AMENDGUARD: CERTIFY-THEN-DECIDE AMENDMENT TRIAGE FOR DRAFTED CLINICAL TRIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

An operationally attractive trial amendment can be scientifically inadmissible, while an admissible package can still be too weak to justify changing a drafted protocol. AmendGuard turns this distinction into a three-action structured decision: it searches coupled eligibility, visit-burden, and endpoint-specification edits, certifies searched packages against reconstruction, power, safety, and local-support gates, and surfaces the highest-utility package only when its lower-confidence benefit clears a separate intervention threshold; the other outcomes are protocol preservation and search-limited refusal. In method-blinded paired adjudication of 72 package, preserve, and refuse decisions, reviewers accept 59 AmendGuard outputs versus 47 retrieval outputs and prefer AmendGuard in 42 cases versus 12. On 117 historically amended trials and 39 matched no-amendment controls, AmendGuard attains 74.2% trial-macro directional concordance, surfaces packages for 3/39 controls, and has no surfaced package fail the encoded-gate recheck. Against decision-matched precedent retrieval, paired gains are 3.6 direction points (95% CI 0.8–6.7), 6.7 exact-match points (1.9–11.4), and 0.06 Jaccard overlap (0.02–0.10). Removing only the power gate gains 1.9 direction points but surfaces 19/156 packages that fail the locked power criterion; restricting the policy to any single edit family loses more than 21 points. The ordering persists in two disjoint later cohorts, where AmendGuard again records no encoded-gate violations.

1 INTRODUCTION

Suppose a trial is enrolling too slowly. Relaxing an eligibility threshold may enlarge the recruitable population, but it can also change event rates; removing visits may reduce burden, but it can alter missingness; moving endpoint collection may improve retention, but it can weaken power. A useful amendment assistant must therefore construct one coherent package while preserving the drafted estimand, hypothesis, analysis population, superiority or non-inferiority formulation, and minimum safety monitoring. It must also recognize that *no change* has two meanings: an eligible package may offer too little benefit to justify intervention, or the searched action set may contain no package that satisfies the scientific checks. Late amendments make these distinctions consequential because they add delay and expense (Getz et al., 2008; 2016; 2011; 2023; Botto et al., 2024) after execution problems are already visible.

AmendGuard represents this committee decision with two explicit boundaries. Its certification boundary searches coupled eligibility, visit-burden, and endpoint-specification edits and removes packages that fail design reconstruction, power preservation, safety commitments, or local empirical support. Its intervention boundary then asks whether the best surviving package has sufficient lower-confidence operational utility to surface. An eligible set can therefore yield PRESERVE; an empty eligible searched set yields REFUSE; and only an eligible, sufficiently beneficial package yields SURFACE. The three outputs expose whether a decision was blocked by a scientific predicate or by insufficient intervention value.

The closest clinical-trial systems solve adjacent prediction or generation tasks. Patient–trial matching, cohort-characterization, and eligibility models support trial cohort assessment (Liu et al., 2021; Hripcsak & Albers, 2013; Jin et al., 2024; Zhang et al., 2020); accrual and operations models support fixed trial designs (Bieganek et al., 2022; Topaloglu & Palchuk, 2018); AMEND++ forecasts future eligibility amendments (Das et al., 2026a); POET guides eligibility-criterion generation (Das et al.,

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.

2026b); and ClinicalReTrial iteratively redesigns failed protocols against an outcome-prediction environment (Xing et al., 2026). Adaptive designs govern randomization, sample size, or stopping during conduct (Pallmann et al., 2018; Thorlund et al., 2018; Chow & Chang, 2008). The unresolved object is a bounded late-amendment decision that couples several edit families, preserves the inferential design, and distinguishes intervention value from searched-set eligibility.

The evaluation tests that object through complementary evidence rather than one retrospective score. Method-blinded experts accept 81.9% of AmendGuard decisions versus 65.3% for retrieval and prefer AmendGuard 42-to-12. Historical packages then provide a stricter structural target: on 117 amended trials, AmendGuard reaches 74.2% trial-macro directional concordance and improves direction, exact identity, and Jaccard overlap over retrieval with paired intervals above zero. Matched controls test unnecessary intervention, the output audit tests execution of the encoded predicates, and two disjoint later cohorts test whether the same ordering survives temporal separation. The power-gate probe exposes the mechanism directly: a 1.9-point direction gain without that gate comes with 19 surfaced power violations.

The technical contribution is a certify-then-decide formulation for structured amendments, together with a concrete policy that combines verified protocol graphs, bounded multi-family search, uncertainty-penalized multi-outcome utility, and candidate-specific inferential checks. The accompanying evaluation aligns all implemented policies to the same package/preserve/refuse vocabulary and separates four claims: expert decision quality, historical package fidelity, restraint on no-amendment controls, and encoded-policy compliance. Together, these results show that package ranking and permission to intervene are distinct learnable objects in high-stakes structured decisions.

2 Related Work

Clinical-trial informatics spans protocol retrieval (Park et al., 2020), rule-based design review (Nammuni et al., 2004), eligibility representation (Kury et al., 2020; Dobbins et al., 2022), enrollment forecasting (Bieganek et al., 2022), and amendment prediction (Das et al., 2026a). POET generates eligibility criteria along clinician-selected semantic axes (Das et al., 2026b); ClinicalReTrial uses multi-agent search, safety-aware validation, and a learned outcome environment for de novo redesign of failed trials (Xing et al., 2026). The amendment decision studied here begins after an interventional design has been drafted: the estimand and inferential commitments remain fixed, edits may interact across three represented families, and the action vocabulary includes both preservation and finite-search refusal. This is a structured decision over a bounded package rather than a criterion-level suggestion or de novo protocol.

