WHAT HANDOFF LOGS HIDE: PAIRED ACCOUNTING FOR LLM-AGENT WORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Lower successor work can reflect useful inherited state, correct automation, or unfinished work hidden by incorrect closures. We introduce an all-request paired protocol that separates these sources by executing successors from nested checkpoints and from the original state, including requests the base agent later closes. Across 1,062 base runs and 9,398 takeovers in two environments, successor work generally decreases, yet four primary τ-bench pairings increase total actions by +0.52–+4.02 per incoming request; the sign depends on the successor. A contemporaneously interleaved interface experiment obtains 93/96 planned triplets. Interpreter preservation with corresponding instructions reduces successor work by 3.44 actions (95% interval 2.80–4.08), with quality equivalence unresolved. In an all-request study of held-out task variants from 126 known AppWorld families, proportional allocation underestimates incorrect-closure work in five of six settings under simultaneous intervals. A frozen source correction reduces allocation error by 0.31 and 0.41 actions per incoming request in two secondary gpt-oss settings; exploratory stratum-matched comparisons retain both gains. The prespecified Gemma test remains unresolved. The transfer condition characterizes when the source correction helps and why a positive association alone is insufficient. These measurements distinguish sources of apparent savings and delimit transfer; they do not measure error-repair cost or establish a better deployment policy.

1 INTRODUCTION

An agent that closes more requests can still increase whole-system work. It may spend more actions than it saves for a successor, leave a costly handoff interface, or close requests incorrectly so that unfinished work disappears from the log. These explanations suggest different interventions: change the allocation of actions, preserve handoff state, or verify closures. Success and total cost describe the outcome but do not fully identify its source.

We study attribution using model generations in a bounded workflow, reported jointly with failure, tokens and prices. For every incoming request we execute a successor from nested checkpoints of one budget-agnostic base trajectory, including the original state even when the base agent later closes the request. An exact accounting separates inherited-state work changes, original work removed by correct closure, and work hidden by wrong closure. Independent policy cohorts can measure total cost; matched takeovers additionally identify inherited work and the combined work of closed requests. We isolate the information still missing: how that closed work is associated with correct versus incorrect termination. When same-request original runs are already available, this needs a join of their labels, not additional executions.

Across 1,062 base runs and 9,398 takeovers in public service and application-operation environments, less successor work often coexists with greater total work. In four primary τ-bench pairings, total actions rise by +0.52–+4.02 per incoming request, consistent with ordinary relative execution efficiency. Selection changes two AppWorld takeover-work contrasts by 3.68 and 4.20 actions per takeover, or 0.96 and 0.68 per incoming request; these are attribution differences, not policy savings.

Two interleaved studies examine the handoff interface and closure attribution. On 93 valid triplets from 96 planned scenarios, preserving the interpreter with corresponding instructions uses 3.44 fewer successor actions; missing validation cases and failure uncertainty limit the deployment inference. The all-request study executes both bases and a shared original-state successor on held-out variants

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

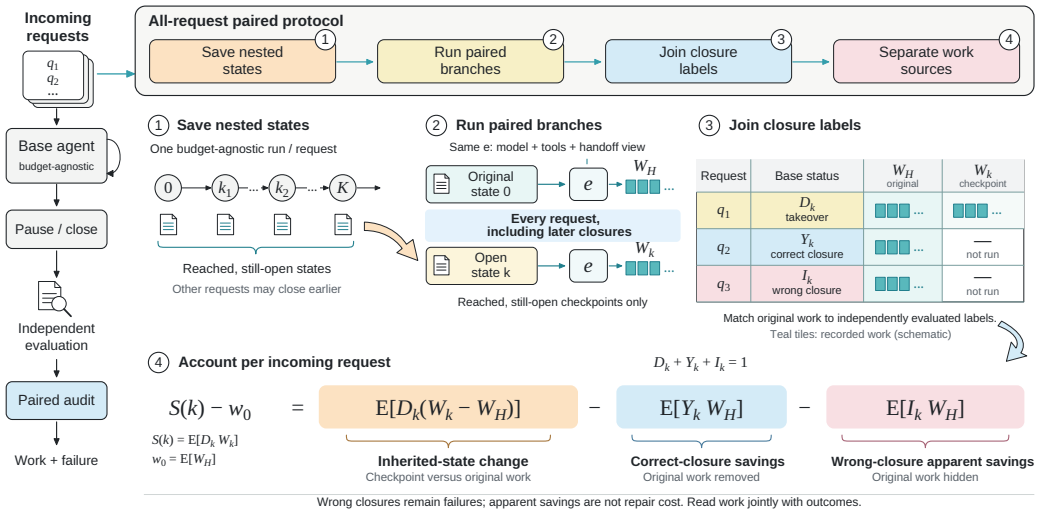

Figure 1: **All-request paired accounting.** (1) One budget-agnostic base run supplies nested, still-open states; other requests may close earlier. (2) The same successor configuration e executes from the original state for every request (W_H), including later closures, and from reached, still-open checkpoints (W_k). (3) Records are joined by request to independently evaluated closure labels; rows and teal work tiles are schematic, and dashes denote unexecuted branches. (4) Per incoming request, the signed sum of inherited-state work change (amber), correct-closure savings (blue), and wrong-closure apparent savings (rose) equals $S(k) - w_0$ (Equation 2). Wrong closures remain failures.

from 126 known AppWorld families. Linking original work to closure labels detects underallocation of incorrect-closure work in five of six settings. A frozen source association improves that allocation in two secondary gpt-oss settings, while the prespecified Gemma test remains unresolved. We derive the transfer condition: an incorrect/correct work association must be strong enough to justify the source correction, which can otherwise overshoot. The paired measurements test whether the source correction transfers.

Contributions. We contribute (1) an executable all-request paired protocol that identifies inherited-state and closure-associated work beyond complete aggregate evaluation; (2) cross-model work diagnoses with a contemporaneous interface intervention; and (3) an explicit allocation-transfer condition tested with frozen source information on held-out task variants. These are measurement and diagnostic contributions. They do not establish a general cost law or a superior deployment controller.

2 Related Work

Deferral, routing and escalation. Learning to defer chooses between a model and an expert (Madras et al., 2018; Mozannar & Sontag, 2020), including unseen experts (Strong et al., 2026). Cascades allocate queries across models, while routers select a model before generation (Chen et al., 2023; Ong et al., 2024). Inside a trajectory, ReDAct defers single steps to a larger model under high uncertainty (Piatrashyn et al., 2026), and TACIT-Switch learns a permanent handoff time from teacher-annotated, interval-censored intervention times with a mixture-cure threshold model on ALFWorld and DABench (Lei & Huang, 2026). These methods use outcome or uncertainty signals to allocate model calls. We keep TACIT-Switch's one-way handoff in a TACIT-style baseline adaptation; our target is the work and failure a specified successor incurs from the state the first agent leaves behind, per incoming request, including requests the agent closes.

What a handoff costs the receiver. Handoff Debt interrupts coding agents at deterministic points and runs successors under four handoff conditions, from repository state only to structured notes (KC

& Budathoki, 2026); our AppWorld interface control is the same kind of comparison. The Handoff Tax varies handoff direction, timing and interface between a low-cost and a high-cost model and finds that escalating a full trajectory recovers less than half of the quality gap at a substantial cost premium (Ganz et al., 2026). Do Not Restart proposes a residual contract for resuming stateful tasks (Deng et al., 2026); multi-turn experiments find persistent errors after early wrong turns (Laban et al., 2026). The Handoff Tax evaluates switched requests and its baselines on the same matched subset; Handoff Debt pairs alternative views at an interrupted state, including already-solved states. Those designs answer conditional handoff questions. We account for the entire incoming population and execute original-state successors for correctly and incorrectly closed requests. This separates closure selection from inherited-state effects. Our unmatched survivor comparison is a diagnostic of incomplete logs, not a description of either prior study's matched comparison.

Automation and the work it leaves; tools we reuse. That automation removes the easy work and leaves the hard remainder is a classical observation (Bainbridge, 1983), echoed by reports of AI "workslop" (BetterUp Labs & Stanford Social Media Lab, 2025; Baltes et al., 2026) and contested productivity estimates (Becker et al., 2026); we measure a narrow, executable version. We reuse a DeepHit-style competing-risks likelihood (Lee et al., 2018), rollout over committed options (Bertsekas et al., 1997) and betting confidence bounds (Waudby-Smith & Ramdas, 2024), without claiming these components as contributions. The environments are τ-bench (Yao et al., 2024; Barres et al., 2025) and AppWorld (Trivedi et al., 2024).

3 Recovery Accounting

Setting and objective. A request arrives and a *base agent* acts on it. An *action* is one model generation sent to the environment or the user, complete once the responses it triggers have been delivered. After each action the request is either open or closed by the agent; an independent evaluator decides whether a closure is correct. An open request can be handed to a *successor configuration* e: a model, its tools, and a fixed deterministic rule for what it sees of the inherited interaction. The base agent may also hand off on its own (a *native transfer*). A policy π decides from the public state when to hand off, and is charged per *incoming* request,

$$J(\pi) = \alpha\, B(\pi) + \nu\, S(\pi), \qquad J_q(\pi) = J(\pi) + \lambda\, \mathrm{Fail}(\pi), \tag{1}$$

where B and S are the expected numbers of base and successor actions, Fail is the probability that the request ends unresolved or wrongly resolved (an incorrect autonomous closure is a failure), and α, ν, λ are declared prices; tokens and dollars are reported separately. The fixed-depth policy C_k lets the base agent act for at most k actions; C_0 is immediate handoff and C_K, with K the action cap, is the *status quo*: the agent runs until it closes the request, transfers it, or is cut off.

Three terms. Let D_k, Y_k and I_k indicate that under C_k the request is taken over, closed correctly, or closed incorrectly, let W_k be the successor's work when it takes over at depth k, and let $W_H = W_0$ be its work under immediate handoff, which is defined for every request. Because $D_k + Y_k + I_k = 1$,

$$\underbrace{\mathbb{E}[D_k W_k]}_{S(k)} - \underbrace{\mathbb{E}[W_H]}_{w_0} = \underbrace{\mathbb{E}[D_k(W_k - W_H)]}_{\text{inflation}} - \underbrace{\mathbb{E}[Y_k W_H]}_{\text{skim}} - \underbrace{\mathbb{E}[I_k W_H]}_{\text{false savings}}. \tag{2}$$

Inflation is what continuation does to the requests the successor still inherits. A negative value means less successor work on those requests; a positive value means more. Calling that change useful progress or damage also requires checking successor outcomes and the handoff interface. *Skim* is successor work that correct automation removes. *False savings* is successor work that disappears only because a request was closed wrongly; it is paid back as failure, since $\mathrm{Fail}(k) = \mathbb{E}[I_k] + \mathbb{E}[D_k Z_k]$ with Z_k the successor's failure indicator. The identity holds for any successor-side quantity (failure, tokens, dollars).

Base logs alone do not identify the three terms. Table 1 separates complete policy evaluation from recovery attribution. The closed-work identity is a population relation; independent cohorts need not balance exactly in finite samples. If original-state runs already exist on the same closed requests and retain request IDs, level 3 requires only a join, with no new model calls. Ordinary single-path deployment does not jointly observe these branches. Appendix G.1 gives populations with identical success, total work, original-work marginals and matched takeovers, but different $\mathbb{E}[I_k W_H]$. Identification implies no predictive or deployment advantage.

Table 1: Cumulative information levels under a common incoming population and fixed execution/evaluation rules. All expectations are per incoming request; $n_D + n_Y + n_I = n_k$.

Evaluation data	Identified quantities
1. Complete independent randomized C_0 (n_0 requests) and C_k (n_k) cohorts: actions, closure labels, terminal outcomes.	$B(k) + S(k) - w_0$, $\mathrm{Fail}(k) - \mathrm{Fail}(0)$; $\mathbb{E}[I_k]$, $\mathbb{E}[Y_k]$.
2. Add original-state runs for the same n_D takeovers in the C_k cohort; retain work and terminal outcomes, weighted by n_D/n_k.	$\mathbb{E}[D_k(W_k - W_H)]$; closed-work total $w_0 - \mathbb{E}[D_k W_H]$.
3. Add original-state W_H linked to closure labels on the same $n_Y + n_I$ closed requests.	Separate $\mathbb{E}[Y_k W_H]$ and $\mathbb{E}[I_k W_H]$.

Why unpaired comparisons mislead. The natural observational comparison, handed-off requests after long runs against all requests, equals

$$\mathbb{E}[W_k \mid D_k = 1] - w_0 = \iota_k + \frac{1 - p_k}{p_k}\big(w_0 - \mathbb{E}[W_H \mid D_k = 0]\big), \qquad \iota_k = \mathbb{E}[W_k - W_H \mid D_k = 1], \tag{3}$$

with $p_k = \mathbb{E}[D_k]$ (Proposition 1). The second term is selection. It is positive whenever the requests an agent closes are cheaper for the successor than average, whatever continuation does to the state. The paired term ι_k measures the inherited-state work change on the same requests. A post hoc audit of the reported tables measures both sides of this identity with clustered uncertainty (Appendix F). The survivor comparison understates the reduction in successor work by 3.68 and 4.20 actions *per takeover* at depth 16 in two AppWorld cells, or 0.96 and 0.68 per incoming request. These are attribution differences, not policy savings.

When autonomy pays. Since $J(k) - J(0) = \alpha B(k) - \nu\{S(0) - S(k)\}$, for $B(k) > 0$ and $\nu > 0$, depth k is cheaper than immediate handoff if and only if

$$\alpha/\nu < r^*(k) := \frac{S(0) - S(k)}{B(k)}, \tag{4}$$

the *break-even price of autonomy*: successor actions saved per base action spent. If $r^*(k) < 1$, total actions rise and a work-cost reduction requires cheaper base actions; if $r^*(k) \leq 0$ there is no strict work-cost saving even when the base agent is free. r^* is always read next to $\mathrm{Fail}(k)$.

4 Nested Same-State Branch Tables

Construction (Figure 1). Run each base trajectory once, without revealing budgets or handoff decisions. At every reached, still-open grid depth $k \in \mathcal{G}$ and any native-transfer or cap event, save the public interaction and environment state and execute each successor configuration from that snapshot. A row records the evaluated outcome, successor work up to $B_{\max}$, tokens by role and typed status. Budget-agnostic prefixes yield all C_k from one trajectory, with same-request contrasts across depths and identical inherited states across successors.

One boundary rule. Checkpoints, caps, training labels and live execution all stop before a new agent generation, after pending environment and user responses arrive. The native-transfer action is charged to the base agent; its transfer call is excluded from the successor's interaction because it is the handoff itself. Scripted-agent tests check these rules (Appendix B).

What counts as a closure. An episode ending without handover or cap is closed. The τ-bench simulated user can also end it: of 108 failed closures, 17 end after consent but before the requested action runs (12 are Gemma's), and 16 use transfer or out-of-scope markers. Main results keep benchmark labels; Section 5.1 and Appendix E report handing these episodes over or excluding them.

State fidelity. τ-bench strict replay raises if a state-changing call fails to reproduce its logged result. AppWorld restores checkpointed app databases into a fresh process and interpreter; predecessor

variables are not inherited. Requests enter the cohort before execution. The sequentially collected primary branch tables retry infrastructure failures; none remained unresolved there. The reconstructed-survivor interface study retains failed validation scenarios as missing, with all attempts and its separate restart rule reported in Appendix H. The all-request study executes base trajectories and rotates interleaved branches, with its distinct state gate and retained partial continuation in Appendix I. Interpreting the sequential-cohort paired differences as inherited-state effects assumes comparable successor configurations across branches. Missing served identifiers and execution times prevent verifying this assumption for the sequential collections (Table 2).

Evaluating stopping rules. A public-state stopping rule fixed independently of the evaluation table does not perturb the budget-agnostic run. Its selected row is therefore a deployment-outcome draw (Proposition 3). The table supplies accounting terms, model constants, controller labels and baseline values without additional execution or off-policy correction. Live execution agrees within wide intervals (Appendix E).

5 Empirical Diagnoses

Setup. *Environments.* τ-bench `airline` and `retail` (164 individually authored requests; LLM user simulator) and AppWorld (203 tasks from 105 scenarios; ReAct code agent). *Agents.* Four open-weight base agents on free endpoints: Gemma-4-31B, gpt-oss-120b, Llama endpoint, Nemotron-3-Super-120B. By a *pairing rule* fixed in advance a successor must beat the base agent's solo success by ten points on a probe of training requests: GPT-5.6 Luna qualifies for every base agent and is the primary successor; GPT-5.6 Terra qualifies for three and, for cost, is run with Gemma on the 65 test requests; gpt-oss-120b is kept as a successor *not* stronger than most base agents. Names refer to API registry labels; retained provider identifiers and snapshot gaps are listed in Table 2. *Scale.* 1,062 base runs and 9,398 successor takeovers in the main tables (including the capability-probe rows), plus replicate, control and prospective runs described below. The artifact includes the API spend ledger, which also includes exploratory studies outside the paper. *Inference.* Contrasts are paired within request; unless otherwise stated, intervals are one-sided 95% bootstrap bounds that resample requests (scenarios in AppWorld), simultaneous over depths where the depth is chosen from the data. The failure price is $\lambda = 2w_0$ with w_0 measured on training requests. Prices, thresholds, splits and the primary analyses were fixed before the corresponding outcomes were seen. Dated amendments record additions and the data already available. Analyses chosen after their outcomes are marked *post hoc* (Appendices B and H). Robustness to the user simulator (the joint rise persists: +3.51 [+2.31, +4.75] for gpt-oss-120b $\rightarrow$ Luna with another simulator), to live execution of stop rules, and to the judge is in Appendix E.

5.1 Total work and inherited-state work can move differently

Autonomous completion is nondecreasing along a nested base trajectory and is zero at immediate handoff. The substantive endpoint test is whether base actions exceed the saved successor actions. All four primary pairings increase total actions by +0.52–+4.02 per incoming request (Figure 2); three meet the prespecified joint-rise criterion, while Llama's lower completion bound is zero. This is compatible with ordinary relative model efficiency. A post hoc audit of all 26 adjacent *positive*-budget transitions supports a joint rise in only one under global 95% bands: Nemotron from 8 to 12 actions gains 28.05 [15.60, 40.50] completion points and costs 0.70 [0.13, 1.26] additional actions. Its failure change is +1.22 [-9.86, 12.30] points, so quality equivalence is not established.

Successor work itself decreases in 165 of 178 contrasts, and increases detectably in none; the total rises when base work costs more actions than it removes. The sign depends on the successor: Gemma with Terra reduces total actions by 0.77. Sensitivity to simulated-user endings, common-failure comparisons, alternate user simulation and token prices is reported in Appendix E. The main failure intervals do not establish preserved quality: increases of up to 15 percentage points among takeovers remain compatible with the data.

A second successor family. Claude Opus 5, run after its plan was frozen († points of Figure 2; Appendix E), qualifies for all three capable agents. On the 145 requests outside its probe the joint rise is supported for two (gpt-oss-120b +0.99 [+0.56, +1.45], Nemotron +2.99 [+2.41, +3.57]) and

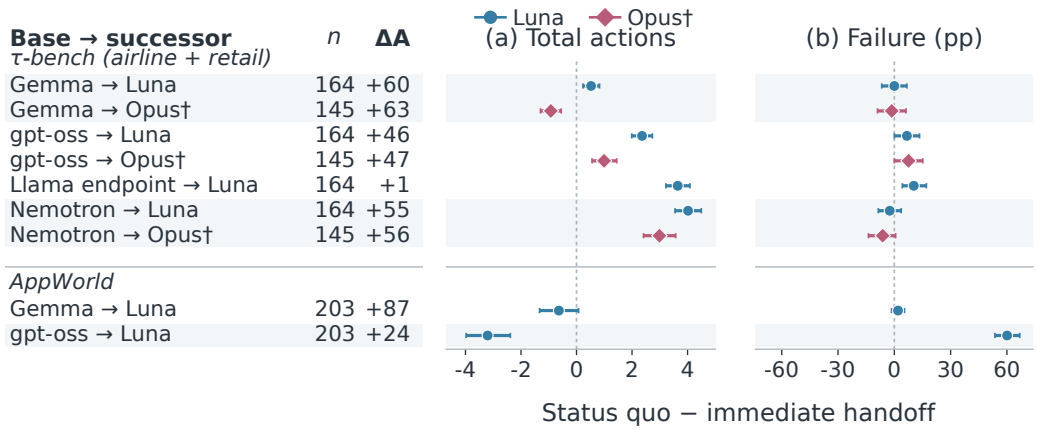

Figure 2: **Status quo minus immediate handoff: primary and second-family pairings.** Points and segments: means and paired one-sided 95% bootstrap bounds for total actions and failure (pp). n: requests; ΔA: autonomous completion (pp). Color/shape identifies the successor (blue circles: Luna; rose diamonds: Opus); gray bands separate base agents. †: 145-request second-family cohort. Full 16-pairing contrasts, including Terra and weaker-successor controls, are in Tables 3 and 4.

reversed for Gemma (-0.92 [-1.30, -0.55]), as the exchange form predicted in advance from Opus's from-scratch work (10.8) and Gemma's break-even (9.00). A no-progress model predicts the same signs (post hoc).

Prospective domain replication. We froze a replication on τ-bench `telecom` before running any of its 114 evaluation tasks (tasks share atomic faults, so intervals are likely too narrow). Luna resolves only 25% of a 20-task probe from scratch, so it fails the pairing rule for the three agents that solve 45–80% and only Llama (0%) qualifies; there the declared hypothesis is supported: the status quo costs +4.24 [+3.55, +4.94] whole-system actions more ($n = 114$; Table 8), the overhead of an unsuccessful predecessor, not the joint rise.

