

Separate Records, Shared Processing

An Experiment in AI-Assisted Investment State Maintenance

John Komkov¹

Working paper October 2026

Abstract

Shared processing can reuse evidence while preserving separate investor records. A paired experiment compares isolated and shared interpretation of 64 synthetic investment histories under retrieval alone and persistent records with validation. Records reduce repeated retrieval. Successful primary contexts reach the measured semantic-extraction ceiling; most exact-output improvement reflects normalization of abstention. Shared processing has lower recorded inference costs, but the tested improvement saves more separately. Positive primary model-implied regret occurs only after service failures. The study measures interpretation-cost incidence and operational consequences, with complete archived receipts, rather than the total cost or economic value of investment organization.

Keywords: artificial intelligence; investment records; retrieval; validation; cost incidence.

JEL: G11, G23, C72, D47, D85.

¹John Komkov has commercial interests in Megalith and Argosy. AI tools assisted with research, drafting, code and analysis.

1 Introduction

An investment record is useful only if it preserves the distinctions on which an action depends. Cash is not deployable capital when it is needed for an accepted call. A stated preference is not authority to transact. A superseded term is not a current condition. The same facts can be maintained separately for each investor or processed together with shared evidence. This experiment compares the costs and errors of those organizational modes.

The comparison matters because technological improvement can favor either arrangement. A common source can be interpreted once and reused across investors; persistent individual records can also remove repeated work from separate processing. The relevant contrast is the change in cost for equivalent work, rather than the mere fact that a shared call is cheaper than several isolated calls. The experiment therefore pairs both organizational modes with the same primary technology comparison.

The findings are strongest on repeated work and narrowest on semantic quality. Records reduce retrieval. All successful primary contexts have the scored state fields correct. Validation improves literal output matching mainly by converting an amount encoded as null under review to zero under the same review status. A post-hoc representation-equivalence check removes that apparent advantage. All positive primary downstream regret comes from service failure. These results describe a bounded processing system with frozen histories; they do not establish that AI has improved actual portfolio returns or reduced the cost of a complete investment organization.

2 Design and measurement

2.1 Matched organizations and a frozen experiment

The technological question is whether the same investor becomes cheaper to serve separately or within a shared arrangement that preserves his individual state. The completed experiment compares those modes using identical economic histories. It measures interpretation and record maintenance, one component of the organizational cost in the participation condition. It does not measure investment forecasting or the full cost of a legal vehicle.

There are 64 held-out economic worlds, each with four investors and four chronological checkpoints. Separate processing isolates each investor’s context and sums its costs. Shared processing reuses common offering evidence while maintaining separate states and decisions. Both perform equivalent required work. The primary comparison is retrieval without persistent semantic records or deterministic validation (R0V0) against the complete workflow (R1V1). A full factorial on 16 predetermined worlds separates records and validation. Four predetermined worlds are repeated. The evaluation contains 1,344 synthetic book-processing episodes: 1,024 primary, 256 factorial-completion and 64 repeat episodes. Twelve public-document cases add 48 episodes.

An episode can require several inference calls. The archived synthetic evaluation has 5,293 calls; the public slice has 72, including charged record preparation. Development is separate.

Model-produced records incur construction, update and retrieval costs; they are not corrected by an oracle. Each checkpoint supplies only then-available evidence. Synthetic truth comes from executable state transitions and deterministic documents. The public cases are an exploratory, previously used slice with source-linked, agent-audited labels. They are not newly held-out or human-validated cases.

The source catalog and authenticated permission registry are common trusted inputs. Validation tests provenance scope, dates, grant expiry, remaining authority, expenses and contractual lots. Documents cannot create permission by asserting it. Validation does not establish that every plausible cited amount was extracted correctly. Acquisition and authentication of the trusted inputs are unmeasured. Human review prices remain assumptions because no expert review time was observed.

