

28.3 Costs, reconstruction and the next scientific boundary

The original finite writer exits normally after 17,863.53s (4h57m43.53s) on the shared local host, not a dedicated machine-speed benchmark. Both original cold audits and complete cost exports match. The first waiting readout stops at its existing external-output guard because the launch destination is source-local. That stopped record remains. A separate read-only export uses the same frozen reader code, same raw data and matching audit/cost files at a fresh external destination; both report/CSV/manifest sets then match. No scientist, critic, proposal, candidate or selection is repeated for this correction.

All eleven literal submitted/reference systems, frozen selection, cross-scores, paths and source/hash receipts are retained. Both final seven-file figure exports match, with inspected layouts; an initial overlapping null annotation is preserved in its earlier exports and corrected only in presentation. The observed zero contrast does not demonstrate useful local invention from this source package. It does not show that source information never helps, that the reference is optimal or that whole-system design cannot improve. With only two proposals per history, a reused condition and this local author/interface, no general method ranking is identified. The next research decision concerns effective search and strong restricted comparators, not automatically enlarging a flat result. Important differentiated findings, appropriate independent histories and selection-frozen unused evaluation remain necessary; this closed pilot is not main-track readiness.

29 Restricted and free methods: a finite native pilot

29.1 A finite search question, not an increase in world scale

We compare uniform priorities, adaptive priorities and free code/graph design under the same declared $L_U = -U$ and $L_C = -U + 0.1K$. The institution is optimized; scientist weights and measurement functions remain fixed. This finite pilot uses earlier development evidence, without reinterpreting preceding studies.

All methods keep the original four-year/five-founder world, reused development seed 61001, 103,107-record offline corpus, fixed Qwen3-8B researchers, native production, resources and catalog durations. Free proposals use the approved local 35B author and the existing v6 authority over Python, parameters, nodes, ports, protocols, wires, schedules and bounded processes. They are not confined to the priority comparator or required to introduce a novel mechanism. U remains an external abstract-content opinion rather than scientific truth; K remains collected native activity cost rather than requested expense, inference compute or welfare. No unpaid-cost, participation or novelty veto is added to either loss.

Three fresh references precede search: explicit zero priorities, the known terminal-consolidation version of that system, and the literal related-topic rule. The first two are visible selectable parents for all methods; the third is comparison-only. First-minimum reference selection gives the zero system to U and the terminal system to C. All three yield $L_U = -28.6667$ and $L_T = -36$; the terminal system has $K = 160$ and $L_C = -12.6667$, versus $K = 200$ and $L_C = -8.6667$ for the other two. Its inherited four-unit accounting advantage is known before search and cannot count as an automatic discovery.

The restricted family varies 53 topic priorities in $[-1, 1]$. Minimum original duration takes precedence, then decreasing priority, then lexical topic. Every other source, graph, schedule and eligibility field remains literal. Uniform search uses two uniform proposals. Adaptive search uses one uniform proposal followed by the existing bounded mutation with initial scale 0.2 and its inverse-square-root schedule. Both start from the exact measured canonical reference, rather than assigning an old loss to a new anchor. Their first samples are deliberately paired within each objective/history; the physically separate evaluations are not independent policy samples. Free design retains its greedy measured-parent/history feedback procedure. These are end-to-end method packages, not equal-compute interventions solely on representational freedom or a benchmark against evolutionary program search.

Two optimizer histories per method/objective use seeds 79011 and 79019, with two own slots each: 24 own slots plus three references, at most 27 new native worlds. Generation zero interleaves history, objective and method; generation one reverses the complete order. Each world has 1,024 scientist HTTP attempts, 900s zero-retry scientist transport and 7,200 safe-point wall seconds. The fixed content observer permits 6,000 critic POSTs. Free design has 16 proposal calls per session, initial and actual-request readiness, at most 488 combined readiness/proposal POSTs and 512 component probes; role loads remain separate. Shared-host time is not a dedicated throughput measurement.

Known technical nulls consume slots with unknown loss, not a hidden penalty or refill. Changed bindings or unknown dispatch stop the experiment. At budget end, selection is the earliest tied complete own submission for each cell, or none, without reference fallback. Two optimizer seeds do not provide two independent researcher populations. No unused-condition evaluation belongs to this pilot.

Slot	Method	h/g	Target	L_U	L_C	L_T
1	R	0/0	U	-27.67	-7.67	-36
2	A	0/0	U	-27.67	-7.67	-36
3	F	0/0	U	-28.67	-8.67	-36
4	R	0/0	C	-25.67	-9.67	-36
5	A	0/0	C	-25.67	-9.67	-36
6	F	0/0	C	-28.67	-12.67	-36
7	R	1/0	U	-28.67	-8.67	-36
8	A	1/0	U	-28.67	-8.67	-36
9	F	1/0	U	-28.67	-8.67	-36
10	R	1/0	C	-25.67	-9.67	-36
11	A	1/0	C	-25.67	-9.67	-36
12	F	1/0	C	-28.67	-12.67	-36
13	F	1/1	C	-28.67	-12.67	-36
14	A	1/1	C	-28.33	-12.33	-36
15	R	1/1	C	-28.67	-12.67	-36
16	F	1/1	U	-28.67	-8.67	-36
17	A	1/1	U	-28.33	-8.33	-36
18	R	1/1	U	-28.33	-8.33	-36
19	F	0/1	C	-28.67	-12.67	-36
20	A	0/1	C	-26.33	-10.33	-36
21	R	0/1	C	-26.33	-10.33	-36
22	F	0/1	U	-28.67	-8.67	-36
23	A	0/1	U	-28.67	-8.67	-36
24	R	0/1	U	-28.33	-8.33	-36

Table 43: All 24 complete own slots at normal writer closure. R denotes uniform priorities, A adaptive priorities, F free code/graph; h/g is optimizer history and generation. Lower losses are better. Rounded values are for display only; literal full-precision observations and feedback are retained. Frozen selection is summarized separately, not inferred from rounded table ties. Subsequent paired closure confirms every row. The interface limitation below prevents an isolated clean method/objective attribution.

29.2 All 24 submitted outcomes and frozen own-only selection

The original writer and its supervisor exit normally at November 16, 2026, 14:05 UTC. All 24 own native observations and three references are recorded; the own-only earliest-tie selections are frozen. At the 14:11 UTC verification, loss recomputation and literal selection checks agree with those records. The full original-source paired cold audits, campaign cost closure and paired all-slot readouts are still running; this terminal snapshot is not that audit. Table 43 retains every submitted slot in order, not a selected subset: 12 tie and 12 worsen their strong visible reference, with zero improvements beyond it. There are no generation or execution nulls in these 24 slots. The earlier dated 15- and 20-slot receipts remain unchanged.

At 14:21 UTC the original chain has subsequently closed: both full native cold audits and cost exports agree, both all-slot readouts agree, and both all-eight author/outcome exports agree. A separate guarded visibility export also runs twice with identical three-file outputs. Every original continuation exits zero, with no new model calls, candidate repair or native-world restart. The terminal-only evidence above and this later full closure are distinct dated records. The final counts and frozen selections do not change.

Across the 12 frozen method/history/objective selections, eight tie and four worsen the common visible reference; none improves it. Free design ties in all four cells. Adaptive priorities tie in both U histories; their C losses remain $7/3$ and $1/3$ worse in histories 0 and 1. Uniform priorities are $1/3$ worse and tied in U, and $7/3$ worse and tied in C, respectively. Each cell selects only its own two submissions, without falling back to the stronger reference. These small, deliberately coupled histories on one reused researcher condition are not independent populations, confidence intervals or method-superiority evidence. In particular, the later adaptive U recovery to the reference does not establish an advantage over free design or uniform search in expectation.

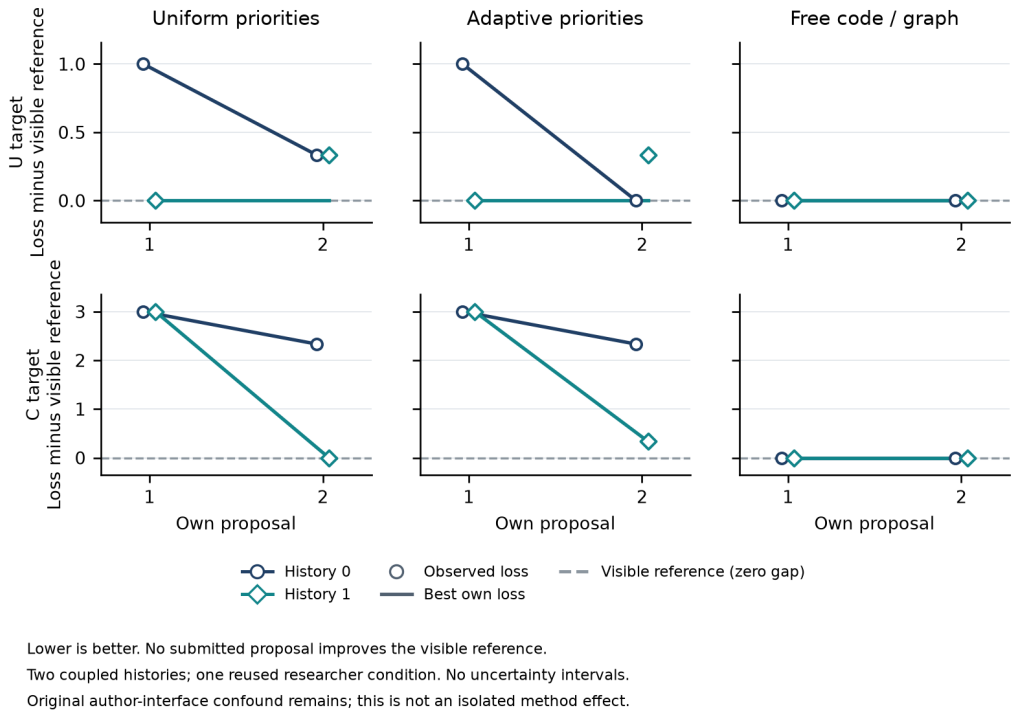


Figure 45: All 12 two-proposal optimization paths in the finite native pilot. Symbols show actual submitted losses; solid lines show best own losses so far. Gaps are relative to the common visible reference for the assigned objective, not different retrospectively substituted losses. No proposal improves on that reference. History 1/U numeric generation one worsens while its own incumbent remains tied. Small horizontal offsets only separate overlapping symbols. Both histories reuse the same researcher condition, and first numeric proposals are paired. The original author-phase confound remains; these are not isolated method effects, independent-population intervals or corrected-inventor results.

The initial U numeric proposals change two founders' actual topics while preserving durations/deadlines; their complete 186 scientist inputs differ from the reference's 179, and assigned loss worsens by one. This is operated parameter variation, not an unused coordinate. The paired first samples do not constitute an adaptive step. In history 1/C, adaptive generation one improves its previous own loss by $8/3$ but remains $1/3$ worse than the known terminal reference. Uniform generation one improves its previous own loss by three and ties that reference. Own-to-own recovery and a best-known gain are different, and these observations establish no adaptive advantage.

Two literal edits with different internal but inert outside effects. Free history 1/U/generation zero writes a feature extractor that adds one to an internal base-utility feature. All four annual features/scores change from zero to one, but the native accept-all control does not use that feature as the external content measurement. Native decision identities/final scores, all 179 complete scientist inputs and literal response choices remain equal to the parent. Complete paper records still have different submission times; this is not whole-record or world byte identity. Assigned loss ties the parent.

Free history 1/C/generation one instead writes an actual new catalog-agenda function. It keeps minimum duration first and replaces parameter priorities with nine hardcoded biases over real catalog topics, including artificial intelligence and deep learning. Fourteen recorded actions contain one source write, eleven reads, seven duplicate unchanged reads and explicit finish. It adds no node, wire or schedule. On all 18 actually logged eligible actor/year inputs, its ranking selects the same direction and duration as the lexical parent: preferred tied choices already coincide, or a shorter duration wins. All 18 logged agenda records differ only in reason text; all 179 complete scientist

Actual complete document	Supplied bytes	Spaced bytes	Phase after 3 reads
Zero-priority U parent	61,892	65,479	Implementation
Terminal C parent	62,755	66,342	General
U after feature-source write	62,032	65,619	General

Table 44: Different serializations, same numeric cap. All three actual native documents fit the supplied-body limit. The 140-byte U edit crosses only the legacy spaced-JSON boundary, changing its tool-phase schedule despite continued full visibility. This is an interface confound, not native loss corruption.

inputs and literal choices match the reference, while full provider envelopes differ. U, K and all three losses tie. The function was generated and executed, but this observed-path edit is not a useful new institution or evidence of equivalence on unobserved inputs.

29.3 A real author-interface confound, not a repaired result

Source and saved-request inspection identify an interface inconsistency. The active V45 renderer supplies the whole compact JSON document when it fits 65,536 UTF-8 bytes. Its inherited V44 phase function independently uses spaced JSON against the same cap. V45 then replaces only the displayed visibility flag, after phase selection. A fully supplied document can therefore remain in the general tool phase after three inspection attempts rather than reaching the declared implementation checkpoint. The counter records attempts, not only successful reads; the three initial boundary inspections in these examples all succeed.