5.2 Wrong closures and the handoff interface

Apparent savings. With a weak base agent (gpt-oss-120b) successor work falls from 14.3 to 1.0 actions and whole-system actions fall too; but 67% of requests end in a wrong closure (74% of its closures), most of the saving is *false savings* (Figure 7, right), and final failure rises from 11% to 71%. With a capable base agent (Gemma, 87% autonomous completion) the two extremes cost about the same, but a takeover after 8 base actions puts the total at 19.4 actions, +4.49 [+3.70, +5.28] above immediate handoff (Figure 5, upper right).

A contemporaneous interface intervention. The interface test was frozen on a reconstructed Gemma depth-8 survivor cohort: one eligible task per scenario, with original-state, fresh-interpreter and preserved-interpreter Luna branches interleaved in randomized rotating order. Application states are checked before calls; actual UTC times, served identifiers and raw responses are retained. Of 96 planned triplets, 93 pass validation; three missing cases and 13 whole-triplet infrastructure restarts remain explicit (Appendix H). In these complete cases, preserving the interpreter with truthful availability instructions changes successor work by -3.44 $[-4.08, -2.80]$ actions and failure by +2.15 $[-4.42, 8.45]$ percentage points (Figure 3). The intervention bundles state availability and instructions; it does not identify a pure variable effect or quality equivalence. Adding the eight predecessor actions leaves +6.11 $[5.30, 6.91]$ excess work with a fresh interpreter and +2.67 $[1.74, 3.59]$ with preservation, relative to contemporaneous original-state runs. Excluding all restarted scenarios gives -3.52 $[-4.21, -2.84]$ actions ($n = 80$). The roughly 73% penalty difference from sequentially collected branches is a time-confounded descriptive comparison (Appendix E); the interleaved conditional study does not validate that fraction or a whole-population curve.

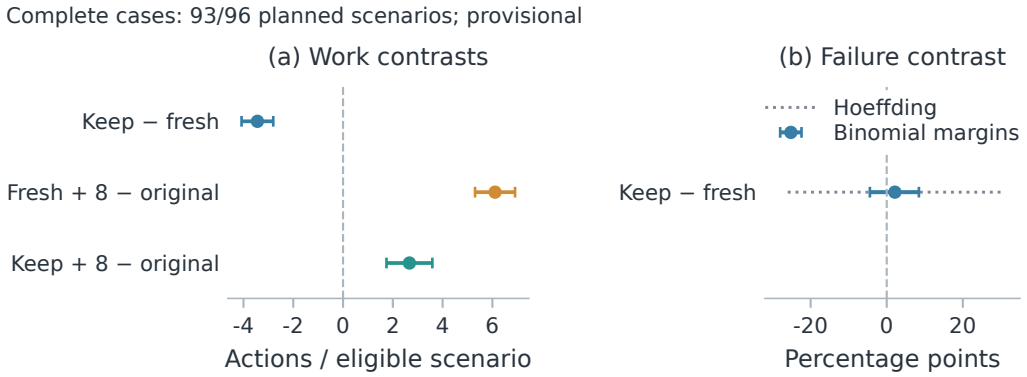

Figure 3: **Interleaved handoff-interface intervention.** Conditional on the reconstructed Gemma depth-8 survivor cohort; 93/96 planned scenarios have complete triplets, so inference remains provisional. (a) Paired work differences: blue, preserved minus fresh; amber/teal, eight base actions plus fresh/preserved work minus original-state work. Bars: pointwise 95% paired t intervals. (b) Failure: blue, conservative 95% binomial-margin interval; dotted gray, independent-bounded-outcome Hoeffding interval. Neither establishes quality equivalence. Full values and missingness bounds are in Appendix H.

5.3 Selection hides part of the inherited-state effect

The full post hoc audit reports every nonzero depth of the six primary pairings: 32 cells, with simultaneous 95% bands across all 96 contrast coordinates (Figure 8; Appendix F). Selection bias is distinguishable from zero in 6 cells. In 25 cells both the paired and survivor-to-population contrasts support less successor work; there is no supported sign reversal. Pairing usually changes the amount of measured reduction rather than its direction.

One cell changes the directional inference and has detectable selection. At AppWorld gpt-oss depth 16, the paired contrast is -6.45 $[-11.01, -1.89]$ actions per takeover, while the survivor comparison is -2.25 $[-5.72, +1.22]$. The selection term is $+4.20$ [1.23, 7.18]: inherited requests would already have required more original-state work. Pairing resolves a reduction that the survivor comparison leaves inconclusive. For Gemma at the same depth, selection changes its magnitude by $+3.68$ [0.50, 6.87], while both comparisons remain below zero. With 53/203 and 33/203 requests taken over, these gaps are 0.96 [0.09, 1.83] and 0.68 [0.06, 1.31] actions per incoming request, respectively. These joint intervals resample the takeover rate and workload contrast (Table 25). This evidence changes attribution, not the measurable policy total or its optimal deployment choice. Quality preservation still requires a separate test. It identifies which quantity an interface or efficiency intervention should target, without establishing that intervention's deployment value.

5.4 Can the closure–work association transfer?

Complete evaluation and matched takeovers identify closed requests' total original-state work, but not its correct/incorrect allocation. We test whether a frozen source association improves prediction of that allocation.

Transfer condition. Let $t = \mathbb{E}[(I+Y)W_H]$, $q_c = \mathbb{E}[I]/\mathbb{E}[I+Y]$, and $F = \mathbb{E}[IW_H]$. The aggregate comparator is $F_0 = tq_c$; the source prediction is $F_\kappa = t\kappa q_c/[1+(\kappa-1)q_c]$, where κ is the source incorrect/correct conditional-work ratio. Both use identical target t, q_c; only F_κ adds the source association. All moments share one allocation-valid distribution. With $t > 0$, $0 < q_c < 1$, $\kappa > 1$, and positive target conditional means, write $\rho = \mathbb{E}[W_H \mid I]/\mathbb{E}[W_H \mid Y]$. Then

$$|F - F_\kappa| < |F - F_0| \quad \Longleftrightarrow \quad \rho > 1 + \frac{\kappa - 1}{2 + (\kappa - 1)q_c}. \tag{5}$$

A positive association alone is insufficient: the correction can overshoot. This boundary concerns the aggregate comparator, not separately pooled within-stratum predictions.

Interleaved all-request test. Before collection, we fixed Gemma/depth 24 as primary and $\kappa = 1673/1060$ from the 40-task prospective source cohort. We planned one held-out variant from each of 126 known AppWorld families. Within each task, Gemma and gpt-oss base runs, a shared original-state Luna run on every request (including later closures), and fresh-interpreter Luna continuations at open checkpoints 8/16/24 or native transfer are interleaved. Evaluated autonomous closures require no continuation.

The sole primary test requires the lower endpoint of the two-sided 95% interval for $M = F - (F_0 + F_\kappa)/2$ to exceed zero. $G = |F - F_0| - |F - F_\kappa|$ measures allocation-error reduction, not deployment savings or repair cost. Appendix I, "Inference and missing records," specifies fixed stratum weights, whole-task resampling, population assumptions, observation masks, and full-cohort completion bounds: missing-tuple inference conditions on observed records. The same appendix retains all six cells, simultaneous secondary bands, work/quality contrasts, state/provenance checks, and structurally uninformative boundaries.

Results. All 126 tasks supply allocation tuples; one failed gpt-oss depth-8 continuation leaves 125 work/quality pairs in that cell. Proportional allocation understates incorrect-closure work in five of six cells under joint 95% bands. At Gemma/24, the measured value is 5.42 actions per incoming request versus 4.64, a gap of 0.78 [0.13, 1.42]. The source correction predicts 6.16: its primary margin is 0.02 [-0.42, 0.48], so the declared test remains unresolved. Its error reduction is 0.04 actions, or 0.22% of original work. Secondary gpt-oss/16 and /24 margins are 0.35 [0.06, 0.64] and 0.39 [0.05, 0.73] under the joint family, with error reductions of 0.31 and 0.41 (Figure 4).

Task-stratum and original-outcome sensitivity. A post hoc comparison gives both predictors the same normal/challenge counts and work totals, retaining the frozen source coefficients (Appendix J). The pooled error reductions are 0.30 [0.03, 0.47] and 0.41 [0.06, 0.61] for gpt-oss/16 and /24 under simultaneous bands; Gemma/24 descriptively overcorrects and remains unresolved. For gpt-oss/24, the residual allocation gap is 0.5867 actions per request: 0.4785 within original-outcome groups and 0.1083 from their composition. These coarse categories therefore do not exhaust the finite-table discrepancy. The decomposition identifies neither a causal mechanism nor workload savings; conditioning on realized outcomes changes the estimand. The all-request work/quality table also reports gpt-oss's failure increase of 50.00 [35.09, 64.91] percentage points at depth 24 (Table 29).

Prediction and policy scope. Accounting diagnoses observed work; prediction and control require additional assumptions. On the 40-task prospective AppWorld evaluation, refreshing the original-work baseline with evaluation outcomes reduces five-constant curve error from 0.237 to 0.078. Given the same information, exchange and no-progress models reach 0.060 and 0.069; this diagnostic does not rescue the frozen forecast (Appendix G.3). The joint-profile controller DOVE does not reliably beat per-agent fixed depths. A direct mean regressor trained on the same labels improves on the carried-over depth on held-out tasks (-2.62 [-4.52, -0.68]), so these tests do not show that state information is useless. Full prediction, acquisition-cost and policy comparisons remain in Appendices C, D and E.

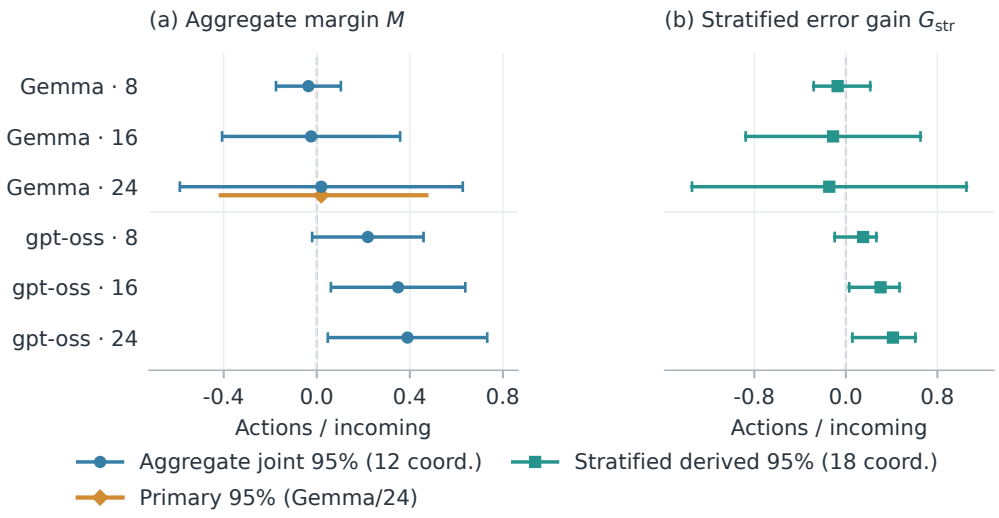

Figure 4: (a) Blue points and intervals show aggregate M with joint 95% bands over 12 coordinates; amber marks the prespecified Gemma/24 pointwise interval. (b) Teal shows exploratory pooled error reduction G_{str}, with intervals mapped from the joint $M_{\text{str}}, d_{\text{str}}$ bands in an 18-coordinate family. Both stratified predictors use the same normal/challenge information; the source coefficient remains fixed. Zero on the right means no prediction-error reduction. All six cells have 126 allocation tuples. Units are actions per incoming request; these are prediction comparisons, not policy work savings.

6 Limitations and Conclusion

Limitations. Successor actions measure bounded model execution, not human labour or the eventual repair of wrong closures. The sequentially collected records lack immutable provider versions and complete timing metadata. The interleaved interface study controls collection order and reconstructed state but covers one reconstructed Gemma survivor cohort and an interpreter–instruction bundle, with three validation failures and amended infrastructure restarts; its complete-case work contrast does not establish quality equivalence. The additional closure association has modest measured magnitude and no demonstrated policy value. The all-request 126-family study uses held-out variants within known families; its primary transfer test is unresolved. Two secondary gains survive exploratory stratum matching, which does not control latent difficulty or establish a causal mechanism. Stable reported identifiers do not certify immutable weights. The five-constant restriction partly fails; target-outcome refreshes explain prediction error without rescuing prospective forecasts. Small calibration sets and limited representations constrain controller conclusions. Second-family evidence covers τ-bench endpoints, and a training rule excluded one domain.

Conclusion. All-request paired execution separates inherited-state work changes, correct-closure savings and work hidden by wrong closure. The audits quantify what these measurements add to complete aggregate evaluation and matched takeovers. Contemporaneous interface branches locate a work difference that survives exclusion of restarted tasks. The interleaved all-request table identifies closure–work associations and tests when their source-based correction transfers. The findings separate measured attribution from its predictive and deployment uses, with explicit quality and configuration limits.

Reproducibility statement

The uploaded supllementary zip file with accompanying code package provides CPU replay entrypoints and figure inputs; response-level replay also requires the separately retained frozen scientific bundles. The entrypoints reconstruct stored-output scores and statistics; they do not recreate provider inference.

AI use statement

Generative AI assisted literature search and synthesis, implementation of the stimulus generators and analysis code, and manuscript proofreading.

References

Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, November 1983. ISSN 0005-1098. doi: 10.1016/0005-1098(83)90046-8.

Sebastian Baltes, Marc Cheong, and Christoph Treude. "an endless stream of AI slop": How developers discuss the burden of AI-assisted software development. arXiv preprint arXiv:2603.27249, 2026.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. arXiv preprint arXiv:2506.07982, 2025.

Joel Becker, Nate Rush, Tom Cunningham, David Rein, and Khalid Mahamud. We are changing our developer productivity experiment design. METR, February 2026. Published February 24, 2026.

Dimitri P. Bertsekas, John N. Tsitsiklis, and Cynara Wu. Rollout algorithms for combinatorial optimization. *Journal of Heuristics*, 3(3):245–262, December 1997. ISSN 1572-9397. doi: 10.1023/a:1009635226865.

BetterUp Labs and Stanford Social Media Lab. Workslop: The hidden cost of AI-generated busywork. BetterUp Labs, 2025. Report on a September 2025 survey; accessed 2026-09-14.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Runzhi Deng, Yiming Zhong, Fang Zhao, and Pan Zhou. Do not restart: Residual completion for stateful agent handoffs. *arXiv preprint arXiv:2609.13800*, 2026.

Roy Ganz, Mor Shpigel Nacson, Adi Kalyanpur, and Ron Litman. The handoff tax: Continuing non-native trajectories in llm agents. *arXiv preprint arXiv:2608.24358*, 2026.

Dipesh KC and Anjila Budathoki. Handoff debt: The rediscovery cost when coding agents take over interrupted tasks. arXiv preprint arXiv:2606.02875, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations (ICLR)*, 2026. Outstanding Paper Award.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11842.

Ji'an Lei and Jian Huang. Tacit-switch: Cost-aware model escalation for llm agents from censored supervision. *arXiv preprint arXiv:2608.27911*, 2026.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Dzianis Piatrashyn, Nikita Kotelevskii, Kirill Grishchenkov, Nikita Glazkov, Ivan Nasonov, Ilya Makarov, Timothy Baldwin, Preslav Nakov, Roman Vashurin, and Maxim Panov. ReDAct: Uncertainty-aware deferral for LLM agents. arXiv preprint arXiv:2604.07036, 2026.

Joshua Strong, Pramit Saha, Yasin Ibrahim, Cheng Ouyang, and J. Alison Noble. Identity-free deferral for unseen experts. In *The Fourteenth International Conference on Learning Representations*, 2026.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16022–16076, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.850.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, February 2024. ISSN 1467-9868. doi: 10.1093/jrsssb/qkad009.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

A Formal Statements and Proofs

Setting. A request X is drawn from a population P. A budget-agnostic base agent acts on it; its public state after j completed actions is s_j, where an action is complete once its pending environment and user responses have been delivered. Let L be the number of actions after which the base run ends, either by closing the request (correctly or incorrectly, as judged by the independent evaluator), by a native transfer, or at the cap K. For a depth k on the grid $\mathcal{G}$, option C_k lets the base agent act for at most k actions and hands the request to the successor configuration e if it is still open. Write D_k, Y_k, $I_k \in \{0,1\}$ for takeover, correct autonomous closure and incorrect autonomous closure within k actions; exactly one of them equals one. For a successor-side quantity G (work W, failure Z, tokens), G_k is its value when the successor takes over at $\min(k, L)$ and $G_H = G_0$ its value under immediate handoff, which is defined for *every* request. Let $g_0 = \mathbb{E}[G_H]$, $p_k = \mathbb{E}[D_k]$, $A(k) = \mathbb{E}[Y_k]$, $c(k) = \mathbb{E}[I_k]$ and $S_G(k) = \mathbb{E}[D_k G_k]$.

Proposition 1 (Recovery accounting and the survivor contrast)**.** *For every depth k,*

$$S_G(k) - g_0 = \underbrace{\mathbb{E}[D_k(G_k - G_H)]}_{\textit{inflation}} - \underbrace{\mathbb{E}[Y_k G_H]}_{\textit{skim}} - \underbrace{\mathbb{E}[I_k G_H]}_{\textit{false savings}}, \qquad \mathrm{Fail}(k) = c(k) + \mathbb{E}[D_k Z_k]. \quad (6)$$

If $p_k > 0$, let $\iota_k = \mathbb{E}[G_k - G_H \mid D_k = 1]$ and $\bar{g}_k^{\mathrm{cl}} = \mathbb{E}[G_H \mid D_k = 0]$ (any value if $p_k = 1$). Then the unpaired survivor contrast satisfies

$$\mathbb{E}[G_k \mid D_k = 1] - g_0 = \iota_k + \frac{1 - p_k}{p_k}\left(g_0 - \bar{g}_k^{\mathrm{cl}}\right). \quad (7)$$

Proof. Since $D_k + Y_k + I_k = 1$, $G_H = D_k G_H + Y_k G_H + I_k G_H$. Subtracting the expectation of this identity from $S_G(k) = \mathbb{E}[D_k G_k]$ gives the first display. A request fails under C_k either because the base agent closed it incorrectly or because it was taken over and the successor failed, and these events are disjoint, which gives the second. For (7), $\mathbb{E}[G_k \mid D_k = 1] = \iota_k + \mathbb{E}[G_H \mid D_k = 1]$ and $g_0 = p_k \mathbb{E}[G_H \mid D_k = 1] + (1 - p_k)\bar{g}_k^{\mathrm{cl}}$, so $\mathbb{E}[G_H \mid D_k = 1] - g_0 = (1 - p_k)(g_0 - \bar{g}_k^{\mathrm{cl}})/p_k$. □

The second term of (7) is a pure selection effect: it is positive whenever the requests the base agent closes are cheaper for the successor than the average request, whatever continuation does to the inherited state. Comparing handed-off requests after long runs with all requests therefore overstates the inherited-state work change under skimming and understates it under reverse skimming. Interpreting the paired term ι_k as an effect of continuation additionally assumes faithful state reconstruction and a stable successor configuration.

Proposition 2 (Break-even price of autonomy)**.** *Let $J(k) = \alpha B(k) + \nu S_W(k)$ with $\alpha, \nu > 0$ and $B(k) = \mathbb{E}[\min(k, L)] > 0$. Then $J(k) < J(0)$ if and only if $\alpha/\nu < r^*(k) := \{S_W(0) - S_W(k)\}/B(k)$. In particular, if $S_W(k) \geq S_W(0)$, no positive price ratio makes depth k cheaper than immediate handoff.*

Proof. $B(0) = 0$, so $J(k) - J(0) = \alpha B(k) - \nu\{S_W(0) - S_W(k)\}$; divide by $\nu B(k) > 0$. □

The comparison is an economic preference only among policies that satisfy the quality requirement; r^* is reported next to $\mathrm{Fail}(k)$ for that reason.

Proposition 3 (Evaluation of stop rules by table lookup)**.** *Let the nested branch table contain, for every request and every checkpoint $j \in \mathcal{G}$ reached by the base run as well as the event at which the base run ends without closing, one successor outcome R_j drawn from the conditional law $\kappa_e(s_j, \cdot)$ of the successor configuration given the inherited state, independently of the base agent's later actions. Let π be a stop rule that (i) decides at checkpoint j from $s_0, \ldots, s_j$ only and (ii) is fixed independently of the table. Let τ_π be the checkpoint at which π stops, or ∞ if the base run closes first. Then the outcome read from the table, R_{τ_π} if $\tau_\pi < \infty$ and the evaluated closure otherwise, has the same distribution as the outcome of deploying π.*

Proof. Because the base agent is budget-agnostic and π only ever stops it, the law of $(s_0, \ldots, s_{\tau_\pi})$ is the same in the table run and in deployment, and the event $\{\tau_\pi = j\}$ is determined by $(s_0, \ldots, s_j)$. In deployment the successor outcome given $\{\tau_\pi = j\}$ and the prefix is a draw from $\kappa_e(s_j, \cdot)$; so is R_j, which by assumption does not depend on what the base agent did after j. Integrating over the prefix gives equality in distribution; the closure branch is identical by construction. □

The statement covers neither rules that look at branch outcomes (the hindsight oracle, which we label an in-sample diagnostic) nor checkpoints outside $\mathcal{G}$. One row per checkpoint gives one draw of the deployed outcome, not its expectation: sampling uncertainty is that of a single deployment per request.

