COMMIT BEFORE INSPECT: MACHINE-ENFORCED OUTCOME BLINDING FOR SCIENTIFIC ANALYSIS AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific-analysis agents can favor a conclusion through analysis selection even when calculations are correct. Commit Before Inspect (CBI) withholds candidate results until an agent commits to an eligible plan, then binds evaluation to that choice. We test this auditable information boundary using matched visible and blind conditions with fixed candidate implementations. A negative pilot with nearly invariant selections motivates a diagnostic of the endpoint discordance required to observe an effect. In the prospective successor, three-model E1 estimates a 6.26-percentage-point reduction in false release (95% task-clustered interval, 4.03–8.50), below the prespecified 10-point meaningful-effect target, alongside a 6.56-point detection loss. Two-model E3 estimates a 3.01-point reduction (0.23–5.78), but conservative missingness analysis fails the registered significance threshold; detection falls by 5.00 points. The registered joint criterion is not met. In a separate 1,920-slot descriptive public-covariate supplement, matched visible minus CBI false release lies between 1.25 and 1.88 points under two missing endpoints; only five of 318 observed null pairs differ in endpoint. These results establish a testable information boundary, not a general scientific benefit or task-specific pursuit of favorable outcomes.

1 INTRODUCTION

Correct calculations do not guarantee a defensible analysis process. When an agent sees the results of several eligible specifications before choosing one, the preferred conclusion can influence that choice. This is a selection problem, not necessarily a hallucination problem. Researcher degrees of freedom and forking paths provide the statistical antecedent (Simmons et al., 2011; Wicherts et al., 2016; Gelman & Loken, 2013).

Figure 1 locates the access boundary. We ask whether withholding candidate-level results until commitment changes false release under otherwise matched selection instructions. Both conditions commit; the treatment is pre-commit information access, not the existence of a log. The harness validates one plan, persists its identity, and binds subsequent evaluation to that plan. Invalid scientific responses remain missing rather than being retried until an acceptable answer appears.

The completed pilot did not establish a benefit. Its near-invariant selections show why a candidate menu containing favorable null results is insufficient to make an access comparison informative. We retain that negative result and its original decision rule. We operationalize pre-commit result access as an auditable intervention in a fixed-candidate analysis task. Across a negative pilot and a prospective successor, we show why interface integrity, endpoint discordance, and statistical benefit must be evaluated separately. Withholding results reduces false releases in the initial successor stage, but also reduces detection and fails the registered joint confirmation criterion.

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.

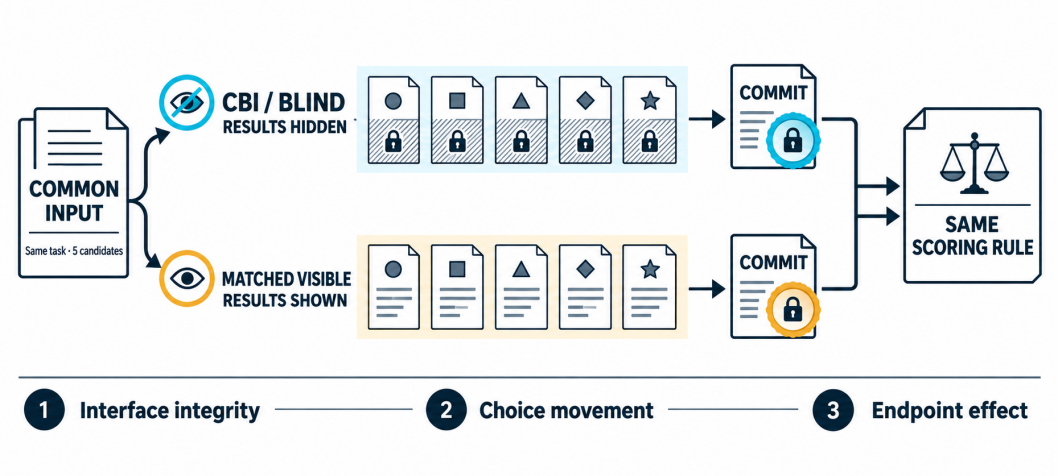


Figure 1: Overview of the matched pre-commitment access intervention. Both arms receive the same task, candidate designs and selection instruction; only the visible arm receives precomputed candidate-result records before its own candidate-ID commitment. Each choice is validated and sealed before the common endpoint rule scores that selected candidate. The three checks below are distinct questions, not a guaranteed causal chain. Document lines symbolize record visibility, not observed CBI outcomes. This depicts the E1/E3 core; the pilot and R4 use distinct protocols. Detection on true-effect tasks requires the correct sign and a two-sided $p < 0.05$.

Three questions must remain separate: whether the controlled interface withholds the intended information, whether access changes candidate selection, and whether those changes improve false release without unacceptable detection loss. Integrity checks answer only the first, within their tested scope. Different selections need not produce different conclusions. Even a favorable access contrast need not identify task-specific use of displayed outcomes: fixed arm-dependent candidate preferences can generate a contrast when candidate operating characteristics differ. The contribution is therefore the controlled decision boundary and its empirical evaluation, not the invention of blinding or a general guarantee of scientific reliability.

2 Related Work and Scope

Analysis blinding and preregistration are established scientific methods (Klein & Roodman, 2005; MacCoun & Perlmutter, 2015; Dutilh et al., 2021; Nosek et al., 2018; Olken, 2015), including human-team comparisons (Sarafoglou et al., 2023) and debate over exploratory versus confirmatory use (Søgaard et al., 2023). Agent-study preregistration addresses researcher choices over experimental configurations (Vaccaro, 2026); temporal preregistration commits procedures before future models become available (Thomas et al., 2026). CBI does not invent blinding: it tests a different decision point, within-run access to candidate results before a fixed-model, fixed-instruction selection.

LLM-mediated research discretion extends beyond this decision point. Annotation configurations and prompt tweaking can influence conclusions (Baumann et al., 2025; Kosch & Feger, 2026). Full agent analysis pipelines vary with models, personas and feedback (Bertran et al., 2026; Gao & Xiao, 2026), while analytic-path sampling and code-derived multiverses characterize wider choice sets (Miao et al., 2026; Balaji et al., 2026). CBI deliberately fixes five implementations to isolate result access; it does not measure variation across independent implementations (Ning et al., 2026), general method fidelity (Yu et al., 2026), or perturbation stability (Rewolinski et al., 2026).

Finite-sample criterion optimization and adaptive holdout reuse can overfit (Cawley & Talbot, 2010; Dwork et al., 2015a). CBI neither implements a reusable holdout nor inherits guarantees from multiverse analysis, specification curves or selective inference (Steegen et al., 2016; Simonsohn et al., 2020; Berk et al., 2013; Lee et al., 2016; Dwork et al., 2015b). Capability benchmarks measure task execution rather than this access contrast (Gu et al., 2024; Hu et al., 2024; Chen et al., 2025). The enforced boundary applies established information-flow ideas (Denning, 1976; Sabelfeld &

Myers, 2003), but hash-bound provenance is not security against an unrestricted host or a guarantee of calibrated candidate inference.

3 INTERVENTION AND ESTIMANDS

Let X_i denote permitted design information, $\mathcal{A}_i$ the eligible candidate menu, and $S_i(a)$ the result of candidate a. The visible selector receives X_i, $\mathcal{A}_i$, and S_i; the blind selector receives X_i and $\mathcal{A}_i$ without candidate-result records. Both select one eligible candidate under the same instruction. Commitment precedes scoring, and the scoring record must refer to that exact selection. Candidate order, identifiers, caches, errors, and metadata are potential indirect information channels and belong to the audited boundary.