This defect must not be projected onto every earlier study. A separate retrospective check covers all eight completed author sessions and 101 original requests in the preceding source-information pilot. Every request supplies the whole document; actual sizes range from 60,210 to 60,885 compact bytes and 63,690 to 64,365 spaced bytes, fitting both limits. No instance of this specific visibility/phase mismatch is observed there. Its original seven complete outcomes, one native null and four tied selections are unchanged. This scoped absence is not a certificate that its entire interface is bug-free or proof of model comprehension. It does not extend to uninspected studies.

Three complete initial sessions supply 33 saved backend requests, including 14 affected general-phase calls. The subsequent C/generation-one session has nine further affected calls. Exact loss arithmetic and measurement definitions are present in the first 33 inspected wires; in particular, institution-written `final_score` is not an input to external U. Their observed prompt counts 32,070–37,382 fit the reserved 8,192 output tokens within 65,536, and replies end in stop. These diagnostics do not establish model understanding or absence of backend truncation. Read bytes were not silently lost. Fourteen-action finishes at the three-call submission reserve must not be called freely chosen early strategic stops. Actual own scalar feedback reaches the next-generation backend request, which proves delivery rather than useful use.

The four original scientific/analysis sources remain pinned and unchanged. Native measurements, candidates, nulls and costs are not rewritten, refilled or converted to failures because of this defect. However, parent and edited document sizes also changed the allowed-tool schedule. This old pipeline is therefore source-specific exploratory evidence, not a clean isolated comparison of objectives, intrinsic model ability or numerical versus free design. No conclusion is drawn about whether an unobserved corrected proposal would have improved the loss.

Prospective repair and its narrower evidence. The corrected renderer determines actual document inclusion before computing the workflow and passes a source-owned boolean into phase selection. Legacy V44 behavior without that override, the byte cap, the 16-call budget and the three-call submission reserve are preserved. This restores the declared technical interface without mandatory edits or a novelty/merit constraint. The new context contract is explicitly named `matched-bounded-author-64k-v2-literal-visibility`; older model certificates cannot authorize the changed source.

Complete native document fixtures reproduce two pre-fix failures. The fixed author source passes 87 light/scripted boundary, wire and replay checks with one unavailable optional archive skipped and one explicit historical case deselected. A further source-contract preparation stop is retained, rather

than bypassed. Prospective routes then bind the designer-only correction as an explicit Git-versioned adapter, leaving the original byte guard and its exclusions unchanged; 43 scoped light cases include rejection of injected author/scientist/executor drift. The older free-topic driver keeps its narrower contract and is tested against its immutable original bytes. These are software and static bindings, not new real-model readiness, full native/public execution or an observed benefit from the repair. None is established by these checks.

A subsequent complete public/native software check at source `adfd6b94` passes in 1,245.16 shared-host seconds. It uses the actual SDK, executor, checkpoints and paired cold/cost/readouts, including 27 development, two upstream-fixture and 18 fixed-evaluation native worlds. Its HTTP responses, corpus and orchestration are explicitly scripted; software seeds 43003/43011 do not consume reserved scientific conditions. All 30 logical evaluation rows and 108 contrast rows are retained. This is a real execution-path check, not real-model readiness, outer subprocess-CLI proof or useful institution search.

29.4 What remains a scientific question

All eight proposal sessions have generated with 102 committed proposal steps; a read-only journal projection finds 45 instances of the specific phase mismatch and no different loss views. These are repeated steps within eight sessions, not 102 independent author trials. All eight native outcomes are now recorded, but the original all-eight author/outcome join and the guarded paired visibility export are not complete at the terminal snapshot; their later 14:21 UTC closure confirms all eight parent-matching input paths. Generated edits and measured improvement remain distinct. For example, the final free U source emits the same 18 actual project commands despite removing its own agenda-assignment bookkeeping; its 179 full researcher inputs and literal choices equal the reference and its loss ties. A self-reported zero counter does not mean that no researcher acted or that the edit failed to execute.

The final all-eight action summary records five actual source writes in four changed submissions, four unchanged submissions, 52 duplicate unchanged reads and explicit finish in all sessions. Every submitted native outcome is complete and tied with its measured parent; all eight observed full researcher-input paths match a complete parent path. No node, wire or schedule changes are submitted. Neither a valid finish nor five successful writes establish useful institution invention. The specific phase confound remains, rather than being erased by full audit closure.

Actual inference accounting, not an equal-compute experiment. The original full campaign records 4,888 complete researcher responses, all with reported usage: 7,392,125 prompt tokens and 442,516 output tokens; no saved researcher input lacks a committed reply. Separate physical content-critic requests number 213. The author performs 256 readiness and 102 proposal POSTs, with 17 designer-load and 27 scientist-load records kept separate. The writer takes 35,144.9 seconds on a shared host, not dedicated model throughput, energy, money or the sum of nested call durations. These totals include references and calibration, not just the 24 optimization observations. Their non-overlapping role counts do not imply equal measured compute across methods.

Further inference should not automatically enlarge a flat pilot. A corrective finite study requires appropriate exact-source native verification and fresh real readiness before dispatch, with a frozen protocol and no failed-slot refill. Separately, unused-condition evaluation of old selected policies would answer their transfer question, not validate the corrected inventor. No such evaluation is completed by this snapshot. Important differentiated findings, adequate independent evidence and a final main-track submission judgment remain open.

Literal sources and dated evidence are retained under scientific commit `705b5259`; prospective author checks use `195739f6`, and scoped source-binding tests use `1e3b114b`. All 24 literal outcomes and frozen selection checks are in `research/results/native-priority-method-terminal-snapshot-v1.json`. The later paired cold/cost/readout/author/visibility closure is retained in `research/results/native-priority-method-comparison-closed-v1.json`, with compressed literal reports and their original-file hashes. The earlier 15-slot prefix is in `research/results/native-priority-first-three-second-generation-outcomes-v1.json`; the defect and original failures are in `research/results/native-author-workflow-visibility-correction-v1.json` and the

additional bindings in
`research/results/native-author-context-source-binding-v2.json`. The separate successful public/native software verification is in
`research/results/native-priority-corrected-public-native-qa-v1.json`. The earlier pilot's complete author-side scope check is in
`research/results/native-consumer-source-visibility-diagnostic-v1.json`. Normal writer completion is not the paired original cold audit, corrected-model proof or a publication-ready differentiated finding.

30 Corrected four-step institution search: design and scope

This section specifies a separate experiment, not repaired observations from Section 29. Its scientific source is `57b35144`. These methods were specified before the original dispatch; Section 31 reports its later complete native outcomes and frozen selections. Normal writer closure, original-source paired analysis and selection-fixed unused-condition evaluation are separate evidence stages. Source preparation is not a newly built or reviewed PDF.

30.1 Fixed researcher world, changed search package

The study retains the original four-year/five-founder native world, reused development condition 61001, 103,107-record offline corpus, fixed Qwen3-8B researcher weights, approved local 35B designer and v6 institutional authority. Researcher weights are fixed, not behavior: institutional code can change real plans, information, resources and later prompts through declared native consumers. The objective definitions remain

$$L_U = -U, \qquad L_C = -U + 0.1K_{\text{collected}}.$$

Here U is the fixed three-position abstract-content opinion, not validated scientific knowledge. Collected native activity resources are not requested expense, inference compute, money or real work. Requested/uncollected costs, throughput and participation remain readouts, not extra loss terms or hidden candidate gates. Novel wiring or improvement is not required for validity.

There is one optimizer history per method and objective, with four chronological own slots: three methods times two objectives times four proposals gives 24. The global optimizer seed is 68011 and the single history seed is 79011; neither it nor world seed 61001 is an unused condition or an independent researcher population. Each generation interleaves objectives then methods, reversing the complete order on odd generations. Adaptive and free methods therefore have three opportunities to use preceding measured feedback; the uniform baseline remains four initial random samples with loss-based selection. This is not a convergence claim, equal-compute comparison or identification of a single interface, representation or objective-term effect.

30.2 Actual proposal spaces and reference information

The study freshly measures three references: the literal zero-priority rule, its known terminal consolidation variant and the literal core-plus-related rule. Only the first two are visible selectable references. Every method starts from the first minimum complete visible reference under its assigned loss. An old measurement is not assigned to a new program or source binding; reference technical failure stops rather than substituting a fallback. The supplied terminal rule's accounting effect is not autonomous invention.

Both restricted methods retain the same 53 original topic-priority fields in $[-1, 1]$, with minimum original duration first, then decreasing priority and lexical topic. All other fields remain literal. Uniform sampling uses the existing numeric sampler throughout. Adaptive search uses the same first sample and then a declared rank kernel: choose uniformly among fixed initial actors' nontrivial shortest-topic groups, swap the group's current highest ranked topic with a uniformly selected alternative, and encode the resulting global order with binary-exact bounded priorities. At scale s and G groups, the move performs $\max(1, \lfloor sG \rfloor)$ swaps, with the existing cooling schedule initialized at 0.2. The groups come from the fresh reference's exact year-one pre-agenda actor projection and original catalog/eligibility, not favorable measured losses, new rollouts or a behavior-change filter. Repeated groups, undo and repeated observed behavior remain possible. This conditional initial

Design dimension	Earlier finite pilot	Corrected study
Own slots per method/target	Two histories, two proposals each	One history, four proposals
Total own/reference ceiling	$24 + 3$ native worlds	$24 + 3$ native worlds
Adaptive proposals	Bounded numeric mutations	Fixed initial-choice rank moves
Author workflow	Recorded serialization/phase mismatch	One literal visibility decision
Component context	Recorded historical parameters	Declared current parameters on recorded inputs
Primary endpoint	Earliest best complete own	Earliest best measured reference or own
Secondary endpoint	Same own-only endpoint	Separate own-only endpoint, possibly absent

Table 45: Changes in an end-to-end experimental package, not an isolated representation intervention. The native world and total slot ceiling do not grow. The corrected primary endpoint labels returned references explicitly; it cannot turn a supplied policy into an invented one.

neighborhood does not exhaust later native policies or guarantee loss locality. The rank move is not new optimizer theory. Existing exact-program uniqueness checks remain; they do not reject repeated observed scientific trajectories. The internal sampler index can consequently exceed the number of evaluated own slots: duplicate-vector draws create no additional native evaluation. They are part of the declared proposal algorithm, not scientific failure retries or replacement slots.

The seed alone does not define a common random-number intervention across objectives. The sampler also binds the full numeric space and parent program; objective-dependent reference selection can therefore change its RNG namespace. Within an objective, the uniform/adaptive initial literal programs are paired. Separate physical executions of that pair are not independent policy samples.

Free proposals retain arbitrary executable source, mechanisms, protocols, ports, connections, schedules and bounded process composition under the same v6 native effect contracts. They need not stay inside the priority family or add a particular node. Their greedy measured-parent procedure retains actual own observations and complete declared loss/history information; a source change, explanation or successful probe is not a measured native gain. Retaining history does not mean repeating every historical raw body in every model message: compact views and explicit on-demand record descriptors are part of the interface, not proof of model understanding.

30.3 Technical readiness is not downstream scientific success

The prospective visibility correction uses one literal document-presence decision for display and workflow phase. Historical requests remain unchanged. Each actual free request needs its own fresh source/model/backend/request-bound readiness certificate; prior assay certificates are not transplanted. These technical checks are distinct from native loss evaluation and useful design.

The component option explicitly projects the current node's non-reserved parameters onto attributed recorded inputs, state, period, phase and randomness. It uses the existing isolated executor and labels the derived context as counterfactual. It does not simulate a new candidate trajectory, change historical fixtures or establish the applicability of inputs after new wiring/state. Probe errors do not veto explicit submission, and the host does not repair source or implement the author's prose. Local code return and compatibility with a later native consumer remain separate properties.

Before the real dispatch, a clean source-bound software check executes the public native path for all 24 own attempts and three references, including a prescribed execution null. It uses the actual SDK, containers, checkpoints, paired cold audits, costs and readouts. HTTP replies, corpus and synchronous orchestration are scripted. Its 351-file native source closure matches the scientific source, with separately checked support helpers. This is engineering evidence, not new real-model capacity, scientific improvement or outer subprocess-CLI verification.

30.4 Charged slots, endpoints and interpretation

The native budget remains 1,024 scientist HTTP attempts per world, a 900-second read allowance with zero retries, 7,200 safe-point wall seconds per world and 6,000 fixed-observer POSTs. The

designer ceiling is nine readiness sessions of at most 40 POSTs plus eight proposal sessions of at most 16 POSTs: 488 combined POSTs. Component probes are capped separately at 512; role loads are separate. Shared-host wall time is not a dedicated throughput benchmark, and equal proposal counts do not imply equal compute.

A slot is charged and journaled before dispatch, not when its result becomes available. Known generation/execution/measurement failures remain terminal nulls with unavailable loss, never artificial penalties, free retries or replacement candidates. A pending response is not an observed failure or a completed world; uncertainty does not authorize redispatch. Observation, actual optimizer tell and final selection are separately retained.

For method m and target j, the primary endpoint is the first minimum-loss complete policy in the ordered pool of visible measured references followed by chronological own proposals. Its origin remains explicitly reference or own. The secondary endpoint is the first minimum complete own proposal, or absent when none completes. The primary is not a disguised own fallback; the secondary never acquires a fabricated parent-policy label. Observed-own, best-own and best-known trajectories are separate, so recovery from a poor own sample is not confused with a gain beyond an already measured reference.