Selective prediction, structured abstention, and learning to defer trade coverage against error or route decisions to experts (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Garcia et al., 2018; Mozannar & Sontag, 2020; Xin et al., 2021). Multi-objective deferral controls expected system-level constraints (Charusaie & Samadi, 2024); risk-controlling and conformal structured prediction provide finite-sample guarantees for set-valued outputs (Bates et al., 2021; Zhang et al., 2025). In AmendGuard, the selective object is a searched structured action: domain predicates define which packages may compete, and a distinct benefit threshold determines whether the best eligible package should alter the protocol. Preserve and refuse consequently identify different states of the decision process. Calibrated regression (Guo et al., 2017; Kuleshov et al., 2018; Kompa et al., 2021) supplies uncertainty estimates for the conservative operational score, while protocol-specific predicates remain external to that score.

The three implemented controls span memory-based transfer, rule-constrained search, and learned operational prediction. Precedent retrieval maps edits from nearest historical protocols into the bounded candidate vocabulary; rule-constrained enumeration applies encoded design and operational rules; and the predictive utility ranker uses the same learned outcome contrasts as AmendGuard without its uncertainty-aware package objective. Each is evaluated natively and through the common certification boundary, separating ranking quality from access to AmendGuard's scientific predicates.

Table 1: Decision objects in clinical-trial amendment support. Implemented comparators share the package/preserve/refuse vocabulary, target-trial information cutoff, and temporal evaluation.

Policy	Coupled package	Preserve	Refuse	Certify	Package score	Temporal
Precedent retrieval (Park et al., 2020)	✓	✓	✓	–	✓	✓
Rule-constrained enumeration (Nammuni et al., 2004)	✓	✓	✓	partial	✓	✓
Predictive utility ranker	✓	✓	✓	–	✓	✓
Feasibility / prediction (Liu et al., 2021; Das et al., 2026a)	–	–	–	–	–	✓
POET / ClinicalReTrial (Das et al., 2026b; Xing et al., 2026)	partial	–	–	partial	✓	✓
AmendGuard	✓	✓	✓	✓	✓	✓

3 METHODS

3.1 TWO-BOUNDARY SELECTIVE DECISION RULE

For trial i, the input is the human-authored protocol and operational information available at its decision index. The output space contains exactly three mutually exclusive actions: *surface* a ranked bounded amendment package; *preserve* the current protocol when eligible searched candidates have low expected benefit; or *refuse* when no searched candidate passes the encoded scientific checks. Eligibility edits may change represented thresholds or categorical criteria; visit edits may change assessment count, spacing, or burden; endpoint-specification edits may change measurement windows, collection timing, missingness handling, or adjudication triggers while preserving the original endpoint scale, estimand, hypothesis, and analysis population. Candidates that alter those identities, minimum safety monitoring, or the superiority/non-inferiority formulation are rejected before ranking.

Let $\mathcal{C}_i$ be the bounded grammar-valid candidate universe and let $\mathcal{S}_i \subseteq \mathcal{C}_i$ be the canonicalized packages actually evaluated under the fixed search budget. The candidate boundary defines

$$\mathcal{E}_i = \left\{ \boldsymbol{\theta}' \in \mathcal{S}_i : r_i(\boldsymbol{\theta}') = s_i(\boldsymbol{\theta}') = q_i(\boldsymbol{\theta}') = 1,\ p_i(\boldsymbol{\theta}') \geq \max\{p_i(\boldsymbol{\theta}_{\text{orig}}), 0.80\} \right\}, \tag{1}$$

where r_i, s_i, and q_i denote successful design reconstruction, safety validation, and local-calibration support, respectively. Writing $\boldsymbol{\theta}_i^\star = \arg\max_{\boldsymbol{\theta}' \in \mathcal{E}_i} \underline{U}_i(\boldsymbol{\theta}')$ for the package with greatest lower-confidence utility, the intervention boundary is

$$\pi_i = \begin{cases} \textsc{Refuse}, & \mathcal{E}_i = \varnothing, \\ \textsc{Surface}(\boldsymbol{\theta}_i^\star), & \underline{U}_i(\boldsymbol{\theta}_i^\star) > 0.10, \\ \textsc{Preserve}, & \text{otherwise}. \end{cases} \tag{2}$$

Thus, searched-set eligibility means $\mathcal{E}_i \neq \varnothing$ and is independent of whether the intervention boundary ultimately selects surface or preserve. Refusal also applies when the relevant calibration stratum contains fewer than 20 reference trials. Reason codes follow the fixed precedence: unreconstructable design, safety violation, power violation, calibration insufficiency, historical-support insufficiency, then operational bottleneck. Because $\mathcal{E}_i$ is defined over $\mathcal{S}_i$, refusal has an operationally reproducible meaning: no package evaluated by the locked finite search passes the candidate boundary. It does not silently assert that every combinatorially possible edit in $\mathcal{C}_i$ has been exhausted. The 155/156 trials with at least one eligible searched package, one search-limited refusal, 39 control decisions, and zero surfaced encoded-gate violations therefore describe distinct stages of the policy.

Table 2: Cohort flow and overlap. Decision metrics use trial-level outputs; replay alone uses records.