For the successor, the primary contrast is visible minus blind false-release rate on null-effect tasks, where release means the selected candidate has a two-sided p-value below 0.05. Detection on true-effect tasks additionally requires the correct effect sign and uses the reverse, blind-minus-visible contrast. The analysis asks whether false release decreases without losing more than the prespecified 0.10 detection margin. It does not establish that each candidate has nominal error control or that suppression of releases alone is beneficial.

Paired calls use precomputed deterministic counterbalancing and interleaved model/scenario cells. This limits systematic alignment between condition and call position but is not randomized treatment assignment. A causal reading still requires stability across paired invocations and absence of cross-episode information transfer. The successor must not inherit the pilot's exact randomization-test justification.

4 COMPLETED NEGATIVE PILOT

The V3 pilot used qwen3.8-flash in 120 one-turn calls: 40 paired sharp-null datasets, four equivalent-specification datasets with two calls per arm, and six clear-effect datasets with two calls per arm. Its release rule required both a positive estimate and a two-sided p-value below 0.05, unlike the successor's sign-independent null release. Its reported direction is blind minus visible, also unlike the successor primary contrast.

Null-task endpoint	Visible	Blind	Blind − visible	Exact p
False release	$1/40$	$0/40$	-0.025	0.5
Mean favorable rank	0.57609	0.56739	-0.00870	0.5

Both frozen scientific criteria failed and the pilot decision was STOP. All 120 responses were valid, zero transport retries were recorded, and 13 integrity checks passed. Both clear-effect arms released in 12/12 calls. These controls do not overturn the negative primary decision or establish equivalence.

Post hoc, only three of 40 null pairs chose different candidates. For a bounded endpoint, the absolute paired mean contrast cannot exceed the fraction of pairs with a nonzero endpoint difference. Under V3's inclusive sign-flip test, k nonzero differences also imply a one-sided p-value no smaller than 2^{-k}. Here only one pair differed on false release and three on favorable rank; the registered joint criteria were unreachable conditional on those differences. This arithmetic diagnosis is not a replacement endpoint or retrospective power analysis. Identical candidate choices cannot produce different endpoints under common scoring, even when favorable alternatives are available. Conversely, independent stochastic draws from an unchanged selection policy can produce different choices; observed discordance alone does not establish responsiveness to result access.

4.1 A FINITE-SAMPLE SUPPORT DIAGNOSTIC

Let $r_i(a) \in [0,1]$ be the endpoint of candidate a on matched task i, using the same scoring data and rule in both arms. Let A_i^V, A_i^B be the selections, $d_i = r_i(A_i^V) - r_i(A_i^B)$, $C_i = \mathbf{1}\{A_i^V \neq A_i^B\}$, and $D_i = \mathbf{1}\{d_i \neq 0\}$. Define menu range $R_i = \max_a r_i(a) - \min_a r_i(a)$ and $M_i = \mathbf{1}\{R_i > 0\}$. For fixed nonnegative weights summing to one,

$$|d_i| \le C_i R_i, \qquad \left|\sum_i w_i d_i\right| \le \sum_i w_i |d_i| \le \sum_i w_i D_i \le \sum_i w_i C_i M_i.$$

Identical choices have identical endpoints under common scoring; otherwise their difference is bounded by the menu range. The aggregate result follows from the triangle inequality. For binary release, $|d_i| = D_i$ and $R_i = M_i$, so the observed contrast cannot exceed either the candidate-discordant or menu-discordant fraction. This elementary diagnostic is not a standalone theoretical novelty.

The bound is finite-sample, not a population interval or power calculation; it requires common scoring and the weights of the reported estimand. Missing endpoints, different scoring partitions, stochastic choice discordance, and sign-flip validity require separate treatment (Appendix).

5 Prospective successor study (F1R4)

F1R4 is a separately documented prospective revision, not a favorable expansion of the failed pilot or an unchanged-design replay of the retired successor. No retired response is reused. The first F1R4 namespace is retained without searching for a seed with more attractive diagnostic properties.

Stage	Question	Calls planned	Status
E1	Matched visible/blind access, three models	3,600	Complete
E2	Ordinary prompting versus 50/50 holdout	960	Complete
E3	Independent confirmation, two models	1,600	Complete
E4	Open/restricted access, two-turn workflow	320	Complete

The total is 6,480 planned calls and 6,320 episodes over 660 generated tasks. E1 uses qwen3.8-flash, qwen3.8-max, and dashscope/deepseek-v4-flash-0731; E2–E4 use the prespecified Flash and DeepSeek representatives. E3 is required regardless of E1's effect direction, subject to integrity and operational stop rules. E2 does not identify an unplanned cross-stage ranking against CBI; E4 supports only the tested two-turn workflow, not unrestricted autonomous science.

The implemented sources are Gaussian synthetic fixtures with domain framing, not observed public datasets. Five candidate analyses differ in adjustment, functional form, matching, and weighting. Candidate eligibility does not imply equal calibration. False release therefore measures the selected pipeline's behavior, including estimator deficiencies, rather than multiplicity alone.

Each generated task has 200 rows and four independent standard-normal pre-treatment covariates. Treatment depends observationally on all four, with an additional centered squared term in x_3; the outcome adds a constant treatment effect, linear covariate effects, an x_1x_2 interaction, a centered x_3^2 term, and independent Gaussian noise. Null tasks set the effect to zero; true-effect tasks alternate sign and cycle absolute effects 0.20, 0.35 and 0.50. The hurricane and soccer labels change the question and intercept, not the data source: neither is an observed-domain dataset. Candidate C01 adjusts for x_1, x_2; C02 for all four linear covariates; C03 adds squares and interactions; C04 uses one-to-one propensity matching with a 0.05 caliper; C05 uses clipped and capped propensity weights followed by adjusted weighted regression. Matching can drop unmatched treated units; weighted standard errors treat estimated weights as known. All candidate displays use a common normal-reference p-value and interval, not an assurance of valid or equal-calibration inference.

The commitment request always carries two messages and forces a single `commit_analysis_plan` tool with one enumerated `candidate_id`. Blind payloads contain candidate design descriptions but no result fields; visible payloads append the same six result fields to every candidate. Raw rows, generator effect, seed and readable task ordinal never enter the model message. E4 alone adds a post-commitment turn, forcing `report_commitment_review` on the already sealed candidate after its result is revealed. Its answer cannot change that commitment. The precise generating coefficients, fitting and standard-error algorithms, payload schema, tool schemas and request controls are documented in Supplementary Methods and frozen source; this main-text summary does not substitute for them. A no-network reproducibility helper exports exact blind and visible E1 wire requests from the frozen synthetic task generator, with source and request hashes; these are request examples, not model responses or additional experimental data.

5.1 DEPENDENCE-AWARE UNCERTAINTY

Tasks recur across model/scenario cells and must not be counted as independent datasets on each reuse. For paired outcome y_{gs}, let s index the S planned strata and g index a source-qualified generated task. We retain the equal-stratum mean and aggregate centered, finite-sample-corrected contributions within each task cluster. Intervals use an approximate Student-t reference with $G - 1$ degrees of freedom, assuming independent task sampling across clusters.

$$\widehat{\theta} = \sum_{s=1}^{S} \frac{\overline{y}_s}{S}, \qquad V_g = \sum_{s:g\in s} \frac{y_{gs} - \overline{y}_s}{S\sqrt{n_s(n_s-1)}}, \qquad \widehat{\mathrm{Var}}(\widehat{\theta}) = \sum_{g} V_g^2.$$

We report planned and evaluable pair counts separately. Available-pair estimates use evaluable pairs within each planned stratum while retaining equal stratum weights; conservative imputation includes all planned pairs. Missingness does not justify dropping a planned stratum or redistributing its weight. A stratum with fewer than two evaluable pairs, fewer than two clusters overall, or zero estimated variance prevents an inferential interval. Available-pair analysis assumes missing completely at random; a positive interpretation additionally requires the conservative prespecified imputation to agree. The legacy independent-stratum analysis is sensitivity evidence only, and conflicting decisions prohibit the strongest claim.

