SEC CDP 001 Pilot Protocol

Draft v0.4.1 | 11 October 2026 | Companion to ER-13 Draft v0.6.4 | Package v1.3

NOT FINAL / NOT LOCKED / NOT OPEN FOR RECRUITMENT

This external pilot is intended to determine whether a controlled human review task is feasible and whether its measurement and analysis procedures work. It compares later flaw detection after exposure to conspicuous harmless errors with detection after corrected versions of those outputs. The settings below are concrete proposals for review. No participants have been enrolled and no behavioral outcomes are supplied.

1 Purpose and study status

The pilot estimates completion, timing, missingness, scorer agreement, hit rates and false alarms, and explores whether the intended mixed model is estimable. It is not powered or authorized to establish efficacy, deception, AGI behavior, human indispensability, or an Archive safety benefit. The crosswalk remains the evidence map; this protocol governs the proposed task once an execution-ready successor is reviewed and frozen.

Proposed enrollment is 60 randomized adults, approximately 30 per arm. This is a feasibility allocation chosen to expose workflow and measurement problems, not a power-derived sample size or assurance of precise variance estimates. Recruitment stops at 60 randomized participants; post-randomization withdrawal or technical failure does not trigger replacement. Report every randomized person and every missing item. Screening failures before randomization may be replaced and are counted separately. No internal efficacy stopping or sample-size adaptation is planned.

The pilot is external to any later confirmatory study. Pilot participants and pilot item instances will not be reused in the confirmatory dataset. Changes informed by the pilot require a new version, new frozen materials and an independently reviewed confirmatory sample-size justification.

2 Question and inference boundary

The question is whether prior exposure to conspicuous errors changes detection of harder flaws in later, separate outputs under this review policy. Exposure and Control differ in whether selected exposure outputs contain an obvious error. This contrast includes observable error presence, its salience, any confidence change, and the experience of finding errors. It cannot identify which of those mechanisms caused an effect. The challenge phase itself has 50% target prevalence in both arms and provides repeated experience of finding flaws. Results apply to this prevalence and may understate a short-lived carryover effect. Early, middle and late windows are reported separately as specified in section 8. Materials are labeled Fictional review material without AI attribution; reactions specific to AI-attributed outputs require a separately reviewed design.

The pilot does not test within-output co-occurrence. Post-exposure target items contain one intended target flaw, with no added conspicuous decoy. A conspicuous-error versus matched subtle-error experiment would require a new comparator and protocol. S19 model wrong-answer agreement is not an endpoint here, so its chance-agreement baseline is not required for this pilot.

3 Proposed settings

Setting	Proposal	Interpretation
People	60 randomized adults; 1 to 1 allocation	Feasibility proposal only; no confirmatory N

Setting	Proposal	Interpretation
Exposure	8 outputs, 30 seconds each	4 conspicuous-error outputs in Exposure; corrected counterparts in Control
Challenge bank	48 outputs viewed by every participant	24 target-flaw and 24 benign; same items in both arms
Strata	3 task families; 8 target and 8 benign items per family	Equal target and benign weights of one third per family
Challenge deadline	60 seconds per item	Voluntary final submission allowed; no revisiting
Session budget	About 65 minutes including instructions and debrief	Maximum 48 minutes for challenge items, plus exposure and setup
Feedback	No correctness feedback during exposure or challenges	Same response task and source labels in both arms
Materials	Frozen local non-adaptive item bank	Fixed bank retained; SEC-CDP-002 dynamic implementation is a separate proposed extension

4 Participants and arrangements

Proposed eligibility is age 18 or older, ability to read the study language and consent, and ability to use the tested desktop interface. Recruit adults who routinely read written instructions or reports, including a range of review experience. This convenience sample supports conclusions about recruited participants; population generalization requires a justified sampling frame. Do not recruit people whose employment evaluation depends on the investigator or their supervisor. This draft sets no minimum recruitment quota by expertise stratum. Under cyclic allocation of six lists, all arm-by-expertise-by-list cells are occupied if and only if each arm has at least six randomized participants in each included expertise stratum. Twelve total in a stratum is necessary but does not guarantee six per arm under an incomplete allocation block. Therefore the prespecified band contrasts may be unavailable; this is accepted in advance for this feasibility pilot. Missing outcomes can further prevent point estimation even when allocation cells are occupied; retain assigned-denominator bounds. Do not recruit extra participants beyond the cap or change strata after inspecting outcomes to recover a band contrast.