Track	Size	Relationship and role
Core decision audit	156	117 amended + 39 matched controls; cross-fitted or temporally held-out outputs
Calibration hold-out	31	Latest 20% of core; excluded from model fitting
First later cohort	52	37 amended + 15 controls; disjoint from the core audit
Second later cohort	96	60 amended + 36 controls; disjoint from the core and first later cohort
Replay stress test	119→89	Core-linked records; 12 unstable pre-trends and 18 incomplete windows removed

3.2 Cohorts, decision targets, and chronology

The core corpus draws completed Phase I–III interventional trials from clinical-trial registries and governed sponsor repositories. It contains 156 trials across 23 countries and 8 therapeutic areas: oncology (n=34), cardiology (28), neurology (22), immunology (19), infectious disease (17), metabolic disorders (16), respiratory (12), and rare disease (8). The final decision audit comprises 117 historically amended trials and 39 matched no-amendment controls. Controls had at least one year of post-index follow-up without an in-scope protocol amendment and were frequency matched to amended trials on phase, therapeutic area, calendar period, site-count band, geography, and pre-index enrollment and screen-failure stress. The decision index is the information cutoff for the target output: the protocol version and operational snapshot available before the subsequent amendment-status interval. For amended trials, the reference package is the canonical difference between that indexed protocol and the first subsequent in-scope amendment. Each difference is mapped to signed eligibility, visit-burden, and endpoint-specification operators; logically identical edits collapse to one canonical identifier, and changes outside the bounded vocabulary do not enter direction, exact-match, or Jaccard targets. For controls, the target is no observed in-scope amendment during the one-year follow-up interval, making package surfacing the predeclared restraint endpoint.

All protocol fields, site trajectories, and operational covariates were truncated at each trial's index date; post-amendment fields, committee decisions, and downstream outcomes were unavailable to extraction, fitting, search, and baselines. The earliest 125 core trials formed the model-development era, with grouped five-fold cross-fitting producing out-of-fold predictions for decision evaluation; the latest 31 trials were withheld from fitting and used for temporal calibration assessment. The action taxonomy, utility transformation, candidate grammar, low-benefit rule, veto thresholds, search budget, and metrics were specified on 30 June 2022. The reported temperature-zero LLaMA-3 front end, available after the model's 2024 release, instantiates that fixed graph schema under deterministic validators. Baseline mappings and predictive models are fit only within their permitted development folds; the held-out target's amendment label remains unavailable until its output is fixed. A disjoint 52-trial later cohort (37 amended, 15 controls) tests the same decision rule, followed by a second disjoint cohort of 96 trials (60 amended, 36 controls). The 31-trial calibration set is a subset of the 156-trial core audit, whereas both later cohorts are mutually disjoint and disjoint from the core; the 89 replay-eligible records are a record-level subset linked to amended core trials and may include multiple endpoint windows from one trial. The split and bootstrap unit is the trial, so all lever decisions and records belonging to a target trial remain together. "Unavailable committee decisions" refers to the target trial: precedent retrieval may use amendment labels only from chronologically eligible source trials inside its training fold, never the held-out target's eventual action.

3.3 Protocol representation and feasibility models

We represent each protocol by a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes encode eligibility criteria, scheduled visits, and endpoint-specification fields and whose edges encode logical and temporal dependencies. A temperature-zero LLaMA-3 70B Instruct front end (Grattafiori et al., 2024), served with vLLM (Kwon et al., 2023), populates this graph; it is used only for structured extraction, never package ranking. Deterministic validators check units, threshold polarity, visit chronology, and endpoint-window consistency before labels are opened. Low-confidence fields

receive bounded clinical verification, and unresolved ambiguity invokes refusal. Protocols contain a median 42 nodes (IQR 28–61) and 67 edges (IQR 41–98).

The enrollment model minimizes log-normal negative log likelihood with a gradient-boosted ensemble (maximum 500 trees, depth 4, learning rate 0.03) over site activation, pre-index accrual, therapeutic area, phase, geography, and protocol-burden features. Screen failure combines criterion-level logistic losses with a Gaussian-copula correction learned from criterion rarity, exclusion overlap, and historical site pass-through. Dropout uses piecewise-exponential survival likelihood over visit density, procedure burden, follow-up duration, and time-varying burden. Missing continuous values are median-imputed within development folds with missingness indicators; categorical unknowns receive a dedicated level. Hyperparameters minimize grouped validation loss, and quantile regression fits conservative predictive intervals on out-of-fold residuals.

Candidate support is assessed locally, not inferred from aggregate MAE. Each candidate inherits a pre-specified phase $\times$ therapeutic-area $\times$ site-count reference stratum; adjacent strata are pooled by a locked hierarchy until minimum support is reached. The candidate passes only when the upper 95% bootstrap bound on leave-one-trial-out MAE is at most 6% for enrollment and 4% for screen failure. These empirical error checks define the local-support gate. Predictive bounds propagate into utility through the lower-confidence enrollment term and upper-confidence screen-failure and dropout terms; dropout uncertainty affects ranking conservatively but is not itself a hard local-error predicate.

3.4 Constrained package ranking and inferential vetoes

Each edit operator acts on a represented graph field: numerical or categorical eligibility criteria, visit count/spacing/burden, or endpoint timing/window/missingness/adjudication. A package is a canonical set of nonconflicting operators: no field receives two values, graph dependencies must remain valid, semantic no-ops are removed, edit identifiers are sorted, and duplicate edit sets collapse before scoring. For each operational outcome, the candidate contrast is obtained by evaluating the same trial and prediction horizon under the edited graph and the original graph, holding non-protocol decision-time covariates fixed. The predicted candidate-minus-original change is centered by the development-fold median, divided by its interquartile range, and mapped through $\tanh$ to $[-1, 1]$. This robust normalization makes equal weighting operationally meaningful across enrollment, screen-failure, and dropout units. Writing $\underline{\Delta\lambda}$ for the lower confidence bound on enrollment change and $\overline{\Delta f}$ and $\overline{\Delta D}$ for upper bounds on screen-failure and dropout change, the intervention score is

$$\underline{U}(\boldsymbol{\theta}') = w_1 \underline{\Delta\lambda}(\boldsymbol{\theta}') - w_2 \overline{\Delta f}(\boldsymbol{\theta}') - w_3 \overline{\Delta D}(\boldsymbol{\theta}'). \tag{3}$$