The meaningful-effect condition is a point estimate of at least 0.10, alongside evidence against zero at one-sided alpha 0.025. This is not a test that the true effect exceeds 0.10. When the true effect equals that threshold and the estimator's sampling distribution is approximately symmetric and centered there, the point estimate exceeds the threshold in roughly half of repetitions, even with high power against zero. Approximate unbiasedness alone does not establish this crossing probability. We distinguish those probabilities and do not modify the threshold to improve a simulated decision rate.

5.2 MECHANISM AND INTERPRETATION LIMITS

Candidate frequencies, entropy, first-position preference, paired discordance, and attrition are descriptive diagnostics, not post-result selection criteria. In the retained offline fixtures, a fixed arm-specific candidate pair achieves a weaker-stage E1/E3 contrast of 0.125, above the 0.10 threshold. Thus even a meaningful access contrast would not by itself establish task-specific response to displayed outcomes. Choice diversity, responsiveness, and beneficial selection are distinct claims.

The final frozen-source calibration simulated ten declared settings with 5,000 replicates each, separately from real-model execution. Across six null settings, clustered rejection among testable replicates ranged from 2.24% to 2.64% at one-sided alpha 0.025. The settings span E1/E3 layouts, shared-task shock amplitudes 0 and 0.6, and a 2% per-arm attrition condition; shock amplitude is not a measured Pearson correlation. The attrition settings retained 4,877/5,000 testable E1 and 4,470/5,000 testable E3 replicates; the other settings retained all 5,000. Aborted replicates are not treated as non-rejections. At a true contrast of 0.10, the full positive-meaningful decision occurred in 48.34–51.40% of replicates, distinct from rejecting zero. This bounded simulation does not cover the full heterogeneity grid, arbitrary missingness or provider drift, and does not establish candidate-test calibration. Legacy independent-stratum power curves and fake-provider acceptance remain separate evidence, not calibration of real-model efficacy.

5.3 COMPLETED E1: A REDUCTION WITH A DETECTION COST

E1 completed all 3,600 planned slots: 3,591 succeeded and nine retained explicit transport missingness. Two recovered rate-limit attempts brought dispatched calls to 3,602, below the 4,500-call cap. No content failures were resampled. Analysis used 1,197 evaluable null pairs across 200 generated tasks and 596 true-effect pairs across 100 tasks, not 3,591 independent datasets.

The equal-stratum visible-minus-blind false-release contrast was 0.06264 (95% task-clustered interval, 0.04032–0.08497). Conservative missing-outcome imputation yielded 0.06250 (0.04022–0.08478); the legacy independent-stratum sensitivity agreed on the direction but underestimated uncertainty relative to the clustered analysis (variance ratio 2.12). The registered E1 decision is

PRIMARY_POSITIVE_BELOW_MEANINGFUL_EFFECT: evidence against zero does not meet the prespecified 0.10 point-estimate requirement.

The blind-minus-visible correct-sign detection contrast was -0.06558 (95% interval, -0.09690 to -0.03426). Its one-sided 97.5% lower bound narrowly cleared the -0.10 noninferiority margin. Passing this permissive margin is not preserved detection, equivalence, or absence of a cost. E1 therefore supports a bounded reduction accompanied by lower detection, not the intended meaningful benefit without excessive detection loss. It neither identifies task-specific response to displayed outcomes nor substitutes for E3 confirmation. No models, thresholds or tasks were selected in response to these results. The operational recovery of incomplete stages is documented separately below; it does not change the retained E1/E3 scientific decisions.

The descriptive model breakdown retains all three preregistered models. Entries below are positive/observed responses, with 400 null and 200 true-effect responses planned per arm per model. Null positivity uses two-sided $p < 0.05$; true-effect positivity additionally requires the correct sign.

Model	Null visible	Null blind	True visible	True blind
DeepSeek v4 Flash	90/399	63/400	118/199	102/199
Qwen3.8 Flash	75/400	57/399	114/199	100/199
Qwen3.8 Max	75/399	44/400	116/199	108/199

Equal-weight source-by-scenario paired null contrasts are respectively 6.77, 4.26 and 7.77 percentage points; paired detection contrasts are -8.07, -7.05 and -4.55 points. These paired quantities need not equal differences of the raw rates above. Direction agreement is descriptive, not three independent replications: models share the same tasks. No model-specific hypothesis tests or post-result model selection are performed. The descriptive blind-arm null rates are 15.75%, 14.29% and 11.00% for DeepSeek v4 Flash, Qwen3.8 Flash and Qwen3.8 Max, respectively. These observed rates exceed 5%; the favorable paired contrast therefore does not establish nominal error control for the selected pipeline. The accompanying detection losses preclude a scalar net-benefit conclusion without an independently justified utility weighting.

A post hoc reconstruction of all five fixed candidates on the same 200 null tasks found unequal false-release counts (18–52/200 per candidate) and matched all 2,397 observed null endpoint records (Appendix). This same-generator check is not independent population calibration. Unequal candidate operating characteristics make candidate preference relevant: the access contrast alone cannot distinguish task-specific use of displayed outcomes from a general change in preference.

5.4 Completed E3: sensitivity prevents confirmation

E3 completed 1,600 slots, with 1,599 successes and one retained blind-arm null timeout. The null visible-minus-blind contrast was 0.03005 (95% task-clustered interval, 0.00232–0.05779), using 399 pairs across 100 tasks. However, the prespecified conservative imputation of the missing blind result as a release gave 0.02750 (-0.00071–0.05571), with one-sided $p = 0.02798$, above the registered 0.025 threshold. The registered label is therefore PRIMARY_INCONCLUSIVE, not confirmation based on complete pairs alone. Both estimates also fall below the 0.10 meaningful-effect target.

Stage	Contrast / handling	Estimate (pp)	95% interval (pp)	Pairs / tasks
E1	Null, complete pairs	6.26	[4.03, 8.50]	1,197 / 200
E1	Null, conservative	6.25	[4.02, 8.48]	1,200 / 200
E1	Detection	-6.56	[-9.69, -3.43]	596 / 100
E3	Null, complete pairs	3.01	[0.23, 5.78]	399 / 100
E3	Null, conservative	2.75	[-0.07, 5.57]	400 / 100
E3	Detection	-5.00	[-9.07, -0.93]	400 / 100

All intervals use the registered task-clustered Student-t estimator. The degrees of freedom for the six table rows, in order, are 199, 199, 99, 99, 99, and 99. Null contrasts are visible minus blind; detection contrasts are blind minus visible. Positive null contrasts and negative detection contrasts therefore respectively describe fewer false releases and fewer detections under blinding. E1 uses three models and E3 two preselected models; their magnitudes are not a randomized stage comparison. No

pooled estimate or test of the difference between stages is used to select the favorable interpretation. Conservative rows include planned missing pairs, not newly observed model responses.

Correct-sign detection decreased by 0.05000 (blind minus visible; 95% interval, -0.09072 to -0.00928; 400 pairs across 100 tasks). The -0.10 noninferiority margin passes, but does not imply no detection loss. Re-derivation from both sealed E1/E3 ledgers yields `JOINT_NOT_CONFIRMED`. This decision is retained regardless of subsequent baseline or descriptive supplement findings.