Full-bank availability: the standardized main summary requires at least one randomized participant per arm in every included expertise stratum with positive common weight. Its proposed bootstrap interval requires at least five randomized participants per arm in every such stratum. These are minimum cell-count conditions, not guarantees of point identification, precision or interval coverage; outcome availability and the section 8 analysis rules also apply. No recruitment floor is adopted, so either the main point summary or its interval may be unavailable. This possibility is accepted in advance. No alternative arm contrast, including an unstandardized randomized-arm difference, is selected as a fallback. Report cell counts, missingness and quantities defined under section 8 without changing strata, weights or the recruitment cap after inspecting outcomes.

Before randomization, record a pseudonymous participant ID, years of document-review experience, frequency of reviewing AI-generated text, and the first-attempt score on six task-comprehension questions. Retain first-attempt and final-pass scores separately. Do not collect names inside response logs. Participants must answer all six comprehension questions correctly after at most two instruction attempts; those unable to do so do not enter randomization. Record screening counts and reasons without publishing identifiers. The comprehension items must be benign and disjoint from experimental materials.

Use baseline review experience for blocked allocation: less than two years versus at least two years of weekly document review. Record the underlying answer before assignment. Within each stratum, use a concealed

computer-generated sequence of randomly selected blocks of four and six. The custodian holds the assignment key; participants, scorers, and the masked analyst do not receive it. Record allocation timestamp, software version, and sequence hash before first use. Do not create the live allocation sequence until execution approval.

A responsible investigator, recruitment channel, compensation amount, data custodian, retention period, consent text, and applicable ethics determination must be named before recruitment. The session uses fictional non-operational materials, yet informed participation, withdrawal, and debriefing arrangements still require review. This document does not determine exemption or issue ethics approval. Compensation and withdrawal-payment terms must be fixed before enrollment and identical across arms. Payment must not depend on speed, flag count or accuracy unless a separate incentive design is registered. A fixed payment can still reward shorter participation in practice; acknowledge this residual incentive and report actual review duration.

5 Materials and independent validation

Use three families: arithmetic consistency in fictional inventories, consistency with an explicitly supplied fictional rule, and internal consistency across a short fictional report. Each challenge is approximately 100 to 180 words, contains all information needed for review, and uses no real medical, legal, financial, security, or industrial-control decision. Target items contain exactly one intended flaw. Benign items meet the same rubric and have no known flaw. Build matched structural templates, but show only one variant from each template to any participant; a benign correction of a target shown to that person must not appear elsewhere.

Construct 24 independent target templates and 24 distinct benign templates, eight of each per family. The Exposure set uses eight additional templates not appearing in the challenge bank. Four have an obvious arithmetic or explicit-rule violation in Exposure and a corrected counterpart in Control; four are identical benign fillers. Match wording and display length as closely as possible and document unavoidable changes. Use the same neutral source label, Fictional review material, in both arms; do not assert actual AI provenance.

Before pilot enrollment, two reviewers who are not pilot participants independently inspect every item, verify its answer key and flaw count, and assign difficulty and salience ratings. A third reviewer adjudicates disagreements without access to any participant outcome. Replace ambiguous or multiply flawed items before freezing. Conduct a small separate usability check with people excluded from the randomized pilot, to verify that obvious exposure flaws are recognizable and challenge flaws are neither universally trivial nor effectively impossible. Record the check as material development, not efficacy evidence; freeze its criteria before collecting its results.

The complete bank, accepted localization aliases, scoring examples, interface build, display settings and file hashes must be frozen before enrollment. No validated bank is included in this draft. A static bank reduces versioning complexity, but hashes establish file identity rather than representativeness, concealment, or correctness. Keep answer keys inaccessible to participants and allocation inaccessible to scorers.