Economic complexity counts distinct, source-linked conditions whose perturbation changes feasible participation. Presentation difficulty and document length are separate. Frozen tiers contain 38 low-, three middle- and 23 high-complexity worlds. The economic world is the statistical unit; its investors and checkpoints are repeated observations. Paired world bootstrap intervals condition on this simulator and its few presentation families. The three-world middle tier is descriptive.

2.2 What changes between configurations

The records treatment changes persistence, not access to perfect state. In R0, the model emits a current state but that state is discarded before the next checkpoint. In R1, the next checkpoint receives the immediately preceding model-produced record, including source identifiers and conditions. Both configurations receive current documents, an archive index, supplied authenticated grants and offering terms. Both have the same allowance for one historical-evidence retrieval. The cost of creating and updating the record appears in model output; construction is not supplied free by the simulator. A failed context leaves no usable new record, and the evaluator does not repair it from ground truth.

The validation treatment operates after interpretation. It checks source existence and permitted provenance, knowledge and effective dates, principal identity, available authority, expiry, expenses and contractual lot sizes. It can prevent an incompatible proposal from proceeding through the simulated action interface. It cannot prove that a cited passage entails an extracted number merely because that passage exists. Correct provenance and correct interpretation are distinct requirements. This distinction matters when reporting the measured semantic ceiling: the absence of a scored error here is an observation about the test, not a property guaranteed by validation.

Shared mode receives one four-investor context at a checkpoint. Separate mode receives four isolated contexts, each performing its investor's required work. Common offering evidence is repeated when isolation requires it. Investor identity, constraints and authority remain separate in either mode. A conventional administrator serving several separate accounts could use shared

processing; the experiment does not equate computational batching with common ownership or a new legal institution.

Table 1: Evaluation schedule and units

Component	Worlds or cases	Episodes
Primary paired comparison	64	1,024
Factorial completion	16	256
Predetermined repeats	4	64
Synthetic evaluation total	64 unique	1,344
Public-document slice	12 cases	48
Development, three recorded versions	8 worlds	128

A synthetic episode processes four investors at one checkpoint. A public episode is one case/configuration pair. Separate mode uses four processing contexts and shared mode one. The synthetic evaluation contains 3,360 contexts and 5,293 recorded calls, including 1,933 historical-evidence followups. Counts of calls, contexts and repeated decisions are not counts of independent economic worlds. Development results are not pooled across versions.

2.3 Chronology, difficulty and the observation unit

Each world supplies an exogenous sequence of four checkpoints. An accepted call is partly paid; a distribution settles; a competing reservation appears; amendments become effective; grants are revoked or expire; new observations may contradict old ones. The simulated investment book updates by explicit state transitions. Neither a model recommendation nor a validator changes later simulator facts. This is a study of maintenance along fixed histories, rather than a closed-loop experiment in which an investment policy changes the world it subsequently observes.

Every arm sees the same economic history at the same information boundary. A future-effective amendment may be visible without yet changing a current decision. An accepted obligation remains on the book when an investor changes a preference. An untrusted source can describe a permission but cannot add it to the authenticated registry. These rules allow the scoring system to distinguish missing evidence, an expired grant, a cash constraint and a genuine authorization. They also make abstention meaningful: review is the correct response to some histories, rather than an error to minimize indiscriminately.

The complexity measure counts consequential conditions, not prose length. A bounded change to each unique investor-condition pair is tested for its ability to change feasible or authorized lot participation at a checkpoint. Conditions appearing repeatedly count once. The resulting world counts span 40–48. Desired investment amount is excluded because it is a utility kink, rather than a feasibility constraint. Split accounts, duplicated facts and hostile instructions are separately recorded presentation features. These limited ranges and only three presentation families restrict what the experiment can establish about complexity outside the simulator.

The paired comparison is computed at the world level, retaining all four investors, checkpoints and configurations within a resampled world. The primary intervals use 2,000 paired percentile

bootstrap draws. The 16-world factorial estimates component effects on its matched subset; its interactions are exploratory. The four repeated worlds measure response consistency and do not enlarge the primary independent sample. Leaving a presentation family out is a useful sensitivity check, but three families cannot support precise population inference over financial documents.