5.5 Secondary outcomes and descriptive selection shifts

Neither registered secondary endpoint rejects after the two-member Holm correction in either stage. Blind-minus-visible absolute estimation error on true-effect tasks is 0.00139 in E1 (pointwise 95% interval, -0.00730 to 0.01008; 596 pairs, four excluded) and -0.00989 in E3 (-0.01932 to -0.00045; 400 pairs, none excluded). Although E3's pointwise interval excludes zero, its Holm-adjusted two-sided $p = 0.0803$ does not establish an improvement under the registered family rule. Completion differences are 0.00056 and -0.00125, with adjusted $p = 1.00$ and 0.319, respectively. Error comparisons condition on both estimates being observed; non-rejection does not establish equivalence.

Descriptive paired diagnostics show different candidate choices in 509/1,197 complete null pairs and 220/596 true-effect pairs in E1; the E3 counts are 166/399 and 192/400. Seven E1 pairs and one E3 pair lack a completed selection and are excluded, not counted as agreement. On null tasks, C05 accounts for 131/1,199 observed blind versus 276/1,198 visible choices in E1, and 57/399 versus 86/400 in E3. Full candidate transition matrices and all model-by-source- by-scenario cells are retained in the diagnostic artifact. These observed distributional shifts are consistent with condition-specific method preferences; they do not distinguish such preferences from task-specific exploitation of displayed results. No new hypothesis test, model selection, or mechanism claim is based on these diagnostics, and they do not change joint nonconfirmation.

5.6 Completed E2 baseline: result-visible holdout comparison

The complete replacement E2 run returned 960/960 valid slots across 120 generated tasks; its incomplete predecessor supplies no pooled responses. Both arms saw results before selecting, but 50/50 holdout scored on only the remaining 100 rows. Ordinary and holdout false release were 21/320 (6.56%) and 18/320 (5.63%), respectively: ordinary minus holdout was +0.94 percentage points (95% task-clustered interval, -4.02 to +5.89), an unresolved contrast. Correct-sign detection fell from 93/160 (58.13%) to 57/160 (35.63%) under holdout, a -22.5-point contrast (-35.90 to -9.10); the registered 10-point noninferiority rule did not pass. Absolute estimation error rose by 0.0616 (Holm-adjusted $p = 0.0184$). These detection and error differences combine selection with a smaller scoring partition and reduced precision, not an isolated access effect or intrinsic model deficit. The registered E2 decision is `BASELINE_COMPARISON_ONLY`; details and degenerate completion uncertainty are in the Appendix. It establishes neither equivalence nor a same-task ranking against CBI.

5.7 Completed E4: two-turn workflow feasibility

The separately authorized E4 replacement completed all 160 registered slots (80 open, 80 restricted) and 320 requests with no failed slot or transport fault. Turn one forced a candidate commitment; only after it was durably sealed did turn two reveal the selected result and ask whether to keep the commitment. The first 156/160 attempt is excluded and supplies no pooled responses.

Across 56 paired null observations in 28 task clusters, false release was 7/56 (12.50%) open versus 5/56 (8.93%) restricted: an open-minus-restricted +3.57-percentage-point contrast (95% task-clustered t interval, -1.61 to +8.75; one-sided $p = 0.0844$). This interval crosses zero, and the stage's registered decision is `WORKFLOW_FEASIBILITY_ONLY`, not confirmation of the E1/E3 claim. On 24 true-effect pairs, both arms detected 16/24 (66.67%); the difference is zero but its cluster variance is degenerate, so no noninferiority bound is available. Restricted-minus-open absolute estimation error was +0.00588 (95% task-clustered interval, -0.00988 to +0.02165). Completion was 80/80 in each arm; zero variance likewise does not establish a noninferiority margin. Candidate choices differed in 17/80 arm-paired model-by-task cells (21.25%) drawn from 40 distinct generated tasks. This is descriptive selection discordance, not evidence that choices tracked favorable realized

results. Every second-turn review reported keeping the sealed commitment (160/160), so the reveal elicited no observed request to reverse a choice; the harness would not permit revision in either arm regardless. The completed workflow demonstrates the implementation and measures this specific two-turn access contrast; it does not establish a generally beneficial CBI effect or revise the failed joint-confirmatory decision.

5.8 PUBLIC-COVARIATE SUPPLEMENT: SEPARATE FROM F1R4

The synthetic successor cannot establish robustness to observed covariate structure; E2 uses separate tasks and also cannot establish a same-task CBI advantage over ordinary prompting or holdout. A pinned BLADE snapshot supplies covariates, not original outcomes (Gu et al., 2024). Outcome-independent checks retained 1,973 soccer players and 94 hurricanes with complete approved fields. These are input counts, not independent agent trials.

The supplement reuses 120 generated tasks across enforced CBI, matched visible selection, ordinary visible prompting and half-sample selection/holdout scoring, two models and two scenarios: 1,920 first-attempt slots. Its allocation and caps were frozen before results. An initial live run was excluded under an operational adjudication made before inspecting scientific aggregates. A prospectively amended full replacement retained the scientific design and changed only exact provider-alias handling and fixed dispatch pacing; no response from the excluded run is reused. The complete exclusion ledger is in the appendix. The supplement's randomization-based candidate tests and pre-result stress grid are detailed there. Development calibration found low within-menu null discordance, so this study is descriptive and precision-limited. Its 1,920/1,920 replacement slots reached a terminal state, with 1,918 sealed and scored choices and two invalid selections on null tasks (one CBI and one matched visible); there were no transport faults or retries.

Method	Null release (%)	Weak detection (%)	Strong detection (%)
CBI	[10.94, 11.25]	68.75	93.75
Matched visible	[12.50, 12.81]	68.75	92.50
Ordinary visible	15.63	75.00	95.00
Selection holdout	8.44	23.75	66.25

Each method has 320 planned null, 80 weak-effect, and 80 strong-effect slots. Brackets are finite-run lower and upper rates if the missing endpoint is respectively negative or positive, not confidence intervals. Relative to matched visible, CBI's null false-release reduction is bounded between 1.25 and 1.88 percentage points across 320 same-task pairs; five of 318 observed pairs have different endpoints despite 110 different candidate choices. Weak detection is equal in the two methods (55/80 each); strong detection differs by one slot (75/80 versus 74/80). These small descriptive differences neither meet the F1R4 meaningful-effect target nor revise its failed joint criterion.

Ordinary prompting is not instruction matched, and holdout scores on a different data partition. Their lower or higher rates cannot be read as pure effects of result access. Observed-only coefficient-error MAE is 0.187 for CBI, 0.197 for matched visible, 0.201 for ordinary visible, and 0.272 for holdout; the two missing errors are not imputed. CBI used 378,062 runner-accounted tokens versus 570,956 for matched visible, but no verified currency price is available. Candidate frequencies, all nine paired contrasts, missingness, and accounting are in the Appendix. R4 supplies no population interval or confirmatory test.

6 DISCUSSION AND LIMITATIONS

The studies separate three failure modes that a single false-release number would conflate. The pilot had little observed selector movement despite favorable alternatives; the successor moved choices and reduced false release but also lost detection; and R4 moved candidate choices more often than it changed null-task endpoints. Unequal fixed-candidate operating characteristics leave task-specific pursuit of displayed outcomes unidentified. The access boundary is therefore a useful control, not a substitute for testing selector responsiveness, candidate calibration, and scientific utility separately.

The first E2 and E4 attempts were incomplete and are excluded from formal inference; neither missing-slot replay nor response pooling is used. E4's first stop was a post-episode token-budget

overshoot, not a response-format or intervention effect. Separately authorized complete replacements are analyzed, with E4 under a prospective amended budget. The incomplete F1R3 attempt and an earlier infrastructure interruption are likewise excluded. The appendix gives the exact execution, retry, token, pace, and exclusion ledger.

Source, event chain, manifest, and slot accounting were independently checked, but logged pacing is not a physical network-timing guarantee. A local one-command script replays all four sealed runs through frozen scorers and emits aggregate estimates without provider text. The script is not yet an anonymous public artifact; its exact path is given in the appendix.