6 Procedure and blinding

After consent, baseline questions and comprehension, assign the participant. During exposure, show each of eight outputs for exactly 30 seconds. Participants can submit localized issue flags or choose No issue found, using the same interface in both arms. Exposure responses are logged only as manipulation/process measures and are not part of the primary challenge endpoint. No correctness feedback is given. At 30 seconds the next output appears regardless of whether a response was entered.

The challenge phase follows immediately, using the same neutral instructions. Present all 48 items once, with a maximum of 60 seconds each. A participant can submit one or more localized flags during review and can finish early with the Final submit button. Final submit displays a confirmation that the item will close and cannot be revisited. Timeout also closes the item. Use a monotonic client timer and server acknowledgments; retain timing diagnostics. Pause only between items, for at most a total of five minutes; longer interruptions are logged as

protocol deviations. Do not alter item deadlines after assignment. A Retract flag action is available until item closure: it records the flag ID and time without deleting the original submission or reversing the cumulative detection/false-alarm endpoint. No issue found is not a benign false flag. Study withdrawal has a separate control and confirmation. Define a trusted event-ordering rule and clock mapping before deployment; client timestamps alone cannot certify timely receipt. Participant instructions shall state: Flags accepted before item closure are recorded and scored even if retracted. Flags received after closure are recorded as late and are not scored. Include this wording in the consent/instructions and ethics submission, with comprehension checked before randomization.

Prepare six fixed order lists before enrollment. Each contains all 48 items and balances target and benign counts within position bands 1-8, 9-24 and 25-48: respectively 4/4, 8/8 and 12/12. For each item type, family counts within a list are a permutation of (2,1,1), the complementary (2,3,3), and (4,4,4); each family therefore retains eight items of each type overall. Rotate the extra early slot across families and lists. Require each item to appear once in the early band, twice in the middle band and three times in the late band across six lists, and avoid runs longer than three of the same type. These count constraints are arithmetically compatible, not a delivered or validated set of lists. Assign lists cyclically within arm and expertise stratum independently of outcomes; no item set is unique to an arm. Verify exact identity, quotas and order coverage before enrollment. Cyclic assignment can leave unequal list counts between arms. For the descriptive position-band contrasts, give all six lists equal weight within each arm and baseline-expertise stratum as specified in section 8; do not use raw list frequencies as weights.

After challenges, ask participants which exposure condition they believe they received, confidence in that guess, perceived error frequency, and perceived task difficulty. Do not show condition labels or directional hypotheses earlier. Explain the manipulation at debriefing. Guesses are diagnostics only and are never exclusion criteria, covariates in the primary effect model, or grounds to discard adverse results.

Scorers receive item ID, response and event times, but not exposure responses, allocation, assignment guesses or participant names. The analyst receives randomized labels A and B. Both condition labels and scorer decisions are locked before unmasking. The materials custodian may know allocation but must not adjudicate outcomes. Record any accidental unblinding.

7 Scoring and terminal outcomes

A valid target detection must identify the intended flawed statement or calculation and describe the inconsistency using the frozen rubric. A vague assertion that the output is wrong does not qualify. Two blinded scorers independently score each accepted flag; a third adjudicates disagreements. Preserve initial scores, adjudication, reason and rubric version. Report raw agreement and disagreement counts; any reliability coefficient is secondary and accompanied by its limitations.

Observed event	Target binary outcome	Timing and benign outcome
Valid target flag submitted before closure	Detection equals 1	First valid submission is the detection time; retain a later retraction separately
Early Final submit without valid detection	Detection equals 0	No detection through 60 seconds under the terminal policy; not early censoring
Deadline reached without valid detection	Detection equals 0	No event; administrative end at 60 seconds
Benign item completed early or at deadline	Not a target endpoint	False alarm equals 1 if any flag was accepted before closure; otherwise 0; multiple accepted flags count once

Observed event	Target binary outcome	Timing and benign outcome
Technical loss or withdrawal before terminal outcome	Missing unless a valid detection is already known	Retain a known target event or known benign false alarm; otherwise binary outcome unknown
Item never presented	Missing	Keep in assigned denominators for bounds; log reason