Proposition 4 (Rollout improvement over committed options)**.** *At an open checkpoint state s_j let $\mathcal{O}_j = \{H\} \cup \{C_{j\to j'} : j' \in \mathcal{G},\ j' > j\}$, where $C_{j\to j'}$ lets the base agent continue to j' and then hands over, and let $L_o(s_j)$ be the conditional expected cost-to-go of committing to o under a fixed cost convention (base price α, successor price ν, failure price λ). Let $\hat{L}$ satisfy $|\hat{L}_o(s) - L_o(s)| \le \eta$ for all states and options, and let π stop at s_j iff $\hat{L}_H(s_j) \le \min_{o\neq H} \hat{L}_o(s_j) + \tau$ and otherwise advance one checkpoint. If N_j checkpoints remain from j on, the expected cost-to-go of π satisfies*

$$V^\pi(s_j) \le \min_{o\in\mathcal{O}_j} L_o(s_j) + N_j\,(2\eta + |\tau|).$$

Proof. Induct backwards over checkpoints. At the last checkpoint only H is available and the claim holds with equality. Suppose it holds at $j{+}1$. If π stops at s_j, then $\hat{L}_H \le \min_{o\neq H} \hat{L}_o + \tau$ implies $L_H \le \min_o L_o + 2\eta + |\tau|$. If π advances, let $o^* = C_{j\to j^*}$ minimise $\hat{L}$ over $\mathcal{O}_j$ if the minimiser is not H; if it is H, then $\hat{L}_H > \min_{o\neq H} \hat{L}_o + \tau$ gives an option $o^* \neq H$ with $L_{o^*} \le L_H + 2\eta + |\tau|$. By the tower property the cost of committing to o^* decomposes through the next checkpoint: the base cost of the interval, the closure cost if the run closes in it, and otherwise $L_{C_{j+1\to j^*}}(s_{j+1})$, with the convention $C_{j+1\to j+1} = H$. Advancing incurs the same interval and closure costs and then $V^\pi(s_{j+1}) \le L_{C_{j+1\to j^*}}(s_{j+1}) + N_{j+1}(2\eta + |\tau|)$ by the induction hypothesis, because $C_{j+1\to j^*} \in \mathcal{O}_{j+1}$. Hence $V^\pi(s_j) \le L_{o^*}(s_j) + N_{j+1}(2\eta + |\tau|) \le \min_o L_o(s_j) + N_j(2\eta + |\tau|)$. □

The bound concerns the scalarised liability only, requires the option family to be closed under continuation (the event rows and the boundary rule provide this), and is conditional on η, which the training loss does not certify. With $\eta = \tau = 0$ the controller is never worse than the best committed option.

State-conditional accounting. Conditioning Proposition 1 on a state explains the controller's statistic. Over one option horizon let $1-h$, h^c, h^i be the conditional probabilities of takeover, correct closure and incorrect closure, a the expected base actions, and $\ell = \nu W + \lambda Z$. Then

$$\begin{aligned} L_C(s) - L_H(s) = \alpha a + (1-h)\,\mathbb{E}[\ell' - \ell_H \mid \text{takeover}, s] \\ - h^c\,\mathbb{E}[\ell_H \mid \text{correct closure}, s] - h^i\big(\mathbb{E}[\ell_H \mid \text{incorrect closure}, s] - \lambda\big), \end{aligned}$$

that is, base cost plus inflation minus skim minus false savings. The profile model estimates the two sides of the difference, not the four terms: marginal option profiles do not identify $\mathbb{E}[\ell_H \mid \text{future closure}, s]$, which depends on the joint law of the immediate-handoff outcome and the base agent's later behaviour. The terms are estimated only from paired records.

Non-identification from base-only logs. Fix any distribution of base histories, tool returns and closure labels. Two successor procedures that receive exactly these histories, one finishing correctly after one action and one after four, induce the same likelihood for the base-only data and different successor work. Identifying recovery work therefore requires executing (or validating against the execution of) a specified successor.

B Protocol Details

Environments and agents. τ-bench: package version 1.0.1 of the public repository (commit `b7ea9074`, recorded with the AppWorld version in the artifact's `VERSIONS.md`), domains `airline` (50 requests) and `retail` (114 requests); the base agent is the benchmark's standard tool-calling agent with the unmodified domain policy; the user is the benchmark's LLM user simulator. AppWorld: version 0.1.3.post1, the official ReAct code-agent prompt, one Python code block per action. All models are black-box APIs. In τ-bench every rollout call (agents, user simulator, and the judge that scores natural-language assertions) requests temperature 0, and the run seed 0 is passed to agents and user simulator; the re-judging by Terra in Appendix E used that endpoint's default temperature and no seed. In AppWorld the open-weight models are called at temperature 0 and the

GPT-5.6 models at their provider default, because we did not send a temperature to those endpoints. No model is fine-tuned.

Handoff view. The successor receives the same system prompt, tools and policy as the base agent, the inherited interaction (messages, tool calls and tool results) as history, and one fixed takeover note stating that a previous agent produced the history, that it may contain mistakes, and that the current state should be checked where needed. In AppWorld the note also says that the interpreter is fresh while changes to the apps persist. No model-written summary is involved, so the view is a deterministic function of the inherited state.

Native transfers. The transfer-call view showed the transfer call and its "transfer successful" result to the successor; in 97% of those 226 rows (Luna; 96% for gpt-oss-120b) it then followed the domain policy's post-transfer script, told the user to hold, and the episode ended after one action with no takeover (success 0.42 for both successors, against 0.77 and 0.71 from the last checkpoint before the transfer). Those rows were excluded and retained for audit. All were re-collected under the pre-transfer view before computing the reported results.

Grids and caps. τ-bench: $\mathcal{G} = \{0, 2, 4, 6, 8, 12, 16\}$ base actions, $K = 16$, $B_{\max} = 30$. AppWorld: $\mathcal{G} = \{0, 4, 8, 16, 24\}$, $K = 24$, $B_{\max} = 40$. In AppWorld a base agent that declares failure is treated as a native transfer: the declaring action is charged to the base agent and the successor inherits the state before it. AppWorld scores the final database state whether or not the agent declared completion; we apply the deployment convention that a request not declared complete within the cap is unresolved, and report the evaluator's state verdict separately.

Boundary-rule tests. Scripted, non-LLM agents and users on the benchmark's mock domain verify that (i) a cap equal to the realised work does not censor the run; (ii) a label taken from a long run with $T = t$ agrees with a live run capped at $b = t$, while $b = t - 1$ is censored after $t - 1$ actions; (iii) pending environment responses are delivered at a cap; (iv) the action that triggers a native transfer is charged to the base agent and is absent from the prefix the successor inherits; (v) resuming from a checkpoint prefix counts only post-prefix successor actions. A sixth test runs two separate AppWorld processes and checks that a predecessor's write to an app persists for the successor while the predecessor's variables do not.

Scoring model substitution. Natural-language assertions in `retail` were scored with GPT-5.6 Luna in place of the benchmark's default judge. Agreement with a stronger judge is audited in Appendix E.

Splits and inference population. τ-bench `airline` and `retail` requests are individually authored, so the request is the independent unit; a frozen, hashed assignment gives 66 training, 33 calibration and 65 test requests, with every request used in pilots placed in training. `banking_knowledge` was excluded after a capability probe on training requests only (11 requests): the six models that completed every probe row solved between 0 and 2 of the 11, Luna and the base agent both solved 2, so no successor met the pairing rule; four other models did not complete the probe (rate limits and rejected calls). A depth-0 row there takes seven to eight times the input tokens of an `airline` or `retail` row (median 317k against 42k for Luna). `telecom` requests are compositions of shared atomic faults; the request-level resampling nevertheless treats them as independent, so those intervals are likely too narrow. AppWorld scenarios have three near-duplicate variants; the test set is one variant of each of the 56 `test_normal` scenarios. Development uses the 147 tasks of the 49 train and dev scenarios: the work-curve models fit on all of them, the controllers on 111 training tasks with 12 calibration tasks (one variant of each of a seeded quarter of the scenarios) for selection, and the matched fixed depths use the same 12. Inference treats test requests as independent draws from the distribution that generated them and makes no claim beyond it.

Pairing rule. A successor is paired with a base agent only if, on a fixed probe of training requests, its solo success exceeds the base agent's by at least ten points. One deliberate negative control uses a successor no stronger than the base agent.

Unresolved rows. Rows that failed for infrastructure reasons (rate limits) were retried from the same inherited state. In τ-bench no row remained unresolved. In AppWorld, 29 rows of the Gemma table failed in the first pass and were retried from kept snapshots; in the gpt-oss-120b table 25 rows failed, and the snapshots of 15 of its requests had already been deleted, so those requests were re-executed from the initial state (base trajectories and successor branches); all attempts and collection logs remain in the artifact. The failures were HTTP 429 responses. We have not analysed whether exposure to rate limits is related to task or run characteristics.

Spend control. Cost calculations use recorded tariffs and returned usage, not invoice verification. The artifact retains the ledgers, including exploratory studies outside this paper and uncertain charges.

Prespecification and amendments. The artifact preserves the analysis plan, all nine dated amendments, and the separate prospective plans. Before calibration/test outcomes, but after recorded training pilots and capability probes, we fixed splits, prices, pairing, model thresholds, controllers, baselines and the quality rule. Material later changes were: (1) controller features and clustered inference, before variant test evaluation but after a debugging printout of observed outcomes; (2) per-agent depth, after the first τ-bench test table and before AppWorld test outcomes; (3–5) implementation/statistical corrections and post hoc exchange, held-out-request, oracle, interface, matched-selection and failure-frontier analyses; (6) correcting missing task/output text in the controller representation and testing controllers on held-out tasks; (7) matching the same-model control's remaining action budget; (8–9) second-family replication and its infrastructure changes. The paired work-model estimator replaced curve regression after an interim look. Dates without supporting logs/hashes are author-attested. The second-family plan was hashed after one retained smoke probe and before remaining probe/evaluation rows. All deviations and excluded attempts remain available.

We regard as confirmatory only the endpoint contrasts, prespecified work-model tests (subject to the interim estimator caveat), prespecified controller rows and declared prospective hypotheses. Analyses labeled post hoc remain exploratory. The closure/transfer audits in Appendix G fixed their analysis choices after the sequential-cohort results were known. The interleaved interface experiment has a separately frozen cohort and recorded amendments (Appendix H).

Resampling. Bootstrap intervals use 4,000 resamples of requests (scenarios in AppWorld) for paired contrasts and policy differences, 2,000 for the oracle analyses, the common-failure frontier and the prospective AppWorld depth selection, 1,000 joint resamples of fit and test requests for the held-out-request test, and 500 joint resamples for the prespecified work-model tests, where constants are refitted in every draw.

Exact controller configuration. The controller has 18 trace counters in τ-bench and 14 in AppWorld, listed in `analysis/states.py`. Every option additionally receives its horizon, the successor's training-probe success rate and mean work, and an airline-domain indicator. AppWorld text contains the full task instruction followed by the last three visible action–output pairs (the last 300 characters of each assistant response and first 300 of each environment output). At a native-transfer event, the give-up action is excluded. The frozen embedding endpoint is requested as `nomic-ai-nomic-embed-text-v1`, with at most 6,000 input characters; the cached vectors have 768 dimensions. PCA uses 16 components, no whitening, automatic solver selection, and seed 0, fitted only on training states. No held-out evaluation state enters this PCA fit. The embedding cache contains 1,654 states, including all 1,182 source-cohort states.

All profile heads use scikit-learn 1.8.0 histogram gradient boosting: 200 iterations, learning rate 0.06, at most 15 leaves, at least 20 samples per leaf, and ℓ_2 regularisation 1.0. The event-hazard and conditional-failure heads are classifiers; the base-work head is a regressor. The direct mean baseline uses regressors with the same settings and option labels. Seed 0 gives the reported operating point; seeds 1–4 are the reported sensitivity check. Successor caps are 30 and 40 actions. Thresholds are selected from $\{-4, -2, -1, -0.5, 0, 0.5, 1, 2, 4, 8\}$. AppWorld training has 111 requests (639 open states, 5,024 option labels across both training successors); calibration has 12 requests; the source-cohort test has 56 requests, and the held-out evaluation has 40. The calibration and test sets each contain one variant per scenario. The post hoc agent descriptor contains four training-only summaries: correct-closure rate, wrong-closure rate, native-transfer rate, and mean base run length.

Model identifiers and logging coverage. Table 2 distinguishes requested API identifiers from identifiers actually retained in provider-response logs. The main run records retain the registry name and harness fingerprint. Raw provider payloads were not retained; AppWorld trajectories also lack absolute execution timestamps. Served-model logging began on September 22, after the source-cohort branch collection. Its snapshot identifiers cannot identify models used for trajectories without response logs. These identifiers remain unknown, including the weights behind the generic `llama-3` endpoint alias. Local plan dates, file modification times and analysis stdout corroborate the local sequence but are not independent preregistration. The interleaved interface collection has complete timestamp/response records (Appendix H) that identify its own calls only.

Table 2: API model identity retained in the artifact. Registry labels identify the requested experimental role, not a verified immutable checkpoint. Times are UTC in 2026. Response identifiers apply only to the logged calls.

Registry label	Requested identifier	Recorded served identifier	Log coverage
Gemma-4-31B	`google-gemma-4-31b-it`	`google/gemma-4-31b-it`	Sept. 22 14:18 to Sept. 24 19:49
gpt-oss-120b	`openai-gpt-oss-120b`	`openai/gpt-oss-120b`	Sept. 23 20:46 to Sept. 25 10:00
Llama-3.3-70B	`llama-3`	unknown	no retained served-ID log
Nemotron-3-Super	`nvidia-nvidia-nemotron-3-super-120b-a12b-fp8`	`nvidia/nemotron-3-super`	Sept. 23 20:43 to Sept. 25 08:51
GPT-5.6 Luna	`gpt-5.6-luna` (logged request)	`gpt-5.6-luna-2026-07-09`	Sept. 22 14:18 to Sept. 25 10:00
GPT-5.6 Terra	deployment read from configuration; collection-time request ID unknown	unknown	no retained served-ID log
Claude Sonnet	`claude-sonnet-4-6`	`claude-sonnet-4-6`	Sept. 22 14:21–14:45
Claude Opus	`Claude-opus-5`	`claude-opus-5`	Sept. 22 14:52–16:17

C Work Models and Decision Probes

A work-curve model to test. We test a restriction of Equation 2: requests a base agent closes cost the successor a fixed fraction of the average, $\mathbb{E}[W_H \mid Y_k] = \rho^c w_0$ and $\mathbb{E}[W_H \mid I_k] = \rho^i w_0$ for all k, and inflation among inherited requests is linear in what the base agent did to them,

$$\iota_k = \beta\, m_k - \gamma\, n_k, \qquad \text{so} \qquad S(k) = w_0\big[1 - \rho^c A(k) - \rho^i c(k)\big] + p_k\big(\beta m_k - \gamma n_k\big), \tag{8}$$

where m_k and n_k are the mean numbers of base actions with and without a state-changing call among requests taken over at depth k, $A(k) = \mathbb{E}[Y_k]$ and $c(k) = \mathbb{E}[I_k]$. Everything on the right except the five constants $(w_0, \rho^c, \rho^i, \beta, \gamma)$ comes from base runs, so constants fitted on some agents predict the work curve of an agent whose successor branches were never executed. The coefficients are predictive associations of action counts that the agent chose, not effects of an action. We fixed acceptance thresholds for this five-constant model in advance (Section 5) and compare it with two nulls, a flat curve and the law without its second term. A post hoc exchange comparator sets $\rho^c = \rho^i = 1$ and $\beta = -\gamma = -g$:

$$S(k) = w_0\, p_k - g\,\mathbb{E}[D_k B_k], \qquad J(k) - J(0) = (\alpha - \nu g)\,\mathbb{E}[D_k B_k] + \mathbb{E}\big[(1 - D_k)(\alpha B_k - \nu w_0)\big], \tag{9}$$

with B_k the base actions spent on the request: every inherited base action saves g successor actions, and a closed request saves the average w_0. In this *exchange form* whole-system work ($\alpha = \nu$) rises with autonomy when the loss $(1 - g)$ on inherited actions outweighs the margin of the requests the agent closes for fewer than w_0 actions.

C.1 Decision probes

Choosing an autonomy budget from the work-curve model. For a held-out base agent, a labelled sample of its own runs gives $A(k)$, $c(k)$, p_k, m_k, n_k and $B(k)$; with the constants of Equation 8 or 9 fitted on *other* agents, whose branch tables exist, no successor has to be executed for the held-out agent:

$$\hat{J}_q(k) = \alpha B(k) + \nu\, \hat{S}(k) + \lambda\big[c(k) + p_k\, \bar{z}\big], \tag{10}$$

with $\bar{z}$ the successor's pooled failure rate on inherited requests in those tables; the budget is $\arg\min_{k \in \mathcal{G}} \hat{J}_q(k)$, and it responds to a change of prices without new data.

State-conditional control (DOVE). DOVE (Downstream Outcome and Value Estimation) predicts option liabilities from terminal-outcome and successor-work profiles, then decides by receding-horizon rollout (Appendix D). Proposition 4 bounds its loss to the best committed option under uniform prediction error; Appendix A gives the state-conditional accounting and the bound's assumptions.

C.2 ESTIMATION AND PREDICTION TESTS

The paired estimator computes w_0 and closure ratios from original-state means and fits progress slopes by request-level regression. It replaced a collinear curve regression after an interim look; both estimators are reported in Appendix E. Prespecified tests hold out a base agent or depths $\{4, 12\}$ in τ-bench ($\{8\}$ in AppWorld), but share requests with fitting. A held-out depth can also reuse an earlier terminal takeover (60% at $k = 12$). The post hoc held-out-request test fits on other agents' train/calibration requests and scores the held-out agent's test requests (Table 9). The model reaches its τ-bench point-error threshold only with the paired estimator; only 27% of joint resamples are below 0.10. Its closure-ratio restriction varies with depth.

A separate successor-transfer check uses 19 Terra probe rows for w_0 and normalized constants from the two training successors, then scores 65 test requests: nMAE 0.061 for five constants, 0.056 for exchange, versus 0.253 and 0.565 for the nulls. This narrow check does not override the failed prospective test.

C.3 DETAILED DEPTH-SELECTION RESULTS

For each held-out agent, constants come from other agents and base-side curves from its train/calibration requests. Predicted work is clipped at zero; depths minimize Equation 10 and are scored on its test requests over five price ratios (40 decisions in τ-bench, 20 in AppWorld). Table 9 shows that the no-progress and endpoint rules preserve the decision performance. Acquisition comparisons in Table 10 charge the source branch tables as well as target-agent observations.

A prospective test. On the frozen AppWorld `test_challenge` test (Table 15), model-selected depth does not beat the status quo; the carried-over depth and endpoint rule have lower point regret. Original-state work rises from 14.3 to 19.8 for Gemma, but the records cannot separate task composition from endpoint/configuration changes. Appendix G.3 diagnoses work-curve error without assigning the failed decision to one parameter. The secondary mid-depth peak and wrong-closure patterns recur, outside the declared hypothesis.

D DOVE: DETAILS

Options and controller. At an open checkpoint j the options are H (hand off now) and, for every later grid depth j', $C_{j\to j'}$ (continue to j', then hand off); the family is closed under continuation. For an option o the profile gives $\hat{L}_o(s) = \alpha\hat{a}_o + \nu\hat{\mathbb{E}}[\min(T, b)] + \lambda(1 - \hat{\Pr}(U{=}\text{correct}, T \le b))$ for any successor budget $b \le B_{\max}$ and any prices without refitting. The controller stops at j iff $\hat{L}_H(s_j) \le \min_{j'>j} \hat{L}_{C_{j\to j'}}(s_j) + \tau$ and otherwise advances one checkpoint and decides again.

Labels. H at checkpoint j is labelled by the row at j; $C_{j\to j'}$ by the evaluated closure if the base run closes before j', by the event row if it ends in a native transfer or at the cap before j', and by the row at j' otherwise. These cases partition the outcomes of a committed option.

Profile model. For an option with horizon $m = j' - j$ (zero for H) under successor e, $q(u, t \mid s, o, e) = \Pr(T = t \mid x)\Pr(U = u \mid T = t, x)$ with $x = [\phi(s), m, d_e]$, where $\Pr(T = t \mid x)$ comes from a discrete-time hazard on $t \in \{0, \dots, B_{\max}\}$ fitted on person-period data, with a tail class for runs censored at $B_{\max}$ (every row shares the same administrative cap, so the likelihood is exact) and autonomous closures at $T = 0$. A second output predicts the base actions the option consumes. $\phi(s)$: depth, tool calls, state-changing calls, tool errors, repeated calls, user turns, message lengths, tokens; d_e: the successor's success rate and mean work under immediate handoff on a fixed probe subset of training requests. Optional blocks: 16 principal components (fitted on training states) of a frozen

text embedding of a deterministic state digest, and six numbers from zero-shot LLM answers (the base agent's verbalized success probability; a stronger model's estimates of success and remaining work under handoff and under continuation). Nothing from the evaluator, the reference solution or the user simulator's private instructions is used. All outputs are gradient-boosted models with fixed hyper-parameters; training is a small regression or logistic fit.

Operating point and baselines. $\lambda = 2w_0^{\text{train}}$; the offset τ is the value on a fixed grid with the lowest calibration cost among those whose calibration failure does not exceed the status quo's. Every learned baseline with a free parameter is selected by the same rule. Reference fixed depths were tuned without the failure constraint; matched-selection comparisons are in Appendix E. The TACIT-style baseline accumulates a softplus risk over checkpoints and hands off when it crosses a threshold. Its parameters are fitted by an interval-censored likelihood in which the interval is the last depth at which handoff still succeeds and the next depth, read from executed branches rather than from a teacher. A logistic cure component on depth-0 features predicts whether any handoff succeeds. The rule hands off when the risk crosses θ unless the predicted rescue probability is below c, with (θ, c) chosen on calibration requests ($c = 0$ switches the component off). These labels are noisy: branch success is neither deterministic nor monotone in depth (among the 247 training sequences with more than one checkpoint and at least one successful handoff, 74 have a failed handoff followed by a successful one at a later depth), so the baseline is TACIT-*style* and its weak result says little about TACIT-Switch itself. Verbalized confidence hands off when the base agent's stated success probability is below a threshold; LLM-as-profiler computes the two liabilities from a stronger model's zero-shot estimates. The variant reported as DOVE is the prespecified feature set with the lowest calibration cost.