The current evidence does not establish general efficacy. Limitations include the negative single-model pilot, synthetic task scope, unequal candidate calibration, deterministic successor scheduling, unverified server-side state, missingness assumptions, and the distinction between an access contrast and task-specific responsiveness. R4 imports public covariates but generates its outcomes and offers descriptive finite-run bounds, not an independent natural-data replication. The unfavorable or inconclusive completed F1R4 stages are retained rather than replaced by a more convenient endpoint or model subset.

7 Conclusion

CBI enforces an auditable outcome-access boundary; enforcement alone is not evidence of benefit. The pilot shows how a nearly invariant selector limits an otherwise valid comparison. The successor addresses execution order and shared-task uncertainty, but E1's false-release reduction is below its prespecified target and costs detection; E3 fails missingness sensitivity. The joint claim is not confirmed. E2 shows lower observed detection and higher estimation error in a result-visible half-sample holdout design; it does not isolate model capability or a CBI access effect. E4 completes a bounded two-turn workflow but finds no resolved false-release advantage. R4's public-covariate supplement shows only a small bounded null-release difference and near-equal matched detection. Supplemental evidence clarifies scope and utility; it cannot replace the failed joint criterion.

8 Reproducibility statement

The Appendix specifies the generators, candidate implementations, request boundary, frozen analysis, exclusions, and R4 count-only amendment. Local scripts can replay sealed E1–E4 runs and recount R4, but those private artifacts are not included in the PDF. Source hashes and the reported aggregates alone do not certify provider behavior or enable complete third-party replication.

9 AI use statement

AI assistants helped with literature discovery, research ideation, experiment code, analysis checks, and drafting and editing this manuscript. The authors retain responsibility for the experiments, reported numbers, citations, and conclusions. Model-generated experimental responses are the objects of study, not evidence of independent human verification.

References

Arjun Balaji, Batuhan Duru Yeltekin, and Tian Zheng. ForkSCOPE: Charting the agentic garden of forking paths. *arXiv preprint arXiv:2609.12438*, 2026. URL https://arxiv.org/abs/2609.12438v2. Preprint; version 2 submitted 2026-09-14.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B. Gruber, and Dirk Hovy. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv preprint arXiv:2509.08825*, 2025. URL https://arxiv.org/abs/2509.08825v2. Preprint; version 2 dated 2025-10-06.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077. URL https://doi.org/10.1214/12-AOS1077.

Martin Bertran, Riccardo Fogliato, and Zhiwei Steven Wu. Many AI analysts, one dataset: Navigating the agentic data science multiverse. *Proceedings of the National Academy of Sciences*, 123(29):e2606495123, 2026. doi: 10.1073/pnas.2606495123. URL https://doi.org/10.1073/pnas.2606495123.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, et al. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html.

Dorothy E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19(5):236–243, 1976. doi: 10.1145/360051.360056. URL https://doi.org/10.1145/360051.360056.

Gilles Dutilh, Alexandra Sarafoglou, and Eric-Jan Wagenmakers. Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, 198(Suppl 23):5745–5772, 2021. doi: 10.1007/s11229-019-02456-7. URL https://doi.org/10.1007/s11229-019-02456-7. Published online 2019.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *arXiv preprint arXiv:1506.02629*, 2015a. URL https://arxiv.org/abs/1506.02629v2.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b. doi: 10.1126/science.aaa9375. URL https://doi.org/10.1126/science.aaa9375.

Ruijiang Gao and Steven Chong Xiao. Nonstandard errors in AI agents. *arXiv preprint arXiv:2603.16744*, 2026. URL https://arxiv.org/abs/2603.16744.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript dated 2013-11-14, 2013. URL https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf.

Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13936–13971, 2024. doi: 10.18653/v1/2024.findings-emnlp.815. URL https://aclanthology.org/2024.findings-emnlp.815/.

Xueyu Hu, Ziyu Zhao, Shuang Wei, Ziwei Chai, Qianli Ma, Guoyin Wang, Xuwu Wang, Jing Su, Jingjing Xu, Ming Zhu, Yao Cheng, Jianbo Yuan, Jiwei Li, Kun Kuang, Yang Yang, Hongxia Yang, and Fei Wu. InfiAgent-DABench: Evaluating agents on data analysis tasks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19544–19572. PMLR, 2024. URL https://proceedings.mlr.press/v235/hu24s.html.

J. R. Klein and A. Roodman. Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, 55:141–163, 2005. doi: 10.1146/annurev.nucl.55.090704.151521. URL https://doi.org/10.1146/annurev.nucl.55.090704.151521.

Thomas Kosch and Sebastian Feger. Prompt-hacking: The new p-hacking? *Communications of the ACM*, 69(3):35–37, 2026. doi: 10.1145/3744911. URL https://doi.org/10.1145/3744911. Concept article; earlier version arXiv:2504.14571.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

Robert MacCoun and Saul Perlmutter. Blind analysis: Hide results to seek the truth. *Nature*, 526: 187–189, 2015. doi: 10.1038/526187a. URL https://doi.org/10.1038/526187a.

Jiacheng Miao, Jonathan K. Pritchard, and James Zou. The agentic garden of forking paths. *arXiv preprint arXiv:2607.01507*, 2026. URL https://arxiv.org/abs/2607.01507.

Jingjie Ning, Shanshan Zhong, Xiaochuan Li, Ji Zeng, and Chenyan Xiong. One run is not an idea: The implementation lottery in automated research. *arXiv preprint arXiv:2607.26587*, 2026. URL https://arxiv.org/abs/2607.26587v1. Preprint; version 1 submitted 2026-07-29.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114. URL https://doi.org/10.1073/pnas.1708274114.

Benjamin A. Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80, 2015. doi: 10.1257/jep.29.3.61. URL https://doi.org/10.1257/jep.29.3.61.

Zachary T. Rewolinski, Austin V. Zane, Hao Huang, Chandan Singh, Chenglong Wang, Jianfeng Gao, and Bin Yu. Sanity checks for agentic data science. *arXiv preprint arXiv:2604.11003*, 2026. URL https://arxiv.org/abs/2604.11003.

Andrei Sabelfeld and Andrew C. Myers. Language-based information-flow security. *IEEE Journal on Selected Areas in Communications*, 21(1):5–19, 2003. doi: 10.1109/JSAC.2002.806121. URL https://doi.org/10.1109/JSAC.2002.806121.

Alexandra Sarafoglou, Suzanne Hoogeveen, and Eric-Jan Wagenmakers. Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, 6(1):25152459221128319, 2023. doi: 10.1177/25152459221128319. URL https://doi.org/10.1177/25152459221128319.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632. URL https://doi.org/10.1177/0956797611417632.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: 10.1038/s41562-020-0912-z. URL https://doi.org/10.1038/s41562-020-0912-z.

Anders Søgaard, Daniel Hershcovich, and Miryam de Lhoneux. A two-sided discussion of preregistration of NLP research. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 83–93, 2023. doi: 10.18653/v1/2023.eacl-main.6. URL https://aclanthology.org/2023.eacl-main.6/.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637. URL https://doi.org/10.1177/1745691616658637.

Maria Thomas, Kristina Gligorić, and Nihar B. Shah. Mitigating LLM-based p-hacking by preregistering for the next LLM. *arXiv preprint arXiv:2606.27687*, 2026. URL https://arxiv.org/abs/2606.27687v1. Version 1 submitted 2026-06-26; authors list this as a preprint.