Any flag accepted before closure on a benign item counts as a false alarm, even if retracted; count retractions separately. On a target item, unsupported additional flags contribute to flag burden, not the benign false-alarm endpoint. Both retained and retracted target flags use the same first-valid-event rule. Draft text never submitted is not a flag. Explicit study withdrawal is distinct from Final submit and must use a separate interface action. Keep raw known-event Y/F fields distinct from binary_model_Y/F fields. All technical-loss or withdrawal records have the binary-model field missing even when a submitted event is known; bounds and timing retain the event. This is an analysis view, not deletion or rescoring of the raw response. Late flags are excluded from all detection, false-alarm, timing and accepted-flag-burden calculations; preserve them separately. They do not erase earlier accepted events or turn an otherwise complete item into missing data.

Deadline normalization: the reference scorer accepts a deadline closure log within an inclusive 0.001-second tolerance of tau and sets the canonical end to tau, retaining the raw logged end and a normalization indicator. A larger discrepancy is a timing error. This tolerance applies to closure logging only; it does not extend the flag deadline. Each flag input shall contain the authoritative boolean accepted_before_closure. True means the recorder accepted the flag before closure in event order; an exact-tau flag can qualify under that order despite equal displayed times. False identifies a late flag: retain the original record and exclude it from scoring without abandoning the item. An absent or non-boolean ordering value is an input error, not a late flag. Late flags require a finite nonnegative receipt time but need no target adjudication; retain receipt and event-order diagnostics even if their timestamps disagree. Accepted flags require valid adjudication where applicable and a timestamp within the canonical observation window; inconsistent accepted records remain input errors. Return received_flag_count, flag_count (accepted only), late_flag_count and late_flags (original records with input indices). The interface must establish ordering; the function checks the supplied assertion, not its authenticity.

For a terminal miss, the time-to-detection endpoint is undetected through 60 seconds, even when actual review time was shorter. Record actual review duration as a separate process measure. This is a task-policy estimand; it does not estimate a latent detection time under forced continued searching. Early technical loss without an observed event is censoring only under stated assumptions, not automatically an observed miss.

8 Estimands and pilot reporting

The crosswalk defines a marginal, stratum-weighted detection risk difference, Exposure minus Control, with a mandatory analogous benign false-alarm contrast. For this exploratory pilot, the main reporting target is the recruited randomized participants and the frozen finite bank, using one-third weight for each family. This is a deliberately narrower target than an unrestricted evaluator and item population. The signed contrast has the same form as Delta; it must be labeled Delta pilot finite bank and must not silently substitute for any later broader population estimand.

For complete data, compute each participant’s detection proportion over the eight target items in each family, then each arm mean within family, then the weighted arm difference. Use the corresponding eight benign items per family for false alarms. Report all family-specific counts and both overall contrasts. For blocked randomization, standardize both arms to the common empirical distribution of baseline expertise among all randomized participants rather than to their separate arm compositions. For each prespecified position band, first compute each participant's within-family/type rate using that person's assigned band count. Within each arm, expertise stratum and order list, average these rates. Average the six list-specific means with weight 1/6,

then apply the common family and expertise weights and take Exposure minus Control. Use the same nested weights for missing-outcome bounds, preserving assigned denominators. If any required arm-by-expertise-by-list cell is empty, the standardized band contrast is unavailable; report cell counts and raw descriptive rates without silently renormalizing or pooling. The full-bank primary summary keeps its existing weights; this list standardization applies to descriptive bands only. Bands have no selected interval or significance test in this pilot. For the full-bank main summary, if a required arm-by-expertise cell with positive common weight is empty, the standardized contrast is unavailable under this rule; do not silently pool strata. No replacement contrast is selected. Report the empty cells and raw descriptive rates where defined, keeping detection and false alarms together. The selected standardized identification-bound calculation also requires those cell means; if a required cell is empty, mark that composite calculation unavailable and retain any defined component bounds. This is a limitation of the selected procedure, not a claim that no other bound could be constructed. With the six-list cycle, the allocation-cell condition is at least six randomized participants per arm per included expertise stratum. Occupancy alone does not resolve missing outcomes. Recruitment does not guarantee that condition under section 4; unavailable band contrasts are a prespecified possible result, not a reason to change weights or drop strata.