AmendGuard searches candidates according to the corresponding point utility

$$\max_{\boldsymbol{\theta}'} \; U(\boldsymbol{\theta}') = w_1 \hat{\lambda}(\boldsymbol{\theta}') - w_2 \hat{f}(\boldsymbol{\theta}') - w_3 \hat{D}(\boldsymbol{\theta}'), \tag{4}$$

subject to

$$\mathrm{Power}(\boldsymbol{\theta}') \geq \max\{\mathrm{Power}(\boldsymbol{\theta}_{\mathrm{orig}}), 0.80\}, \tag{5}$$

$$\boldsymbol{\theta}' \in \mathcal{C}_{\mathrm{safety}}. \tag{6}$$

We set $w_1 = w_2 = w_3$. Power is analytical for standard two-arm designs using the reconstructed analysis rule and its decision-time variance or event-rate inputs. Complex endpoint structures use 10,000 Monte Carlo replicates under the reconstructed original analysis rule; at power 0.80, the corresponding Monte Carlo standard error is 0.004. The safety set inherits minimum monitoring frequencies and contraindication exclusions from the drafted protocol, so a candidate may not relax those commitments. Endpoint edits are certified only when endpoint scale, observation window, variance or event-rate inputs, and the original analysis rule are reconstructable. Bayesian search has a budget of 200 unique packages per trial and uses seed 202701; trials whose bounded grammar produces fewer candidates stop after exhaustion. Across the core audit, canonicalization and deduplication leave 10842 evaluated packages, so $\mathcal{S}_i$ is the searched set rather than a claim of 200 realized packages for every trial. Ties resolve by larger minimum power margin, then fewer edits, then lexicographic edit ID.

3.5 Comparators and evaluation protocol

Precedent retrieval uses cosine similarity over protocol-graph counts, phase, area, site-count band, geography, and pre-index stress; the nearest 5 amended protocols vote on canonical edits, weighted by

Algorithm 1 AmendGuard certify-then-decide package policy

```
Extract protocol graph; verify low-confidence fields.
Generate, canonicalize, and deduplicate bounded candidate packages.
for each candidate in Bayesian-search order do
    Predict normalized utility and uncertainty.
    Apply design-reconstruction, safety, power, and local-calibration vetoes.
end for
if no candidate is eligible then
    RETURN REFUSE with the highest-precedence reason code.
else if best eligible package clears the low-benefit rule then
    RETURN SURFACE with package, margins, and reason trace.
else
    RETURN PRESERVE with low-benefit reason.
end if
```

similarity. The rule enumerator applies the same candidate grammar in fixed eligibility–visit–endpoint order and ranks packages by the number of satisfied operational rules minus boundary penalties. The predictive utility ranker is the equal-weight sum of clipped predicted enrollment gain, screen-failure reduction, and dropout reduction, with the same low-benefit threshold but without AmendGuard's confidence-bound objective. Native baselines use their own refusal logic; common-certification variants pass native rankings through AmendGuard's reconstruction, safety, power, and calibration checks. All mappings were fit only within training folds and frozen before held-out outputs.

Amended trial i contributes $s_i = m_i/|L_i|$, where L_i is its historically amended lever-family set and m_i counts signed family matches. Preserve and refuse actions score zero on amended trials. The headline directional concordance is the trial macro-average,

$$\text{Direction} = 100 \times \frac{1}{117} \sum_{i=1}^{117} s_i, \tag{7}$$

so it is not the pooled ratio of family-specific match counts. The family-specific denominators are 73 eligibility, 80 visit-burden, and 38 endpoint-specification decisions; a multi-lever trial can contribute to more than one family. Exact match compares the full canonical package; Jaccard overlap compares in-scope edit sets. Control package-surfacing uses the observed event proportion as the point estimate. Coverage is the fraction receiving surface or preserve rather than refuse, and surfaced violations count recommendations that fail the evaluation-reference veto audit. This audit repeats the encoded reconstruction, safety, power, and calibration predicates on the final surfaced package; zero violations therefore measures implementation compliance with the locked predicates, whereas blinded concern ratings test broader method-blinded expert judgment.

Confidence intervals use 1000 paired, stratified trial-level bootstrap replicates with seed 202703. Resampling is stratified by amended/control status and therapeutic area; all lever decisions from a sampled trial remain together; models are not refit because evaluation targets the locked policy. Percentile intervals are reported for marginal metrics and paired system-minus-baseline differences. For blinded adjudication, cases are the bootstrap cluster: 72 cases were fixed before scoring from the pooled core and two later cohorts, stratified by area, lever, control status, endpoint edits, calibration-tail status, and output availability. This action-enriched sample evaluates decisions and is not a prevalence estimate for any one cohort; its five AmendGuard refusals therefore do not contradict the single refusal in the 156-trial core audit. Two clinical reviewers and one statistician independently compared de-identified, left/right-randomized outputs under seed 202709; majority preference defines win/tie/loss, with adjudication after independent ratings. The display included the historical amendment as a separate reference, so this study tests method-blinded expert acceptability and comparative usefulness with committee precedent available.

Analyses ran on Ubuntu 22.04, Python 3.11, PyTorch 2.2.1, CUDA 12.1, and one NVIDIA A100 80 GB GPU. Bayesian search is stochastic only through its fixed seed; model fitting and baseline mappings are deterministic conditional on the grouped folds.

Table 3: Blinded paired adjudication on 72 fixed package, preserve, and refuse decisions. Acceptability intervals use the case-clustered bootstrap; concern counts denote cases with at least one independently flagged power or safety issue.