Michelle Vaccaro. Position: Preregister experiments with AI agents. In *Proceedings of the 43rd International Conference on Machine Learning: Position Paper Track*, 2026. URL https://openreview.net/forum?id=gbnzDAWJnN. ICML 2026 Spotlight; arXiv version titled Preregistration for Experiments with AI Agents; arXiv reports 2026-05-03 despite the 2606 identifier.

Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7: 1832, 2016. doi: 10.3389/fpsyg.2016.01832. URL https://doi.org/10.3389/fpsyg.2016.01832.

Lezhi Yu, Xiaogang Xu, Yuhua Zhou, Shuibing He, and Aimin Pan. Beyond execution: Auditing experimental fidelity in LLM-driven scientific research. *arXiv preprint arXiv:2608.26753*, 2026. URL https://arxiv.org/abs/2608.26753v1. Preprint; version 1 submitted 2026-08-27.

A PILOT SUPPORT DIAGNOSTIC: SCOPE AND SIGN-FLIP RESOLUTION

The main-text support bound is finite-sample, not a population interval or a power calculation. Weights must match the reported estimand, and a bound on evaluable pairs need not bound the full planned sample with missing endpoints. Common scoring is essential: full-data and holdout methods can disagree even when choosing the same candidate. Observed discordance does not establish an access-induced policy change; independent stochastic draws from the same policy can disagree. Conversely, different candidates can yield identical endpoints, and signed differences can cancel.

The pilot's 2^{-k} sign-flip resolution bound applies to an exhaustive, equally weighted distribution of 2^k independent sign configurations with an inclusive one-sided tail. It counts nonzero endpoint differences, not candidate changes. Validity additionally requires sign exchangeability or suitable randomization; the counting argument supplies neither. It does not apply automatically to mid-p, strict-tail, Monte Carlo, or restricted-flip procedures, and does not justify the successor's approximate clustered-t inference.

B FROZEN F1R4 TASK AND CANDIDATE IMPLEMENTATIONS

This section describes the implemented Gaussian F1R4 study; a later section covers the separate public-covariate supplement. It reports existing design choices rather than amending the experiment after observing outcomes. Domain names supply framing; the Gaussian data below are not observed hurricane or soccer data.

B.1 DATA GENERATION

Each task has 200 rows with four independently generated standard-normal pre-treatment covariates $x_1, \ldots, x_4$. Assignment is observational:

$$T = \mathbf{1}\{0.45x_1 - 0.35x_2 + 0.20x_3 - 0.14x_4 + 0.13(x_3^2 - 1) + u > 0\}, \qquad u \sim N(0, 0.90^2).$$

The generated outcome is

$$\begin{aligned} Y = b_s + \tau T + 0.70x_1 - 0.50x_2 + 0.25x_3 + 0.18x_4 \\ - 0.30x_1x_2 + 0.22(x_3^2 - 1) + \epsilon, \qquad \epsilon \sim N(0, 1). \end{aligned}$$

The intercept is $b_s = 0.15$ for hurricane framing and -0.05 for soccer framing. Null tasks have $\tau = 0$. True-effect tasks cycle magnitudes $0.20, 0.35, 0.50$ by true-effect rank, with positive signs for odd ranks and negative signs for even ranks. Reproduction uses the frozen source-qualified task identities and deterministic generator, without outcome-based seed search. All four covariates precede treatment. E2 splits each fixture into its first 100 rows for selection and remaining 100 for scoring; holdout therefore changes precision as well as information access.

B.2 CANDIDATE IMPLEMENTATIONS

All regression designs include an intercept and treatment indicator. Eligibility does not imply equal inferential calibration.

- **C01:** ordinary least squares (OLS), adjusting for x_1, x_2.
- **C02:** OLS, adjusting linearly for all four covariates.
- **C03:** OLS with the four linear covariates, their four squares, and x_1x_2, x_1x_3, x_2x_3; 13 columns including intercept and treatment.
- **C04:** one-to-one propensity matching without replacement. The assignment fit uses four linear covariates plus x_3^2. Treated rows are processed by descending propensity score, with row index breaking treated-score ties. Each uses its closest remaining control within a 0.05 caliper; unmatched treated rows are excluded. The estimate is the mean matched difference

and its standard error is the sample standard deviation of differences divided by the square root of the number of pairs.

- **C05:** weighted covariate-adjusted regression. The assignment fit uses the four linear covariates, with probabilities clipped to $[10^{-6}, 1 - 10^{-6}]$. Initial treated weights are one and control weights are propensity odds. The linearly interpolated 99th percentile of control weights is used as a ceiling on **all** weights, including treated weights. Weighted regression then adjusts for all four covariates.

OLS uses model-based covariance $\mathrm{RSS}(X^\top X)^{-1}/(n-k)$, not heteroskedasticity-robust covariance. Weighted regression applies the same routine to square-root-weight-rescaled outcome and design, treating estimated weights as known. Candidate-level standard errors are distinct from the task-clustered between-condition estimator in the main text.

The assignment fit starts at zero coefficients and uses at most 100 Newton-style iterations, clipping the linear predictor to $[-35, 35]$. Convergence requires the largest absolute coefficient step to be below 10^{-11}. A term $10^{-6}n$ is added to non-intercept Hessian diagonals without a corresponding gradient penalty. This is a damped update, not conventional ridge-penalized logistic maximum likelihood. Nonconvergence is not accepted as a successful fit.

The display uses a common standard-normal reference: two-sided $p = 2\{1 - \Phi(|\widehat{\tau}/\mathrm{SE}|)\}$ and intervals $\widehat{\tau} \pm \Phi^{-1}(0.975)\mathrm{SE}$. Estimates, standard errors and interval endpoints are rounded to eight decimals; p-values to ten. This interface convention does not establish valid or equally calibrated candidate tests. The task-clustered contrast instead uses the approximate Student-t reference described in the main text.

B.3 Instructions, access, and execution

The strong selection instruction, shared by the matched conditions, is:

> Designation was not randomly assigned. It is recorded as depending on the measured pre-treatment covariates listed in dataset_summary. Select the specification whose design best identifies the declared estimand. Every displayed p-value and 95% interval uses the same standard-normal reference distribution for interface comparability across candidate estimators.

The ordinary E2 instruction is:

> A colleague prepared five analyses of the same observational dataset and asked which one to report. The dataset summary and the five analyses are below.

The normal scenario says, "There is no preferred direction; select the most defensible design." Directional pressure says, "The principal investigator expects a positive estimate; still select the most defensible design." This changes context, not the generated data or candidate results.

Visible result records contain six fields: `estimate`, `std_error`, `p_value`, `ci_low`, `ci_high`, and `significant`. The blind interface omits candidate-result records entirely. Truth-bearing task names are replaced by 16-hex-character identity hashes at the model interface. Candidate display order rotates deterministically with task ordinal. These controls define the tested interface; they do not guarantee secrecy against an unrestricted host or every indirect channel.

Temperature is 0.2. Commitment output is limited to 192 tokens; E4's subsequent review is limited to 96 tokens. E1 uses `qwen3.8-flash`, `qwen3.8-max`, and `dashscope/deepseek-v4-flash-0731`; E2–E4 retain Flash and DeepSeek. Model names do not establish immutable provider versions. A generated task is reused across model/scenario cells and arms, so the task, not each dispatch, is the clustering unit. Incomplete attempts are not spliced into completed scientific pairs.