When binary outcomes are unknown, report full-data bounds as the primary transparent missingness analysis: within each participant and family, the lower rate is observed successes divided by all eight assigned items; the upper rate adds unknown items to successes before division. Apply the same expertise and family weights. The contrast lower bound is Exposure lower minus Control upper, and the upper bound is Exposure upper minus Control lower. These are identification bounds, not confidence intervals. A known event stays a success. Report available-outcome proportions separately as descriptive sensitivity estimates; do not call them unbiased randomized effects without missingness assumptions. Available-outcome rates can be biased upward by retaining interrupted successes while omitting interrupted non-events. This selection can distort an arm contrast; no universal numerical bias is asserted. For position-band bounds use the actual assigned family/type count in that band, not eight; for full-bank bounds continue to use eight. Empty data cells do not justify excluding assigned participants or items.

Proposed sampling uncertainty uses 10,000 bootstrap resamples of evaluators within arm and expertise stratum, retaining each selected evaluator's full item vector and the original common expertise weights. Use the 2.5th and 97.5th percentiles for complete-data contrasts; for missing-data bounds report bootstrap uncertainty of each bound separately rather than interpreting their union as a validated 95% confidence region. This conditions on the fixed bank. Small expertise cells and interval coverage require simulation and statistical review before execution. If any required arm-by-expertise cell has fewer than five randomized participants, mark bootstrap intervals unavailable under this draft rule; report point summaries or identification bounds only when their own cell and outcome requirements are met. Five per arm per included stratum is a minimum reporting threshold, not evidence of adequate coverage. Unavailability is accepted in advance; no fallback interval or replacement arm contrast is selected. This proposed interval procedure concerns the full-bank pilot summary; it is not the confirmatory randomization-based implementation selected in crosswalk section 14.2. Position-band outputs remain descriptive without intervals under this draft.

No efficacy hypothesis is accepted or rejected on the basis of this pilot. Report two-sided intervals where validated, effect direction and uncertainty, false alarms, raw flag burden, completion time and missingness. No safety benefit follows from a higher hit rate alone. No meaningful-effect threshold is adopted here. Fix delta_star and its justification before pilot unmasking. If it remains unset, the confirmatory planning group receives only arm-pooled nuisance summaries until it is fixed; retain an access log. If that group has already seen arm effects, disclose this and do not claim effect-independent threshold selection.

9 Model feasibility and timing

Fit the v0.6.4 joint logistic mixed model only as an exploratory feasibility analysis after an executable implementation has been independently reviewed. Use condition, item type and their interaction; family effects and condition-by-family effects within item type; prespecified three-level position-band terms and condition-by-band terms within each item type; baseline expertise and first-attempt comprehension score. If a baseline covariate has zero variance, omit that constant term and record the omission before fitting; this is not outcome-driven variable selection. Do not add post-exposure guesses or confidence as adjustment covariates. Include evaluator intercept and item-type slope, plus item intercept and condition slope with separate covariance matrices for target and benign items. Use binary_model_Y/F, which masks every interrupted record. Label this a complete-item sensitivity fit, conditional on the completion process; it does not identify the full-data contrast without justified missingness assumptions. The primary transparent pilot reporting remains complete-data summaries where available and full-data identification bounds otherwise.

The proposed deterministic fitting sequence is: full correlated covariance structure, then zero covariance while retaining all variances and slopes. Do not silently remove item condition slopes or merge item types to obtain significance. Report singularity, optimizer diagnostics and boundary estimates for every attempt. If the second specification is unstable, report exploratory model summaries unavailable; finite-bank summaries and bounds remain reportable under their stated cell and missingness rules. The pilot may be too small to estimate the full model reliably, which is itself a feasibility finding.

Any integration over fitted evaluator and type-specific item distributions is a bank-referenced hypothetical-population sensitivity summary, even with a stable fit. It must use common baseline and position distributions in both arms; setting random effects to zero is not integration. The primary finite-bank report instead averages the actual bank using the declared family and expertise weights. No probability-sampled item population is created by fitting Gaussian random effects or by two-way resampling. A deterministic quadrature or Monte Carlo implementation must document tolerance and pass simulation checks before use. Broader generalization needs a defined sampling mechanism or explicitly defended transport assumptions, not fit stability.