Outcome	AmendGuard	Precedent retrieval
Majority preference	42 wins, 18 ties, 12 losses	Reference
Acceptable decision	59/72; 81.9% [72.2, 89.6]	47/72; 65.3% [53.6, 76.0]
Power / safety concern	5/72	14/72
Actionable without revision	58/72	46/72

Table 4: Native policies in the matched action space. Direction, exact match, and Jaccard use 117 amended trials; "control surface" is the observed count on 39 matched no-amendment controls; refusal and encoded-gate violations use all 156 trials.

System	Direction (%)	Exact (%)	Jaccard	Control surface	Refuse / viol.
AmendGuard	74.2	31.4	0.62	3/39	1 / 0
Precedent retrieval	70.6	24.7	0.56	4/39	5 / 5
Rule enumerator	62.4	19.6	0.49	2/39	10 / 4
Predictive utility ranker	68.3	22.1	0.53	6/39	3 / 4

4 Results

The results follow the decision hierarchy. Blinded judgments test whether experts accept the complete output; historical amendments test package structure; controls test restraint; the gate audit verifies execution of the declared predicates; and disjoint later cohorts test temporal transport.

4.1 Blinded experts prefer the complete decisions

Across 72 action-enriched cases and 216 independent ratings, majority judgments favor AmendGuard in 42 cases, tie in 18, and favor decision-matched retrieval in 12. Reviewers deem AmendGuard acceptable in 59/72 cases (81.9%, 95% CI 72.2–89.6) versus 47/72 (65.3%, 53.6–76.0) for retrieval. They mark power or safety concerns in 5 versus 14 cases, and inter-rater agreement is $\kappa = 0.68$. Thus acceptability rises by 16.6 points, actionable-without-revision by 16.7 points, and concern incidence falls from 19.4% to 6.9%. These concern ratings evaluate broader clinical-statistical judgment than the encoded predicates, while the gate audit below verifies policy execution.

4.2 Historical decisions expose package fidelity

AmendGuard achieves 74.2% trial-macro-averaged signed lever-family concordance (95% CI 67.8–80.1) on the 117 amended trials. The descriptive family-level ratios are 78.1% (57/73) for eligibility, 71.3% (57/80) for visit burden, and 68.4% (26/38) for endpoint specification; Equation 7 explains why their pooled ratio is not the 74.2% macro-average. Exact identity and Jaccard then ask whether the entire coupled edit set, rather than only its signed families, agrees with the historical package.

The respective non-refusal coverages are 155, 151, 146, and 153 of 156 trials.

Against retrieval, paired gains are 3.6 direction points (95% CI 0.8–6.7), 6.7 exact-match points (1.9–11.4), and 0.06 Jaccard (0.02–0.10). Case-level package utility gives 51/38/28 wins/ties/losses. Direction gains are 11.8 points over rules and 5.9 over the predictive utility ranker; Supplementary Table S2 reports all intervals. Rules surface the fewest control packages, but the trade is an 11.8-point direction deficit, an 11.8-point exact-match deficit, 0.13 lower Jaccard, and four encoded-gate violations.

Common-certification variants isolate package ranking from gate access. Wrapped retrieval reaches 69.4% direction and 0.55 overlap with 8 refusals; wrapped rules reach 61.8% and 0.48 with 12; the

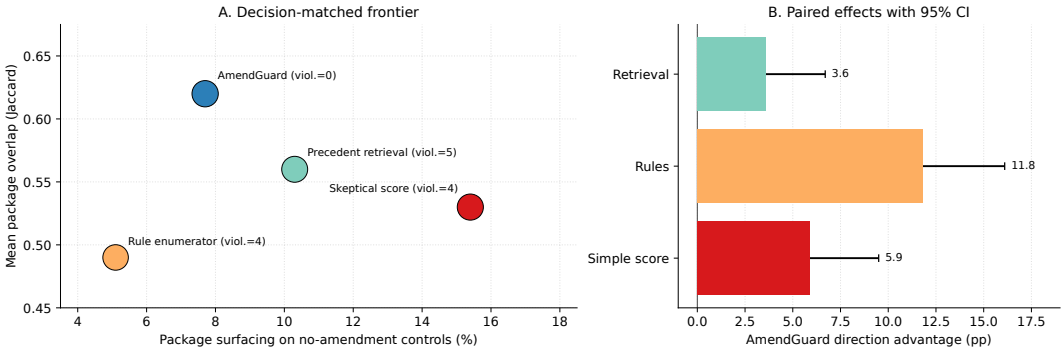

Figure 1: Matched decision frontier. Points plot package overlap against control-package surfacing; marker area encodes non-refusal coverage and labels report encoded-gate violations. Bars give paired direction advantages for AmendGuard with 95% trial-bootstrap intervals.

Table 5: Boundary activation in overlapping hard-case strata. "Filtered" is the percentage of searched candidates rejected by certification or retained below the intervention threshold; actions are surface/preserve/refuse. "No-power viol." counts violations after removing only the power gate.

Protected setting	Trials	Filtered	AmendGuard actions	No-power viol.
Endpoint specification	33	24.8%	24/8/1	5
Margin-sensitive design	21	31.2%	14/6/1	4
Fewer than 10 sites	26	28.7%	17/9/0	4
Calibration-tail error	8	46.4%	3/5/0	3
Closure / regulatory shock	17	41.1%	8/9/0	5

wrapped predictive utility ranker reaches 66.9% and 0.52 with 7. All wrapped policies satisfy the encoded gates, but each remains below AmendGuard in package concordance, locating the remaining difference in ranking quality.