AppWorld state representation. The AppWorld and held-out panels of Table 14 include task instructions, instruction length and recent environment outputs. These fields were added after the first AppWorld test evaluation: training models were refitted, calibration thresholds reselected and embeddings recomputed. The 406 depth-zero states shared one text representation without these fields; with them, they contain 186 distinct task texts. For the selected embedding variant, cost changes from $J_q = 14.20$ without these fields to 14.39 with them, against 13.18 for per-agent depth. Its paired difference is inconclusive, $+1.21$ with a 90% interval $[-0.01, 2.51]$. The post-hoc agent-descriptor advantage on the AppWorld test set (-0.60 $[-1.01, -0.24]$) does not recur in the held-out evaluation ($+0.23$ $[-1.09, 1.41]$). CPU reproduction from cached embeddings exactly reproduces all policy outcomes, selected thresholds and statistics in both evaluations; this checks computation, not prespecification of the representation or test comparisons. AppWorld uses counters and embeddings; the six LLM-critic features were collected only for τ-bench.

E ADDITIONAL RESULTS

Matched quality and tokens. Table 5 asks whether the main finding survives a matched quality criterion and a different unit of cost, for every τ-bench cell (post hoc). The ratio of actions per *resolved* request, status quo over immediate handoff, exceeds one significantly in 3 of the four Luna pairings; the paired action contrast on requests resolved under both policies is positive in all four. The third column is the accounting identity applied to failure, $\mathbb{E}[D_K(Z_K - Z_H)]$: the successor's failures on the requests it inherits at the end of the run, against what it would have done under immediate handoff. Tokens are agent input plus output tokens (user-simulator tokens excluded); the break-even base price is the successor's list-price saving per request divided by the base agent's tokens, so autonomy lowers the token bill only for a base model cheaper than that.

Who ends a closed episode. A τ-bench base run counts as closed when its episode ends without a handover or the cap. We classified every failed closure of the main tables from its last user and agent messages (`analysis/closure_audit.py`): the user's final message carries the transfer or out-of-scope marker; it carries the stop marker and opens with consent ("Yes, please proceed") to an agent message that asked a question and made no tool call, so the action asked about never ran; it opens with a refusal after such a question (user declined); or the agent closed (it reported the request done or declined it, and the user ended). Table 6 gives the counts: 16 of Gemma's 34 failed closures were ended by the simulated user, mostly after Gemma asked for an explicit confirmation. Table 7 recomputes the status-quo-versus-immediate-handoff contrasts with these requests handed over at the last checkpoint before the stop that has an executed successor row (an executed outcome of the same

Table 3: **Complete endpoint outcomes: status quo minus immediate handoff.** Brackets: paired lower/upper one-sided 95% bootstrap bounds; bold action/cost contrasts have positive lower bounds. Actions include base and successor work; J_q adds failure price at $\alpha = \nu$. Row tint identifies the successor: blue Luna, teal Terra, gray gpt-oss, rose Opus; color does not encode significance. †: second-family primary cells, 145 non-probe requests (Appendix E). Component contrasts for the same rows appear in Table 4.

Base	Successor	n	Δ actions	Δ fail (pp)	ΔJ_q
τ-bench (airline + retail)					
Gemma	Luna	164	**+0.52 [+0.23, +0.83]**	$+0\,[-7, +7]$	$+0.52\,[-0.50, +1.56]$
gpt-oss	Luna	164	**+2.36 [+1.99, +2.73]**	$+7\,[+0, +13]$	**+3.46 [+2.40, +4.54]**
Llama endpoint	Luna	164	**+3.65 [+3.22, +4.09]**	$+10\,[+4, +17]$	**+5.41 [+4.35, +6.51]**
Nemotron	Luna	164	**+4.02 [+3.55, +4.49]**	$-2\,[-9, +4]$	**+3.61 [+2.52, +4.70]**
Gemma	Terra	65	$-0.77\,[-1.45, -0.14]$	$+2\,[-9, +12]$	$-0.47\,[-2.49, +1.44]$
Gemma	gpt-oss	164	$-1.15\,[-1.59, -0.70]$	$-11\,[-18, -4]$	$-3.29\,[-4.69, -1.98]$
gpt-oss	gpt-oss	164	$+0.20\,[-0.29, +0.65]$	$-4\,[-11, +2]$	$-0.70\,[-2.08, +0.63]$
Llama endpoint	gpt-oss	164	**+3.01 [+2.49, +3.50]**	$+7\,[+1, +13]$	**+4.50 [+3.13, +5.83]**
Nemotron	gpt-oss	164	**+2.16 [+1.71, +2.62]**	$-18\,[-26, -11]$	$-1.70\,[-3.21, -0.19]$
Gemma	Opus†	145	$-0.92\,[-1.30, -0.55]$	$-1\,[-9, +6]$	$-1.20\,[-2.67, +0.19]$
gpt-oss	Opus†	145	**+0.99 [+0.56, +1.45]**	$+8\,[+0, +15]$	**+2.56 [+1.00, +4.14]**
Nemotron	Opus†	145	**+2.99 [+2.41, +3.57]**	$-6\,[-14, +1]$	**+1.75 [+0.29, +3.18]**
AppWorld					
Gemma	Luna	203	$-0.64\,[-1.33, +0.07]$	$+2\,[-1, +5]$	$-0.06\,[-1.26, +1.18]$
gpt-oss	Luna	203	$-3.20\,[-3.97, -2.38]$	$+60\,[+54, +67]$	**+13.66 [+11.82, +15.48]**
Gemma	gpt-oss	203	**+3.15 [+2.34, +3.93]**	$-60\,[-67, -53]$	$-9.34\,[-10.89, -7.80]$
gpt-oss	gpt-oss	203	$+0.14\,[-0.71, +0.99]$	$+1\,[-4, +6]$	$+0.36\,[-1.01, +1.74]$

Table 4: **Autonomy and successor work for every endpoint pairing above.** ΔA: correct autonomous completion, in percentage points; ΔS: successor actions. Brackets retain the same paired one-sided 95% bounds; the † intervals are in Table 12. r^*: successor actions saved per base action (work rises iff $r^* < 1$). ρ^c: original-state work among correct closures relative to w_0; —: not computed for endpoint-only Opus cells. Counts, row order and successor colors match Table 3.

Base	Successor	ΔA (pp)	ΔS	r^*	ρ^c
τ-bench (airline + retail)					
Gemma	Luna	$+60\,[+54, +66]$	$-7.98\,[-8.32, -7.62]$	0.94	0.98
gpt-oss	Luna	$+46\,[+40, +52]$	$-7.33\,[-7.73, -6.93]$	0.76	0.97
Llama endpoint	Luna	$+1\,[+0, +3]$	$-1.53\,[-1.99, -1.07]$	0.30	0.75
Nemotron	Luna	$+55\,[+49, +62]$	$-7.67\,[-8.05, -7.28]$	0.66	0.95
Gemma	Terra	$+58\,[+49, +69]$	$-9.12\,[-9.97, -8.32]$	1.09	0.98
Gemma	gpt-oss	$+60\,[+54, +66]$	$-9.65\,[-10.21, -9.08]$	1.13	0.95
gpt-oss	gpt-oss	$+46\,[+40, +52]$	$-9.49\,[-10.10, -8.89]$	0.98	0.95
Llama endpoint	gpt-oss	$+1\,[+0, +3]$	$-2.17\,[-2.82, -1.56]$	0.42	0.63
Nemotron	gpt-oss	$+55\,[+49, +62]$	$-9.52\,[-10.02, -9.05]$	0.81	0.94
Gemma	Opus†	$+63$	-9.43	1.11	—
gpt-oss	Opus†	$+47$	-8.71	0.90	—
Nemotron	Opus†	$+56$	-8.70	0.74	—
AppWorld					
Gemma	Luna	$+87\,[+82, +92]$	$-14.38\,[-15.12, -13.67]$	1.05	0.95
gpt-oss	Luna	$+24\,[+18, +30]$	$-13.32\,[-13.96, -12.69]$	1.32	0.83
Gemma	gpt-oss	$+87\,[+82, +92]$	$-10.60\,[-11.54, -9.73]$	0.77	0.93
gpt-oss	gpt-oss	$+24\,[+18, +30]$	$-9.98\,[-10.83, -9.13]$	0.99	0.73

Table 5: The main contrast under matched quality and in tokens (τ-bench, status quo against immediate handoff; request-level 90% intervals).

(a) Resolved-request comparisons

Base agent	Actions per resolved request, ratio	Both resolved	
		Share (%)	Δ actions
Successor: GPT-5.6 Luna			
Gemma-4-31B	1.06 [0.98, 1.15]	62	+0.77 [+0.48, +1.07]
gpt-oss-120b	1.41 [1.28, 1.55]	54	+2.22 [+1.76, +2.70]
Llama endpoint	1.64 [1.51, 1.80]	60	+3.30 [+2.85, +3.76]
Nemotron-3-Super	1.42 [1.31, 1.54]	66	+4.03 [+3.48, +4.56]
Successor: gpt-oss-120b			
Gemma-4-31B	0.76 [0.68, 0.84]	53	−0.76 [−1.13, −0.38]
gpt-oss-120b	0.95 [0.86, 1.05]	53	+0.44 [+0.18, +0.70]
Llama endpoint	1.45 [1.30, 1.62]	49	+2.84 [+2.37, +3.31]
Nemotron-3-Super	0.93 [0.83, 1.04]	53	+1.93 [+1.37, +2.51]
Successor: GPT-5.6 Terra			
Gemma-4-31B	0.94 [0.81, 1.08]	58	−0.05 [−0.68, +0.58]

(b) Successor failure inflation (percentage points)

Base agent	Per incoming request	Given takeover
Successor: GPT-5.6 Luna		
Gemma-4-31B	+0.0 [−2.4, +2.4]	+0.0 [−12.5, +12.5]
gpt-oss-120b	+1.8 [−0.6, +4.3]	+6.5 [−2.2, +15.2]
Llama endpoint	+5.5 [−0.6, +11.6]	+5.9 [−0.7, +12.4]
Nemotron-3-Super	+1.2 [−1.8, +4.3]	+4.0 [−6.0, +14.0]
Successor: gpt-oss-120b		
Gemma-4-31B	−1.8 [−4.3, +0.6]	−9.4 [−21.9, +3.1]
gpt-oss-120b	−6.1 [−9.8, −3.0]	−21.7 [−32.6, −10.9]
Llama endpoint	+5.5 [−0.6, +11.6]	+5.9 [+0.0, +12.4]
Nemotron-3-Super	−8.5 [−12.8, −4.3]	−28.0 [−40.0, −16.0]
Successor: GPT-5.6 Terra		
Gemma-4-31B	+0.0 [−3.1, +3.1]	+0.0 [−16.7, +16.7]

(c) Tokens and prices

Base agent	Tokens (k): immediate → status quo	Break-even base price (USD/Mtok)
Successor: GPT-5.6 Luna		
Gemma-4-31B	46.4 → 59.5 (+28%)	0.173
gpt-oss-120b	45.4 → 57.3 (+26%)	0.171
Llama endpoint	46.5 → 75.0 (+61%)	0.038
Nemotron-3-Super	47.0 → 100.9 (+115%)	0.095
Successor: gpt-oss-120b		
Gemma-4-31B	54.5 → 59.6 (+9%)	–
gpt-oss-120b	58.0 → 57.0 (−2%)	–
Llama endpoint	55.5 → 80.7 (+45%)	–
Nemotron-3-Super	55.1 → 99.8 (+81%)	–
Successor: GPT-5.6 Terra		
Gemma-4-31B	52.5 → 60.0 (+14%)	1.898

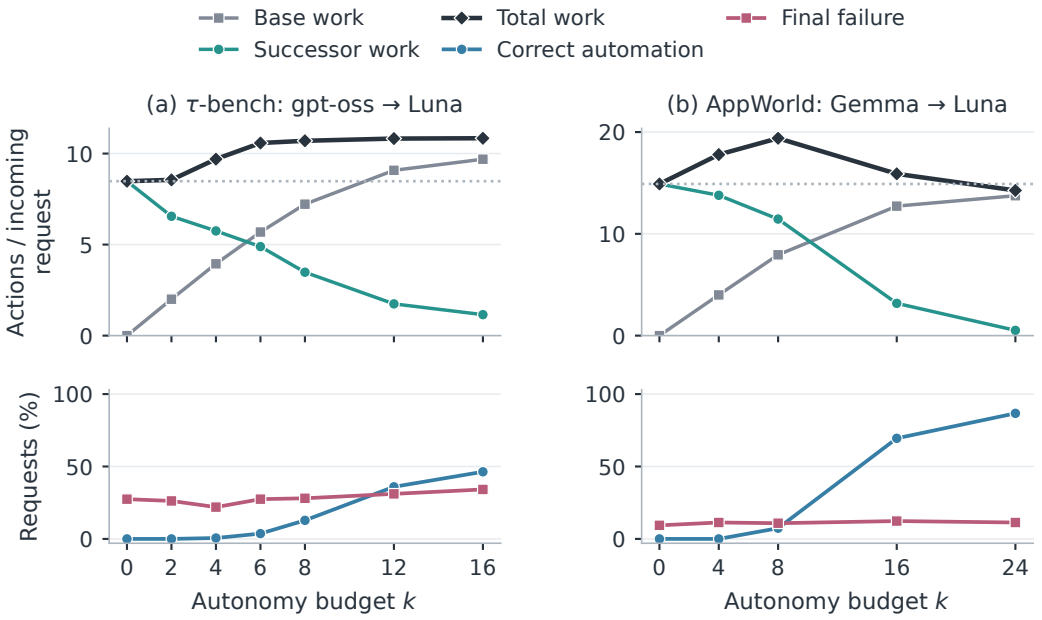

Figure 5: **Autonomous completion and total work are different quantities.** Columns show τ-bench gpt-oss and AppWorld Gemma, both with Luna. Every depth averages all incoming requests under C_k, including evaluated closures with zero successor work. Top: base work (gray squares), successor work (teal circles), and their sum (dark diamonds); dotted lines mark immediate-handoff total work. Bottom: correct automation (blue circles) and final failure (rose squares), as percentages of incoming requests.

request), or excluded from both policies. Every whole-system rise with Luna survives (+0.88 to +4.04 actions when handed over); Gemma's failure contrast becomes −5.5 points [−11.6, +0.6]; Gemma's fall with Terra vanishes (−0.06 [−0.51, +0.38]); and in the second-family replication, collected at the endpoints only, excluding these requests keeps all three signs (Gemma −0.62 [−0.98, −0.27]). Episodes that end this way under immediate handoff count against the successor in every table. All other analyses keep the benchmark's labels.

Table 6: Failed closures of the τ-bench base runs by who ended the episode (main tables, 164 requests per base agent).

Base agent	failed closures	agent closed	consent, not run	transfer marker	out-of-scope marker	user declined
Gemma-4-31B	34	16	12	2	2	2
gpt-oss-120b	42	37	2	1	2	0
Llama endpoint	9	3	0	5	1	0
Nemotron-3-Super	23	16	3	3	0	1

Prospective replication on `telecom`. The plan fixed the 20-task "small" probe and 114-task "base" evaluation before evaluation calls, using the main protocol. Three cost-probe tasks preceded it; a dated correction fixes the plan's initial miscount. Harness/analysis amendments are recorded in its addendum. Probe autonomous success was Gemma-4-31B 80%, gpt-oss-120b 80%, Llama endpoint 0%, Nemotron-3-Super 45%, versus 25% for Luna from scratch (80 rows). Only 1 base agent qualifies; Table 8 retains its 0 unresolved requests. Tasks combine shared atomic faults, although the reported request bootstrap treats them as independent.

Acquisition cost of depth selection. The work model uses other agents' full tables (1,359 successor rows per τ-bench decision). Table 10 compares this with sampling the target agent's own branches

Table 7: Status quo minus immediate handoff with the simulator-ended failed closures as labelled, handed over at the last executed checkpoint, or excluded (one-sided 95% bounds; † endpoint rows only, non-probe requests).

Metric	As labelled	Handed over	Excluded
Gemma-4-31B → *GPT-5.6 Luna*; $n = 164$, relabelled: 16			
Δ actions	+0.52 [+0.23, +0.83]	+0.88 [+0.64, +1.13]	+0.89 [+0.62, +1.15]
Δ failure (pts)	+0.0 [−6.7, +6.7]	−5.5 [−11.6, +0.6]	
Gemma-4-31B → *GPT-5.6 Terra*; $n = 65$, relabelled: 10			
Δ actions	−0.77 [−1.45, −0.14]	−0.06 [−0.51, +0.38]	−0.05 [−0.55, +0.45]
Δ failure (pts)	+1.5 [−9.2, +12.3]	−10.8 [−20.0, −1.5]	
gpt-oss-120b → *GPT-5.6 Luna*; $n = 164$, relabelled: 5			
Δ actions	+2.36 [+1.99, +2.73]	+2.43 [+2.05, +2.81]	+2.36 [+1.99, +2.74]
Δ failure (pts)	+6.7 [+0.0, +13.4]	+5.5 [−1.2, +12.2]	
Llama endpoint → *GPT-5.6 Luna*; $n = 164$, relabelled: 6			
Δ actions	+3.65 [+3.22, +4.09]	+3.72 [+3.30, +4.15]	+3.72 [+3.27, +4.18]
Δ failure (pts)	+10.4 [+4.3, +17.1]	+6.7 [+0.6, +12.8]	
Nemotron-3-Super → *GPT-5.6 Luna*; $n = 164$, relabelled: 6			
Δ actions	+4.02 [+3.55, +4.49]	+4.04 [+3.58, +4.52]	+4.10 [+3.61, +4.60]
Δ failure (pts)	−2.4 [−8.5, +3.7]	−4.3 [−10.4, +1.8]	
Gemma-4-31B → *Claude Opus 5*†; $n = 145$, relabelled: 12			
Δ actions	−0.92 [−1.30, −0.55]	–	−0.62 [−0.98, −0.27]
Δ failure (pts)	−1.4 [−9.0, +6.2]	–	
gpt-oss-120b → *Claude Opus 5*†; $n = 145$, relabelled: 4			
Δ actions	+0.99 [+0.56, +1.45]	–	+1.04 [+0.59, +1.50]
Δ failure (pts)	+7.6 [+0.0, +15.2]	–	
Nemotron-3-Super → *Claude Opus 5*†; $n = 145$, relabelled: 5			
Δ actions	+2.99 [+2.41, +3.57]	–	+3.09 [+2.49, +3.69]
Δ failure (pts)	−6.2 [−13.8, +0.7]	–	

Blank: not tabulated. –: no executed handover contrast.

Table 8: Prospective replication on τ-bench `telecom`: status quo against immediate handoff for the pairing admitted by the pairing rule (one-sided 95% bounds; bold: whole-system actions significantly higher).

Base agent	n	ΔA (pts)	ΔS	Δ actions	Δ fail (pts)	r^*	both resolved (%): Δ actions
Llama endpoint	114	+0 [+0, +0]	−1.25 [−1.89, −0.64]	**+4.24 [+3.55, +4.94]**	+4 [+0, +9]	0.23	18: +2.71 [+2.00, +3.43]

Table 9: **Work forecasts and depth selection.** (a) Error of $\hat{S}(k)$ relative to w_0 (sign accuracy of $J(k) - J(0)$; share with the optimum within one grid step). Column headings name what is held out. Bold: all three thresholds met (≤ 0.10, ≥ 0.80, ≥ 0.70). †Prespecified tests sharing requests with fitting. (b) Regret to the hindsight-best fixed depth, averaged over held-out agents, successors and five price ratios [90% interval]. Five constants: prespecified; exchange and disjoint-request tests: post hoc. Endpoint rule chooses immediate handoff or the status quo by the exchange model.

(a) Work forecasts: nMAE (sign accuracy, optimum within one step)

Model	Agent†	Depths†	Agent + requests
τ-bench: 16 cells; disjoint requests: 8 cells, 65 requests			
Five constants	**0.100** (0.84, 0.88)	**0.097** (0.81, 0.88)	**0.088** (0.90, 0.88)
Exchange (post hoc)	**0.099** (0.83, 0.88)	0.105 (0.81, 0.88)	**0.083** (0.90, 0.88)
No progress	0.229 (0.80, 0.81)	0.260 (0.78, 0.88)	0.215 (0.83, 0.75)
Flat	0.391 (0.80, 0.81)	0.467 (0.78, 0.81)	0.395 (0.83, 0.75)
Own-table reference	–	–	0.023 (0.94, 1.00)
AppWorld: 4 cells; disjoint requests: 4 cells, 56 requests			
Five constants	0.111 (0.75, 0.50)	0.038 (0.75, 0.50)	0.105 (0.88, 0.75)
Exchange (post hoc)	**0.045** (0.81, 0.75)	0.043 (0.75, 0.75)	**0.092** (1.00, 1.00)
No progress	0.142 (0.81, 0.50)	0.186 (1.00, 0.50)	0.107 (0.94, 0.75)
Flat	0.497 (0.75, 0.50)	0.470 (0.75, 0.50)	0.428 (0.81, 0.75)
Own-table reference	–	–	0.074 (0.88, 0.50)

(b) Depth selection: regret (%) on held-out requests

Rule	τ-bench	AppWorld
Five constants	2.7 [2.1, 7.1]	1.1 [0.5, 7.2]
Exchange (post hoc)	2.3 [1.8, 6.7]	1.0 [0.5, 4.4]
No progress	2.6 [2.3, 7.4]	0.9 [0.3, 4.5]
Flat	36.4 [29.2, 62.7]	16.6 [10.7, 27.9]
Endpoint rule	3.7 [2.8, 7.8]	0.4 [0.2, 3.1]
Other-agent depth	5.9 [5.1, 11.7]	34.9 [24.7, 56.9]
Status quo	7.1 [5.6, 11.6]	9.8 [6.5, 15.8]
Immediate handoff	71.8 [56.8, 99.7]	85.4 [61.4, 131.2]
Own-table reference	3.2 [2.6, 8.1]	0.7 [0.2, 3.0]

(post hoc). Using only its endpoints, 28 rows yield 3.8% regret against the model's 2.7% in τ-bench; in AppWorld, 10 rows already beat it (0.8% versus 1.1%). The source-table acquisition cost remains part of the comparison.