For fully observed task-policy outcomes, construct detection curves and restricted mean time without detection through 60 seconds using all target items. A terminal miss contributes the full 60 seconds to time without detection, even if search stopped earlier. A valid detection contributes its event time. Separate actual time spent and stopping time summaries. With unresolved loss to follow-up, provide sensitivity bounds and an explicitly assumed censoring analysis; do not treat voluntary submission as loss to follow-up. Any timing model must retain evaluator and item dependence and be validated before use. Voluntary terminal misses are competing events, and detection-by-time is the cumulative incidence of detection before such closure. No hazard regression is selected for this pilot. A future hazard analysis must state whether it targets the cause-specific or subdistribution hazard and account for competing events and crossed dependence. Do not estimate detection incidence by independently censoring terminal misses in a Kaplan-Meier calculation. [M01]

10 Feasibility decisions and confirmatory planning

The following are proposed operational criteria, not empirically validated thresholds: at least 80% of randomized participants reach the final challenge; technical loss affects no more than 5% of assigned challenge items; at least 90% of accepted flags receive the same initial binary score from both scorers. Each family should avoid both an observed target detection rate below 10% and above 95% when arms are pooled, and benign false alarms above 30%. Report uncertainty and denominators for every criterion. Never use effect sign or statistical significance as a progression criterion. Use the observed point estimates for these provisional decisions, with equality passing at the stated limits; report uncertainty separately, not as an alternative decision rule. Undefined criteria, absent denominators or unusable measurements cannot pass. Let N_R be the actual number randomized, including withdrawals; completion uses N_R people, and technical loss uses 48 times N_R assigned

challenge slots, with withdrawal and unpresented items separately reported. Initial scorer agreement uses all accepted flags with two initial scores; report missing second scores, and any missing required scorer judgment prevents that criterion from passing. Pooled family detection and false-alarm criteria use full assigned denominators when complete; with unknown outcomes, require the entire identification interval to satisfy the relevant range. Report this as a conservative feasibility rule, not a confidence-bound test. No observed arm effect determines progression. The planned cap remains 60; early closure and its reason are reported as a deviation, not hidden by a planned denominator. If N_R is zero, all these rates are undefined and cannot pass. Full assignment of 48 slots per randomized participant remains the denominator even when some slots were never presented. Report late-flag counts separately; they require no adjudication and are outside the scorer-agreement denominator.

Manipulation and delivery reporting: distinguish technical display completion from observed flagging. Report the share of the four intended conspicuous errors correctly flagged in Exposure, with all four assigned slots per person as denominator and unknown outcomes bounded. Report false flags on benign exposure items: four fillers per Exposure participant and eight corrected/filler items per Control participant. Show item-level and participant-level distributions, actual exposure duration, retractions and technical losses. Low flagging indicates weak observed engagement and prompts design review; it does not prove that an error was unseen. No unvalidated numerical manipulation-success threshold is adopted, and these diagnostics cannot be used to exclude randomized participants.

A failed criterion triggers diagnosis and a new reviewed version before confirmatory work. Passing criteria permits planning, not automatic execution. An inability to fit the proposed model requires statistical redesign even if task completion is good. Report all adverse findings, including increased false alarms, shorter search with more misses, unblinding, and model failure. A prespecified criterion that cannot be evaluated also triggers diagnosis; passing this checklist is not proof of adequate precision. The 80% completion proposal permits substantial loss and potentially wide identification bounds; consider the achieved uncertainty before confirmatory design selection without altering the pilot rules retrospectively.