All 24 slots, nulls, literal programs, cross-losses, effects, complete model input/choice classes and role costs must survive original paired analyses. Equal observed paths are not global program equivalence or independent replication. Both endpoint banks must be frozen before a separately declared selection-fixed evaluation with strict unused-condition inventories; proposed seeds alone are not evidence of unusedness. No such evaluation or full closure is supplied by this methods section. After complete traces, any additional research contrast needs its own finite design rather than automatic scaling of a flat result. A consequential differentiated claim, adequate independent evidence and a final main-track judgment remain open.

The immutable protocol is `research/protocols/campaigns/native-corrected-choice-search-v1.md`; actual source/model/budget/seed bindings are in `research/results/native-corrected-choice-real-dispatch-v1.json`. The public software receipt is `research/results/native-corrected-choice-full-public-qa-v1.json`. Parameter-context scope and later readout preparations are documented separately from scientific outcomes. The historical reviewed 111-page artifact excludes this corrected study. A separate source-bound build receipt records the later integrated artifact; an old PDF is never relabeled as containing new results.

31 Corrected search: complete outcomes and retained incumbents

31.1 All four-step paths, including the execution null

The original writer closes on November 17, 2026, at 02:20:21 UTC; its supervisor records exit zero at 02:20:24 UTC. All 24 own attempts and three fresh references are present. Twenty-three own outcomes complete and one remains an original execution null. Eight complete own outcomes tie their assigned visible reference, 15 worsen it and none improves it. No attempt is repaired, replaced, omitted or given an imputed failure loss.

Table 46 contains every own assigned loss, not only selected candidates. Its six rows are coupled search histories on one reused world condition, not six independent scientific populations. Within each objective, uniform and adaptive generation-zero programs are literal matches but separately executed worlds. Three later adaptive moves are actual feedback-conditioned proposals; the four uniform proposals are initial random samples with loss-based selection. The comparisons remain end-to-end search packages, not isolated representation or equal-compute effects.

Several paths recover their own earlier loss without improving best known. Adaptive C moves from -9.6667 to -12.6667 at its first update, matching the supplied terminal reference. Its next update worsens, and its last ties again. Uniform C reaches its best own loss at its fourth sample, still about $1/3$ worse than the reference. Free U recovers from -7 and an execution null to the known -28.6667 reference level. These are observed search paths, not new best-known discoveries or demonstrated

Method	Target	Reference	$g = 0$	$g = 1$	$g = 2$	$g = 3$	Best own
Uniform	U	−28.67	−27.67	−28.33	−27.67	−28.33	−28.33
Adaptive	U	−28.67	−27.67	−25.67	−28.33	−26.67	−28.33
Free	U	−28.67	−7.00	null	−28.67	−28.67	−28.67
Uniform	C	−12.67	−9.67	−10.33	−10.67	−12.33	−12.33
Adaptive	C	−12.67	−9.67	−12.67	−11.67	−12.67	−12.67
Free	C	−12.67	−12.67	−12.67	−12.67	−12.67	−12.67

Table 46: All 24 own attempts in the corrected native study. Lower assigned loss is better. Exact retained floats, not displayed rounding, govern selection. The null has unavailable loss, not zero or an infinite penalty. Best own excludes the measured warm reference; best known includes it and never leaves the reference level in this study. This is not an independent method-ranking sample.

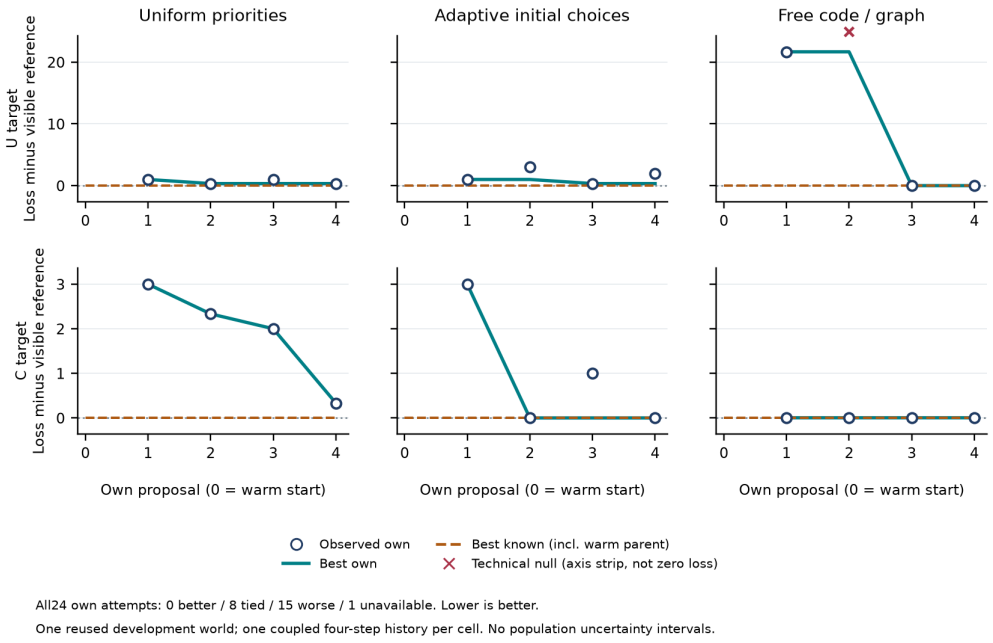


Figure 46: Complete corrected search paths, relative to each assigned measured reference. Open circles show actual own losses, teal lines the best complete own proposal and dashed orange lines best known including the warm reference. The execution null is marked above its panel, not plotted as zero loss. Both rendering runs produce identical PNG/SVG bytes from the original closed read-out. One coupled development condition is shown; no confidence interval or independent method superiority is inferred.

causal benefits of feedback. The adverse and repeated proposals remain informative records rather than reasons to refill the declared budget.

31.2 The returned policy is not necessarily an invented policy

The original writer freezes both endpoint scopes before any unused-condition evaluation. Every primary best-known endpoint returns a measured visible reference: the priority-zero policy for U and terminal-priority-zero policy for C. The six logical labels therefore return two supplied policies, not six independently invented institutions. They are genuine measured incumbents, not fabricated fallback losses, and the separate own-only results remain visible.

The exact frozen selection is retained independently of manuscript prose: its content checksum is `358c6a22e75b540cc8170f3d4e28387640d57bb13ac16bece2b3cdc48b7ab609`. Changing the displayed table cannot change those programs, origins or tie rules. Fixed-condition evaluation must

Method	Target	First best own g	Primary origin	Own minus reference
Uniform	U	1	supplied priority-zero	$+1/3$
Adaptive	U	2	supplied priority-zero	$+1/3$
Free	U	2	supplied priority-zero	0
Uniform	C	3	supplied terminal-zero	$+1/3$
Adaptive	C	1	supplied terminal-zero	0
Free	C	0	supplied terminal-zero	0

Table 47: Literal earliest-minimum selection from the original completed study. All six primary endpoints tie their measured reference by retaining it. Among six separate own-only endpoints, three tie and three worsen the reference. An exploratory sample does not make the best-known incumbent worse, but retaining a supplied incumbent is not useful autonomous invention.

reuse them literally. It would evaluate fixed policies, not create independent optimizer replications or retroactively make a reference an own submission.

31.3 Source edits, invoked rules and researcher behavior are different layers

All eight free-author sessions explicitly submit after 14 operations each. A passive reader verifies their complete original binding, chronological dispatch/completion chains and terminal file hashes twice, with byte-identical outputs. The 112 original completed actions include 25 applied source edits. Nineteen incremental changed-node rows have the same normalized Python AST and six have a different AST; no connection or schedule edit occurs. These are edit counts, not 25 candidate worlds, 19 distinct unchanged programs or six useful invented algorithms. Other supported forms of design remain allowed. The complete visibility export finds zero serializer mismatches, phase interpretation mismatches or missing/different loss views across the eight sessions. This addresses the specific earlier observed interface defect; it does not certify comprehension, every interface or useful invention.

That observation is more precise than claiming that all free edits are formatting. For example, the final free U proposal replaces conditional native acceptance with constant acceptance. Its source AST differs from the parent, and four original annual selection receipts join bijectively to committed native execution events. Its final portable checkpoint contains the submitted candidate exactly. Nevertheless, the parent's actual current rule is already accept-all, and all 179 recorded scientist request payloads and response-choice sequences match the measured priority-zero reference. Provider response bodies differ and their hashes are checked. This is an observed repeated trajectory, not global program equivalence; source change and actual invocation are not sufficient evidence of changed researcher behavior.

The first free U proposal is different in another way: its agenda source actually favors longer catalog projects, despite its accompanying explanation. Its 10 accepted decisions and $U = 7$ are worse than the parent's 36 decisions and $U = 28.6667$. The next submitted U program has a double payload envelope: local component returns do not establish downstream native paper-ID coverage, and the original native outcome has unavailable loss. Neither the explanation nor the probe result repairs that submitted program. In the later C sessions, planned parameter/threshold changes are not implemented; literal source edits preserve other configuration and, in the recorded cases, the parent's AST.

The passive edit analysis adds no production prompt, model call, code execution or candidate gate. Equal AST never forbids a design, and different AST never certifies usefulness. A genuine numerical topic-priority change remains visible despite identical source AST. These observations motivate more consequential proposal procedures; they do not themselves demonstrate that new diagnostic feedback improves invention or identify intrinsic designer competence as the cause.

31.4 Content support and physical cost

The complete content export retains 910 accepted occurrences across 26 complete physical worlds: 734 have positive centered panel utility, 176 have zero and none has negative utility. There are 88 content hashes across worlds and no within-world repeated content. Cross-world sharing is not within-world score inflation or independent replication. Deleting fixed scored records alone there-

Recorded physical component	Calls/replies	Prompt count	Output count
Scientist complete replies	4,650	7,016,180	425,437
Designer execution-readiness	258	5,456,852	12,605
Designer authoring operations	112	4,230,690	27,012
Common-content observer POSTs	264	—	—
Empty-prompt model loads	44	unknown	unknown

Table 48: Physical call components of the original development campaign, not per-method equal-compute budgets. Scientist counts use provider-reported tokens; designer counts use backend-native prompt/evaluation counts. Cached prompt counts are subsets and are not added again. The observer's call count does not assert comparable token accounting. Seventeen designer and 27 scientist loads are separate from generation. Missing costs remain unknown.

fore cannot increase their summed U, holding all other consequences fixed; this is not an attainable dynamic counterfactual or a global policy bound.

The terminal-priority-zero reference has $U = 28.6667$ and collected $K = 160$. Every complete own outcome has no greater U and no lower K. This reference weakly dominates all 23 complete own points on the observed U–K support. Consequently, re-ranking this archive under $-U + \lambda K$ for any $\lambda \geq 0$ cannot place an own point strictly ahead of that reference. The statement concerns these observed points only, not all executable policies. It explains why merely changing the cost coefficient after this study is not evidence of a newly discovered institutional tradeoff. Finding such a tradeoff requires consequential new proposals or a separately justified objective, not retuning a coefficient to claim a retrospective win.

The scientist's reported token total is 7,441,617, including the reported 1,113,278 cached prompt tokens as a subset of prompt counts. No recorded scientist wire input lacks its committed reply. The original supervisor spans approximately 9 hours 19 minutes on a shared host. This is not a dedicated hardware throughput benchmark; nested transport/provider durations must not be summed into total wall time, GPU FLOPs, energy, money or an exclusive-resource speed estimate. Generated proposals, readiness controls, probes and physical world evaluation remain distinct costs.

31.5 Closed development evidence, not yet unused-condition inference

All seven original analysis chains now close normally: paired native cold audits/costs, all-slot readouts, author outcomes, author visibility, native source/consumer/resource effects, loss paths and signed-content exports. Their paired files agree byte-for-byte. These are original-source analyses of the unchanged scientific run, not newly generated candidates or observations.

The selection-fixed public/native software test also completes: actual native execution with SCRIPTED HTTP, corpus and supervision processes 14 physical evaluation worlds, 15 logical labels and 30 logical rows, with 120 paired contrast rows. Its paired audits/costs/readouts match and the single full-path test passes in approximately 23 minutes. It makes no real model calls and does not consume scientific reserved conditions. The outer subprocess CLI routing is outside this check. Software verification is not scientific gain.

The strict unused-condition inventory remains unresolved: exact metadata-bearing SCRIPTED fixtures use one proposed seed, while old fixture stores and directory symlinks produce coverage errors. No directory-name exemption, silent condition switch or assertion of unusedness is applied. No real selection-fixed evaluation is reported here and no original result is reselected using evaluation outcomes. An important supported main-track claim and adequate independent evidence remain research obligations, not consequences of another valid pilot.

32 Discussion: what can be concluded across the program

32.1 The objective-to-system map is the research object

The studies are not another survey of hand-selected acceptance or funding settings. A loss enters a search procedure, which returns a specification and a history; common evaluation then characterizes that returned system. Parameter optimization exhibits two selected designs under three objectives.

Structural cases exercise different organizing algorithms. V2 shows that strong simple rules can remove apparent gains against an editable seed. The meaningful output is the conditional map from objective and permitted design domain to generated mechanisms, paths, costs and consequences.

That map need not be injective: different objectives can select the same institution. It need not be robust: a development winner can change rank on unused conditions. It need not select a novel institution: all-accept or reviewer-free reject may minimize a proxy. These are valid experimental outcomes, but not automatically important general findings. A future paper must substantiate which observed pattern extends beyond elementary incentives and this particular designer/world combination.