Table 10: Depth selection: relative regret to the best fixed depth in hindsight on the held-out agent's test requests, for zero-run rules and for rules that sample n_s fit requests of the target agent (mean over 300 subsamples), with the successor rows each rule consumes.

	τ-bench		AppWorld	
Rule	regret (%)	successor rows of the target agent	regret (%)	rows
Five-constant model, other agents' tables	2.7	0	1.1	0
Exchange form, other agents' tables	2.3	0	1.0	0
No-progress null, other agents' tables	2.6	0	0.9	0
Endpoint rule, other agents' tables	3.7	0	0.4	0
Status quo	7.1	0	9.8	0
Endpoint sampling, 3 requests	9.2	4	3.0	3
Full-grid sampling, 3 requests	17.0	14	12.9	9
Endpoint sampling, 5 requests	6.7	7	1.6	5
Full-grid sampling, 5 requests	14.8	23	7.1	14
Endpoint sampling, 10 requests	4.4	14	0.8	10
Full-grid sampling, 10 requests	10.0	46	2.4	29
Endpoint sampling, 20 requests	3.8	28	0.5	21
Full-grid sampling, 20 requests	6.0	92	0.9	58
Endpoint sampling, 40 requests	3.5	57	0.4	42
Full-grid sampling, 40 requests	4.0	183	0.5	115
Own full table	3.2	453	0.7	423

A second successor family, frozen in advance. The plan tested a second successor family in fixed order: Sonnet 4.6, then Opus 5 only if Sonnet qualified for no base agent. It was hashed after one retained smoke-probe row and before all remaining rows. On 19 training requests, successor solo success had to exceed the corresponding base agent's same-model depth-0 success by ten points. Base counts were 12/19 (Gemma), 13/19 (gpt-oss) and 12/19 (Nemotron); required successor counts were 14/19, 15/19 and 14/19. The primary population was the 145 non-probe requests, with the joint endpoint test of Section 5.1 and an exchange-model forecast frozen before evaluation.

Table 11 retains every probed successor. Sonnet fails all pairings: 9/19 successes (47%); of 10 failures, 8 made no write and 5 ended when the simulated user confirmed and closed after a confirmation question. Luna's 3 failures all made a write. Opus qualifies for all three bases with 15/19 successes (79%) and 10.8 original-state actions. Its forecasts and outcomes appear together in Table 13; plans, amendments and hashes remain in the artifact.

Table 11: Second-family probe: solo success of every probed successor at depth 0 on the 19 probe requests of the Gemma table, its from-scratch work $w_0(S)$, and whether the pairing rule (base agent's own solo success plus 10 points) holds for each capable base agent. Only the second-family successors for which it holds were evaluated in this replication.

Successor	Family	Solo success (depth 0, probe)	%	$w_0(S)$	Rule holds against		
					Gemma	gpt-oss	Nemotron
GPT-5.6 Luna	GPT-5.6	16/19	84	8.7	yes	yes	yes
GPT-5.6 Terra	GPT-5.6	14/19	74	9.6	yes	no	yes
Claude Sonnet 4.6	Claude	9/19	47	7.1	no	no	no
Claude Opus 5	Claude	15/19	79	10.8	yes	yes	yes
DeepSeek-V4-Pro	DeepSeek	11/19	58	8.8	no	no	no
Kimi-K2.6	Kimi	6/13	46	8.5	no	no	no
gpt-oss-120b	open	13/19	68	9.9	no	no	no
Nemotron-3-Super	open	12/19	63	12.9	no	no	no
Gemma-4-31B	open	12/19	63	8.8	no	no	no
Llama endpoint	open	3/19	16	5.9	no	no	no

Second family: outcome. Endpoint collection covered 3 bases $\times$ 164 requests, with one generation failure on a probe request; there is no depth grid or curve fit. In the primary 145-request population, the joint rise is supported for gpt-oss and Nemotron and reversed for Gemma; successor work falls in all three (Table 12). Gemma's reversal is consistent with Opus's original-state work 10.01 exceeding its break-even 9.00. The complete 164- and 65-request sensitivity grids remain in the artifact and preserve all three signs. Their quality bounds, like the primary bounds, are all positive: quality preservation is established in none. For gpt-oss, the 164-request failure interval is $[+1, +14]$ points, extending above the ten-point tolerance.

The frozen forecast gets every sign and misses every magnitude interval (Table 13). A post-hoc no-progress model using the same probe mean also gets the signs and misses all intervals. Its absolute errors are 0.70, 0.89 and 1.17, versus 0.81, 0.93 and 1.55 for exchange; these signs do not validate the progress term. The 19-row original-state mean accounts for the Gemma/gpt-oss forecast errors and most of Nemotron's; terminal successor work accounts for the remainder (full decomposition in the artifact). Final failure is predicted within three points in every cell.

Closure-ratio restriction. With Luna, the pooled correct-closure workload ratio changes from 0.68 [0.63, 0.73] at $k = 6$ to 0.97 [0.94, 1.00] at the τ-bench cap, and from 0.75 [0.69, 0.83] at $k = 8$ to 0.93 [0.90, 0.96] at the AppWorld cap (clustered 90% intervals). This violates the depth-constant assumption. Full fitted coefficients are retained in the artifact; progress slopes remain associations, not effects of chosen actions.

Two estimators of the same constants. The curve-level least-squares estimator regresses the seven-point curve $S(k)$ on $A(k)$, $c(k)$, $p_k m_k$ and $p_k n_k$, which are strongly collinear. The paired estimator was selected after an interim look at coefficient instability and held-out-agent error above the threshold. On the reported data, curve-level regression gives 0.146 for held-out agents (the paired estimator: 0.100) and 0.067 for held-out depths (0.097). We report both because the choice was made after an interim look. Under joint resampling of requests the difference between the five-constant

Table 12: **Second successor family: primary endpoints.** Claude Opus 5 on all 145 non-probe requests per base, with no unresolved primary rows. All contrasts are status quo minus immediate handoff. Brackets: paired lower/upper one-sided 95% bootstrap bounds. Bold action/cost contrasts have positive lower bounds. Supported requires positive lower bounds for both ΔA and total actions; reversed requires a negative upper action bound. U: betting upper bound on excess failure; positive values are inconclusive. Dominates: immediate handoff lowers actions and failure (bootstrap share in parentheses). Rose identifies the Opus successor. Full sensitivity grids remain in the artifact.

(a) Autonomy and work

Base	ΔA (pp)	ΔS	Δ actions
Gemma	$+63\,[+57, +70]$	$-9.43\,[-9.88, -8.98]$	$-0.92\,[-1.30, -0.55]$
gpt-oss	$+47\,[+41, +54]$	$-8.71\,[-9.25, -8.15]$	**+0.99 [+0.56, +1.45]**
Nemotron	$+56\,[+49, +63]$	$-8.70\,[-9.19, -8.19]$	**+2.99 [+2.41, +3.57]**

(b) Quality, cost and declared reading

Base	Δ fail (pp)	ΔJ_q	r^*	U	Dominates (share)	Reading
Gemma	$-1\,[-9, +6]$	$-1.20\,[-2.67, +0.19]$	1.11	$+0.084$	no (0.00)	reversed
gpt-oss	$+8\,[+0, +15]$	**+2.56 [+1.00, +4.14]**	0.90	$+0.173$	yes (0.96)	supported
Nemotron	$-6\,[-14, +1]$	**+1.75 [+0.29, +3.18]**	0.74	$+0.036$	no (0.09)	supported

Table 13: **Forecast frozen before evaluation.** Exchange-form predictions against the primary outcomes: all three signs are correct, and none of the predicted action changes lies within the observed 90% interval. Pred./obs.: predicted/observed terminal successor work and failure. Rose identifies the Opus successor.

Base	Δ actions predicted	Δ actions observed [90% CI]	$S(16)$ pred./obs.	Fail(16), % pred./obs.
Gemma	-1.73	$-0.92\,[-1.30, -0.55]$	$0.61/0.58$	$25/22$
gpt-oss	$+0.07$	$+0.99\,[+0.56, +1.45]$	$1.22/1.23$	$32/32$
Nemotron	$+1.44$	$+2.99\,[+2.41, +3.57]$	$0.59/0.94$	$20/19$

model and the exchange form is [−0.003, +0.006] in τ-bench and [+0.039, +0.095] in AppWorld, and between the law and the null without a progress term [−0.143, −0.114] and [−0.049, −0.014]. Resampling adds noise to an absolute error, so percentile intervals of the errors themselves sit above the point estimates; we report intervals of differences.

Depth selection as prespecified. The prespecified analysis used the held-out agent's same requests for base-side curves and scoring (80 decisions across two domains); its complete table is retained in the artifact. Table 9 distinguishes those results from the post hoc disjoint-request analysis.

Controller comparisons. Table 14 shows no established DOVE advantage over per-agent depth; same-label mean regressors improve on held-out tasks in one pointwise comparison among many. Per-agent train+calibration depth was specified after the τ-bench test and before AppWorld outcomes; the agent descriptor is post hoc. Calibration-only depth uses the controllers' selection split: its decisions are unchanged by the failure constraint, and its AppWorld row matches train+calibration selection. Unconstrained DOVE counters are unchanged; the selected embedding variant was not reselected without the constraint (Appendix B).

The prospective comparison carried source-trained, calibration-selected policies to 40 `test_challenge` tasks without refitting or reselection, under a comparison declared before analysis. Source-test representations were data-informed and include task instructions and environment outputs (Appendix D). The descriptor gain does not recur. Held-out quality bounds were not reported, so cost improvements do not establish quality preservation.

Table 14: **Core controller comparisons at the declared primary prices.** ΔJ_q is relative to the cohort's per-agent train+calibration depth (carried on held-out tasks); negative favors the row. Intervals are pointwise request-clustered 90%, without a simultaneous claim. U bounds excess failure over *status quo*, not per-agent depth ($\delta = .10$): only $U \leq 0$ supports quality not lowered; "—" means unreported. *: lowest calibration-cost prespecified feature set, reported as DOVE. †: the merged τ-bench rows match status-quo decisions on this sample, not policy equivalence. Pale blue identifies fixed references; violet identifies DOVE variants, not significance.

Policy	Fail (%)	J_q	ΔJ_q vs per-agent [90% CI]	U
τ-bench: 65 requests × 4 bases				
Per-agent (train+cal)	28.1	8.35	$+0.00\,[+0.00, +0.00]$	$+0.049$
Per-agent (cal)	25.8	7.92	$-0.43\,[-0.80, -0.08]$	$+0.023$
DOVE*/mean regressors†	28.5	8.19	$-0.15\,[-0.54, +0.22]$	$+0.048$
AppWorld: 56 scenarios × 2 bases				
Status quo / pooled	42.0	14.97	$+1.80\,[+0.27, +3.26]$	$+0.055$
Per-agent (train+cal/cal)	15.2	13.18	$+0.00\,[+0.00, +0.00]$	-0.186
DOVE* (embedding)	26.8	14.39	$+1.21\,[-0.01, +2.51]$	-0.068
DOVE + agent (post hoc)	15.2	12.58	$-0.60\,[-1.01, -0.24]$	-0.186
Mean regressors	27.7	14.11	$+0.94\,[-0.46, +2.44]$	-0.057
Held-out AppWorld: 40 challenge tasks × 2 bases				
Per-agent (carried over)	35.0	21.68	$+0.00\,[+0.00, +0.00]$	—
DOVE* (carried)	35.0	19.63	$-2.05\,[-4.18, +0.11]$	—
DOVE + agent (post hoc)	40.0	21.91	$+0.23\,[-1.09, +1.41]$	—
Mean regressors	36.3	19.06	$-2.62\,[-4.52, -0.68]$	—

Complete controller artifact. `paper/generated/controller_compact_source_data.json` contains all 51 policy rows (20/16/15), both paired cost comparisons, the ten action-price cells and eight failure-price rows. Full and compact tables are available under `paper/generated/`, including `tab_policy_ratios.tex` and `tab_kappa.tex`; seed-0–4 ranges remain in `numbers.tex` and the corresponding `results/policy_tau3_final*.json` / `results/policy_appworld_final*.json` files. The compact-table source sidecar records the selected rows without recomputation.

Same-state replicates. Replicates keep the inherited state, seed and temperature fixed (Figure 6; full grids in the artifact). The oracle knows the realised base trajectory, and its reference depth is selected on the same draws; their contrast is therefore a diagnostic, not a decision-time bound. The variance decomposition accounts for identical options after an early terminal takeover: between-state shares of the cost advantage are 0.57, 0.57 and 0.57 at depths 0, 4 and 8. Depth-0 rows across base-agent tables have different regenerated user greetings and are excluded from same-state groups.

Prospective test on AppWorld `test_challenge`. The plan (`analysis/PROSPECTIVE_PLAN.md` in the artifact, with hashes of the code, the task lists and the frozen constants) was written before any task of this set was run by any agent (author-attested: the records carry no execution timestamps; the hashes of the plan, the code, the task lists and the constants match). Scenarios are sorted by id; variant 1 of the first 30 is the labelled sample (base runs only, no successor), variant 1 of the next 40 the evaluation set (full branch tables with Luna, 0 unresolved rows). For each agent the constants, $\bar{z}$ and the failure price come from the other agent's source table, as in the held-out-request test; nothing was re-estimated on the evaluation population. Table 15 gives every rule. Differences of the model-selected depth to the carried-over depth, the endpoint rule and the no-progress null: [−5.8, +20.5], [−7.1, +13.8] and [−10.8, +9.2] points. The comparison of the descriptive findings with Table 3 reported in the main text was not declared in the plan.

Decision-time oracle. The oracle of Figure 6 conditions on one realised base trajectory. For the decision at depth 0, the base agent was run three more times from the identical state (greeting and

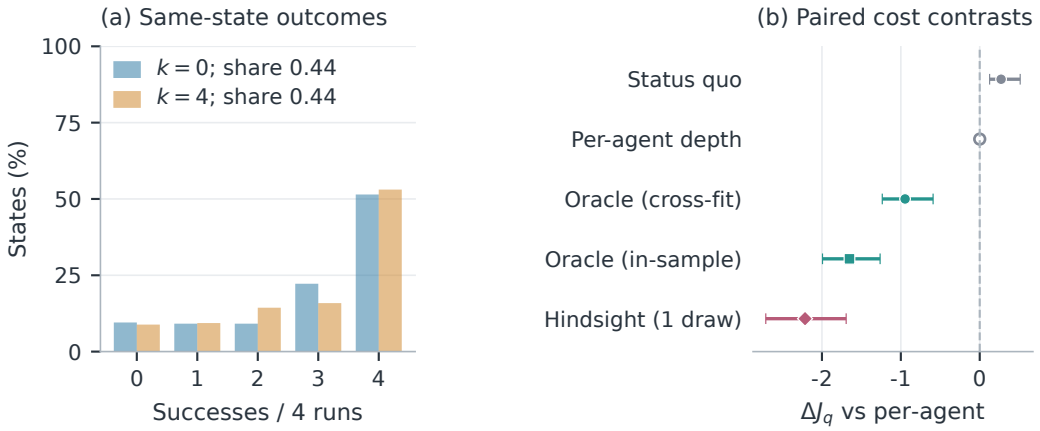

Figure 6: **Taking the oracle gap apart.** Left: successes in four successor runs from the identical inherited state, at depth 0 (blue) and 4 (amber); legend shares are between-state shares of binary failure variance. Right: paired J_q differences from the best per-agent depth on $\{0, 4, 8, \text{never}\}$, with request-clustered 90% intervals. Gray marks fixed references; teal marks cross-fitted and in-sample trajectory oracles (diagnostics, not bounds); rose marks single-draw hindsight. The open marker is the zero-by-definition reference.

Table 15: Prospective test of depth selection on 40 held-out AppWorld tasks (two base agents, five price ratios; 90% intervals from joint resampling of sample and evaluation tasks). The declared hypothesis, model-selected depth below the status quo, is not supported.

Depth chosen by	relative regret (%) on the evaluation tasks
Five-constant model	14.0 [3.3, 28.7]
Two-constant exchange form	13.3 [2.5, 27.7]
Null: no progress term	13.0 [4.3, 32.9]
Endpoint rule (exchange form)	8.7 [3.2, 24.4]
Carried-over depth (selected on the source-cohort branch table)	7.0 [2.2, 16.3]
Status quo (run to the cap)	11.7 [6.1, 23.2]
Immediate handoff	33.6 [21.2, 57.4]

first user message of the table's run, same seed and temperature), to the cap, with the table's boundary rule and transfer view, and the successor took over wherever a run ended without closing. With the table's own run this gives four independent draws of "never stop", and the four successor draws at depth 0 give "hand off now"; no draw of one option shares anything with a draw of the other beyond the state. Table 16 reports the two-option problem. The between-state share of the variance of the handoff advantage is 0.32 here, against 0.57 when the base trajectory is held fixed.

Table 16: Decision at depth 0, hand off now or never stop, with both options re-drawn from the decision-time state (260 request–agent pairs, four draws per option; request-clustered 90% intervals).

ΔJ_q vs best per-agent choice (depth-0 decision)	$\alpha/\nu = 0.1$	0.5	1
Always hand off	+4.93 [+4.37, +5.66]	+1.92 [+1.41, +2.47]	+0.00 [+0.00, +0.20]
Never stop (status quo)	+0.00 [+0.00, +0.26]	+0.57 [+0.28, +0.86]	+3.12 [+2.47, +3.89]
Decision-time oracle, cross-fitted (pessimistic)	+0.18 [+0.00, +0.49]	−0.15 [−0.44, +0.13]	−0.02 [−0.19, +0.23]
Decision-time oracle, in-sample (optimistic)	−0.67 [−0.78, −0.44]	−1.02 [−1.29, −0.79]	−0.91 [−1.11, −0.66]
Single-draw hindsight oracle	−1.23 [−1.53, −0.82]	−1.58 [−2.00, −1.19]	−1.42 [−1.83, −0.98]

Handoff-interface controls. Table 17 retains all depths. The same-model control clips resumed work to the uninterrupted run's remaining budget $K - k$ and counts unfinished runs as failures; unclipped results show the effect of the larger successor cap. Reconstruction compares every replayed output with its recorded value: 3 of 1,323 rows failed and were dropped. After two collectors accidentally overwrote common files, 284 rows were restored from logged work, reward and replay checks and flagged. The logs and replaced records remain available. Luna's fresh arms and preserved arms were collected sequentially, so their 73% difference cannot exclude endpoint or timing effects and is not an identified causal fraction. Appendix H gives the separate contemporaneous experiment.

Table 17: Handoff-interface controls on AppWorld (Gemma base trajectories). In (a), differences are preserved minus fresh interpreter. In (b), differences are resumed minus uninterrupted execution; cap is the uninterrupted run's remaining budget. ΔW_{clip} enforces this cap; ΔW_{raw} uses the full successor cap. Fail_0 is uninterrupted failure. Intervals are scenario-clustered 90%; gray bands group checkpoint depths.

(a) Gemma → Luna

k	n	W_{fresh}	W_{keep}	ΔW [90% CI]	ΔFail (pp) [90% CI]
4	203	13.79	12.06	−1.73 [−2.18, −1.24]	+1 [−1, +3]
8	184	12.58	8.95	−3.63 [−4.06, −3.20]	−1 [−4, +2]
16	53	12.15	7.15	−5.00 [−6.08, −3.96]	+2 [−3, +8]

(b) Gemma → Gemma

k (cap)	Interface	n	ΔW_{clip} [90% CI]	ΔW_{raw}	ΔFail (pp) [90% CI]	Fail_0 (%)
4 (20)	keep	203	−0.33 [−0.60, −0.08]	−0.05	+0.5 [−2.4, +3.4]	13
4 (20)	fresh	203	+0.28 [−0.04, +0.60]	+0.49	−2.0 [−5.0, +1.0]	13
8 (16)	keep	184	+0.05 [−0.16, +0.26]	+0.41	+2.2 [+0.0, +4.4]	13
8 (16)	fresh	184	+1.60 [+1.30, +1.90]	+1.84	+0.5 [−2.7, +3.8]	13
16 (8)	keep	53	+0.30 [−0.04, +0.67]	+2.53	+9.4 [+3.7, +16.0]	34
16 (8)	fresh	53	+2.11 [+1.71, +2.49]	+4.70	+13.2 [+4.2, +21.8]	34

Tokens and dollars. For gpt-oss-120b → Luna in τ-bench ($n = 164$), total tokens per request rise from 45.4k under immediate handoff to 57.3k under the status quo (+26%), about the same relative

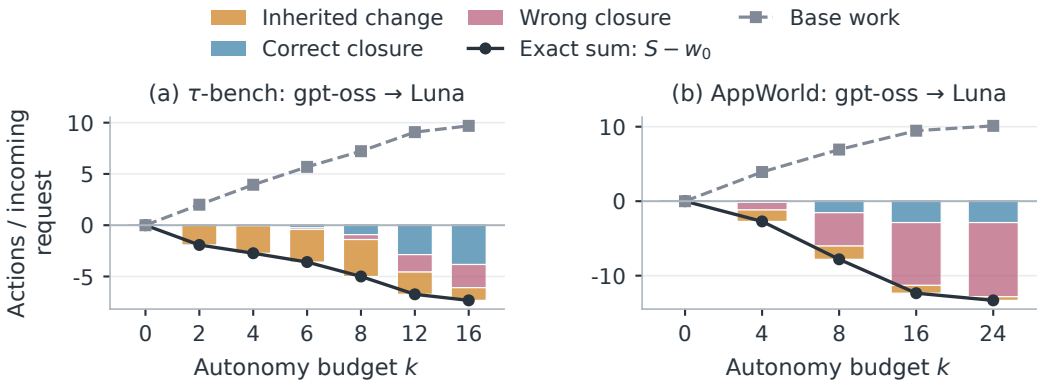

Figure 7: **Where the successor's work goes.** Exact accounting of $S(k) - w_0$ per depth. In τ-bench (left), inherited-state work reduction dominates at small depths and correct automation at the end; the reduction costs more base actions than it saves. With the weak AppWorld base (right), most apparent savings come from wrong closures. Amber: inherited-state change; blue: correct-closure savings; rose: wrong-closure savings; dark circles: their exact signed sum; gray dashed squares: base work.

increase as in actions (+28%). The successor's bill falls. Equation 4 specifies when the base price makes this reduction sufficient to lower total cost.