4.3 The boundaries change which packages reach review

Among 10842 deduplicated searched packages, 614 first fail reconstruction or power, 321 calibration, 173 safety, and 96 endpoint support; another 1287 pass candidate checks but remain below the intervention threshold. The remaining 8,351 packages pass the listed candidate checks and threshold, before one highest-utility output is selected per trial. At the trial level, 155/156 core trials contain at least one eligible searched package, one produces search-limited refusal, and no surfaced package fails the evaluation-time recheck of the encoded predicates; the one-sided 95% exact binomial upper bound for a zero-event rate is 1.9%.

The power-gate removal pinpoints the accuracy–compliance trade. Direction rises from 74.2% to 76.1%, but 19/156 (12.2%) surfaced packages fail the locked power criterion. Restricting the action vocabulary to one lever family yields 41.2–52.8% direction; every restricted policy trails the coupled action space by more than 21 points. Feasibility-only scoring reaches 63.1% and raises the bootstrap-mean control-surfacing rate from 7.4% to 14.8%.

Boundary activation concentrates in difficult strata (Table 5). The sole core refusal is a margin-sensitive endpoint case counted in both rows; calibration-tail and closure/regulatory-shock cases are handled mainly by preservation. On the 31-trial temporal calibration hold-out, enrollment MAE is 4.3% (95% CI 3.1–5.6) and screen-failure MAE is 2.8% (1.9–3.8). Only 3/8 calibration-tail trials surface a package; the other 5 preserve.

Across 21 utility-weight settings, direction remains 72.8–75.6%, control surfacing remains 5.1–10.3%, and the selected-policy gates remain satisfied. The separate local-support sweep changes candidate eligibility: the strict setting gives 69.7% direction, 91.7% non-refusal coverage, and zero violations, whereas the permissive setting gives 76.0% direction, 99.4% coverage, and 4/156 violations (2.6%).

The intervention cutoff has different semantics: raising it to 0.20 lowers control surfacing to 2.6% and the overall surface rate to 94.2% while retaining 70.8% direction; non-refusal coverage remains 99.4% because below-threshold eligible packages become PRESERVE, not REFUSE. Accordingly, the intervention sweep is summarized by direction–surface-rate area: 0.694 (95% CI 0.651–0.731) for AmendGuard versus 0.635 (0.589–0.677) for retrieval.

4.4 DISJOINT LATER COHORTS RETAIN THE ORDERING

On the first disjoint later cohort, AmendGuard attains 69.1% direction (95% CI 58.8–78.4) on 37 amended trials, surfaces 2/15 control packages, refuses 3/52 trials, and records no encoded-gate violations. Retrieval reaches 65.4% with 3 violations, rules reach 59.2% with 1, and the predictive utility ranker reaches 63.8% with 3 violations. In the second disjoint cohort of 96 trials, AmendGuard records 67.8% direction, 0.59 Jaccard, 3/36 control surfaces, 2 refusals, and no encoded-gate violations; its paired direction advantage over retrieval is 4.3 points (95% CI 1.1–7.8).

The record-level replay diagnostic retains 89 of 119 historical records after excluding 12 unstable pre-trends and 18 incomplete windows. Screen failure decreases by a median 22.4% (IQR 11.7–34.8) after the observed aligned amendments; this historical consistency result concerns realized committee amendments, not counterfactual outcomes for AmendGuard-selected packages.

5 DISCUSSION

AmendGuard separates permission to compete from the value of intervention. Common-certification wrappers satisfy the same encoded predicates yet trail in package fidelity, showing that ranking remains consequential after gate access is equalized. Removing the power gate buys only 1.9 direction points at the cost of 19 noncompliant surfaces, while preservation absorbs many calibration-tail and operational-shock cases without converting low benefit into scientific refusal. This is why neither coverage nor agreement alone captures the policy.

The evidence hierarchy assigns a distinct interpretation to each result. The 59/72 versus 47/72 acceptability result measures method-blinded expert judgment on complete decisions; historical direction, exact match, and overlap measure correspondence to committee packages; controls measure restraint when no in-scope amendment was observed; and the output audit verifies execution of the encoded predicates. The five AmendGuard decisions with broader expert power or safety concerns do not fail those encoded checks; they identify review criteria beyond the policy's formal rule set and support presenting the gate trace alongside, rather than in place of, expert review. The two disjoint later cohorts preserve the comparative ordering across a separate temporal slice.

The evaluated regime is retrospective committee-facing support for completed Phase I–III trials in the represented therapeutic-area and site-count strata. Historical amendments provide reference packages rather than universal optima, no-amendment controls encode one year of observed restraint, and replay follows realized committee amendments rather than counterfactual AmendGuard actions. Refusal is deliberately scoped to the locked finite search, and the hard local-error predicates cover enrollment and screen failure while dropout contributes only to uncertainty-penalized ranking. Within this regime, the appropriate decision object is a bounded package for expert review with visible gate margins, utility bounds, protocol preservation, and search-limited refusal.

6 CONCLUSION

Clinical-trial amendment support is a selective structured-decision problem: the system must construct a coupled package, establish that it may be considered, and decide whether intervention is worthwhile. AmendGuard operationalizes those stages with separate certification and intervention boundaries. It improves every package-fidelity measure over decision-matched comparators, is preferred 42-to-12 over retrieval in method-blinded adjudication, and blocks the 19 power-noncompliant surfaces exposed by removing the power gate. The ordering persists in two disjoint later cohorts. The resulting review object carries its inferential margins and can recommend a package, preserve the drafted protocol, or state that the searched set contains no eligible action.

Ethics Statement

The retrospective study used completed-trial documents and de-identified aggregate operational inputs under governing data-use terms. Experts reviewed de-identified outputs without patient-level data, trial or care decisions, or separate compensation.