2.4 Scores and the action interface

The frozen exact-output endpoint compares five fields: financial capacity, recommended amount, authorized capacity, proposed amount and status. A wrong derived amount can fail that endpoint even when no purchase is proposed. Semantic-state scoring separately tests extracted state fields. For unresolved contradictory cash evidence, it excludes the unknown cash truth; the reported ceiling therefore concerns the fields the scorer can resolve. It does not mean that every natural-language explanation or every financial fact has been independently validated.

Unsupported positive actions receive their own predicate, based on readiness, authority, lot size and minimum participation. Review with no positive proposal is not treated as an unauthorized transaction. The study retains raw model decisions and final decisions after the guard, so an error intercepted by validation can be distinguished from one surviving it. In this primary sample the guard’s effects are narrower: all its changes normalize null to zero under unchanged review status. This is why the report displays both the frozen exact endpoint and a labeled post-hoc abstention-equivalence analysis.

Call failures stay in scheduled denominators. A service-capacity error and a timeout are unavailable responses, not negative answers to an investment question. Missing responses produce no actionable decision. Record persistence cannot preserve an update that was never returned; the implementation records that missing state rather than filling it with scorer truth. No outcome-dependent replacement run is used to improve an experimental arm.

3 Interpretation, repeated work and operational failure

Table 2 preserves the frozen complete-output scores. A decision is exact only when capacity, recommendation, authorization, proposed amount and status all match. Every successful primary processing context has all scored semantic state fields correct. No unsupported raw or final action appears in a primary arm.

Most of the exact-output improvement is a narrow encoding correction. Validation converts a null proposed amount to zero while retaining review status. It makes 53 such corrections separately and 36 jointly, adding 52 and 36 fully exact outputs. It blocks no positive proposal and changes no status. A post-hoc equivalence rule that treats null-under-review and zero-under-review as the same abstention yields 1,010 versus 1,007 exact decisions separately and 1,020 versus 1,016 jointly. The apparent improvement disappears under that representation; the original scores remain unchanged.

Table 2: Primary processing results

	Separate		Shared	
	R0V0	R1V1	R0V0	R1V1
Exact final decisions / 1,024	946	1,007	974	1,016
Exact raw decisions / 1,024	946	955	974	980
Inference calls	1,791	1,429	448	353
Failed contexts	3	4	0	1
Recorded dollars / world	0.3311	0.2970	0.1457	0.1398
Mean regret / checkpoint (\$)	106.25	365.23	0.00	285.94

Each arm contains 64 four-checkpoint worlds. Dollar costs price returned usage at \$2, \$0.20 and \$8 per million uncached input, cached input and output tokens. Failed-call usage is missing. Regret uses the common synthetic downstream objective with $\eta = .12$; it is not a realized return or an observed service benefit.

Persistent records reduce repeated retrieval. Primary calls fall by 362 separately and 95 jointly. On the factorial subset, the records main effect is -5.28 calls per four-checkpoint book separately and -1.56 jointly. The validation main effect on complete-output accuracy is 5.86 percentage points in each mode, reflecting the same encoding correction. Accuracy interaction intervals include zero. Fewer calls do not necessarily imply lower summed call time; shared-mode call time rises in the complete configuration.

All 192 public-slice answers are correct, another accuracy ceiling. Predetermined repeats match all five scored decision fields in 248 of 256 investor outputs; reason text is not required to repeat. Eight synthetic calls fail: seven service-capacity errors and one 180-second timeout. All remain in scheduled denominators, with no outcome-dependent retry. A shared failure can withhold several investors’ decisions in one response. The observed timeout is such an operational event; it is not evidence that successful interpretations share a common semantic error or that financial markets have acquired a new common factor.