User simulator. With gpt-oss-120b in place of Luna as user simulator on the test requests, the status quo against immediate handoff gives +3.51 [+2.31, +4.75] whole-system actions for gpt-oss-120b → Luna, against +2.51 on the same requests with the main simulator.

Lookup against live execution. Proposition 3 says that reading a stop rule's outcome from the table is a draw of what deploying it would give. We ran one rule live, "hand off after four base actions", with Gemma-4-31B and Luna on the 65 test requests: fresh base runs, fresh successor runs, the rule in the loop. Live minus lookup: base actions −0.05 [−0.14, +0.00], successor actions −0.09 [−0.46, +0.28], failure +5 points [−3.15, +13.85]. The intervals contain zero but are wide; this is a consistency check, not a demonstration of equivalence. A state-dependent rule that needs no model, "hand off at the first checkpoint whose trace shows a tool error or a repeated call" (chosen for its varied stopping depths, without looking at outcomes; it hands off in 83% of requests by lookup and 91% live), gives live minus lookup +0.17 [−0.17, +0.52] base actions, +0.34 [−0.14, +0.85] successor actions and +2 points [−8, +11] of failure on the same 65 requests. (Five live runs that ended in a native transfer were repeated after the live harness was given the transfer view of the table; the repeated runs are flagged in the result file, and their collection log and excluded attempts remain in the artifact.)

Judge substitution. 100 executed `retail` trajectories (a fixed-seed sample over 38 tasks that carry natural-language assertions, 148 assertions; every row is identified by its run record) were re-judged by GPT-5.6 Luna, the judge used in the paper, and by GPT-5.6 Terra with the benchmark's own prompt. The two agree on 100% of trajectories (Cohen's $\kappa = 1.00$; pass rate 88%).

F Complete Attribution and Held-Out-Population Audits

Full-grid attribution and positive-budget contrasts (post hoc). We include all six primary Luna pairings and every nonzero depth: 24 cells from τ-bench (164 requests per base agent) and 8 from AppWorld (203 requests from 105 scenarios per base agent). The three attribution quantities are ι_k, the survivor-to-population contrast, and their difference in Equation 3. A separate family contains all 26 adjacent transitions between positive budgets, each with autonomous correct completion, total actions and final failure. No transition involving C_0 enters that family.

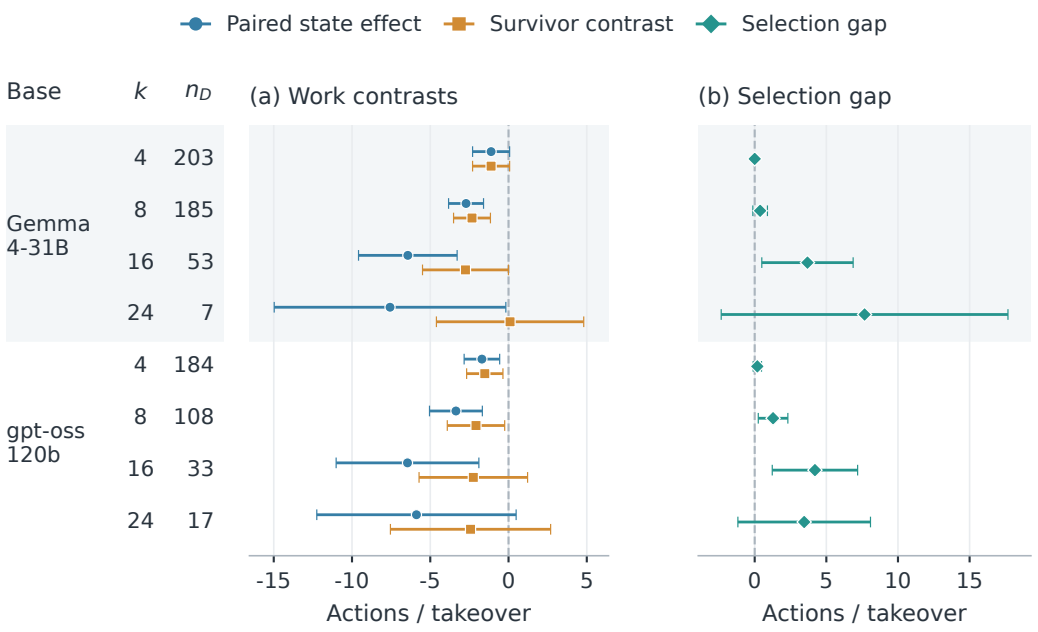

Figure 8: **Selection can hide inherited-state work reduction.** All nonzero AppWorld depths with Luna. Left: paired state effects (blue circles) and survivor-to-population contrasts (amber squares); right: their difference (teal diamonds). Negative left-panel values mean fewer successor actions. Bars retain the post hoc global simultaneous 95% calibration over all 96 coordinates of all 32 audit cells, including τ-bench; AppWorld is resampled by scenario. n_D gives takeover counts; shading separates base models. At gpt-oss depth 16, selection masks a supported reduction. The cap-depth Gemma group has only 7 takeovers.

For each family we use 20,000 bootstrap draws, sharing request weights across agents and depths within an environment and resampling AppWorld by scenario. A global simultaneous 95% band is obtained from the 95th percentile of the maximum absolute centered bootstrap deviation divided by each coordinate's bootstrap standard deviation: 96 attribution coordinates or 78 transition coordinates. This accounts for inspecting the whole reported grid. The cap-depth AppWorld Gemma cell contains only 7 takeovers; 8 draws have no takeover. Their maximum statistic is set to infinity for family calibration, while the ratio's standard deviation uses its defined draws. These rare undefined draws are not discarded from multiplicity calibration. We do not infer a sign reversal from that cell.

Table 18: Complete audit summary with global simultaneous 95% bands. Selection significance concerns the additive bias; joint rise concerns both autonomous correct completion and total actions. All analyses are post hoc.

Environment	Nonzero-depth cells	Bias detected	Changed with detected bias	Positive-budget joint rises
τ-bench	24	3	0	1/20
AppWorld	8	3	1	0/6

Frozen-parameter curve forecasts on held-out AppWorld tasks (post hoc). The prospective plan specified depth-selection regret. This post hoc audit also scores its full predicted work curves without refitting any parameter. The sample forecast uses the 30 base-only sample tasks and scores the 40 independent evaluation tasks. The evaluation-covariate diagnostic instead supplies the evaluation tasks' own base-side curves, excluding successor outcomes; it is a conditional diagnostic rather than a disjoint-sample forecast. Models and nonnegative prediction clipping are unchanged from the frozen experiment.

Table 19: Every primary τ-bench attribution cell. Global simultaneous 95% bands across both environments.

Base	k	Takeovers	Paired effect ι_k	Survivor contrast	Selection bias
Gemma	2	162/164	$-1.71\,[-2.23, -1.19]$	$-1.74\,[-2.27, -1.22]$	$-0.03\,[-0.13, +0.07]$
Gemma	4	158/164	$-3.43\,[-3.93, -2.94]$	$-3.37\,[-3.90, -2.85]$	$+0.06\,[-0.14, +0.26]$
Gemma	6	142/164	$-4.96\,[-5.61, -4.32]$	$-4.70\,[-5.38, -4.03]$	$+0.26\,[-0.05, +0.58]$
Gemma	8	101/164	$-6.21\,[-7.09, -5.33]$	$-5.51\,[-6.31, -4.71]$	$+0.70\,[+0.16, +1.24]$
Gemma	12	44/164	$-5.89\,[-7.59, -4.18]$	$-5.37\,[-6.61, -4.12]$	$+0.52\,[-1.02, +2.06]$
Gemma	16	32/164	$-5.59\,[-7.94, -3.25]$	$-5.61\,[-6.85, -4.37]$	$-0.02\,[-2.05, +2.01]$
gpt-oss	2	164/164	$-1.93\,[-2.43, -1.42]$	$-1.93\,[-2.43, -1.42]$	$-0.00\,[-0.00, +0.00]$
gpt-oss	4	162/164	$-2.72\,[-3.33, -2.10]$	$-2.66\,[-3.29, -2.03]$	$+0.06\,[-0.09, +0.20]$
gpt-oss	6	154/164	$-3.38\,[-4.16, -2.61]$	$-3.27\,[-4.06, -2.49]$	$+0.11\,[-0.07, +0.29]$
gpt-oss	8	132/164	$-4.48\,[-5.34, -3.62]$	$-4.16\,[-4.99, -3.32]$	$+0.32\,[+0.03, +0.61]$
gpt-oss	12	72/164	$-4.94\,[-6.25, -3.64]$	$-4.51\,[-5.59, -3.43]$	$+0.43\,[-0.43, +1.30]$
gpt-oss	16	46/164	$-4.41\,[-6.33, -2.50]$	$-4.37\,[-5.62, -3.13]$	$+0.04\,[-1.45, +1.53]$
Llama	2	164/164	$-0.36\,[-1.00, +0.28]$	$-0.36\,[-1.00, +0.28]$	$+0.00\,[-0.00, +0.00]$
Llama	4	162/164	$-0.55\,[-1.24, +0.14]$	$-0.52\,[-1.21, +0.18]$	$+0.03\,[-0.05, +0.11]$
Llama	6	160/164	$-0.66\,[-1.42, +0.10]$	$-0.60\,[-1.36, +0.16]$	$+0.06\,[-0.05, +0.17]$
Llama	8	159/164	$-0.92\,[-1.73, -0.11]$	$-0.86\,[-1.67, -0.05]$	$+0.06\,[-0.05, +0.17]$
Llama	12	156/164	$-0.99\,[-1.90, -0.09]$	$-0.94\,[-1.84, -0.04]$	$+0.05\,[-0.07, +0.17]$
Llama	16	153/164	$-1.07\,[-2.03, -0.10]$	$-1.02\,[-1.97, -0.07]$	$+0.04\,[-0.09, +0.18]$
Nemotron	2	164/164	$-1.71\,[-2.27, -1.15]$	$-1.71\,[-2.27, -1.15]$	$-0.00\,[-0.00, +0.00]$
Nemotron	4	164/164	$-2.57\,[-3.31, -1.82]$	$-2.57\,[-3.31, -1.82]$	$-0.00\,[-0.00, +0.00]$
Nemotron	6	160/164	$-3.44\,[-4.22, -2.65]$	$-3.37\,[-4.16, -2.58]$	$+0.07\,[-0.06, +0.19]$
Nemotron	8	148/164	$-4.56\,[-5.37, -3.76]$	$-4.35\,[-5.14, -3.56]$	$+0.21\,[-0.01, +0.43]$
Nemotron	12	88/164	$-5.86\,[-7.32, -4.40]$	$-5.10\,[-6.30, -3.90]$	$+0.76\,[+0.05, +1.48]$
Nemotron	16	50/164	$-6.04\,[-8.06, -4.02]$	$-5.45\,[-6.81, -4.09]$	$+0.59\,[-0.85, +2.04]$

Table 20: Every primary AppWorld attribution cell, with the same global simultaneous bands.

Base	k	Takeovers	Paired effect ι_k	Survivor contrast	Selection bias
Gemma	4	203/203	$-1.11\,[-2.30, +0.07]$	$-1.11\,[-2.30, +0.07]$	$+0.00\,[-0.00, +0.00]$
Gemma	8	185/203	$-2.71\,[-3.83, -1.60]$	$-2.33\,[-3.51, -1.16]$	$+0.38\,[-0.13, +0.89]$
Gemma	16	53/203	$-6.43\,[-9.58, -3.28]$	$-2.75\,[-5.50, -0.01]$	$+3.68\,[+0.50, +6.87]$
Gemma	24	7/203	$-7.57\,[-14.97, -0.17]$	$+0.10\,[-4.61, +4.80]$	$+7.67\,[-2.34, +17.68]$
gpt-oss	4	184/203	$-1.70\,[-2.84, -0.57]$	$-1.52\,[-2.68, -0.36]$	$+0.18\,[-0.11, +0.47]$
gpt-oss	8	108/203	$-3.36\,[-5.05, -1.67]$	$-2.08\,[-3.91, -0.25]$	$+1.28\,[+0.25, +2.31]$
gpt-oss	16	33/203	$-6.45\,[-11.01, -1.89]$	$-2.25\,[-5.72, +1.22]$	$+4.20\,[+1.23, +7.18]$
gpt-oss	24	17/203	$-5.88\,[-12.25, +0.49]$	$-2.43\,[-7.55, +2.69]$	$+3.45\,[-1.17, +8.08]$

Table 21: Every adjacent positive-budget contrast in τ-bench. Global simultaneous 95% bands across both environments.

Base	Budgets	ΔA (pp)	Δ actions	Δ failure (pp)	Joint rise
Gemma	$2 \to 4$	$+1.83\,[-1.88, +5.54]$	$+0.21\,[-0.35, +0.76]$	$-2.44\,[-14.71, +9.83]$	–
Gemma	$4 \to 6$	$+4.88\,[-1.12, +10.87]$	$+0.12\,[-0.36, +0.60]$	$+1.22\,[-10.31, +12.75]$	–
Gemma	$6 \to 8$	$+18.29\,[+7.60, +28.98]$	$-0.05\,[-0.41, +0.30]$	$+1.83\,[-9.35, +13.01]$	–
Gemma	$8 \to 12$	$+28.05\,[+15.77, +40.32]$	$+0.11\,[-0.19, +0.41]$	$+0.61\,[-5.82, +7.04]$	–
Gemma	$12 \to 16$	$+6.71\,[-0.20, +13.62]$	$-0.02\,[-0.10, +0.06]$	$-2.44\,[-6.69, +1.81]$	–
gpt-oss	$2 \to 4$	$+0.61\,[-1.54, +2.76]$	$+1.14\,[+0.51, +1.77]$	$-4.27\,[-16.99, +8.46]$	–
gpt-oss	$4 \to 6$	$+3.05\,[-1.75, +7.84]$	$+0.88\,[+0.27, +1.50]$	$+5.49\,[-6.42, +17.39]$	–
gpt-oss	$6 \to 8$	$+9.15\,[+1.09, +17.20]$	$+0.12\,[-0.39, +0.63]$	$+0.61\,[-10.93, +12.15]$	–
gpt-oss	$8 \to 12$	$+23.17\,[+11.43, +34.92]$	$+0.12\,[-0.32, +0.56]$	$+3.05\,[-6.27, +12.37]$	–
gpt-oss	$12 \to 16$	$+10.37\,[+1.92, +18.81]$	$+0.02\,[-0.14, +0.17]$	$+3.05\,[-2.58, +8.68]$	–
Llama	$2 \to 4$	$+0.61\,[-1.54, +2.76]$	$+0.86\,[+0.44, +1.28]$	$-3.05\,[-13.41, +7.32]$	–
Llama	$4 \to 6$	$+0.61\,[-1.55, +2.77]$	$+0.40\,[+0.07, +0.72]$	$+0.00\,[-6.08, +6.08]$	–
Llama	$6 \to 8$	$+0.00\,[+0.00, +0.00]$	$+0.18\,[-0.07, +0.42]$	$+3.66\,[-2.39, +9.71]$	–
Llama	$8 \to 12$	$+0.00\,[+0.00, +0.00]$	$+0.47\,[+0.06, +0.88]$	$+2.44\,[-4.28, +9.16]$	–
Llama	$12 \to 16$	$+0.00\,[+0.00, +0.00]$	$+0.18\,[-0.11, +0.47]$	$+3.05\,[-2.61, +8.71]$	–
Nemotron	$2 \to 4$	$+0.00\,[+0.00, +0.00]$	$+1.15\,[+0.48, +1.81]$	$+0.61\,[-11.34, +12.56]$	–
Nemotron	$4 \to 6$	$+2.44\,[-1.83, +6.71]$	$+1.03\,[+0.44, +1.62]$	$+1.22\,[-10.21, +12.65]$	–
Nemotron	$6 \to 8$	$+6.10\,[-0.48, +12.68]$	$+0.61\,[+0.02, +1.20]$	$+0.00\,[-11.79, +11.79]$	–
Nemotron	$8 \to 12$	$+28.05\,[+15.60, +40.50]$	$+0.70\,[+0.13, +1.26]$	$+1.22\,[-9.86, +12.30]$	yes
Nemotron	$12 \to 16$	$+18.90\,[+8.18, +29.62]$	$+0.25\,[-0.03, +0.53]$	$-3.05\,[-7.85, +1.75]$	–

Table 22: Every adjacent positive-budget contrast in AppWorld, with the same global simultaneous bands.

Base	Budgets	ΔA (pp)	Δ actions	Δ failure (pp)	Joint rise
Gemma	$4 \to 8$	$+7.39\,[-0.52, +15.30]$	$+1.61\,[+0.45, +2.76]$	$-0.49\,[-7.30, +6.31]$	–
Gemma	$8 \to 16$	$+62.07\,[+45.94, +78.19]$	$-3.50\,[-5.11, -1.88]$	$+1.48\,[-4.84, +7.79]$	–
Gemma	$16 \to 24$	$+17.24\,[+4.78, +29.70]$	$-1.63\,[-3.01, -0.25]$	$-0.99\,[-5.33, +3.36]$	–
gpt-oss	$4 \to 8$	$+11.82\,[+3.40, +20.25]$	$-2.08\,[-3.71, -0.45]$	$+20.69\,[+6.72, +34.66]$	–
gpt-oss	$8 \to 16$	$+10.34\,[+1.90, +18.79]$	$-2.01\,[-3.43, -0.60]$	$+24.14\,[+11.83, +36.45]$	–
gpt-oss	$16 \to 24$	$+0.00\,[+0.00, +0.00]$	$-0.32\,[-1.04, +0.40]$	$+4.43\,[-0.49, +9.35]$	–

For model m and depth set $\mathcal{K}$, normalized error is

$$\mathrm{nMAE}_m = \frac{1}{2} \sum_{h \in \{\mathrm{Gemma}, \mathrm{gpt\text{-}oss}\}} \frac{|\mathcal{K}|^{-1} \sum_{k \in \mathcal{K}} |\hat{S}_{mh}(k) - S_h^{\mathrm{eval}}(k)|}{S_h^{\mathrm{eval}}(0)}.$$

Agents and depths are equally weighted; $\mathcal{K} = \{0, 4, 8, 16, 24\}$ for full curves and $\{4, 8, 16\}$ for interior curves. We independently resample sample and evaluation scenarios 4,000 times, sharing each scenario's weight across agents. All intervals condition on the frozen source coefficients, failure rates and prices; they exclude uncertainty from drawing and refitting an independent source table. The normalized denominator is recomputed in each draw. These exploratory percentile intervals are not adjusted for multiple model comparisons. The analysis reproduces the frozen-model decision regrets before evaluating curve-error metrics.

Table 23: **Frozen work forecasts on held-out AppWorld tasks.** nMAE [pointwise 90% scenario-bootstrap interval], conditional on frozen source parameters. Sample columns use 30 base-only tasks and 40 disjoint evaluation tasks; interior depths are 4, 8 and 16. †: the evaluation-covariate column is a conditional diagnostic. This post hoc audit is not multiplicity-adjusted; paired model contrasts are reported in the text.

Frozen model	Sample: all depths	Sample: interior	Eval. covariates†
Five constants	0.237 [0.164, 0.302]	0.273 [0.184, 0.355]	0.244 [0.192, 0.296]
Exchange	0.219 [0.151, 0.286]	0.255 [0.169, 0.336]	0.228 [0.172, 0.279]
No progress	0.162 [0.113, 0.239]	0.166 [0.105, 0.258]	0.180 [0.134, 0.220]
Flat	0.303 [0.277, 0.335]	0.224 [0.193, 0.264]	0.303 [0.277, 0.335]

Both agents contribute to the no-progress advantage. Full-curve sample errors are 0.259 versus 0.170 for Gemma and 0.216 versus 0.154 for gpt-oss. Five constants minus no progress is +0.075 [0.039, 0.081] over all depths and +0.108 [0.064, 0.116] at interior depths; with evaluation covariates the full-curve gap is +0.065 [0.051, 0.083]. The five-constant model's work MAE is 4.59 [2.92, 6.31] actions and all ten point forecasts underestimate work. Original-state work changes from frozen source values 14.31/14.90 to evaluation values 19.75/18.88. The failure predictor's mean absolute error is 9.9 [7.4, 17.6] percentage points with sample covariates and 8.2 [4.3, 14.4] with evaluation covariates. These measurements quantify forecast errors in both coordinates, without isolating a unique cause of the failed decision test.

G Additional Information and Transfer Diagnostics

These audits were specified after inspecting the sequential-cohort results and before computing the diagnostics; they are post hoc, not confirmatory experiments.

G.1 What remains unidentified for a stronger comparator?

Fix a depth and suppress its subscript. Let $C = Y + I = 1 - D$, $a = \mathbb{E}[YW_H]$, and $b = \mathbb{E}[IW_H]$. A comparator with deployment success and total work, w_0, and matched original/continuation branches on $D = 1$ identifies the inherited term and total closed-request original work, $a + b = w_0 - \mathbb{E}[DW_H]$. The missing association is how that known total is divided between correct and incorrect closures. This information cannot change the already observed total cost or failure rate.