Confirmatory planning follows crosswalk section 14.4: hold the planned bank fixed within each design scenario, simulate actual allocation, order assignment, responses and missingness, and use the finite-bank randomization-based interval selected in section 14.2. Examine both effect directions, baseline hit and false-alarm rates, heterogeneity, evaluator counts and alternative prespecified bank designs; hold the selected bank fixed within each scenario. Calibrate to the marginal contrast and assess type-I error, coverage, power, unavailable estimates and Monte Carlo uncertainty for the actual primary pipeline. Model-based power applies only to the separately labeled optional model analysis. The pilot's exploratory evaluator bootstrap in section 8 is distinct from this confirmatory procedure. No such simulation has been run. Threshold selection and planning-team access follow section 8; study or target changes require a new frozen specification.

11 Records and reproducibility

Keep raw event logs append-only, with monotonic elapsed times, server receipt times, item and interface hashes, pseudonymous participant ID, masked arm, order-list ID, family, event type, submitted localization, finalization reason and technical-error code. Retain initial scorer judgments and adjudications in separate tables. Keep identity/contact records and the allocation key separate with access restricted to the named custodian. Define retention and sharing before consent; do not publish raw free text that contains identifiers. Preserve late flags with their ordering field, receipt time, original event/flag IDs and exclusion reason. The scoring schema records accepted_before_closure and scoring_exclusion_reason; no rejected late submission is silently deleted.

The companion settings JSON and blank CSV schemas express proposed fields only. The scoring reference code demonstrates the explicit outcome policy on labeled software fixtures; it is not the study interface, scorer, mixed-model implementation, or a behavioral dataset. Real classification still depends on a validated rubric and blinded scoring. SEC-CDP-002 v0.1.3 is a separate proposed infrastructure extension. The current pilot does not

adopt its dynamic generator, additional review streams or hardware profile. Adding them requires a protocol successor and new independent review; no engineering test is reported in this package.

12 Execution gates and next review

Gate	Current status	Required evidence
Responsible investigator and ethics	OPEN	Named lead; applicable determination; consent, debrief and withdrawal procedures
Recruitment and data governance	OPEN	Recruitment route, compensation, custodian and retention schedule
Validated material bank	OPEN	48 challenge items, 8 paired exposure outputs, scoring key and independent validation record
Interface and order lists	OPEN	Tested authoritative event ordering, 1 ms closure-log tolerance, retractions, terminal submission, loss logging and six verified order lists
Allocation preparation	OPEN	Concealed sequence generated only after approval; version and hash recorded
Analysis implementation	OPEN	Tested estimators, bootstrap and model diagnostics; coverage checks on specified designs
Independent statistical review	OPEN	Written assessment of estimands, missingness, feasibility N and analysis
Protocol registration and execution approval	OPEN	All preceding fields frozen in an execution-ready successor

The immediate review should determine whether the selected comparator answers the intended question, whether the proposed 60-person feasibility allocation is useful, and whether the measurement and missing-data rules are coherent. This draft authorizes no participant contact or Archive change. The next version should close execution gates rather than add unsupported maturity claims.

13 Source and document lineage

Primary evidence and full source register: Aegis Solis ER-13 External Empirical Evidence Crosswalk Draft v0.6.4, especially sections 13 and 14. Relevant evidence includes S27 Berbaum et al. for within-output search, S28 Wolfe et al. for prevalence and stopping behavior, and S29 Dietvorst et al. for reactions to algorithm error. These motivate competing mechanisms rather than establish this pilot's effect. S24-S26 motivate sample-size methodology without validating the exact pipeline.

Additional resampling reference: Owen AB (2007), The pigeonhole bootstrap, Annals of Applied Statistics 1(2):386-411, DOI 10.1214/07-AOAS122; https://arxiv.org/abs/0712.1111. Its crossed-data resampling results are methodological support under assumptions, not validation of this unexecuted protocol.

[M01] Austin PC, Lee DS, Fine JP (2016), Introduction to the Analysis of Survival Data in the Presence of Competing Risks. Circulation 133(6):601-609. DOI 10.1161/CIRCULATIONAHA.115.017719. https://pmc.ncbi.nlm.nih.gov/articles/PMC4741409/ . Supports distinguishing cumulative incidence and hazard targets; it does not validate the proposed crossed timing analysis.

Prepared with AI assistance for Aegis Solis. Concrete design settings are proposed implementation choices. Their appearance in this draft does not imply author approval of recruitment, completed ethics review, independent human review, or successful study execution.