4 Cost incidence

At the token prices in Table 2, the separate-minus-shared reduction in recorded cost is \$0.0281 per four-checkpoint world, with a paired 95% interval [\$0.0173, \$0.0393]. Shared processing is cheaper in level; the tested technology change saves more separately. Leaving any one presentation family out preserves that direction. These amounts are applied prices on recorded usage, not invoices. Missing usage for failed calls makes each recorded bill incomplete; the difference between two savings need not be a lower bound.

The complexity hypothesis is unsupported. Mean incidence is \$0.0290 in the low tier, \$0.0470 in the three-world middle tier and \$0.0242 in the high tier. High-minus-low intervals include zero. The experiment does not show that the most complex mandates receive the largest advantage from separation.

Review can dominate the token bill. Pricing each final review flag at \$5 produces mean incidence of \$0.263 with interval $[-\$0.277, \$0.961]$; pricing it at \$25 gives \$1.200 with interval $[-\$1.527, \$4.711]$. Flags include appropriate abstention and missing outputs as well as correctable errors. These sensitivities neither measure labor nor impose equal achieved quality.

The theoretical bridge is member-specific pass-through. Under equal allocation of shared processing costs, separate savings exceed allocated shared savings for 164 of 256 investor histories; allocation by frozen condition counts gives 166. Both balanced rules produce mean incidence of \$0.0070 per investor history, but individual signs differ. For a fixed contract, a positive incidence $d_i^{\text{sep}} - d_i^{\text{sh}}$ consumes that amount of participation slack. In the continuous model, reductions in costs incurred only upon exit shift f_i downward; a cost change affecting the outside investment retained alongside the program must enter both the retained-portfolio and all-outside valuations. This is the participation comparison developed in the companion theory paper (Komkov, 2026). The experiment observes neither an actual investor’s initial slack nor a provider’s pass-through policy. Its cent-scale contrasts cannot be converted into annual economic gains without additional frequency, quality and service-cost evidence.

4.1 Levels, changes and pass-through

Let $K_{w,c}^m$ be the recorded cost for world w , mode m and configuration c , summed across its four checkpoints. The primary incidence contrast is

$$D_w = (K_{w,00}^{\text{sep}} - K_{w,11}^{\text{sep}}) - (K_{w,00}^{\text{shared}} - K_{w,11}^{\text{shared}}). \quad (1)$$

Positive D_w means that the tested technological change saves more in separate processing. It is compatible with shared processing being cheaper in both configurations. The distinction is essential: a comparison of levels alone does not predict which mode benefits more from a new record technology.

Cost uses separate uncached input, cached input and output counts. Cached tokens are a subset of input and are not charged twice. The middle price illustration gives the four arm means in Table 2; low and high price schedules remain in the archive. System and wrapper input are part of recorded usage, so the comparison is not purely the cost of reading document text. Missing failed-call usage is also retained as missing. Equal imputed costs would cancel in the difference of savings because the signed missing-count contrast is $3 - 4 - 0 + 1 = 0$; unequal missing costs need not cancel.

Calls and time are different resources. The primary contrast saves 4.17 more calls separately per world, with a paired interval $[3.36, 4.94]$. The analogous contrast in summed call time is 48.66 seconds $[30.24, 67.59]$. This sum is not end-to-end service latency: calls can occur in parallel, and queueing, retries internal to the service and operator delays are different quantities. Both complete configurations emit more output tokens while using fewer input tokens. Shared-mode summed call time increases even as its call count falls.

To connect these costs to a participation condition, a provider must allocate shared costs to investors. For investor i , let d_i^{sep} be his separate-route saving and d_i^{sh} his allocated shared saving. Under full pass-through, his previous participation slack s_i becomes

$$s'_i = s_i + d_i^{\text{sh}} - d_i^{\text{sep}}. \quad (2)$$

This is an accounting bridge, not a measured participation response. The report observes neither s_i nor actual pass-through. Equal shares and frozen-condition shares are declared allocation policies; both balance at the book level, but they can assign different signs to the same investor. Provider fees, correction effort and funding obligations would have to be included before the expression could price a real institutional choice.