An elementary nonidentification example. Let $P(D) = 1/2$ and $P(Y) = P(I) = 1/4$. On takeovers, set $W_H = 6$ and $W_k = 4$; base work is two actions on every request. All original-state and takeover successor runs succeed. In world A, $W_H = 2$ on correct closures and 10 on incorrect closures; world B exchanges these assignments. Both worlds have $w_0 = 6$, $S = 2$, total policy work 4, failure $1/4$, inherited term -1, and closed-request original work 3. Their deployment distributions, full marginal W_H distribution, and matched takeover distributions agree. Yet $(a, b) = (0.5, 2.5)$ in A and $(2.5, 0.5)$ in B. Thus even those richer observations do not determine the allocation. This construction establishes neither empirical prevalence nor a better policy; W_H is original-state work, not the work of repairing an incorrectly closed state.

Sharp bounds for a realized table. Suppose the comparator knows the multiset of original workloads among the n_C closed requests, but its association with the m incorrect labels is erased. For N incoming requests, sort these workloads as $w_{(1)} \le \cdots \le w_{(n_C)}$. Its empirical contribution satisfies

$$\frac{1}{N}\sum_{j=1}^{m} w_{(j)} \;\le\; b \;\le\; \frac{1}{N}\sum_{j=n_C-m+1}^{n_C} w_{(j)}. \tag{11}$$

Every compatible labeling selects an m-element subset. Exchanging a selected value with a smaller omitted value cannot increase its sum, so the m smallest and largest values attain the two endpoints. This is the standard rearrangement argument. The bounds are for the observed table, not population confidence limits. They presume no retained covariates or strata constrain admissible assignments; such information can narrow them. A full empirical W_H marginal together with matched takeover workloads already determines the closed multiset by subtraction, so this is additional information only relative to the total-only comparator.

If original-state runs already cover these same requests and retain identifiers, joining them to the closure labels recovers the association by bookkeeping. Additional model calls are then unnecessary. Where original evaluation instead uses independent cohorts or omits branches on closed requests, the cost of collecting the missing association must be charged before claiming an acquisition advantage.

G.2 Complete audit of the additional closure allocation

The count-proportional comparator allocates the known closed-request total as

$$b_{\text{ex}} = \frac{P(I)}{P(C)}(a + b), \qquad \delta_I = b - b_{\text{ex}}, \qquad \delta_Y = -\delta_I. \tag{12}$$

We set both allocation and gap to zero when no closure is observed. The gap measures one degree of freedom: the association of correctness with original workload within closures. The weaker predictions $P(Y)w_0$ and $P(I)w_0$ also confound closed-versus-open selection; their errors are $\mathrm{Cov}(Y, W_H)$ and $\mathrm{Cov}(I, W_H)$.

Population and uncertainty. Table 24 retains all 32 primary nonzero-depth Luna cells and all eight analogous cells from the 40 disjoint prospective evaluation scenarios. The latter constitute post hoc replication: the allocation analysis was specified after their outcomes were observed. Cohorts are complete: 164 requests per τ-bench base, 203 per source AppWorld base, and 40 per prospective-evaluation base. We use 20,000 bootstrap draws (seed 2026092507), jointly over requests in τ-bench or scenarios in AppWorld, sharing weights across agents and depths and drawing source/evaluation AppWorld cohorts independently. We use the centered max-$|t|$ bands defined in Appendix F. The closure family contains δ_I, δ_I/w_0, and the two weaker-comparator covariances at every cell: 128 primary, 32 replication, or 160 combined coordinates. The table uses the *combined* family. Absent empirical closure classes and zero variance yield degenerate empirical gaps, not evidence of population equivalence. Small groups limit bootstrap accuracy.

Magnitude and limits. Median primary $|\delta_I|/w_0$ is 0.205%; the maximum is 3.571%. Three primary gaps exclude zero within the primary family, but only AppWorld gpt-oss at $k = 24$ survives the combined family: $\delta_I = 0.511\ [0.070, 0.952]$ actions per incoming request, or 3.571% $[0.497, 6.645]\%$ of w_0. Here 137 of 186 closures are incorrect (73.656%), accounting for 77.641% of the original work among closures; $b = 9.956$ versus $b_{\text{ex}} = 9.445$. Its multiset bounds are $[8.227, 10.424]$, a width

Table 24: **All 40 closure-allocation cells.** Luna successor; gaps are actions per incoming request. Intervals use the joint 160-coordinate family; bold marks its sole nonzero detection. Normalized gaps are point percentages. Empty empirical closure classes can give zero gaps without establishing population equality. Full component values and finite-table identification bounds remain in the artifact.

Base	Depth	Correct / wrong closures	δ_I [simultaneous 95%]	δ_I / w_0 (%)
τ-bench; $N = 164$				
Gemma	2	0 / 2	+0.000 [+0.000, +0.000]	+0.00
Gemma	4	3 / 3	+0.043 [-0.053, +0.138]	+0.50
Gemma	6	11 / 11	+0.107 [-0.010, +0.224]	+1.25
Gemma	8	41 / 22	+0.138 [-0.004, +0.279]	+1.61
Gemma	12	87 / 33	+0.141 [-0.077, +0.359]	+1.65
Gemma	16	98 / 34	+0.099 [-0.128, +0.326]	+1.16
gpt-oss	2	0 / 0	+0.000 [+0.000, +0.000]	+0.00
gpt-oss	4	1 / 1	-0.006 [-0.023, +0.011]	-0.07
gpt-oss	6	6 / 4	+0.005 [-0.064, +0.073]	+0.06
gpt-oss	8	21 / 11	-0.004 [-0.094, +0.086]	-0.05
gpt-oss	12	59 / 33	+0.063 [-0.132, +0.258]	+0.74
gpt-oss	16	76 / 42	+0.094 [-0.148, +0.336]	+1.11
Llama	2	0 / 0	+0.000 [+0.000, +0.000]	+0.00
Llama	4	1 / 1	+0.000 [+0.000, +0.000]	+0.00
Llama	6	2 / 2	-0.003 [-0.013, +0.007]	-0.04
Llama	8	2 / 3	+0.004 [-0.020, +0.027]	+0.04
Llama	12	2 / 6	+0.014 [-0.023, +0.050]	+0.16
Llama	16	2 / 9	+0.018 [-0.026, +0.063]	+0.21
Nemotron	2	0 / 0	+0.000 [+0.000, +0.000]	+0.00
Nemotron	4	0 / 0	+0.000 [+0.000, +0.000]	+0.00
Nemotron	6	4 / 0	+0.000 [+0.000, +0.000]	+0.00
Nemotron	8	14 / 2	+0.004 [-0.027, +0.035]	+0.04
Nemotron	12	60 / 16	+0.047 [-0.072, +0.167]	+0.55
Nemotron	16	91 / 23	+0.080 [-0.133, +0.293]	+0.93
AppWorld, original; $N = 203$				
Gemma	4	0 / 0	+0.000 [+0.000, +0.000]	+0.00
Gemma	8	15 / 3	+0.030 [-0.079, +0.138]	+0.20
Gemma	16	141 / 9	+0.121 [-0.108, +0.350]	+0.81
Gemma	24	176 / 20	+0.460 [-0.268, +1.188]	+3.09
gpt-oss	4	4 / 15	+0.084 [-0.046, +0.215]	+0.59
gpt-oss	8	28 / 67	+0.231 [-0.046, +0.508]	+1.61
gpt-oss	16	49 / 121	+0.390 [-0.007, +0.787]	+2.73
gpt-oss	24	49 / 137	**+0.511 [+0.070, +0.952]**	+3.57
AppWorld, disjoint tasks; $N = 40$				
Gemma	4	0 / 0	+0.000 [+0.000, +0.000]	+0.00
Gemma	8	3 / 0	+0.000 [+0.000, +0.000]	+0.00
Gemma	16	13 / 5	+0.868 [-0.255, +1.991]	+4.40
Gemma	24	21 / 10	+1.483 [-0.044, +3.010]	+7.51
gpt-oss	4	1 / 4	+0.105 [-0.196, +0.406]	+0.56
gpt-oss	8	2 / 12	+0.279 [-0.459, +1.016]	+1.48
gpt-oss	16	2 / 21	+0.380 [-0.563, +1.324]	+2.02
gpt-oss	24	2 / 29	+0.379 [-0.571, +1.329]	+2.01

of 2.197 actions or 15.353% of w_0. The range measures unresolved attribution before the association is observed; it is not a policy saving.

In replication, one of eight gaps excludes zero under its own family, but it is Gemma at $k = 24$: 1.483 $[0.289, 2.677]$ actions, with combined interval $[-0.044, 3.010]$. The corresponding gpt-oss gap is 0.379 $[-0.571, 1.329]$ under the combined family and has only two correct closures. The conditional wrong-minus-correct workload contrast is undefined in 2,529 of 20,000 draws for that cell. Against the weaker count-times-w_0 comparator, incorrect-closure covariance is detected in 0/32 primary cells and correct-closure covariance in 7/32 under the primary family; nondetection does not prove equality. These allocations neither establish transfer of a configuration-specific effect nor identify otherwise unmeasured failures or a useful repair decision.

Keeping survivor-selection denominators explicit. Let $u = \mathbb{E}[W_k \mid D = 1] - w_0$ and $\iota = \mathbb{E}[W_k - W_H \mid D = 1]$. Their per-incoming attribution difference is $p(u - \iota)$, with $p = P(D = 1)$. Table 25 jointly resamples the rate, conditional means, and w_0; it does not multiply conditional interval endpoints by a fixed rate. Its takeover family contains $(u, \iota, u - \iota)$ over the 32 primary cells (96 coordinates); its incoming family contains $(pu, p\iota, p(u - \iota), p(u - \iota)/w_0)$ over those cells (128 coordinates). Undefined ratio draws contribute infinite max statistics; wholly undefined point coordinates are excluded and reported. The 3.683 and 4.205 conditional gaps become 0.962 and 0.684 actions per incoming request. These are distinct from the extra closure-allocation gaps at the same cells, 0.121 and 0.390, and neither quantity is a policy saving.

Table 25: **Survivor-selection attribution at AppWorld depth 16.** Luna; simultaneous 95% bands in the separately defined takeover and incoming families. The normalized column is a percentage of w_0. Every rate and denominator is recomputed jointly in each draw.

Base	Takeovers / N	Gap / takeover	Gap / incoming	Gap / w_0 (%)
Gemma	53/203	3.683 [0.475, 6.892]	0.962 [0.094, 1.829]	6.454 [1.048, 11.859]
gpt-oss	33/203	4.205 [1.240, 7.170]	0.684 [0.060, 1.307]	4.777 [0.620, 8.933]

G.3 Diagnosing the frozen workload-transfer failure

Using all five depths of both agents, we diagnose the failed transfer in Section 5.4. For the raw frozen prediction $\hat{S}$, the exact residual decomposition is

$$S - \max(\hat{S}, 0) = \Delta w_0 - \Delta C_Y - \Delta C_I + \Delta P + \min(\hat{S}, 0), \tag{13}$$

where each Δ subtracts its frozen predicted component from its measured evaluation component: $C_Y = \mathbb{E}[YW_H]$, $C_I = \mathbb{E}[IW_H]$, and $P = \mathbb{E}[D(W_k - W_H)]$. Predicted components use either the 30 base-only sample tasks or, diagnostically, the evaluation tasks' own base covariates. Averaged over agents and depths, the sample-input signed components are $+4.707$ $[2.919, 6.570]$, -0.184 $[-0.848, 0.496]$, -0.004 $[-1.042, 1.051]$, $+0.128$ $[-0.390, 0.638]$, and -0.056 $[-0.307, 0]$ actions per incoming request, respectively (pointwise 90% intervals), summing to the 4.591-action underprediction. Components can cancel across depths: Gemma's inherited-progress discrepancy ranges from $+1.825$ at $k = 4$ to -1.753 at $k = 24$.

Target-oracle replacements. We retain all eight combinations of three parameter groups in Table 26: w_0, both closure ratios, and both progress slopes. Target w_0 is the evaluation immediate-handoff mean; closure ratios divide cap-closure conditional original-work means by that mean; slopes use the original request-level regression through the origin across positive depths. Refreshing w_0 preserves the selected ratios and therefore rescales both closure means; refreshing ratios uses target ratios at the selected w_0. Slopes retain successor-actions-per-base-action units. All predictions are clipped at zero. Target fits use evaluation outcomes, so these are apparent-fit diagnostics, not adaptation methods or forecasts.

The metric is the equal-agent mean of mean absolute depth error divided by that agent's evaluation w_0. We jointly bootstrap scenarios across agents, independently for sample/evaluation cohorts, with 4,000 draws (seed 20260925); target fits and denominators are recomputed, source quantities remain fixed. Intervals are 90% percentile intervals conditional on the frozen source estimates, without multiplicity

Table 26: **All target-oracle parameter replacements.** Work nMAE; lower is better. Sample columns retain the original 30-task base covariates; the last column also uses evaluation base covariates. Interior depths are 4, 8, and 16. Every refreshed quantity uses target evaluation outcomes; no row revises the prospective test.

Oracle target refresh	Sample: all	Sample: interior	Eval. covariates
None (frozen)	0.237	0.273	0.244
w_0	0.078	0.098	0.091
Closure ratios	0.200	0.235	0.206
Progress slopes	0.228	0.258	0.235
w_0 + closure	0.058	0.075	0.037
w_0 + progress	0.075	0.092	0.079
Closure + progress	0.192	0.219	0.195
All three	0.060	0.075	0.024

adjustment. The w_0-only refresh reduces nMAE to 0.078 $[0.060, 0.128]$, a paired reduction of 0.159 $[0.082, 0.194]$: 67.1% of the point error, with interval $[46.8, 71.9]\%$ for the fraction removed. Thus a majority reduction is not established throughout the interval. Progress-only refresh improves nMAE by 0.009 $[-0.006, 0.025]$. Adding it after baseline and closure refresh slightly worsens the point error, from 0.058 to 0.060. The apparent nMAE with all target parameters and evaluation covariates is 0.024 $[0.017, 0.039]$; this does not validate the restrictions out of sample. The gpt-oss correct-closure class disappears in 497/4,000 draws; the original estimator then uses its fallback ratio of one.

Giving simple baselines the same target information. A comparison specified after the eight-combination audit gives every model the same target endpoint mean, preserving its own source closure ratios and slopes. With the same target w_0, exchange reaches nMAE 0.060 $[0.046, 0.110]$ and no progress reaches 0.069 $[0.049, 0.121]$, compared with 0.078 $[0.060, 0.128]$ for five constants. Neither paired comparison establishes a five-constant advantage. Incomplete source-cohort endpoint/configuration/timing metadata prevents assigning that mismatch uniquely to harder tasks rather than model, configuration, or temporal changes. The parameter $\bar{z}$ affects failure prediction and J_q, not workload, and is absent from this work diagnostic.

The paired five-constant-minus-comparator differences are 0.018 $[-0.010, 0.036]$ for exchange and 0.009 $[-0.036, 0.048]$ for no progress (90% intervals). The refreshed flat model has error 0.350 $[0.313, 0.393]$, against 0.303 frozen. Full numerical comparisons remain in the artifact.

H Contemporaneous Interpreter-Interface Experiment

Frozen scope. Before the interleaved successor calls, we selected the Gemma depth-8 peak from the sequentially collected base cohort: 185 of 203 requests in 98 scenarios remained eligible. Seed 20260925 selects one eligible request uniformly per scenario and shuffles scenarios; two engineering scenarios leave 96 for evaluation, including the known replay-failure scenario. The population is the reconstructed Gemma survivor cohort; base trajectories were not re-executed.

Paired intervention. Each task has original-state, fresh-interpreter and preserved-interpreter Luna branches. The latter two share code/output history and application state but receive truthful variable-availability instructions: an interpreter–instruction bundle. Caps are 40 actions and 1,500 output tokens; sampling is provider-default without a requested seed. Six initial arm orders occur 16 times each. Each round rotates the order and calls each active arm once, including the contemporaneous original branch. Four task blocks run concurrently.

State and provenance checks. Independent processes replay the eight recorded base actions, requiring exact displayed-output agreement, then build preserved and fresh branches. Before any model call, full logical SQLite schemas and sorted table contents must match the reconstructed checkpoint; the fresh interpreter must have no predecessor variables. Python random states and interaction counters also match. AppWorld is 0.1.3.post1 with data 0.1.0. These checks authenticate reconstruction. The source snapshots and transitive dependency versions are unavailable. Actual UTC

times, request/response IDs, models, parameters, usage, payload hashes and sanitized raw responses are retained outside the benchmark clock. A changed reported model halts collection; a stable alias does not certify immutable weights.

Amendments and failures. Two excluded engineering triplets completed (70 calls) before the four-worker amendment and parallel preflight; cohort, arm orders and settings stayed fixed. HTTP 500 errors later caused global stops and interrupted peers. Dated amendments retained all partial attempts/costs and permitted one whole-triplet infrastructure restart per affected task. Unaffected blocks were retained; no task was replaced or restart selected by outcome. Replay/state failures are not retried. The gate includes full-text-search shadow tables: a mismatch need not mean different user-visible contents. Infrastructure-conditioned restarts do not establish uninterrupted operation.

Estimands and intervals. Primary contrasts are preserved minus fresh actions and failure; secondary contrasts are $8 + W_i^{(a)} - W_i^{(0)}$, $a \in \{\text{fresh}, \text{preserved}\}$, on the same survivors. Every generation/code attempt counts, including malformed responses. Failure means no declared, evaluator-correct completion by the cap. Work receives pointwise 95% paired Student-t and 20,000-draw scenario-bootstrap intervals; incomplete-cohort estimates remain provisional.

For failure, the protocol amendment adopted before evaluation calls forms 97.5% Clopper–Pearson intervals $[L_H, U_H]$ and $[L_B, U_B]$ for preserved-only and fresh-only failure probabilities, then reports $[L_H - U_B, U_H - L_B]$. Joint coverage is at least 95% under binomial sampling margins; this exact guarantee is not asserted for arbitrary heterogeneous fixed tasks. The additional Hoeffding interval has half-width $\sqrt{2\log(40)/n}$ for independent paired outcomes in $[-1, 1]$. Nonsignificance and zero discordances do not imply equivalence. Protocols and all attempts remain in the artifact.

Terminal cohort and results. All 96 planned scenarios reached a terminal record: 93 valid triplets, one output-replay failure (`c77c005_1`, actions 5 and 7), and two fresh-state validation failures (`9dabbc9_1`, `ccf4b82_1`). For the latter two, only Venmo full-text-search shadow-table hashes differ; every other table and the preserved branch match. The frozen gate is not relaxed. None of these three tasks received a model call or replacement. Thirteen interrupted triplets were restarted once. Formal calls span 2026-09-25 16:40:18–18:04:47 UTC. All 3,674 successful responses, including engineering and restart attempts, report `gpt-5.6-luna`; two additional attempts returned HTTP 500.

Table 27: **Contemporaneous interface results.** Complete cases only, $n = 93$ of 96 planned scenarios; Luna on the reconstructed Gemma depth-8 survivor cohort. Action intervals are pointwise 95%; paired contrasts share tasks, and bootstrap draws resample scenarios. Failure counts are shown without claiming equivalence.

Successor start	Actions [95% t CI]	Bootstrap CI	Failures
Original state	14.71 [13.77, 15.65]	[13.82, 15.68]	17/93
Fresh interpreter	12.82 [11.72, 13.91]	[11.77, 13.90]	9/93
Preserved interpreter	9.38 [8.21, 10.55]	[8.24, 10.53]	11/93
Paired work contrasts (actions per selected scenario)			
Preserved − fresh	-3.44 [−4.08, −2.80]	[−4.05, −2.81]	—
Fresh +8− original	+6.11 [5.30, 6.91]	[5.32, 6.90]	—
Preserved +8− original	+2.67 [1.74, 3.59]	[1.77, 3.58]	—

Preserved minus fresh failure is +2.15 percentage points: two preserved-only failures and zero fresh-only failures. Its conservative interval is $[-4.42, 8.45]$; the Hoeffding interval is $[-26.02, 30.32]$. A post hoc descriptive partition assigns -3.25 of the -3.44 action difference per complete scenario to the 82 pairs where both interfaces succeed; the nine jointly failed pairs contribute -0.13, and the two discordant pairs -0.06. Conditioning on these outcomes selects a subset and does not identify a quality-preserving deployment effect.

Excluding all 13 restarted scenarios leaves 80 cases and a paired work difference -3.52 $[-4.21, -2.84]$. Allowing every missing arm any work in $[0, 40]$ gives a *finite-planned-cohort* mean range $[-4.58, -2.08]$; allowing any missing failure in $[0, 1]$ gives $[-1.04, 5.21]$ percentage points. These post hoc ranges concern unobserved realized outcomes, not population confidence, and do not restore

full-cohort confirmation. Both total-work contrasts have positive finite-cohort ranges: $[4.92, 7.42]$ fresh and $[1.58, 4.08]$ preserved.

I Interleaved All-Request Execution and Frozen Allocation Test

Population and execution. All installed AppWorld families were represented in the source records. We enrolled one unused, held-out variant from each of the 126 eligible known families (56 normal, 70 challenge), shuffling execution with seed 2026092519. Two exposed engineering tasks were excluded. Gemma and gpt-oss bases have a 24-action cap; one shared Luna original-state branch runs on every incoming request, including closures. Fresh-interpreter Luna branches receive open checkpoints at 8, 16 and 24 actions and native transfers. Successors have a 40-action cap; all routes have a 1,500-token output cap. Bases request temperature zero; Luna uses provider defaults without a requested seed. Successful completion requires declared success and the official evaluator's correct-state verdict. Every generation counts, including malformed code. Native give-up is charged; its successor inherits the pre-action state and ledger. That terminal draw serves budgets at or beyond the charged boundary; earlier budgets use their own checkpoints.