AI Use Statement

Generative AI served two disclosed roles. Within the method, temperature-zero LLaMA-3 70B Instruct structures protocol fields into the fixed graph schema; deterministic validators and bounded clinical verification check those fields before package search, and the model does not rank amendments. Separately, generative writing tools assisted manuscript language and presentation. This statement reports the documented uses of generative AI in the study and manuscript preparation.

References

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Cameron Bieganek, Constantin Aliferis, and Sisi Ma. Prediction of clinical trial enrollment rates. *PLOS ONE*, 17(2):e0263193, 2022. doi: 10.1371/journal.pone.0263193.

Emily Botto, Zachary Smith, and Kenneth Getz. New benchmarks on protocol amendment experience in oncology clinical trials. *Therapeutic Innovation & Regulatory Science*, 58(4):645–654, 2024. doi: 10.1007/s43441-024-00629-2.

Mohammad-Amin Charusaie and Samira Samadi. A unifying post-processing framework for multi-objective learn-to-defer problems. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23705–23755, 2024. doi: 10.52202/079017-0748. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2a5a41a536d3ada8fbf61a9d6fbf18d2-Abstract-Conference.html.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Shein-Chung Chow and Mark Chang. Adaptive design methods in clinical trials – a review. *Orphanet Journal of Rare Diseases*, 3:11, 2008. doi: 10.1186/1750-1172-3-11.

Trisha Das, Mandis Beigi, Jacob Aptekar, and Jimeng Sun. AMEND++: Benchmarking eligibility criteria amendments in clinical trials. *arXiv preprint arXiv:2601.06300*, 2026a. doi: 10.48550/arXiv.2601.06300.

Trisha Das, Katherine Kero, Dorinda Schumann, Tracy Ohrt, Sanjit Singh Batra, Gregory D. Lyng, and Robert E. Tillman. POET: Protocol optimization via eligibility tuning. *arXiv preprint arXiv:2602.00370*, 2026b. doi: 10.48550/arXiv.2602.00370.

Nicholas J. Dobbins, Tony Mullen, Özlem Uzuner, and Meliha Yetisgen. The leaf clinical trials corpus: A new resource for query generation from clinical trial eligibility criteria. *Scientific Data*, 9:490, 2022. doi: 10.1038/s41597-022-01521-0.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Alexandre Garcia, Chloé Clavel, Slim Essid, and Florence d'Alche Buc. Structured output learning with abstention: Application to accurate opinion prediction. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1695–1703, 2018. URL https://proceedings.mlr.press/v80/garcia18a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30*, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Kenneth Getz, Zachary Smith, and Marcy Kravet. Protocol design and performance benchmarks by phase and by oncology and rare disease subgroups. *Therapeutic Innovation & Regulatory Science*, 57(1):49–56, 2023. doi: 10.1007/s43441-022-00438-5.

Kenneth A. Getz, Julia Wenger, Rafael A. Campo, Edward S. Seguine, and Kenneth I. Kaitin. Assessing the impact of protocol design changes on clinical trial performance. *American Journal of Therapeutics*, 15(5):450–457, 2008. doi: 10.1097/MJT.0b013e31816b9027.

Kenneth A. Getz, Rachael Zuckerman, Anne B. Cropp, Anna L. Hindle, Randy Krauss, and Kenneth I. Kaitin. Measuring the incidence, causes, and repercussions of protocol amendments. *Drug Information Journal*, 45(3):265–275, 2011. doi: 10.1177/009286151104500307.

Kenneth A. Getz, Stella Stergiopoulos, Mary Short, Leon Surgeon, Randy Krauss, Sybrand Pretorius, Julian Desmond, and Derek Dunn. The impact of protocol amendments on clinical trial performance and cost. *Therapeutic Innovation & Regulatory Science*, 50:436–441, 2016. doi: 10.1177/2168479016632271.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

George Hripcsak and David J. Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013. doi: 10.1136/amiajnl-2012-001145.

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *Nature Communications*, 15:9074, 2024. doi: 10.1038/s41467-024-53081-z.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4:4, 2021. doi: 10.1038/s41746-020-00367-3.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7:281, 2020. doi: 10.1038/s41597-020-00620-0.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Ruishan Liu, Shemra Rizzo, Samuel Whipple, Navdeep Pal, Arturo Lopez Pineda, Michael Lu, Brandon Arnieri, Ying Lu, William Capra, Ryan Copping, and James Zou. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592:629–633, 2021. doi: 10.1038/s41586-021-03430-5.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Kushan Nammuni, Claire Pickering, Sanjay Modgil, Alan Montgomery, Peter Hammond, Jeremy C. Wyatt, Douglas G. Altman, Robert Dunlop, and Henry W. W. Potts. Design-a-trial: a rule-based decision support system for clinical trial design. In Max Bramer, Richard Ellis, and Ann Macintosh (eds.), *Applications and Innovations in Intelligent Systems XI*, pp. 3–17. Springer, London, 2004. doi: 10.1007/978-1-4471-0643-2_1.

Philip Pallmann, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16:29, 2018. doi: 10.1186/s12916-018-1017-7.

Junseok Park, Seongkuk Park, Kwangmin Kim, Woochang Hwang, Sunyong Yoo, Gwan-Su Yi, and Doheon Lee. An interactive retrieval system for clinical trial studies with context-dependent protocol elements. *PLOS ONE*, 15(9):e0238290, 2020. doi: 10.1371/journal.pone.0238290.

Kristian Thorlund, Jonas Haggstrom, Jay J. H. Park, and Edward J. Mills. Key design considerations for adaptive clinical trials: A primer for clinicians. *BMJ*, 360:k698, 2018. doi: 10.1136/bmj.k698.