5 Financial consequences and limits

A common downstream engine selects allocations using either interpreted or oracle state. It accounts for fixed and proportional expenses, a formation minimum and contractual lots. It then evaluates the selected action against the oracle. Regret is assigned only to financially feasible, individually acceptable choices; unsupported or loss-imposing actions are separately reported. The utility schedule is synthetic, and its parameters are described in the Internet Appendix and all member outcomes are retained in the replication record.

At $\eta = .12$, the oracle forms a program in 182 of 256 checkpoints; interpreted-state decisions form in 181, 181, 182 and 181 across the four primary arms. Every selected action is feasible and individually acceptable. All positive primary regret occurs in books with a service failure. The single shared-mode timeout accounts for all \$882,000 of withheld valid participation in shared R1V1 and its positive regret. No primary arm falsely attains the \$650,000 funding target, so the prespecified joint unsupported-action endpoint is zero. This is an absence of such errors in the tested histories, not a measured reduction in economy-wide correlated risk.

The study therefore supplies cost incidence, a semantic ceiling and an operational failure record. Its strongest positive result is less repeated retrieval. Its main caution is that amortizing work can concentrate the consequences of a missing response. Both quantities belong in the economics of implementation; neither establishes which institutional form investors should adopt. The archived model identifier, access timestamps, prompts, evidence and usage permit exact rescoring of the recorded experiment. They do not identify immutable proprietary model weights or guarantee identical future responses.

5.1 A shared failure and an external-document ceiling

The shared timeout occurs in world W023 at checkpoint T2 under R1V1. The oracle's valid proposals are \$309,000, \$272,000, \$301,000 and zero. Their total is \$882,000. One missing shared response withholds all four outputs; the corresponding downstream loss is \$73,200 at $\eta = .12$, or

\$285.94 averaged across the arm’s 256 checkpoints. This is one failure event with several affected investors. Counting four affected outputs as four independent demonstrations would overstate the evidence.

The seven remaining failures report service capacity unavailable. Their distribution across arms does not identify a causal treatment effect on reliability. A successful-only analysis would answer a different question and omit the only source of positive primary regret, so the principal results retain all scheduled contexts. The downstream engine recalculates a permitted allocation from interpreted state in every arm; it does not directly execute the model’s proposed number. Its ability to correct displayed arithmetic is therefore common infrastructure, not a validation-treatment benefit.

The public slice has three cases each from credit agreements, indentures, tender documents and interval-fund filings. Two credit cases use the same filing, leaving 11 unique filings. Questions refer to scoped passages with agent-audited exploratory labels. The same slice appeared in an earlier edition and is not newly held out. One ambiguous economic-rights-equivalence question was replaced before freeze by an explicit conversion-ratio question; that amendment is retained in the record. All 192 answers are correct. The slice supports passage-level feasibility of the task but supplies no measured accuracy ranking. Its single-dossier cases also do not independently identify the value of persistent records over time.

6 Conclusion

Maintaining separate investor state does not require interpreting every common source separately. In this experiment, shared processing is cheaper in recorded inference-cost levels, while persistent records reduce duplicated work more in isolated processing. The difference between these two facts is the central measurement result. An institutional comparison needs both levels and incidence.

The accuracy result is equally specific. The measured semantic task is at a ceiling among successful primary contexts. Literal output gains chiefly concern abstention encoding, not the recovery of missing obligations or the interception of unauthorized purchases. Operational failures nevertheless matter: one shared timeout withholds several otherwise valid decisions. A system can amortize work and concentrate the consequences of an unavailable response at the same time.

The next empirical step is equivalent-quality comparison on harder, independently sampled histories, with observed correction time and documented delivery costs. The current archive supports exact rescoring of what occurred. It does not supply immutable proprietary model weights, measured human labor costs, or actual investment outcomes. A cost frontier for financial organization requires those additional observations.

References

Komkov, John, 2026, The Organization of the Investment Frontier: Independent Mandates and the Formation of Shared Investments, working paper, refinement 4.