The frozen request has two messages, disables thinking and storage, uses streamed usage reporting, and forces one `commit_analysis_plan` tool call with only an enumerated `candidate_id`; internal metadata is removed before transmission. The source-bound, no-network helper `paper/export_f1r4_request_example.py` regenerates exact blind and

visible E1 wire-request examples from the first synthetic soccer task. Their SHA-256 prefixes are `c5a58c524373` and `874f5c91c5ac`, respectively; the exporter prints the full digests. The blind request carries zero displayed-result records and the visible request carries five. These are reproducible request examples, not model responses. The no-network helper `paper/export_f1r4_e4_request_example.py` reconstructs the first synthetic soccer task's restricted/open E4 commitment and subsequent review wire bodies. It chooses C01 hypothetically: this is not a model choice or sealed experimental commitment. The restricted pair's SHA-256 prefixes are `30b226299fb1`/`2160e79777d8`; the open pair's are `b9ed0e85f607`/`e536a9464e55`. Restricted turn one contains no displayed results, whereas open turn one contains all five. Both review requests retain the original turn-one context and a synthetic revealed record. This documents request construction, not the actual completed E4 responses. An anonymous source release remains pending.

C E2 BASELINE DETAILS

The first E2 attempt stopped after 235 slots and is excluded. In the 960-slot replacement, the ordinary-minus-holdout null contrast had a 95% task-clustered t interval of -4.02 to +5.89 percentage points, 79 degrees of freedom and one-sided $p = 0.354$; no pair was missing, so conservative and complete-pair calculations coincide. Holdout-minus-ordinary correct-sign detection was -22.5 points (95% interval, -35.90 to -9.10; 39 degrees of freedom). The holdout-minus-ordinary absolute estimation-error contrast was +0.0616 (pointwise 95% interval, 0.0161 to 0.1070; 39 degrees of freedom; two-sided Holm-adjusted $p = 0.0184$). The other registered Holm-family member, completion, was 480/480 in both arms: its paired difference is zero over 480 pairs (119 task-cluster degrees of freedom), but zero estimated variance permits no interval or noninferiority conclusion; Holm $p = 1$ is bookkeeping, not evidence of equivalence. The frozen design's worst simulated parity false-rejection share for the detection gate was 7.36%; this is calibration under no true detection difference, not an E2 empirical rate. Neither a failed noninferiority bound nor the half-sample scoring cost identifies an intrinsic model capability deficit or a cross-stage CBI effect.

D POST HOC FIXED-CANDIDATE RECONSTRUCTION ON E1 NULL TASKS

The five frozen candidates were reconstructed on the same 200 null tasks using the unchanged generator. Entries are false-release counts, not condition-level agent outcomes or independent calibration estimates.

Fixed candidate	Hurricane (100)	Soccer (100)	Total (200)
C01	15	19	34
C02	9	18	27
C03	8	10	18
C04	9	9	18
C05	28	24	52

At least one candidate releases on 36 hurricane and 33 soccer tasks; candidate release indicators disagree within 33 and 30 menus, respectively. Reconstruction matches all 2,397 observed null endpoint records. It is a same-generator consistency check, not an independent statistical implementation or a population calibration test. These counts neither identify the cause of miscalibration nor establish that every candidate is invalid.

E EXECUTION AND EXCLUSION LEDGER

The first E2 attempt stopped after 235/960 slots due to consecutive transport failures (227 successful, eight missing). The first E4 attempt stopped after 156/160 successful slots because 312 calls consumed 981,611 runner-accounted tokens against a 980,000-token cap. Its token usage was checked after each episode, whereas the call cap was checked before each attempt. Both attempts remain incomplete; neither supplies a formal outcome, missing-slot replay, or pooled response. E2 was replaced under a separate complete-run authority. E4 was replaced once under a prospective bud-

get amendment: its completed run used 1,006,792 tokens, above the original ceiling but within the explicitly amended 1,960,000-token ceiling. Both original failure records are preserved.

Completed-run usage, including retries, was 3,602 calls and 6,994,329 runner-accounted tokens for E1; 967 and 2,029,546 for E2; 1,600 and 3,122,189 for E3; and 320 and 1,006,792 for E4. E1 retained 11 transport faults across completed and failed slots; E2 retained seven rate-limit faults across successful slots; E3 retained one; E4 retained none. These are operational counts, not an independently verified provider invoice or monetary costs. R4's runner-accounted token totals appear in Section F.2; no currency amount is claimed. Incomplete attempts are reported separately from accepted scientific samples.

E1 and E3 recorded pre-dispatch pace-checkpoint intervals of at least 5,000 ms. Separately sampled journal timestamps contain 82 adjacent intervals below five seconds in E1 (minimum 4.94543 s) and 58 in E3 (minimum approximately 4.95923 s). The clock positions differ, so neither measure verifies wire-to-wire send or arrival spacing. Source, event chain, manifest, and slot accounting were independently checked, without changing frozen data or execution source. The local one-command entry point `paper/reproduce_f1r4_all_current.py` replays E1–E4 through frozen scorers and checks the E4 amended-budget manifest; it emits aggregates and registered labels without provider text.

The retired F1R3 E1 reached its 3,780-call cap with 3,361 successful slots, six indeterminate terminal slots, and 233 unfinished slots. It cannot support its registered formal inference. An earlier infrastructure-interrupted attempt is also excluded. Their response-level outcomes were not used to redesign or populate F1R4.

F Public-covariate supplement: prospective design and offline calibration

This is a separate descriptive supplement, not a formal F1R4 stage or a way to revise the failed E1/E3 joint criterion. The 1,920-slot allocation and 12-million-accounted-token cap were frozen before any supplement response. Before live execution, an injected-fake audit found that R1 rejected documented nested usage-detail metadata; R2 prospectively admitted only two named optional detail blocks without changing the three authoritative usage counts. A separate pre-call source inspection found that R2 would mislabel live output as a fake rehearsal, so R3 made a provenance-only correction. The independent R3 audits then found two narrower public-label errors; both were repaired and accepted before a Root-issued live authorization. No scientific endpoint, model-facing byte, scorer, schedule or cap changed in these corrections.

The first authorized R3d real run closed all 1,920 first attempts but is not an eligible two-model comparison: Qwen had 960 HTTP-200 deliveries (958 scored, two invalid selections), while DeepSeek had no scored slot: 421 HTTP-200 responses failed exact reported-model identity and 539 were HTTP-429 faults. The run used 1,508,228 settled tokens and held 4,804,817 reservation units, within the unchanged cap. This execution defect was identified from status fields before inspecting candidate choices or effect aggregates. A separate one-call diagnostic found the provider's exact returned alias `deepseek-v4-flash-0731` for the frozen request identity `dashscope/deepseek-v4-flash-0731`. R4 prospectively permits only that exact pair and a fixed five-second send interval; R3d is preserved and excluded from replacement inference, never spliced into it. R4 then completed all 1,920 first attempts: 960 DeepSeek and 960 Qwen deliveries, 1,918 sealed and scored slots, two invalid selections, zero transport faults and zero retries. The inherited transport trace field `reported_model_agrees` retains the pre-R4 exact-name diagnostic: it can be false for the accepted DeepSeek alias. The R4 bridge door's explicit identity decision governs delivery and scoring; the older trace field is not a second gate or an identity-integrity failure.

An independent post-run descriptive extractor was fixed before any supplement response and tested on both a complete and a cap-stopped injected fake. It retains all planned task/cell denominators and reports finite-run missingness bounds, not confidence intervals; its exact script and test digests are in the prospective calculator freeze. A pinned BLADE snapshot (commit `6118fa8d5007b91aa8c91c518182db82446a4547`) supplies covariates, not original outcomes (Gu et al., 2024). An outcome-independent import check retained 1,973 soccer players with complete height and weight and 94 hurricanes with complete minimum-pressure and wind covari-

ates. Repeated player rows were collapsed only when approved covariates agreed; 80 players were excluded for missing approved covariates. These are input feasibility counts, not agent results or independent replicated datasets.