32.2 Expressivity, exercise and optimization must be measured separately

The completed controls in Section 21 do exercise different timing and information routes with consequential native outcomes. They are hand controls in a stopped experiment, not own optimized programs. The same-grant resource contrast is useful mechanism evidence, but it does not cure the missing automatic objective-to-institution result. Likewise, directed thinking-mode calibration succeeds while free proposals reach an output limit or violate graph causality. Platform expressivity, a compiler-valid submission, an operated change, a loss improvement and independent validation remain separate claims.

The later open-code T0 in Section 22 actually changes subsequent native scientist inputs and selected content, while tying parent throughput and worsening the fixed common-content loss. Changed source and exercised information flow therefore do not by themselves imply a useful institution. Its following calibration stop does not erase that unfavorable observation or complete the missing objective-conditioned search.

The V8/V12 cases show that source-defined nodes, routes and plan processes can execute; V13 shows observed free-code gains with an external author. The later matched local study does not exercise declared topology changes at all. These are compatible results with different provenance, not a contradiction to be resolved by discarding the earlier cases or attributing them to the local designer. The platform can support a behavior that a particular search does not find within its finite budget.

The subsequent actual staged study in Section 23 completes every own slot but still does not add nodes or wires. One agenda program changes native scientist inputs and worsens utility, while the other five own outcomes tie R7. Component observations, successful submission and whole-world completion therefore cannot substitute for an identified benefit. Unequal inline objective-definition coverage is an observed presentation limitation, not a demonstrated causal explanation. Future information/authority comparisons must not be described as repairs of this closed experiment.

Section 18 strengthens that distinction. All six local proposals execute, but only rewrite an already all-accept selector. The permitted planning and allocation freedom is not exercised. Complete inputs show three U programs preserve their strong parent's observed behavior, while two T programs change inputs and outside U despite equal T. The realized content objective refines count ties without strict preference conflict; all accepted content contributions happen to be nonnegative. A lack of generated advantage here cannot establish that no better institution exists, nor can different objective-conditioned outputs establish an unavoidable trade-off.

Section 19 examines an allocation parameter's actual consumer without mistaking conditional commands for new worlds. Its 240 component evaluations produce nonmonotone transfer responses on four fixed historical inputs, while the original ledger connects one real command path to cost, award and attrition timing. This explains an exercised operation and motivates a separate process-design question; it neither supplies a counterfactual loss frontier nor establishes a useful newly invented system.

The finite numeric comparator in Section 20 then completes all six native worlds. Its own U path improves while its reference-inclusive incumbent does not; its T proposals all remain worse than the known reference. This is why trajectories must distinguish best-known from best-generated outcomes. The cross-study comparison to the earlier free-program pilot is post-hoc and not an isolated effect of design freedom.

The concrete agenda comparison in Section 17 supplies a useful planning path, rather than just an unconsumed interface: ten advice choices match their actual agendas, and direct assignment separately validates ten native effects. Both effectful arms raise accepted decisions from 12 to 18 on one reused seed while improving P as well. This is developer intervention evidence, not a locally generated institution or a distinction between the systems selected by different objectives. Its later full model-input histories differ, so it does not isolate duration as the sole causal mediator.

Similarly, technical validity rates are scientifically relevant to end-to-end optimization, but they confound any claim about intrinsic design-space value. In V2, all local LLM numeric proposals failing while uniform numeric proposals execute is evidence about the tested package. It is not proof that parameters are intrinsically harder than code or that code/graph freedom is superior. Future interface interventions should be prospective, not repairs/refills of the V2 record.

The original-choice controls in Section 25 extend this distinction: advice is delivered and output improves relative to unadvised choice, yet direct assignment remains stronger. None of the twenty native direction requests chooses a one-year topic despite an unrestricted topic-string schema. This does not identify why researchers choose as they do, nor justify calling a hand-authored information office an autonomous discovery. The new passive ledger instrument exposes a finite output–resource contrast that T/U alone did not express. Retrospective cost measurement and fixed-pair loss algebra are not a new optimization study, an independent replication or a reason to replace the old objectives after seeing their outcomes.

The same-authority topic contrast in Section 26 subsequently improves both observed losses at equal collected cost by changing a hand rule's candidate set, without shortening original topic durations. It identifies an actionable developer contrast, not a useful automatically generated institution or a duration-only causal effect. Its conditional calendar relaxation is saturated on the recorded fixed founders/support, while every accepted content contribution is nonnegative; neither diagnostic proves a world-level optimum. The separate three-objective free search uses both strong references visibly. All six own slots close: an observed-path-inert tensor edit, source reproducing the weaker known core rule, an ineligible two-hop assignment, an AST-preserving submission, explicit parent retention and accept-all code with a conditional empty-case incompatibility. Five native outcomes complete, one is null and none improves the strong parent. U's own best improves by returning to a known rule, not by moving the reference-inclusive incumbent. These cases and their paired reconstruction separate available freedom, exercised actions and actual loss without identifying a general causal explanation. Adding a resource preference is a declared new study, not retrospective reinterpretation of old T/U or their selections. The separate terminal-cost study in Section 27 now realizes a four-unit L_C gain: identical recorded scientist inputs and accepted content accompany 40 fewer collected units and 40 more unpaid units. This source-only hand rule establishes reachable objective opportunity, not local-designer discovery, lower requested expense or better science. New non-veto component diagnostics remain software preparation, not evidence that they improve free-system search.

32.3 Source knowledge does not establish effective search

The subsequent source-information pilot in Section 28 completes seven of eight own worlds and retains one original eligibility null. Neither information condition's selected program improves its strong visible reference. Literal source access does not suffice for useful discovery in this finite package; it is not a general negative effect of source information. The adverse utility path's recovery must not be sold as a new best-known institution, and full-input ties must not be sold as proof that arbitrary programs are equivalent. Next method comparisons need effective operative proposals and a named strong parameter family, not only a larger budget or more evaluations of the same incumbent.

32.4 The author interface can depend on the institution being proposed

Section 29 completes a restricted/free comparison with all 24 own outcomes and paired audits. Its saved requests expose a document-visibility/phase inconsistency: parent and edited document byte sizes change whether the author reaches an implementation checkpoint, despite the whole actual document being present. This is not corrupted native scoring, but it prevents attributing the old

pipeline's differences solely to objective, search method or intrinsic designer competence. Null or adverse outcomes still belong in the record; a prospective interface fix does not replace them.

The same evidence distinguishes two kinds of inert edit. An internal feature can change while its native selection consumer ignores it. A newly generated agenda rule can be consumed but select exactly the same directions on every observed input. Neither new source nor successful execution establishes useful invention. Strong reference-relative trajectories, actual researcher inputs and complete costs are needed alongside ask/tell feedback. Conversely, flat finite outcomes do not prove the general design space has no better system. The prospective fix's light checks are necessary technical evidence, not a corrected-model result or the missing important main-track finding.

The later corrected four-step experiment (Sections 30 and 31) does measure the original researcher world again: 23 complete own outcomes and one execution null. No complete own proposal beats its strong visible reference. All six primary returned policies are supplied references; separate own-only endpoints keep three ties and three adverse results visible. The outcome is not evidence that the whole policy space is exhausted, that unrestricted programs cannot help, or that another loss, proposer, model or budget would behave identically. Nor can warm-start retention be described as successful new institutional invention. The next research decision needs a consequential change in proposal effectiveness or an identified objective opportunity, not automatic scaling of this flat result.

32.5 Gradients are an option, not an ontological requirement

The most general target is free scientific-system optimization using available evaluation and gradient information. Exact autograd is appropriate inside a represented differentiable institution; response surrogates require native fidelity checks; likelihood-ratio estimation can surround a discrete/opaque world if its sampled design distribution is known. No universal differentiable replacement of language, retrieval and science is required.

Our measurements do not show a practical estimator advantage. The V9 fitted operator fails an unused finite-path check, V10's moving methods tie the known incumbent, V11's updates do not discover a better first best, and neither V13 gradient followup improves its previous best. These observations should guide better estimator targets and uncertainty accounting, not be hidden behind the fact that a backward pass exists. They also do not prove all gradient methods ineffective: targets, information and tiny budgets differ.

32.6 From institutional revision to recursive improvement

There are at least four different claims: a program is revised; the revised program has a better observed loss; repeated revision improves expected performance on new conditions; the revision procedure or inventor itself becomes better. Our external-author cases sometimes support the second, and local pipelines implement the first. They do not establish the third or fourth. Scientist/designer weights are not trained here. A history-aware author is not thereby a recursively self-improving optimizer.

A persistent society that changes its own organizing system is another axis. It raises state migration, policy-selection bias and counterfactual evaluation questions not answered by resetting worlds. Checkpointed designs, explicit adoption and source-bound trajectories provide infrastructure, not a completed recursive-improvement result. Keeping these levels explicit preserves the long-term motivation without selling it as current evidence.

33 Related work

ScienceUtopia studies a persistent research ecosystem and institutional interventions [Jin et al., 2026]. Our design object is the executable organizing system returned by a loss-conditioned search. This is a different question, not a claim that the source world is realistic enough to validate policy or that its reported large-model experiments are reproduced locally. Science-of-science and research-assessment work motivate plural objectives and skepticism about proxy optimization [Fortunato et al., 2018, Hicks et al., 2015].

Programmatic mechanism design already synthesizes multi-agent transition and reward components in a domain-specific language, combining stochastic program search with Bayesian parameter op-

timization [Narayanaswami et al., 2022]. Thus neither representing institutions as programs nor separating structural and numeric search originates here. Our setting asks how such search behaves when persistent, fixed-weight language-model researchers interact with scientific institutions; that application is not itself a new synthesis algorithm or evidence of greater generality.

Optimization through generated programs is prior art. OPRO uses evaluated solutions as feedback [Yang et al., 2024]; ADAS searches code-defined agentic workflows [Hu et al., 2025]; AMD-LLM evolves mechanism code with problem-specific fixing [Liu et al., 2025]. Our no-host-repair contract is a comparison choice, not a superiority claim. SGA combines LLM outer expression search with continuous inner optimization [Ma et al., 2024]. GPTSwarm optimizes agent-graph connections with score-function methods [Zhuge et al., 2024]. Thus code generation, graph search and mixing discrete design with gradients are not first contributions here.

FunSearch couples code generation to an evaluated program pool [Romera-Paredes et al., 2023]. AlphaEvolve samples prior programs and uses an evolutionary database for exploration and exploitation [Novikov et al., 2025]. ShinkaEvolve studies parent/inspiration sampling and novelty rejection [Lange et al., 2025]. Retaining our complete history is not implementing all these search policies, and our local pilots are not benchmark comparisons against them. We do not import a novelty veto into the assigned institutional objective or infer that these methods transfer successfully to our world. A prospective method comparison must hold authority and information fixed, retain nulls and report actual computational costs. Neither an evaluated archive nor ordinary feedback-driven code search supplies our missing scientific-system finding by itself.

Behavioral collapse and design cost are also not new observations: El et al. [2025] compare archive-curation strategies, evaluation- vector similarity and economic viability in code-defined question-answering agents. Our exact native input/choice classes are a different diagnostic in a persistent world, not a priority claim or evidence that our search learns from history. Those results do not establish the cause of our local stagnation.

AFlow already searches code-represented workflows with Monte Carlo tree search and execution feedback [Zhang et al., 2025]. ADAPT for Science evolves the optimization procedure as well as where to search, evaluated on unseen physical designs [Feng et al., 2026]. Mining the Design Space of Multi-Agent Workflows retains intermediate workflows and tests mined design principles on held-out searches [Cadet et al., 2026]. Thus retaining trajectories, using errors as feedback or revising an optimizer is not a standalone novelty claim. The AI Scientific Community also precedes our broad vision of interacting virtual laboratories, while describing its working implementation as under development [Braga-Neto, 2026]. Our remaining empirical question concerns which observations, actions and decision times matter when an objective selects an institution in a persistent researcher world; the current pilots do not yet supply a general answer.

SciMARS adapts scientific multi-agent harnesses through separate role and coordination editors, with an additional memory-policy editor in its extended variant [Wu and Liu, 2026]. Candidate executions inform composition and selection; revision is task-local and the outer optimizer remains fixed. Scientific harness adaptation, component-specific editing and evidence-aware revision therefore are not first contributions here. The longitudinal-information comparison being run in this program varies which native parent trajectories reach an institution author, not the researchers' memory service. Its results are not reported here before the frozen analysis closes. It is neither a reproduction nor a matched performance comparison of SciMARS. The candidate distinction is optimization of the organizing system of a persistent researcher society under declared losses; that setting alone does not establish an important empirical answer.

More direct institutional precedents also narrow the novelty claim. MACC proposes a scientific blackboard with reproducibility rewards and an adaptable, potentially differentiable incentive mechanism; it is a Blue Sky architecture and research agenda [Oyama et al., 2026]. A Society of Researchers describes persistent laboratories, competitive funding and independent review, while leaving matched organizational comparisons open [Asaria et al., 2026]. Institutional AI specifies executable governance manifests, policy programs, notices and replayable logs for multi-agent markets; its fixed-topology policy screening precedes locked independent evaluation [Bracale Syrnikov et al., 2026]. An executable institution artifact or screened social-policy selection is therefore not new here. The inspected implementation does not establish free LLM-authored topology/code search, but that scoped distinction is not a field-wide priority claim or a reproduction of its empirical results. The Station executes open-ended agent research in persistent rooms with scorable tasks [Chung and Du,

2025]; EurekAgent studies permissions, artifacts and budgets for executable discovery [Xin et al., 2026]. Neither an institutional vision nor adding verified computational tasks alone is new. Our empirical contribution must be the objective-to-institution relationship and its consequences, not priority for these general ideas.