Internet Appendix

I. Frozen evidence and computational scope

The replication package separates evidence from executable policy. The synthetic simulator supplies economic state, dated source documents and authenticated permission records. The interpretation system sees the documents, catalog and permitted registry fields; the scorer alone sees the complete synthetic state. The public slice preserves its source URLs, excerpts, question definitions and exploratory labels. Source text is not an authority channel.

The primary evaluation has 64 economic worlds, each with four principals and four checkpoints. A complete book episode contains all four principals even when they are processed separately. The primary comparison contains $64 \times 4 \times 2 \times 2 = 1,024$ book episodes. Completing the factorial on 16 predetermined worlds adds $16 \times 4 \times 2 \times 2 = 256$ episodes; repeating four predetermined worlds adds 64. The total is 1,344 synthetic episodes, requiring 3,360 processing contexts because separate mode uses four contexts and shared mode one. A context can invoke the model once more after requesting dated historical evidence. Twelve public cases in four configurations add 48 episodes and 72 underlying calls, including preparatory extraction. Development used 128 episodes in three separately recorded versions.

The model product identifier, reasoning setting, CLI version, timestamps, prompt hashes, schema hashes, request and response text, source retrieval, tool-use status and reported token usage accompany each call. Calls exceeding the prescribed time limit, using prohibited tools or failing principal-coverage requirements count as failures. Replaying a completed local receipt is an idempotent resume, not a new inference attempt. A future call to the same product identifier can return a different answer because the available interface does not expose an immutable weight snapshot.

II. State transitions and capacity calculations

A principal’s financial capacity is the greatest permitted lot quantity after cash calls, protected liquidity, competing reservations, expenses and concentration are accounted for. In the synthetic task, the displayed formula is

$$C_i = 1,000 \left\lfloor \frac{1}{1,000} \max \left\{ 0, \min \left[\frac{B_i - O_i - L_i - R_i - 5,000}{1.02}, M_i - X_i \right] \right\} \right\rfloor.$$

Here B_i is cash, O_i outstanding calls, L_i protected liquidity, R_i competing reservations, M_i the permitted exposure limit and X_i existing exposure. Unresolved contradictory cash evidence makes the displayed financial capacity unknown. The authorized projection additionally accounts for grant budget, expiry, scope, eligibility and current instrument terms. Financially feasible capital can therefore exceed currently executable capital.

At the first update, the simulator settles an announced distribution and pays part of a previously accepted call. A 100,000 distribution less a 20,000 payment raises cash by 80,000. The remaining call stays on the book. Later updates add a competing reservation, permit an eligibility certificate to expire, change terms, make a preference amendment effective, revoke one grant and allow another to expire. Conflicting current observations require review. These transitions are exogenous and common across experimental arms; no model proposal rewrites the subsequent economic history.

Economic complexity is counted through bounded interventions in source-linked conditions. A condition counts when changing it can alter financial or authorized lot capacity at a checkpoint. Repeated appearances of the same principal-condition pair do not increase the count. Preferred investment amount is a utility kink and is excluded from this feasibility measure. Document length, split accounts, duplicate attribution and hostile instructions remain separate presentation features. Complexity bands are fixed from development worlds with ties retained together.

III. Exact downstream allocation

The downstream comparison uses one objective and the same authenticated registry in every arm. For a positive allocation q_i , net incremental CE is

$$v_i(q_i) = \eta \min\{q_i, d_i\} - 0.02q_i - 5,000, \quad \eta \in \{0, 0.04, 0.08, 0.12\}.$$

Zero participation has zero incremental value and no expense. Positive quantities lie on 1,000-dollar lots, are at least the 25,000-dollar participant minimum and respect the interpreted capacity. Each member must have nonnegative net value under the book used to choose the allocation. A formed program needs at least 650,000 dollars of aggregate participation. The planner can instead choose no program.