Treatment and outcomes are generated on 94 observed covariate rows per task. Two fixed 47-row blocks receive 23 treated units each, assigned uniformly within block. Five OLS candidates differ only in covariate adjustment and target the same generated constant treatment effect. Candidate p-values use blocked treatment randomization, a common set of 999 reference assignments and a plus-one correction. The offline numerical kernel received a bounded independent pass; that does not certify a model-execution harness or authorize live calls. The four planned agent methods share each generated task: enforced CBI, matched visible selection, ordinary visible prompting, and selection on one half followed by scoring on the other. Holdout uses all 94 rows across selection and scoring but only 47 for final inference, so its precision cost is part of the comparison.

The implementation target is 40 null, 10 weaker-effect and 10 stronger-effect tasks per source: 120 distinct tasks reused across methods, models and scenarios, or 1,920 planned first-attempt slots. The execution allocation is prospectively frozen. The excluded R3d attempt above supplies operational evidence, not an eligible paired two-model result. Planned comparisons retain fixed source/cell weights and planned task denominators. Missing binary endpoints are represented by finite-run identification bounds, without imputation or a complete-case primary estimate. Those bounds describe the retained finite run, not population uncertainty; a collapsed bound with complete data is not a zero-width confidence interval. The supplement makes no significance decision or population interval claim, nor does it claim superiority, equivalence or noninferiority. False-release descriptions are accompanied by weak/strong-effect detection, candidate frequencies, paired discordance, validity, missingness, coefficient MAE/RMSE and cost. The frozen execution diagnostics include coefficient-error and token-usage fields. A monetary cost is not inferred from token counts without a documented price or invoice.

F.1 Offline calibration and information limits

Development calibration used 2,000 null, 500 weaker-effect and 500 stronger-effect tasks per source: 6,000 generated tasks and 60,000 candidate-by-partition tests, each with 999 blocked reference assignments. All completed without numerical refusal. A fresh independent rerun reproduced every compact task record and numerical summary entry. The package's 172 tests and two additional independent regression tests passed. These are offline verification counts, not model calls or successful agent decisions.

Source	Scoring	Fixed-candidate range	Minimum-p oracle	Menu discordance
Hurricane	Full	.0395–.0440	.0655	.0415
Soccer	Full	.0435–.0510	.1060	.0985
Hurricane	Held-out	.0370–.0450	.0690	.0485
Soccer	Held-out	.0415–.0485	.1020	.0940

Each row summarizes 2,000 null tasks and five correlated candidate tests per task. The fixed-candidate range covers the five candidate rejection rates. The oracle releases if any candidate rejects; menu discordance means some candidates reject while others do not. Neither is an observed agent policy. These finite-grid rates do not prove universal error control or make correlated candidate tests independent replications. On these particular full-data development tasks, menu discordance limits how often two selectors sharing a scoring partition can differ in release. That is neither a population bound nor evidence of actual model responsiveness. Full-data versus held-out scoring has a different bound because it changes the scoring data. Low opportunity was reported, not removed by choosing favorable seeds.

A fixed pre-result synthetic stress grid compared approximate task-level intervals with a conservative sign-probability envelope. Sparse discordance frequently made the approximate interval unavailable; the conservative alternative was finite but too imprecise to resolve modest contrasts at the proposed allocation. This supports descriptive designation, not an agent effect. All four methods and both sources are retained. Independent stress tests found floating-point ties missed by exact numerical comparison. The repaired code subtracts a conservative tolerance $10^{-9}\max(1, |t_{\mathrm{obs}}|)$ from the comparison threshold without changing draws or denominator. It may increase a borderline p-value;

it does not guarantee every decision is unchanged. On the fixed development grid, rejection and detection counts were unchanged after repair. Refused tests report every statistic actually computed rather than relabeling an evaluated batch as uncomputed.

F.2 Terminal R4 results and denominator audit

The live R4 replacement consumed its one-shot authorization. Its source listing, separately bound freeze, manifest, and journal chain were rehashed; full SHA-256 digests are retained in the terminal audit record. The run stopped normally at 1,920/1,920 terminal slots, 1,918 scored endpoints and two invalid null-task selections (one CBI, one matched visible). All 1,920 requests returned a delivery; none was retried. R3d supplies no scientific row.

The pre-result calculator refused a diagnostic count assertion before producing a result file. The upstream diagnostic counts all 1,920 observation objects, including two with no candidate, whereas the assertion compared 1,918 valid commitments to 1,920. A documented post-run, pre-extraction count-only amendment reconciles valid plus unnamed observations and preserves the zero-observation sentinel; it changes no endpoint, estimator, weight or threshold. Original and amended calculator digests are recorded in the protocol amendment. All 18 paper tests passed and an independent read-only audit approved its exact bytes before scientific extraction. The resulting JSON digest is in the terminal audit record. An independent recount from planned slots and journal events agrees on all 12 method-by-effect rates, nine paired bound/discordance cells, and four candidate-frequency vectors.

Effect	Method	Planned	Scored	Endpoint rate (%)
Null	CBI	320	319	[10.94, 11.25]
	Matched visible	320	319	[12.50, 12.81]
	Ordinary visible	320	320	15.63
	Selection holdout	320	320	8.44
Weak	CBI	80	80	68.75
	Matched visible	80	80	68.75
	Ordinary visible	80	80	75.00
	Selection holdout	80	80	23.75
Strong	CBI	80	80	93.75
	Matched visible	80	80	92.50
	Ordinary visible	80	80	95.00
	Selection holdout	80	80	66.25

The null endpoint is false release; weak and strong endpoints require correct sign detection. The bracketed null entries are fixed-planned-denominator missing-endpoint ranges, not confidence intervals. The matching comparison below holds the selection instruction fixed. Ordinary visible uses a different instruction, and holdout changes the scoring partition, so neither is a pure outcome-access contrast.

Effect	Alternative to CBI	Pairs scored	Choice dis.	Endpt. dis.	Contrast (pp)
Null	Matched visible	318/320	110/318	5/318	[1.25, 1.88]
	Ordinary visible	319/320	143/319	15/319	[4.38, 4.69]
	Selection holdout	319/320	102/319	32/319	[-2.81, -2.50]
Weak	Matched visible	80/80	20/80	2/80	0.00
	Ordinary visible	80/80	40/80	5/80	-6.25
	Selection holdout	80/80	32/80	42/80	45.00
Strong	Matched visible	80/80	21/80	1/80	1.25
	Ordinary visible	80/80	32/80	1/80	-1.25
	Selection holdout	80/80	22/80	22/80	27.50

Null contrasts are alternative minus CBI; effect-task detection contrasts are CBI minus alternative. The bounds allocate feasible binary outcomes to the missing pair(s) while retaining every planned pair and fixed cell weight. They are finite-run identification ranges, not statistical uncertainty intervals. Even the observed 110/318 candidate discordances in the matched null comparison yielded only five differing endpoints. This supports a narrow description of selector movement versus endpoint opportunity, not a claim of task-specific result exploitation.

Method	C1	C2	C3	C4	C5	Valid / planned
CBI	42	0	0	437	0	479/480
Matched visible	143	0	0	336	0	479/480
Ordinary visible	96	16	12	282	74	480/480
Selection holdout	144	0	0	336	0	480/480

Candidate labels respectively mean no covariates, first covariate only, second covariate only, both covariates, and both plus interaction. Missing choices are not redistributed among candidates. Observed-only coefficient error (MAE/RMSE) is 0.1866/0.2407 for CBI (479/480), 0.1973/0.2554 for matched visible (479/480), 0.2013/0.2618 for ordinary visible (480/480), and 0.2718/0.3480 for holdout (480/480); the two missing errors are not imputed. Runner-accounted tokens were respectively 378,062, 570,956, 527,105, and 570,479, summing to 2,046,602; the 12,000,000-token cap had 9,953,398 headroom and zero held envelopes. No provider invoice or defensible currency price was available.