The AI Economist studies learned social policy [Zheng et al., 2020], and LLM Economist adapts workers and a planner through context under bounded tax actions [Karten et al., 2025]. AgentSociety 2 integrates composable environments and specification-to-code [Piao et al., 2026]; SocialSystemArena compares governance specifications [Fei et al., 2026]; PolicySim optimizes exposure actions in an LLM-based social environment [Huang et al., 2026]. We focus on editable scientific organization and its longitudinal native effects. Domain specificity alone does not establish greater generality, scientific significance or priority.

Likelihood-ratio, reparameterization and control-variate methods are existing theory [Parmas and Sugiyama, 2021, Parmas and Seno, 2022, Parmas et al., 2023]. Our contribution is the explicit placement and evaluation of these boundaries in this design environment, not a new unbiased estimator for arbitrary LLM-generated institutions.

34 Limitations and responsible interpretation

The native world is itself simplified. Literature retrieval stands in for many real scientific activities; it does not conduct physical experiments, verify novel knowledge or reproduce the full social organization of science. Our extension preserves that world but does not eliminate its assumptions. The native action registry is extensible, not an already implemented semantics for every possible laboratory, government or hypothesis-testing institution.

Experiments are small and exploratory. Historical objectives and coefficients are engineering choices informed by development, not preregistered hypotheses about real science. The parameter study's five evaluation seeds and V2's five unused seeds are small; their source/design spaces differ. Other cases share one development seed, some use common-prefix replay, and external-author provenance differs from local 35B generation. These distinctions prevent pooled causal method rankings or broad model generalization. The new information comparison has three four-year development seeds; its single acceptance and variable paired differences are descriptive, not a reliable mechanism ranking or an objective-conditioned search result. The separate closed T/E pilot has only two proposals per arm, one development seed and three unused evaluation seeds. It completes objective-conditioned selection and comparison, not a reliable method ranking. Its five free-source runtime failures and absence of actual rewiring remain part of the result.

Review scores are endogenous LLM outputs, not truth. Better count, review utility or resource snapshots do not validate scientific value, equity or harm. Proxy-seeking institutions should be examined, not deployed by default. Section 13 makes the observer dependence explicit: unchanged measurement code does not imply institution-independent panels. Its fixed-record readouts neither change actual review work nor establish that any submitted institution manipulated its panel successfully. The later routing-authority campaign stops before its planned evaluation; its separately completed thirty-world fixed-submission evaluation has a different protocol and retains both unavailable graph arms. Its numerical T selection ties accept-all, and numerical E loses to one-reviewer accept-all on all three seeds. Section 14 also shows that conditional proxy incentives can coincide and that an earlier completed cohort's saved activity stock is flat. Neither establishes a global optimum or an absence of transient resource pressure. The subsequent count/content development closes eight submissions with six completions and two technical nulls. Its selected programs tie strong rules; four candidates have a reference's identical observed model inputs, while other tied scores accompany changed inputs. Its separately closed eighteen-world fixed evaluation retains all three unused seeds: T-selected ties connected accept-all rules on publication count, while its cross-objective U differences against no-review accept-all are mixed. U-selected is literal reference retention. More completed world seeds do not turn one design history into independent optimizer replication or a stable strong-rule advantage. The three advice-timing controls in Section 15 establish literal message delivery without a topic/duration or T/U gain at their reused seed. They do not establish complete behavioral equivalence: full requests, replies and some resource trajectories differ. These developer contrasts are not new optimizer histories or evidence against information institutions in general. The separately closed capacity pilot in Section 16 submits four more programs, with three completions

and one invalid virtual-source transfer. All complete programs tie strong reserve allocation rules in both T and P without changing wiring. Its two observed score points have no trade-off under any nonnegative linear mixture. This describes the finite archive, not all reachable institutions: three complete submissions do not distinguish objective alignment from a designer that preserves an incumbent's behavior. Its prepared unused-seed evaluation is not dispatched merely to enlarge that flat comparison. The later three-arm planning comparison changes actual topic/duration choices but remains a single reused-seed developer intervention; it supplies neither independent optimizer histories nor a held-out benefit claim. The subsequent free planning pilot in Section 18 is normally closed, but its own assigned losses tie its best references. The separate structural compiler-feedback search and its unused evaluation also close; their selected systems retain the hand's observed behavior. The open-code pilot instead stops after one own completion. These different closure and information conditions must not be pooled as independent method comparisons. No study of actual researchers, human intervention, independent truth verifier or real scientific-policy deployment is reported.

The source-information pilot in Section 28 has one openly reused development condition and two own proposals per isolated history. Additional source tokens and the fixed implementation checkpoint are part of the tested interaction package. Equal proposal counts do not establish equal compute or isolate comprehension, and no new unused evaluation is attached. Its source-grounded execution/selection reconstruction does not turn that limited sample into an independent method comparison. The separate priority control's static shortest-topic choices and SCRIPTED native checks are software preparation, not new real-model gains or replacements for this null result.

Reproducibility has layers: numerical reconstruction from retained reports, full raw/source re-audit, and fresh inference. The first is demonstrated for V2, the new closed information study and the T/E pilot, with explicit historical receipts for other studies; fresh inference is not guaranteed by seeds. The local 35B upstream revision/conversion recipe remains unverified. Some usage is unknown and total project FLOPs, energy and monetary cost are not measured. Appendix G states these boundaries.

This manuscript uses a retained NeurIPS 2025 preprint style for readability, without a page cap; it is not compliance with a selected future conference cycle. Human authorship/responsibility and public redistribution rights are not finalized. Experimental LLM roles and AI assistance in implementation, analysis, figures and prose must be disclosed, not represented as independent human peer review. No actual submission, public upload, model replacement is performed by preparing this manuscript. The new source-specific readiness and information-path study have their own explicit protocols and current statuses in Section 12; they do not amend historical experiments.

35 Conclusion

The system that organizes science can be an executable optimization object containing the original persistent researcher world. Institution code, mechanisms, connections and processes are not restricted to a menu of acceptance/funding knobs, and black-box evaluation can coexist with scoped gradient information without making the entire world differentiable.

The integrated evidence is mixed and informative: objective-conditioned parameter selection yields uncertain trade-offs; structural programs can alter real plans and information routes; external source redesign can improve an observed proxy while gradient followups worsen; the matched local study reveals compliance failures, no exercised topology search and no advantage over strong simple references. A canonical game-like interface makes those same successes, ties and failures inspectable.

The completed four-year T/E pilot adds an observed numerical feedback-history decrease and selection-fixed evaluation on three unused seeds. It exercises source algorithms but not new wiring; the small paired difference from reviewed accept-all does not establish superiority. These results target organizing freedom's native effects separately from interface burden and elementary proxy incentives, not scale alone. The program separates implemented capability, exercised mechanism, observed optimization and stronger claims still requiring evidence.

The closed authoring and panel diagnostics sharpen the remaining boundary: static submission is not full native execution, offered explanation is not received explanation, and an endogenous score is not independent knowledge. These are explicit experimental distinctions, not new optimizer superiority or a substitute for an important institutional-design result.

The separate thirty-world evaluation preserves the stopped search rather than completing it retrospectively. Numerical T ties accept-all's throughput; numerical E loses to a one-reviewer rule on each tested seed, and no working graph policy exists for evaluation. These outcomes motivate distinguishing the objective's incentives, available authority and actual program execution before treating larger-scale search as the missing ingredient.

Planning authority makes the distinction concrete. Advice and direct native assignment increase output under a fixed developer agenda, but the following six-proposal free search only rewrites strong all-accept selection code. Its two objectives have overlapping observed minima, and its content instrument supplies no negative accepted contribution. Full-input comparisons separate actual behavior retention from equal scalar scores. Useful autonomous invention and general optimizer advantage remain open; a larger run alone need not resolve narrow exploration.

The later two-objective source-information pilot makes the remaining limitation explicit: all four selected own systems still tie their strong parent, and no new wiring is exercised. Readiness, valid source changes and complete exports are necessary infrastructure but not the missing important scientific result. This working paper retains that boundary rather than declare the project publication-ready on the basis of implementation or a correctly measured null.

References

Ali Asaria, Deep Gandhi, and Tony Salomone. A society of researchers: Designing institutions for populations of autonomous research agents. *arXiv preprint arXiv:2610.10468*, 2026. URL https://arxiv.org/abs/2610.10468v1.

Marcantonio Bracale Syrnikov, Federico Pierucci, Marcello Galisai, Matteo Prandi, Piercosma Bisconti, Francesco Giarrusso, Olga Sorokoletova, Vincenzo Suriani, and Daniele Nardi. Institutional AI: Governing LLM collusion in multi-agent cournot markets via public governance graphs. *arXiv preprint arXiv:2601.11369v1*, 2026. URL https://arxiv.org/abs/2601.11369v1.

Ulisses Braga-Neto. The AI scientific community: Agentic virtual lab swarms. *arXiv preprint arXiv:2603.21344*, 2026. URL https://arxiv.org/abs/2603.21344v1.

Xavier Cadet, Junyan Cheng, Peter Chin, and Kyle Richardson. Mining the design space of multi-agent workflows. *arXiv preprint arXiv:2610.31709*, 2026. URL https://arxiv.org/abs/2610.31709v1.

Stephen Chung and Wenyu Du. The station: An open-world environment for AI-driven discovery. *arXiv preprint arXiv:2511.06309*, 2025. URL https://arxiv.org/abs/2511.06309v1.

Batu El, Mert Yuksekgonul, and James Zou. Inefficiencies of meta agents for agent design. *arXiv preprint arXiv:2510.06711*, 2025. URL https://arxiv.org/abs/2510.06711v1.

Chao Fei, Hongcheng Guo, and Yanghua Xiao. When agents evolve, institutions follow. *arXiv preprint arXiv:2604.27691*, 2026. URL https://arxiv.org/abs/2604.27691v1.

Zhaoran Feng, Yuchen Ni, Jianan Liu, Lemao Liu, Dejing Dou, and Kairui Feng. ADAPT for science: A multi-agent system for self-evolving scientific optimization. *arXiv preprint arXiv:2610.30851*, 2026. URL https://arxiv.org/abs/2610.30851v1.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018. doi: 10.1126/science.aao0185.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. Bibliometrics: The leiden manifesto for research metrics. *Nature*, 520:429–431, 2015. doi: 10.1038/520429a.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, 2025. Official proceedings.

Renhong Huang, Ning Tang, Jiarong Xu, Yuxuan Cao, Qingqian Tu, Sheng Guo, Bo Zheng, Huiyuan Liu, and Yang Yang. PolicySim: An LLM-based agent social simulation sandbox for proactive policy optimization. *arXiv preprint arXiv:2603.19649*, 2026. URL https://arxiv.org/abs/2603.19649v1.

Yiqiao Jin, Yijia Xiao, Yiyang Wang, and Jindong Wang. SciEvo: A 2 million, 30-year cross-disciplinary dataset for temporal scientometric analysis. *arXiv preprint arXiv:2410.09510*, 2024. URL https://arxiv.org/abs/2410.09510.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Bing He, Siheng Xiong, Yijia Xiao, B. Aditya Prakash, Josiah Hester, Srijan Kumar, James Evans, and Jindong Wang. Science utopia? closed-loop LLM simulation of academic research ecosystems. *arXiv preprint arXiv:2610.01257*, 2026. URL https://arxiv.org/abs/2610.01257v1.

Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. LLM economist: Large population models and mechanism design in multi-agent generative simulacra. *arXiv preprint arXiv:2507.15815*, 2025. URL https://arxiv.org/abs/2507.15815.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. *arXiv preprint arXiv:2509.19349v1*, 2025. URL https://arxiv.org/abs/2509.19349v1.

Jiayuan Liu, Mingyu Guo, and Vincent Conitzer. An interpretable automated mechanism design framework with large language models. *arXiv preprint arXiv:2502.12203*, 2025. URL https://arxiv.org/abs/2502.12203.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B. Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. LLM and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33940–33962, 2024. URL https://proceedings.mlr.press/v235/ma24m.html.

Sai Kiran Narayanaswami, Swarat Chaudhuri, Moshe Vardi, and Peter Stone. Automating mechanism design with program synthesis. In *Proceedings of the Adaptive and Learning Agents Workshop (ALA) at AAMAS*, 2022. URL https://ala2022.github.io/papers/ALA2022_paper_30.pdf.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131v1*, 2025. URL https://arxiv.org/abs/2506.13131v1.

Satoshi Oyama, Yuko Sakurai, and Hisashi Kashima. MACC: Multi-agent collaborative competition for scientific exploration. In *Proceedings of AAMAS, Blue Sky Ideas Track*, pages 3900–3904, 2026. doi: 10.65109/JLGE7606. URL https://ifaamas.org/Proceedings/aamas2026/pdfs/JLGE7606.pdf.

Paavo Parmas and Takuma Seno. Proppo: a message passing framework for customizable and composable learning algorithms. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/bbc9d480a8257889d2af88983e8b126a-Abstract-Conference.html.