Contemporaneous state controls. Within each task, worlds start with matching application states, counters and Python random states. A frozen randomized role order rotates each round: one action per active branch, new successors joining next round, up to four concurrent tasks. Successors inherit snapshots from the interleaved base runs, without replaying recorded trajectories. Continuation requires matching standard database contents, schema, sequence state, FTS rows/configuration, fresh interpreters and preserved counters/random states. SQLite integrity and FTS external-content consistency are checked; only physical FTS shadow storage is excluded, and contentless FTS is rejected. This gate does not alter the reconstructed-survivor interface study's missingness. Actual UTC times, requested/reported identifiers, raw sanitized responses, usage and request hashes are retained outside the benchmark clock. Identifier changes stop collection; stable identifiers do not certify immutable weights.

Frozen source comparison. Gemma at $k = 24$ is primary, chosen from the source-cohort results. Its incorrect/correct conditional original-work ratio is frozen at $\kappa = 1673/1060$ from the 40-task prospective source cohort (21 correct, 10 incorrect closures), also used at other Gemma depths. gpt-oss uses its 203-task source ratio, 1.241992; a Gemma sensitivity uses 1.369125 from its source cohort. No target refit or source reselection is allowed. Let $W = W_H$, $a = \mathbb{E}[IW]$, $t = \mathbb{E}[(I+Y)W]$, $i = \mathbb{E}[I]$, and $y = \mathbb{E}[Y]$. Both predictors receive the same target (t, i, y); only the source predictor receives the frozen association:

$$F = a, \qquad F_0 = t\frac{i}{i+y}, \qquad F_\kappa = t\frac{\kappa i}{\kappa i + y}, \qquad M = F - \frac{F_0 + F_\kappa}{2}.$$

For $\kappa > 1$, $i, y, t > 0$, put $d = F_\kappa - F_0 > 0$. The absolute-error improvement satisfies

$$G = |F - F_0| - |F - F_\kappa| = \mathrm{clip}(2M, -d, d).$$

Place F below, between, or above the predictions to obtain this identity. Hence $G > 0$ iff $M > 0$: we test smooth M and report G as effect size. Absent closure classes or $t = 0$ are uninformative; retain coincident allocations and set no-closure allocations to zero. W is capped attempt work, not repair effort. We report original cap/failure counts by closure group, descriptive within-stratum proportional/fixed-κ predictions, and their sums under frozen mixture weights.

Transfer boundary. For the aggregate comparator, set $q_c = i/(i+y)$ and $\rho = \mathbb{E}[W \mid I]/\mathbb{E}[W \mid Y]$ on the same allocation-valid distribution. When $0 < q_c < 1$, $t > 0$ and both conditional means are positive, $F/t = \rho q_c/[1 + (\rho - 1)q_c]$, so

$$\frac{M}{t} = \frac{q_c(1-q_c)\{(\rho-1)[2+(\kappa-1)q_c] - (\kappa-1)\}}{2[1+(\rho-1)q_c][1+(\kappa-1)q_c]}.$$

The positive denominator gives Equation 5, restating the frozen M test. With missing tuples it concerns the fixed-weight observed distribution, not the full cohort; required empty strata are undefined. Separately pooled stratum predictors have a different boundary. For this correction, $|G| \le F_\kappa - F_0 \le t(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$: the second bound maximizes $(\kappa-1)q_c(1-q_c)/[1+(\kappa-1)q_c]$ at $q_c = 1/(1+\sqrt{\kappa})$. This bounds the predictor's error change, not all paired information or policy value.

Inference and missing records. The finite 126-task table is descriptive. Population intervals assume independent, exchangeable scenario executions within strata, conditional on configuration and fixed 56/126 and 70/126 weights; shared provider disturbances can violate independence. The 20,000 stratified bootstrap draws (seed 2026092521) retain each task's two bases, shared original and depths together, holding source ratios fixed. Gemma/24 is the sole primary setting: support requires a positive lower endpoint of the two-sided 95% percentile M interval. An interval below zero favors proportional allocation; crossing zero is unresolved. G measures allocation-error improvement in actions per incoming request; the delta-method interval is only a check. Secondary centered max-$|t|$ bands with fixed bootstrap standard errors cover six M and six $F - F_0$ coordinates. A separate 36-coordinate family covers six cells' total-work change, $\mathbb{E}[D(W_k - W_H)]$, correct/incorrect closure work, failure change and correct automation. A pointwise 95% failure interval is $\widehat{\Delta}_Z \pm \sqrt{2\log(40)\sum_s \omega_s^2/n_s}$, clipped to $[-1, 1]$, where n_s counts observed paired quality outcomes and $\omega_s \in \{56/126, 70/126\}$. The Hoeffding bound requires independent bounded scenario differences; nonsignificance is not equivalence.

Allocation requires known base status at that depth and terminal original work; decomposition uses a common valid mask, quality its own paired mask. Original evaluator failure retains terminal work. Closed-base evaluator failure makes I, Y unknown from closure onward; native/cap handoff status is known independently of that evaluator. Later base failure preserves verified earlier checkpoints. Complete-case intervals condition on observed tuples, without recovering the full cohort. Draws recompute observed stratum denominators with fixed weights. Fewer than two observations in a required stratum disables regular inference; zero-observation draws make the primary interval unbounded and family maximum infinite. Absent closure classes retain the allocation convention and count as boundary draws. For q observed allocation tuples, let $m = 126 - q$ and $\mathbf{v} = (A, T, J_I, J_Y)$ sum their (IW, CW, I, Y), where $C = I + Y$. Outer moment endpoints are $\ell = \mathbf{v}/126$ and $u = [\mathbf{v} + m(40, 40, 1, 1)]/126$. Writing $h(i, y) = \frac{1}{2}[i/(i + y) + \kappa i/(\kappa i + y)]$, the finite-cohort completion range is

$$M \in [a_L - t_U h(i_U, y_L),\ a_U - t_L h(i_L, y_U)].$$

Here h increases in i and decreases in y; zero-denominator corners allow $h \in [0, 1]$. These conservative ranges retain denominator 126 and are not confidence intervals. Other completion bounds use action caps and the binary failure range.

Inherited-only infrastructure/state failures leave that branch missing; bases and original continue. Base/original infrastructure failure permits one entire-task restart: retain the earlier attempt and select the final attempt wholly. Evaluator, semantic and outcome failures do not authorize restarts. No tasks are replaced; no-restart results are descriptive sensitivities. Money/time limits are fixed, with no interim scientific analysis. All attempts, masks, coefficient sums, technical amendments and the full frozen protocol remain in the artifact.

Audited execution. All 126 formal blocks terminated without restart. The raw audit reconciled 11,149 requests, including 136 engineering requests, with one failed request; $18.9789648 was accounted, with an additional $0.0225572 allowance for uncertain billing. Formal calls began at 19:09:44 UTC and the final block terminated at 23:29:17 UTC on September 25, 2026. Requested/reported identifiers and state checks reconciled. These results concern the configurations served during collection and held-out variants of known families; identifiers do not establish immutable weights.

Observed allocation and scope. All six allocation cells retain 126 requests. The primary margin is $M = 0.0194$ (pointwise 95% CI $[-0.4214, 0.4800]$), with $G = 0.0388$ actions/request (0.22% of w_0): the frozen primary is unresolved. The 12-coordinate simultaneous family supports positive M only for gpt-oss at 16 and 24; five of six $F - F_0$ bands are positive. These allocation advantages do not establish work savings or quality equivalence: both supported cells have higher failure and total-work bands crossing zero. Pooling separately computed stratum predictors gives primary $G = -0.1452$; the source-cohort Gemma sensitivity coefficient gives $G = 0.5291$. Both are descriptive checks, with no replacement of the frozen aggregate test or source coefficient.

Realized missingness. Only the gpt-oss/8 continuation for `6b6ca61_2` failed; its 25 partial actions are excluded from terminal work. Its base class and original work remain known, preserving allocation. Common-work and paired-quality masks contain 125 requests (55 normal, 70 challenge);

Table 28: All-request allocation levels and contrasts. n counts observed tuples out of 126; $Y/I/D$ are unweighted correct-closure, incorrect-closure and takeover counts. Means use the fixed normal/challenge mixture. Gray F_0 and teal F_κ headers identify predictors sharing target aggregates. The pale-blue row marks the prespecified Gemma/24 setting. Panel (b) brackets are joint 95% bands over 12 allocation coordinates; the primary pointwise interval appears below.

(a) Allocation levels

Base/k	n/N	$Y/I/D$	F	F_0	F_κ	G
Gemma/8	126/126	10/4/112	0.43	0.39	0.53	-0.07
Gemma/16	126/126	53/22/51	3.07	2.64	3.56	-0.05
Gemma/24	126/126	72/35/19	5.42	4.64	6.16	0.04
gpt-oss/8	126/126	8/33/85	4.28	3.98	4.14	0.16
gpt-oss/16	126/126	15/64/47	8.66	8.15	8.47	0.31
gpt-oss/24	126/126	19/82/25	11.46	10.86	11.28	0.41

(b) Allocation contrasts

Base/k	n/N	M [joint 95%]	δ [joint 95%]	$100G/w_0$
		-0.04	$+0.03$	
Gemma/8	126/126	$[-0.18, 0.10]$	$[-0.11, 0.18]$	-0.41
		-0.02	$+0.44$	
Gemma/16	126/126	$[-0.41, 0.36]$	$[0.02, 0.85]$	-0.28
		$+0.02$	$+0.78$	
Gemma/24	126/126	$[-0.59, 0.63]$	$[0.13, 1.42]$	0.22
		$+0.22$	$+0.30$	
gpt-oss/8	126/126	$[-0.02, 0.46]$	$[0.01, 0.58]$	0.89
		$+0.35$	$+0.51$	
gpt-oss/16	126/126	$[0.06, 0.64]$	$[0.16, 0.85]$	1.78
		$+0.39$	$+0.60$	
gpt-oss/24	126/126	$[0.05, 0.73]$	$[0.20, 1.00]$	2.35

Primary Gemma/24 pointwise 95% CI for M: $[-0.42, 0.48]$.

the other five cells retain 126. Granular closure work retains 126 and must not be mixed into the 125-request common-mask decomposition. Finite-cohort completion ranges are $[-0.3095, 0.5159]$ for total-work change, $[-2.5397, -1.9048]$ for inherited-work change, and $[0.1746, 0.1905]$ for failure change. These are bounded missing-outcome ranges, not CIs; complete allocation makes its completion ranges point values, not population guarantees. Tables 28–29 retain all cells, masks and inferential families.

Table 29: All-request work and quality. Each work row shares one observation mask and count; B, w_0, S and $B + S$ are mean base, original, successor and total actions. Work contrasts and quality/automation bands form one joint 95% family of 36 coordinates. Quality and automation retain separate counts. Failure differences and automation are percentage points/percent. The separate Hoeffding interval is pointwise 95% under independent bounded scenario differences. Nonsignificance does not imply equivalence. Means condition on observed components; completion bounds are distinct from CIs.

(a) Common-mask work means

Base/k	n/N	B	w_0	S	$B+S$
Gemma/8	126/126	7.90	17.61	14.36	22.25
Gemma/16	126/126	13.44	17.61	5.99	19.43
Gemma/24	126/126	15.71	17.61	2.13	17.84
gpt-oss/8	125/126	7.23	17.56	10.32	17.55
gpt-oss/16	126/126	11.23	17.61	5.62	16.85
gpt-oss/24	126/126	13.06	17.61	3.06	16.13

(b) Work decomposition [joint 95%]

Base/k	n/N	Total $-w_0$	Inherited	Correct closure	Incorrect closure
Gemma/8	126/126	$+4.64$ $[2.89, 6.40]$	-1.87 $[-3.30, -0.45]$	0.95 $[0.02, 1.88]$	0.43 $[-0.28, 1.14]$
Gemma/16	126/126	$+1.82$ $[-0.59, 4.22]$	-2.63 $[-4.23, -1.04]$	5.91 $[3.89, 7.93]$	3.07 $[1.13, 5.01]$
Gemma/24	126/126	$+0.23$ $[-2.24, 2.70]$	-1.29 $[-2.61, 0.02]$	8.77 $[6.46, 11.08]$	5.42 $[2.84, 8.00]$
gpt-oss/8	125/126	-0.01 $[-2.52, 2.50]$	-2.24 $[-3.58, -0.91]$	0.68 $[-0.05, 1.40]$	4.32 $[2.06, 6.58]$
gpt-oss/16	126/126	-0.76 $[-3.81, 2.28]$	-1.93 $[-3.32, -0.53]$	1.40 $[0.33, 2.48]$	8.66 $[5.97, 11.34]$
gpt-oss/24	126/126	-1.48 $[-4.65, 1.68]$	-1.17 $[-2.33, 0.00]$	1.92 $[0.61, 3.23]$	11.46 $[8.72, 14.21]$

(c) Quality and correct automation

Base/k	n_Z/N	ΔZ [joint 95%]	Hoeffding 95%	n_Y/N	$100Y$ [joint 95%]
Gemma/8	126/126	-2.38 $[-13.17, 8.41]$	$[-26.58, 21.82]$	126/126	7.94 $[0.36, 15.51]$
Gemma/16	126/126	$+7.14$ $[-5.09, 19.38]$	$[-17.06, 31.34]$	126/126	42.06 $[28.62, 55.51]$
Gemma/24	126/126	$+11.11$ $[-2.17, 24.40]$	$[-13.09, 35.31]$	126/126	57.14 $[43.63, 70.66]$
gpt-oss/8	125/126	$+18.53$ $[3.46, 33.60]$	$[-5.77, 42.82]$	126/126	6.35 $[-0.40, 13.10]$
gpt-oss/16	126/126	$+42.86$ $[28.01, 57.71]$	$[18.66, 67.06]$	126/126	11.90 $[2.95, 20.86]$
gpt-oss/24	126/126	$+50.00$ $[35.09, 64.91]$	$[25.80, 74.20]$	126/126	15.08 $[5.10, 25.06]$

J Stratum-Matched Allocation and Original-Outcome Sensitivity

This post hoc analysis was specified after observing aggregate outcomes; it does not replace the frozen primary test. All six allocation cells retain 126 requests. Both predictors receive identical stratum information. Write $\mathbb{E}_s$ for expectation within normal/challenge stratum s, $W = W_H$, $C = I + Y$, and $\omega_s \in \{56/126, 70/126\}$. Define

$$(F_s, t_s, i_s, y_s) = \mathbb{E}_s[(IW, CW, I, Y)], \quad P_{0s} = t_s \frac{i_s}{i_s + y_s}, \quad P_{\kappa s} = t_s \frac{\kappa i_s}{\kappa i_s + y_s},$$

using the frozen model-level coefficients in Appendix I. Then

$$F = \sum_s \omega_s F_s, \qquad P_0^{\mathrm{str}} = \sum_s \omega_s P_{0s}, \qquad P_\kappa^{\mathrm{str}} = \sum_s \omega_s P_{\kappa s},$$

$$G_{\mathrm{str}} = |F - P_0^{\mathrm{str}}| - |F - P_\kappa^{\mathrm{str}}|,$$
$$H = \sum_s \omega_s \big(|F_s - P_{0s}| - |F_s - P_{\kappa s}|\big).$$

Their point estimates coincide in all six cells; equality is not general. Let $\delta_{\mathrm{str}} = F - P_0^{\mathrm{str}}$, $M_{\mathrm{str}} = F - (P_0^{\mathrm{str}} + P_\kappa^{\mathrm{str}})/2$ and $d_{\mathrm{str}} = P_\kappa^{\mathrm{str}} - P_0^{\mathrm{str}}$.

The 20,000 stratified bootstrap draws (seed 2026092601) share task weights across all cells, holding mixture weights and coefficients fixed. Centered max-$|t|$ bands, using fixed bootstrap standard errors, cover 18 coordinates: six each of $M_{\mathrm{str}}, d_{\mathrm{str}}, \delta_{\mathrm{str}}$. Intersecting rectangles with $d_{\mathrm{str}} \geq 0$ and mapping through $G_{\mathrm{str}} = \mathrm{clip}(2M_{\mathrm{str}}, -d_{\mathrm{str}}, d_{\mathrm{str}})$ gives conservative bands relative to those rectangles, with approximate population coverage under independent, exchangeable within-stratum executions. Source uncertainty is excluded. Absent closure classes retain coincident predictions; no closures give zero allocations. Sparse classes limit bootstrap accuracy. H percentile intervals are descriptive and pointwise because absolute errors are nonsmooth.

For disjoint original-outcome groups g (success, noncapped failure, capped failure), set

$$A_{sg} = \mathbb{E}_s[IW\mathbf{1}_g], \quad T_{sg} = \mathbb{E}_s[CW\mathbf{1}_g], \quad q_s = \frac{i_s}{i_s + y_s}, \quad q_{sg} = \frac{\mathbb{E}_s[I\mathbf{1}_g]}{\mathbb{E}_s[C\mathbf{1}_g]},$$

with zero-denominator ratios zero. All moments retain the stratum denominator. Exactly,

$$\delta_{\mathrm{str}} = \underbrace{\sum_{s,g} \omega_s (A_{sg} - q_{sg} T_{sg})}_{\text{within-outcome residual}} + \underbrace{\sum_{s,g} \omega_s (q_{sg} - q_s) T_{sg}}_{\text{outcome composition}}. \tag{14}$$

For each base class $Y/I/D$, columns n, fail and cap give request, original-failure and cap counts; caps are included in failures.

The gpt-oss/16 and /24 error improvements retain positive joint lower bounds with matched strata. Gemma/24 descriptively overcorrects and remains unresolved. For gpt-oss/24, $\delta_{\mathrm{str}} = 0.5867$, with within/composition terms 0.4785 and 0.1083; the cap contributes 0.0560. The sole capped original is in D at every Gemma depth. These finite-table residues exclude complete explanation by the recorded coarse categories, without identifying latent difficulty or a causal mechanism. The success-only cohort contains 98 requests (48 normal, 50 challenge); conditioning on this post-treatment outcome changes the estimand, can induce collider bias, and supplies no deployable covariate. Empty required strata remain undefined without weight renormalization.

Table 30: Stratum-aware allocation predictions, using all 126 requests and the fixed 56/70 normal/challenge mixture. Gray P_0^{str} and teal P_κ^{str} use the same stratum information; the source coefficient is unchanged. Brackets for δ_{str} are joint 95% bands from the exploratory 18-coordinate family; those for G_{str} map the joint $M_{\mathrm{str}}, d_{\mathrm{str}}$ rectangle. F, predictors and contrasts are actions per incoming request. The mean within-stratum error reduction H equals pooled G_{str} at these observed points; the two estimands need not agree in general. All cells are shown.

Base/k	F	P_0^{str}	P_κ^{str}	δ_{str} [joint 95%]	G_{str} [derived 95%]
Gemma/8	0.429	0.395	0.534	0.034 $[-0.115, 0.183]$	-0.072 $[-0.281, 0.213]$
Gemma/16	3.071	2.681	3.573	0.390 $[-0.011, 0.791]$	-0.112 $[-0.876, 0.652]$
Gemma/24	5.421	4.753	6.234	0.668 $[0.048, 1.288]$	-0.145 $[-1.345, 1.054]$
gpt-oss/8	4.278	4.008	4.159	0.270 $[-0.021, 0.561]$	0.151 $[-0.098, 0.267]$
gpt-oss/16	8.659	8.198	8.501	0.461 $[0.103, 0.818]$	0.303 $[0.029, 0.469]$
gpt-oss/24	11.460	10.874	11.285	0.587 $[0.170, 1.003]$	0.412 $[0.057, 0.608]$

Table 31: Original-successor outcomes and allocation-gap components for all six cells. (a) Blue Y, rose I and gray D identify correct closures, incorrect closures and takeovers. $\bar{W}$ is original work; $\bar{W}_f$ restricts its mean to failures (— when none). Counts are $n/f/c$: requests, original failures and 40-action caps; caps are included among failures, so these counts overlap. Means pool requests across both task strata. (b) Success, noncapped failure and capped failure are disjoint original outcomes; their contributions sum to δ_{str}. The alternative within-outcome association and outcome-composition components also sum to this gap before rounding. Components are actions per incoming request. This post hoc decomposition is descriptive: conditioning on realized original outcomes neither identifies causal effects nor validates a deployment policy.

(a) Original work and outcome counts by base status

	Y			I			D		
Base/k	$\bar{W}$	$\bar{W}_f$	$n/f/c$	$\bar{W}$	$\bar{W}_f$	$n/f/c$	$\bar{W}$	$\bar{W}_f$	$n/f/c$
Gemma/8	12.00	15.00	10/1/0	13.50	14.00	4/1/0	18.26	22.65	112/26/1
Gemma/16	14.06	15.00	53/1/0	17.59	18.75	22/8/0	21.31	23.84	51/19/1
Gemma/24	15.35	20.33	72/3/0	19.51	20.40	35/15/0	22.68	25.10	19/10/1
gpt-oss/8	10.50	—	8/0/0	16.33	28.40	33/5/1	18.78	20.70	85/23/0
gpt-oss/16	11.80	15.00	15/1/0	17.05	24.82	64/11/1	20.23	20.62	47/16/0
gpt-oss/24	12.74	15.00	19/2/0	17.61	22.50	82/18/1	21.32	22.88	25/8/0

(b) Two decompositions of the stratum-aware gap

Base/k	Success	Noncapped failure	Capped failure	Within-outcome association	Outcome composition	δ_{str}
Gemma/8	-0.0113	0.0454	0.0000	0.0221	0.0119	0.0340
Gemma/16	-0.3289	0.7190	0.0000	0.2710	0.1190	0.3901
Gemma/24	-0.6736	1.3415	0.0000	0.4225	0.2454	0.6679
gpt-oss/8	0.0843	0.1459	0.0397	0.1755	0.0945	0.2699
gpt-oss/16	0.2123	0.2041	0.0441	0.3814	0.0792	0.4605
gpt-oss/24	0.1947	0.3360	0.0560	0.4785	0.1083	0.5867