For each subset of active principals, the algorithm first finds each member's individually optimal admissible lot quantity, choosing the largest quantity in a value tie. If their sum meets formation scale, these quantities maximize the subset's separable objective. Otherwise, the algorithm adds the best available next marginal lot until the threshold is reached, provided individually acceptable capacity suffices. Marginal increments weakly decrease because each value schedule is concave. Merging the largest available increments therefore minimizes the utility sacrificed to reach the threshold. Enumerating all active subsets and the no-program alternative yields the global finite-lot maximum. Remaining objective ties are resolved by the lexicographically greatest quantity vector in principal order.

The requirement to consider quantities above the preferred kink is material. Four principals each preferring 150,000 dollars with 200,000 dollars of capacity and $\eta = .12$ have unconstrained individual maxima totaling only 600,000 dollars. An algorithm that stops there incorrectly rejects formation. An admissible 650,000-dollar allocation has total net CE of 39,000 dollars, with every

member weakly better off. This case is included in the regression suite. An independent exhaustive audit checked 240 small economies with no optimizer mismatch.

The extracted-book choice is then evaluated against oracle financial and authority constraints and the oracle utility schedule. A financially invalid allocation receives an infeasibility flag. A financially admissible allocation that makes a true member worse off receives a separate participation-failure flag and member values. Neither is assigned a fictitious monetary penalty. Regret is reported only for admissible, individually acceptable actions, as oracle-optimal CE less the oracle value of the action.

IV. Cost incidence and uncertainty

The paired technology contrast is computed within each economic world before aggregation. Let $K_{w,c}^m$ be an observed processing resource for world w , organizational mode m and configuration c , summed over its four checkpoints. For each resource separately, the contrast is

$$D_w = (K_{w,00}^{\text{sep}} - K_{w,11}^{\text{sep}}) - (K_{w,00}^{\text{shared}} - K_{w,11}^{\text{shared}}).$$

A positive cost contrast means the tested change reduces that resource more in separate mode. The same algebra applied to accuracy is an accuracy contrast, not a cost saving. Input tokens, cached tokens, output tokens, prompt characters, model-call time and review flags remain distinct observations. Prices applied to those resources are declared sensitivities.

For a member-level bridge, isolated calls can be attributed directly to their principal. Shared inference costs need a rule. We compare equal shares with shares proportional to frozen consequential-condition counts, using the same weights at every checkpoint and in every configuration. Review flags are assigned directly to the affected principal; a per-flag review price assumes no shared economy in human review. At a stated set of inference and review prices, the pivotal-member test then compares his separate cost reduction with his allocated shared reduction. The allocation rule is part of the institution; it is not estimated from book-level usage.

The primary intervals resample whole worlds 2,000 times. Principal-checkpoint observations within a world are not independent cases. The finite family sensitivity and the repeated-world results describe robustness within this benchmark; three presentation families and one common renderer per split cannot identify uncertainty over the population of financial documents. The factorial comparison uses only the predetermined 16 worlds with all four configurations. The 12 public cases span 11 filings and previously appeared in the earlier research edition; they are an exploratory external slice, not a new held-out financial-document sample.

Joint-error frequencies are reported with aggregate evidence-supported capital and false attainment of the 650,000-dollar target. Excess joint errors over a product-of-marginals reference can reflect common case difficulty as well as shared interpretation failure. These statistics are not estimates of a causal algorithmic-monoculture effect or actual closing probabilities.

V. Model access and analysis provenance

The recorded experiment used the product identifier `gpt-6.1-sol`, low reasoning effort, and Codex CLI 0.159.2 in read-only ephemeral sessions, accessed October 4–5, 2026. Exact UTC timestamps, prompts, rendered evidence, schema hashes, retrieved excerpts, returned usage and failures are preserved in the evaluation archive. The identifier is not an immutable model-weight snapshot. This report replays archived records and recomputes the common downstream allocation; it makes no new model inference calls for the evaluation. Frozen primary scores are preserved, and abstention-equivalence scoring is labeled as a post-hoc diagnostic. The synthetic renderer and scoring ground truth are executable, independent of the evaluated model’s outputs.