Paavo Parmas and Masashi Sugiyama. A unified view of likelihood ratio and reparameterization gradients. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4078–4086, 2021. URL https://proceedings.mlr.press/v130/parmas21a.html.

Paavo Parmas, Takuma Seno, and Yuma Aoki. Model-based reinforcement learning with scalable composite policy gradient estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 27346–27377, 2023. URL https://proceedings.mlr.press/v202/parmas23a.html.

Jinghua Piao, Jun Zhang, Haoyu Huang, et al. AgentSociety 2: An integrated research environment for executable social science. *arXiv preprint arXiv:2607.11895*, 2026. URL https://arxiv.org/abs/2607.11895.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 2023. doi: 10.1038/s41586-023-06924-6. URL https://www.nature.com/articles/s41586-023-06924-6. Online-publication citation follows the author repository.

Weimin Wu and Han Liu. SciMARS: Multi-agent recursive self-improvement for closed-loop scientific discovery. *arXiv preprint arXiv:2611.12552*, 2026. URL https://arxiv.org/abs/2611.12552v1.

Amy Xin, Jiening Siow, Junjie Wang, Zijun Yao, Fanjin Zhang, Jian Song, Lei Hou, and Juanzi Li. EurekAgent: Agent environment engineering is all you need for autonomous scientific discovery. *arXiv preprint arXiv:2606.13662*, 2026. URL https://arxiv.org/abs/2606.13662v1.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024. Official proceedings.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025. Official proceedings.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher. The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020. URL https://arxiv.org/abs/2004.13332.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. GPTSwarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62743–62767, 2024. URL https://proceedings.mlr.press/v235/zhuge24a.html.

A Complete parameter-study context

The fixed-topology study and V2 use different sources, horizons, design spaces and losses. They are retained because they answer different questions, not combined into one leaderboard. Table 49 shows both distinct designs, rather than imply that three objective names produce three different institutions.

Table 49: Full frozen parameter settings. Topic and burden independently select the right-hand design. Funding application cost is fixed at 1 in both.

Parameter	Participation-selected	Topic/burden-selected
Acceptance rate	.496777	.113159
Funding application cost	1	1
Funding slot fraction	.390677	.327413
Resubmission cost	12.8905	16.3863
Review policy	persona	standardized
Reviewer capacity	8	2
Reviewer matching	topic	random
Reviews per paper	3	4

The retained intervals are exact nonparametric percentile bootstrap intervals for the complete LossSpec aggregate over the five declared evaluation seeds. They are not confidence guarantees for a different source or a broader model population. All nine paired loss-contrast intervals include zero. Two labels share a design but have separate stochastic held-out trajectories; comparing them as if they were distinct institutions would exaggerate diversity.

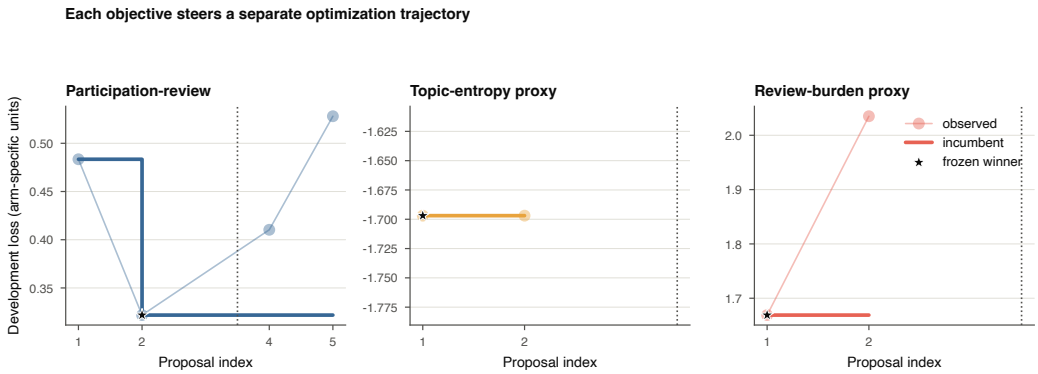


Figure 47: **Complete parameter-study development paths.** Three fixed objectives, five candidates per arm and common development conditions. Technical nulls do not become numeric penalties. Stars are frozen selections; these are not additional held-out results or an optimizer comparison.

Declared gaming flags in that study are diagnostic alerts, not invalidity or adoption vetoes. They do not override the loss. Recorded attempt runtime is at least 52.35 h across that historical study, with one interrupted attempt's wall time unknown; it is neither the later V2 runtime nor new compute for this integrated manuscript. Logical rows, physical executions, adopted trajectories and retries remain separate in the original paper slice.

B Complete structural case tables and interpretation

An initial V8 authoring attempt fails before native inference because its parent normalization does not match the exact saved request. That zero-world failure remains in provenance, not as a fabricated scientific loss or a successful attempt. The separately declared corrected-file workflow does not replace an unsuccessful native trajectory. External authors may inspect their own source and traces, richer information than a scalar-only oracle.

Table 50: All eight V8 structural-case outcomes. Graph is nodes/edges. $U = 3.25A - Q + 0.05R$ differs from V2's C. Each row is a development execution; cross-scores reuse its checkpoint. Values are display-rounded only.

Arm	Generation/draw	Graph	A	Q	R	T	U
Throughput	G0/0	16/22	9	29	27	-9	1.6000
	G0/1	16/22	9	29	27	-9	1.6000
	G1/0	17/24	10	31.3333	30	-10	2.6667
	G1/1	16/22	10	33.6667	30	-10	.3333
Utility	G0/0	16/22	0	0	0	0	0
	G0/1	16/22	0	0	0	0	0
	G1/0	18/24	1	3.6667	18	-1	.4833
	G1/1	18/24	1	3.6667	18	-1	.4833

Table 51: All six V12 free-redesign observations, including the worse clinic revision and return to its actual incumbent parent. Repetition is not new invention; Q uses original review means.

Arm	Generation	Graph	A	Q	R	T	U
Throughput	G0	17/25	10	$95/3$	30	-10	$7/3$
	G1	17/25	10	$95/3$	30	-10	$7/3$
	G2	17/25	10	$95/3$	30	-10	$7/3$
Utility	G0	16/22	3	$32/3$	15	-3	$-1/6$
	G1	17/24	1	4	15	-1	0
	G2	16/22	3	$32/3$	15	-3	$-1/6$

In V12, 32 source/context/result executions join committed native events; throughput's10 amendments per world match generated outputs and annual project state. Utility G1 emits two year 1 amendments and no additional year 2 clinic amendment. The historical code ABI does not retain a fully bound upstream plan, so a stronger complete applied-plan hash cannot be reconstructed for that slice. V13 prospectively records the full upstream plan and ordered commands, enabling the stronger join. Later provenance improvements do not retroactively repair old records.

For V13, actual eligible reviewer identities pass through a typed routing capability to the original review prompt/model/parser. The source chooses a subset of tasks, including none; the host does not choose favorable scores. Original eligibility/COI and identity legality remain explicit native semantics. A zero-review rule must author its own compatible selection consumer. The new G1 ranking uses prior reviews to predict task desirability, not to replace actual reviewer evidence in the measurement.

C Gradient assumptions, diagnostics and complete followups

C.1 What each derivative covers

For a categorical choice with logits α, the sampled score is the one-hot choice minus its probability vector. For a Gaussian latent with mean μ and fixed scale σ, the mean score is $(z - \mu)/\sigma^2$. These derivatives concern the retained sampling law. A fixed sigmoid mapping the latent to a bounded cutoff does not require a native world derivative in the LR estimator. Replacing the sampling law, truncating its support without accounting, or changing source semantics changes the estimator's target.

Leave-one-out baselines rely on conditional independence of the other design draws; they are not a license to reuse arbitrary correlated outcomes as a baseline. Controls in Eq. (5) are fixed from eligible prior history for the current scored batch. Their expectation correction must be computed for the same distribution. If the world's conditional law depends on θ beyond the sampled executable design, the stated simple score identity is incomplete. Missing technical outcomes and adaptively selected histories require their own estimand analysis.

The local response representation takes a different route: it approximates sensitivities inside the native path. A faithful hard replay verifies value accounting, not the correctness of its soft derivative. A parameter disconnected inside the explicitly represented subgraph can have derivative zero there; unknown external researcher/host responses must remain unknown. Neither LR nor the response operator differentiates arbitrary LLM source invention.

Table 52: All V9 coupled-history endpoints. C: calibration; U: unused coordinate; G: frozen gradient proposal. The proposal at .50 is an exact no-op at stored precision. The better observed .80 is not substituted for it.

Role	θ	A	Q	R	Native U	Predicted U
C0	.20	0	0	18	.9000	.9000
C1	.50	0	0	18	.9000	.9000
C2	.80	1	$10/3$	18	.8167	.8167
U0	.35	0	0	18	.9000	.9000
U1	.65	0	0	18	.9000	.8168
G	.50	0	0	18	.9000	.9000

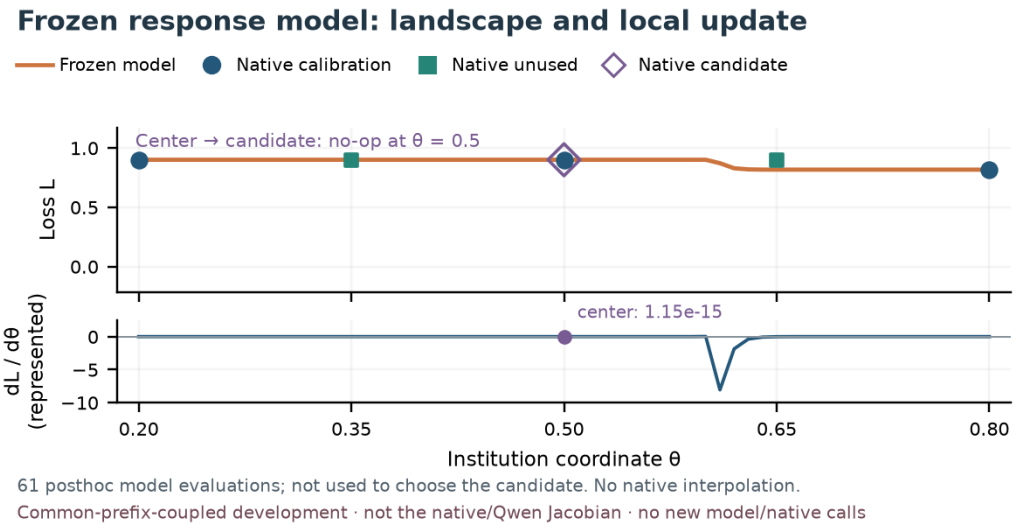


Figure 48: **Frozen represented landscape and actual native endpoints.** The plotted model evaluations are offline diagnostics, not new Qwen calls. Calibration fit does not imply unused-coordinate or finite-path fidelity. The precommitted gradient candidate remains the center rather than the known lower calibration endpoint.

Table 53: Offline temperature diagnosis at $\theta = .5$ on the same historical calibration data. These are not four native experiments or prospective temperature selection using future success.

Temperature	Center represented derivative	Calibration-loss RMSE
1	$-.24092$	.52717
.2	$-.09426$	.32974
.05	$+.01156$	.01522
.01	1.15×10^{-15}	1.16×10^{-5}

The V10 gradient steps use a fixed temperature schedule $(1, .2, .05)$ and $\theta_{k+1} = \mathrm{clip}_{[.2,.8]}[\theta_k + \mathrm{clip}_{[-.1,.1]}(-.5\widehat{g}_k)]$. They refit on only the three shared calibration endpoints and their own earlier completed endpoints. The black-box policy instead refines the widest interval adjacent to its observed incumbent, choosing the lower interval on ties. Each arm's new losses are $(.9, 49/60, 49/60)$; neither beats the shared historical best. Their six forwards contain 504 logical replies, 318 prefix replays and 186 new calls; support from shared historical calibration is counted once, not once per method or as six independent societies.

C.2 Score/control measurements and the previously missing forward

V11's two arms each have four native losses $(.7833333, 1.2, 1.2, 1.2)$. The first draw has $A = 1, Q = 11/3, R = 24$; each remaining draw has $A = Q = 0, R = 24$. The G0 score step changes Gaussian mean from 0 to $-.0407768$ and logits to $[-.0495833, .0495833]$ before G1. The final prior-control estimate has mean gradient $.742395$ and logit gradient $[.109754, -.109754]$; plain LR on the same batch gives $.945250$ and $[-.570274, .570274]$. These single noisy estimates are not empirical variance measurements.

The saved final mean $-.1150163$ and logits $[-.0605587, .0605587]$ are not reevaluated within V11. Table 54 records the separate complete two-world followup, not a retroactive extra V11 generation. Historical outcomes are provenance, not a randomized concurrent control.

Table 54: Complete fresh followup of the exact saved V11 proposal. The bounded executable cutoff differs from its Gaussian latent. Lower is better under the same U, but two draws do not estimate comparative expected improvement.

Aggregation	Cutoff	A	Q	R	U	Fresh replies
Minimum	3.549508	0	0	21	1.05	91
Mean	3.080488	2	7	21	.55	89

D V2 all-candidate cross-scores and failure accounting

The two matrices retain all 60 scheduled rows per primary objective, including three histories and four generations for each method. G/X marks denote generation/execution nulls, not zeros. Full-precision values, raw metric components and missingness remain in the numerical JSON/CSV; display rounding never affects selection. Loss scales are independent and not combined.

The 31 execution failures comprise scalar-shape misuse 12, world-envelope errors 7, decision-identity coverage 6, missing outputs 4, an extra output 1 and invalid decision/score/reference 1. These are literal observed categories, not isolated causes of designer cognition. The 25 generation-null slots exhaust the declared action budget without valid explicit finish. A final provider envelope or HTTP 200 is not necessarily a submitted executable institution.

The distinction between permission and use is also numerical: 40 of 64 completed specifications change selection source; existing parameter-change row counts overlap (bias 26, memory cost 24, threshold array 25, process count 20, router threshold 24). No declared node, wiring, ABI, state or schedule changes are observed. String/source identity, AST identity and arbitrary-input behavioral equivalence are separate notions, none a hidden novelty gate.

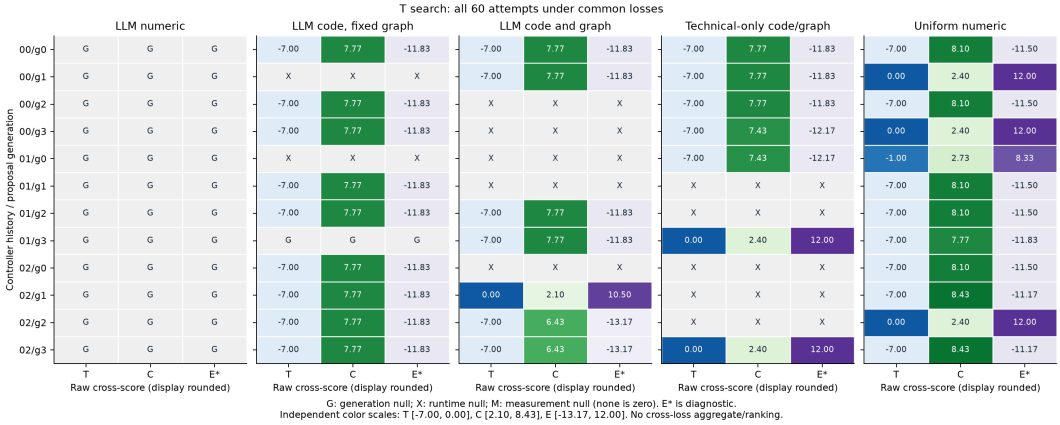


Figure 49: **Every V2 throughput-arm attempt under T/C/E.** These are development cross-scores, not E-optimized selections or extra worlds. A favorable E observation in a T arm is not relabeled an E-search result.

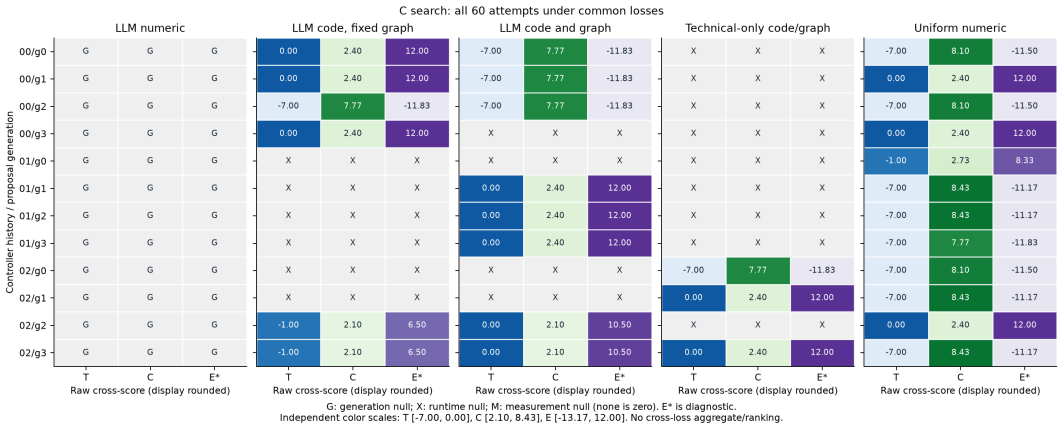


Figure 50: **Every V2 C-arm attempt under T/C/E.** All failed attempts, initial selections, ties and worse generations remain; there is no best-run-only or successful-proposal-only reporting.

E V2 costs, resources and separate longer horizon

Table 55: Search-only V2 components, pooled across three four-slot histories. Designer seconds are host model-call time, not end-to-end wall or total compute. Scientist replies include partial worlds. Zero author calls in uniform sampling does not make native evaluation free.

Loss	Method	Complete/12	Author calls	Author s	Scientist replies	World dirs
T	Numeric LLM	0	192	2497.2	0	0
T	Code/fixed	9	172	2219.1	859	11
T	Code/graph	7	168	2413.4	739	12
T	Technical-only	7	168	2318.9	745	12
T	Uniform	12	0	0.0	1097	12
C	Numeric LLM	0	192	2580.9	0	0
C	Code/fixed	6	168	2519.7	682	12
C	Code/graph	8	167	2770.8	826	12
C	Technical-only	3	160	2361.9	459	12
C	Uniform	12	0	0.0	1107	12

Search records 6514 scientist replies and 640 institution-code executions. Its 95 world directories, plus 9 fresh references and 90 physical evaluations, yield 194 response journals, not 194 completed forwards: completed forwards total 163. Generation-null slots never dispatch a world. Proposal requests 1387 have 1385 final signals; validity/submission is checked separately.

Table 56: Measured V2 response-token scopes. Cached input counts are subsets, not extra tokens. Model-load usage and total-project compute remain unknown.

Scope	Responses	Input	Cached input subset	Output
Scientists incl. partial worlds	16043	15547758	2976386	1473665
Designer proposals	1387	25969867	3102175	267057
Designer calibration	84	808832	22464	3819

The 1764 auxiliary intents comprise 1387 proposal, 84 calibration and 293 role-load intents. Loads separate 99 designer and 194 scientist contexts; their native tokens are unreported, not zero. Native/provider/host durations overlap and cannot be summed as project wall time. Overall writer wall is known, but exact end-to-end per-trial wall allocation is not reconstructed from component data. Retrieval, container orchestration, model loading, prior development and this coding assistant's compute are not a measured project FLOP/energy/money total.

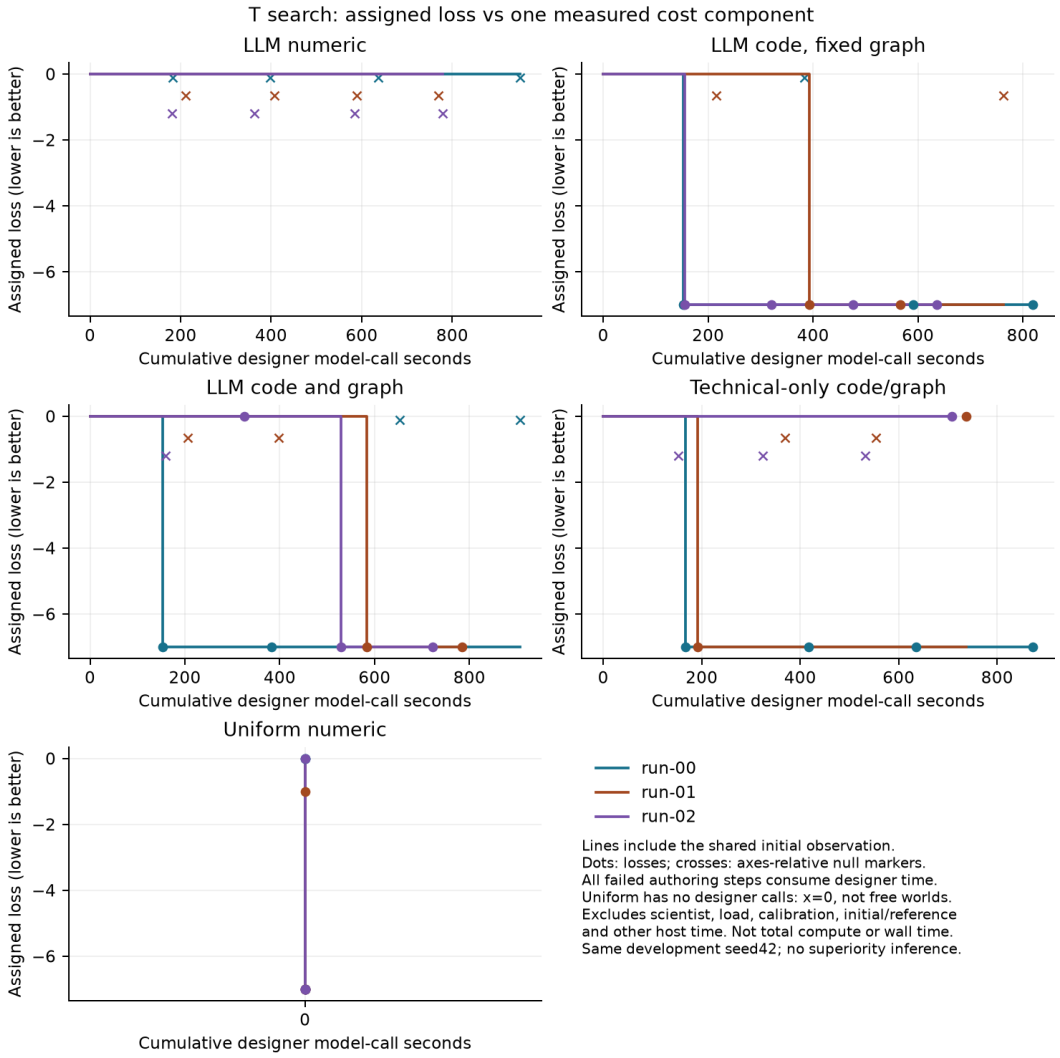


Figure 51: **V2 T paths against cumulative designer host model-call time.** All failed author steps are charged. Uniform is at zero designer time, not zero world cost. Shared initial observations/calibration have separate excluded costs; this axis is a measured component, not equal-compute superiority.

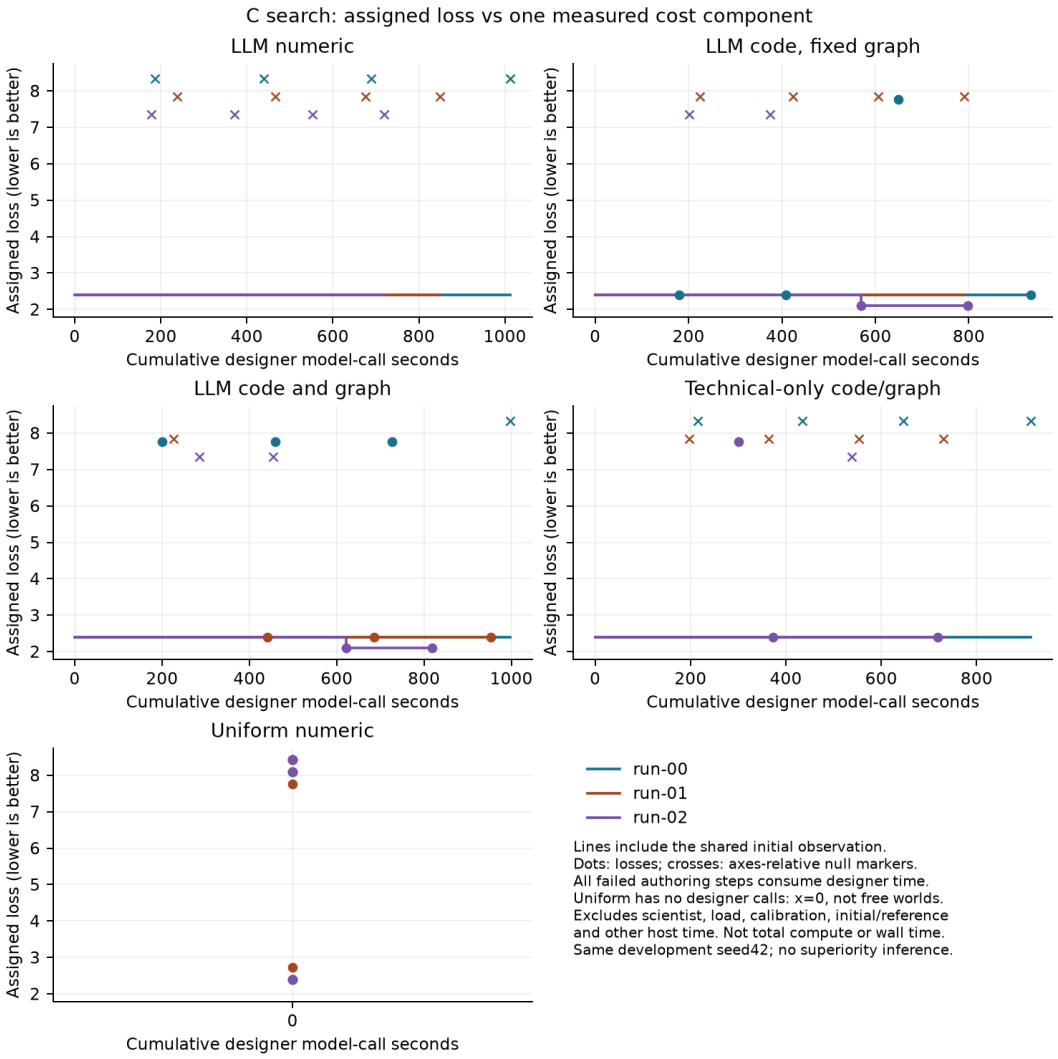


Figure 52: **V2 C paths against the same measured cost component.** Three histories are shown without treating them as independent environments. Missing component times would stop the cumulative axis, not be imputed as 0.

Outside metrics retain accepted decisions, original accepted peer-score sum, review records, final original active authors, final author resources, unreviewed acceptances, rejections and institutional-context deliveries. Two-year resource means/ranges include initial 616(585–630),C code 627(595–640) and T uniform 632(620–640). Descriptive overlap is not an equivalence test; differences are not isolated causal benefits. Four-year initial 720,C code 780 and T uniform 760 remain single observations.

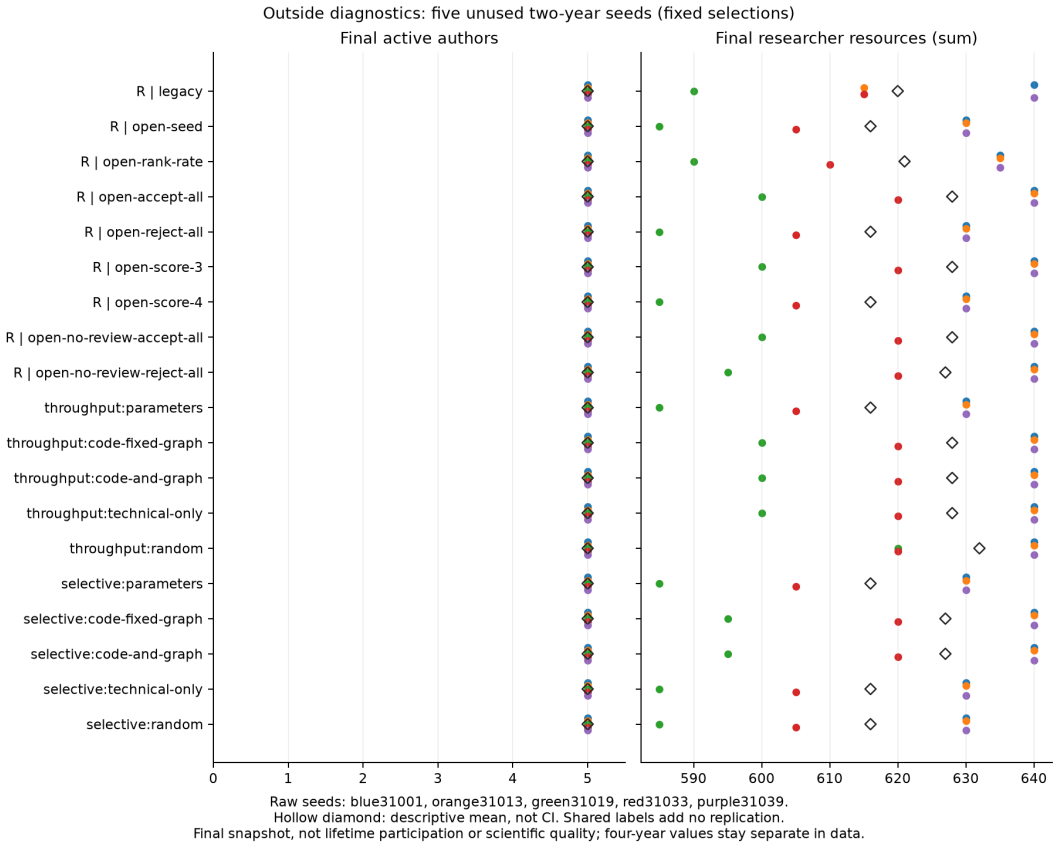


Figure 53: **V2 outside activity/resources on all selected/reference labels.** The full 0–5 activity axis shows no final active-author separation. Five points and descriptive means retain shared-policy identities. Native resources are not money, welfare, scientific truth or an additional selection constraint.

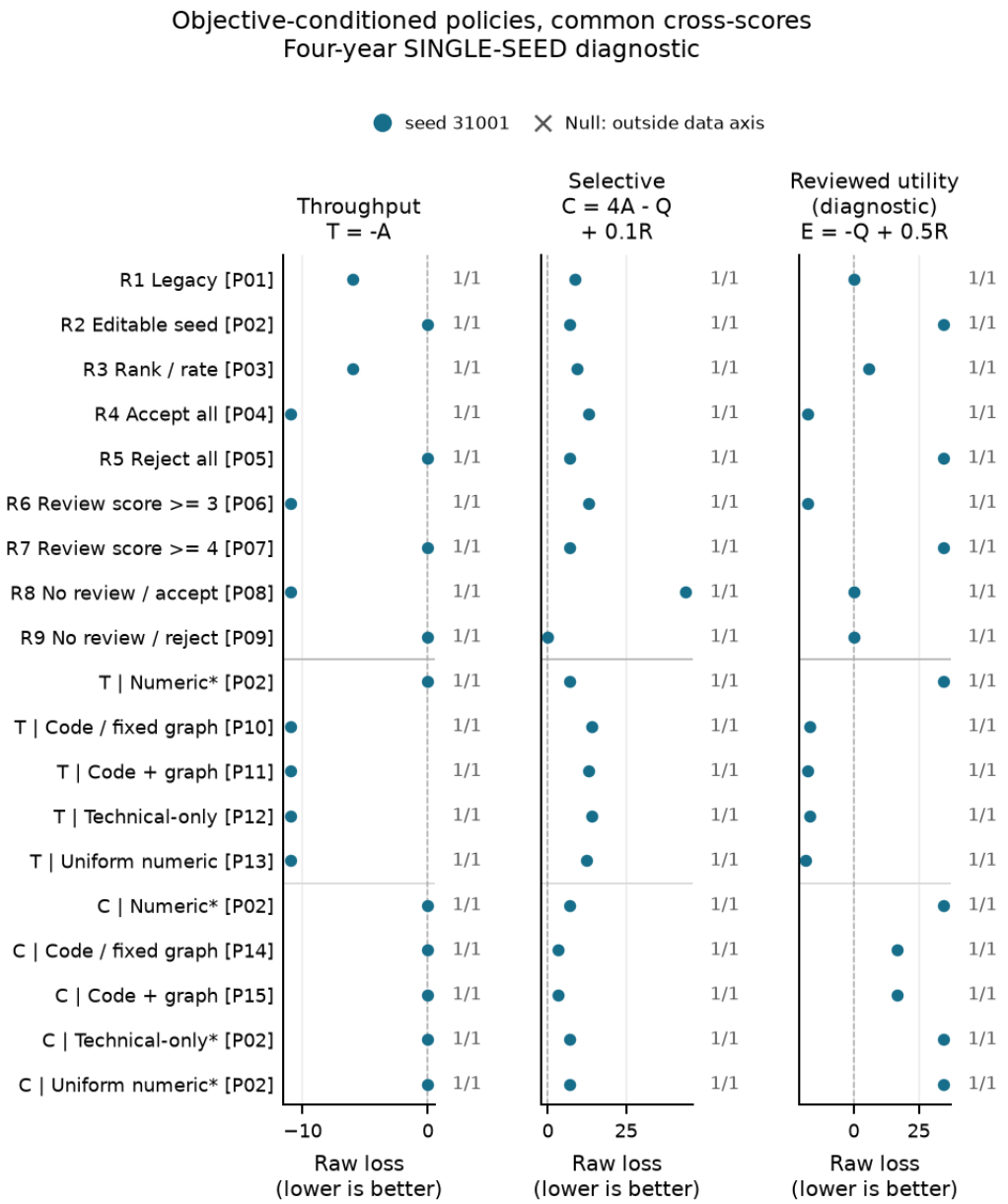


Figure 54: **V2 separate four-year diagnostic.** One seed, no confidence band or extra independent five-seed sample. Same labels/sharing/loss definitions as the two-year plot. A repeated exact policy is not another observation.

F Fixed-record incentives versus dynamic policy optima

For V2, freeze N realized records, original reviewer means q_i (zero without review) and review count R. If $z_i \in \{0, 1\}$ only changes acceptance on those records, then

$$T^{\mathrm{rec}}_{\min} = -N, \tag{17}$$

$$C^{\mathrm{rec}}_{\min} = 0.1R + \sum_i \min(0, 4 - q_i), \tag{18}$$

$$E^{\mathrm{rec}}_{\min} = 0.5R + \sum_i \min(0, -q_i). \tag{19}$$

Zero coefficients remain ties. For a full measured trajectory with $A > 0$, $C < 0$ requires $Q/A > 4 + 0.1R/A$. These elementary relations explain declared incentives; they are not a new theorem or an optimal policy for the dynamic world. Decisions affect later resubmissions, reviewer work, memory and resources. The original reviews are endogenous rather than a fixed true-quality oracle.

The fixed-record projection contains 219 identities: 163 complete forwards and 56 unavailable entries. It does not execute 219 new worlds. Missing scores can make a relaxation unavailable without invalidating a different measurable loss. Neither this algebra nor AST/source inspection establishes equivalence on unobserved inputs.

G Reproducibility, provenance and editorial scope

G.1 Distinct execution identities

The parameter study pins source `1078279d1ce8` and a simultaneous winner seal `254c3366e921`. Structural V8 uses `425ec10396b1`; free redesign V12 uses `ecab4b45c3ec`; nested V13 uses `fdf4d1c16964`. The V10 method-path source is `5ea52a520d7f`, and V11's native source is `acf3a2e`. V2 uses `b6654adefc42`. Full hashes, protocols and original audit identities remain in machine-readable evidence records. Historical reset worlds, coupled prefixes and source-equivalence bridges are not relabeled a same-source matched experiment.

The scientist alias is `qwen3:8b`; the designer alias is `qwen3.6:35b-mlx-64k`. Actual designer context is 40960 despite the alias; the manifest/digest and model-load observations are retained. The upstream revision and conversion command are unknown. Approved five HEAD-only registry checks return advertised descriptor matches for three sampled local tensors and two controls, with no body retrieval. They do not establish tensor contents, all 986 descriptors, conversion provenance or fresh-machine reproducibility.

V2 uses 103107 CS 2016–2019 SciEvo records with pinned corpus/retriever binding. The corpus is a literature-grounding substrate, not a license to redistribute every underlying document or proof of scientific validity. Current public model tags are not assumed equivalent to local retained weights.

G.2 Audit and reconstruction tiers

V2 closes with a matching normal terminal record, actual process exit 0 and closed progress. Selection precedes first evaluation. Two read-only full audits match byte-for-byte and verify 278 source blobs and 7038 raw files. A pinned order-only adapter reconstructs equal-valued plan/selection mapping order lost by sorted JSON. It preserves values, hashes, list chronology and original guards; the unmodified frozen audit CLI is not claimed to pass. Replaying an auditor does not create independent scientific evidence.

The exact retained closed-report archive supports numerical reconstruction: 120 statuses/ties and 90 evaluations rejoin, 462 T/C/E losses recompute from measured $A/Q/R$ with zero residual, and 22 PNGs match the reviewed originals. Exact auxiliary/scientist cost projections separately rebuild the full supplement. This is not raw inference or independent replication. Paper-only 13-path subsets rebuild offline with cached tooling; rendered pages and text match their reviewed variants, not necessarily binary PDF bytes. A cache-free clean-machine build and fresh-model inference have not been established.

The separate12-world information study at source `215311c` closes with two byte-identical full original-source audits. Its retained report, cost and selector-trace projections are not pooled with

V2. A numerical rebuild recomputes all36 T/E/C cross-scores, means and nine paired differences from all12 retained observations, regenerates identical PNG/SVG/report bytes and produces the manuscript table. The original tensor compiler also replays all48 recorded annual selector outputs and memory transitions without interventions. Neither operation is new inference, an independent scientific replication or a counterfactual loss evaluation. The separate T/E pilot at source `62388a7` also closes: its two full audits match all 49 physical worlds and 1,859 raw files. The retained-data rebuild recomputes 132 available physical cross-scores, all 16 proposal slots, fixed selections and 96 paired reference comparisons, regenerating the actual report, two PNG/SVG figures and component CSV. Its three evaluation seeds and shared exact policies are not pooled with V2 or the information intervention.

The repository exposes reproduction commands, compact JSON/CSV projections, source-bound receipts and figure provenance. Large caches, reply databases, corpus and model weights remain outside the publication subset. Complete numerical projections include operational locators and are not certified anonymous supplements. Author-name removal alone is not anonymity, rights clearance or permission to upload. Human authors must review and accept responsibility before an actual submission.

G.3 What was deliberately not expanded into the scientific narrative

Earlier local-author failures, copy controls, transport fixes and loop-retention pilots are retained in the repository. They inform literal execution, context-budget and completion checks. We do not count test-suite sizes or a repaired editor protocol as a scientific discovery. S1's strong controls and V1's eight-slot feedback pilot motivated V2 but are not pooled with its fresh references. Numerical miniature/unit worlds verify implementations and do not replace native-world results. Repeated UI screenshots are omitted in favor of one actual evidence view.

These editorial choices remove duplication and weak evidence for broad claims, not entire research tracks. This integrated manuscript restores the design, gradient, structural, objective and UI arguments while keeping every retained study's limitations. A longer paper is not new evidence or an automatic remedy for the significance gap identified in the prior submission review.