Umit Topaloglu and Matvey B. Palchuk. Using a federated network of real-world data to optimize clinical trials operations. *JCO Clinical Cancer Informatics*, 2:1–10, 2018. doi: 10.1200/CCI.17.00067.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.84. URL https://aclanthology.org/2021.acl-long.84/.

Sixue Xing, Kerui Wu, Xuanye Xia, Meng Jiang, Jintai Chen, and Tianfan Fu. ClinicalReTrial: Clinical trial redesign with self-evolving agents. *arXiv preprint arXiv:2601.00290*, 2026. doi: 10.48550/arXiv.2601.00290.

Botong Zhang, Shuo Li, and Osbert Bastani. Conformal structured prediction. In *Proceedings of the Thirteenth International Conference on Learning Representations*, pp. 8563–8583, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1868a3c73d0d2a44c42458575fa8514c-Abstract-Conference.html.

Xingyao Zhang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. Deepenroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, pp. 1029–1037. Association for Computing Machinery, 2020. doi: 10.1145/3366423.3380181.

A SUPPLEMENTARY MATERIAL

A.1 CERTIFICATION-AWARE DECISION SPECIFICATION

Table S1 collects the action grammar, finite searched set, normalization, candidate boundary, and intervention boundary used by every AmendGuard evaluation. The key distinction is between the bounded candidate universe $\mathcal{C}_i$ and the canonical packages $\mathcal{S}_i$ actually evaluated under the locked 200-package budget: all reported refusal semantics refer to the latter.

Table S1: Locked action space, certification rules, and implementation settings used throughout the evaluation.

Setting	Fixed value or rule	Operational role
In-scope amendment space	Represented eligibility thresholds and categories; visit count, spacing, and burden; endpoint windows, timing, missingness, and adjudication triggers	Generates coupled packages while preserving endpoint scale, estimand, hypothesis, and analysis population
Forbidden amendment space	Estimand, endpoint family, superiority/non-inferiority hypothesis, analysis population, and minimum safety-monitoring changes	Rejects identity-changing candidates before utility evaluation
Utility normalization	Development-fold median/IQR scaling followed by $\tanh$; equal weights over enrollment, screen failure, and dropout	Places heterogeneous operational outcomes on a common bounded scale
Hard gates	Power $\geq \max\{\mathrm{Power}(\theta_{\mathrm{orig}}), 0.80\}$; safety set; local upper MAE bound $\leq 6\%$ enrollment and $\leq 4\%$ screen failure; minimum support 20	Screens searched candidates before final action selection
Two-boundary policy	Filter $\mathcal{S}_i$ through reconstruction, safety, power, and calibration; surface if the best eligible package has normalized multi-outcome utility lower bound > 0.10; preserve otherwise; refuse if no searched package is eligible	Separates searched-set eligibility from intervention value
Extraction verification	LLaMA-3-assisted structuring followed by deterministic validators and bounded manual confirmation before labels open	Resolves protocol semantics; unresolved fields trigger refusal
Search and tie break	Budget of 200 unique Bayesian evaluations, seed 202701; stop after grammar exhaustion; deduplicate canonical edit sets; then power margin, package size, edit ID	Defines the finite searched set and deterministic reported outputs conditional on the seed
No-amendment audit	39 matched controls with at least one year of post-index follow-up	Measures package surfacing when no in-scope amendment was observed
Compute environment	Ubuntu 22.04, Python 3.11, PyTorch 2.2.1, CUDA 12.1, one NVIDIA A100 80 GB GPU	Median runtime 4.7 minutes per trial (IQR 2.1–8.3); 14.2 GPU-hours for the core audit

This specification makes the two boundaries independently inspectable: a candidate can fail an encoded predicate, pass those predicates but remain below the intervention threshold, or survive both and compete for the single surfaced output.

A.2 Paired Comparative Effects

The paired intervals in Table S2 use identical sampled trials for both systems, so they test the within-trial ordering rather than comparing overlapping marginal intervals.

Table S2: Paired full-system minus baseline effects on the 117 amended trials. Intervals use the paired trial bootstrap. Win/tie/loss uses package utility with gate violations ordered below any certified output.

Baseline	Direction difference	Exact-match difference	Jaccard difference	Win/tie/loss
Precedent retrieval	3.6 [0.8, 6.7] pp	6.7 [1.9, 11.4] pp	0.06 [0.02, 0.10]	51/38/28
Rule enumerator	11.8 [7.6, 16.1] pp	11.8 [6.4, 17.3] pp	0.13 [0.08, 0.18]	62/32/23
Predictive utility ranker	5.9 [2.4, 9.5] pp	9.3 [4.2, 14.4] pp	0.09 [0.04, 0.14]	56/34/27

All three comparators have genuine case-level wins, but every paired interval for direction, exact package identity, and overlap remains above zero.

A.3 Common-Certification Baselines

Applying the same candidate checks to baseline rankings removes differences in access to the encoded gates. Table S3 therefore isolates whether each original ranking places high-concordance packages early enough to survive a shared boundary.

Table S3: Baseline rankings after the common reconstruction, safety, power, and calibration boundary. Control surface is the observed count on 39 matched no-amendment controls; refusal and violations use 156 trials.

Wrapped policy	Direction	Jaccard	Control surface	Refuse	Violations
Precedent retrieval + veto	69.4	0.55	3/39	8	0
Rule enumerator + veto	61.8	0.48	2/39	12	0
Predictive utility ranker + veto	66.9	0.52	4/39	7	0

The wrappers eliminate encoded-gate violations for every policy, while the direction and overlap ordering remains unchanged.

A.4 Blinded Decision Adjudication

The 72 cases were fixed from the pooled core and two later cohorts before review and balanced across area, lever family, no-amendment controls, endpoint edits, calibration-tail cases, and output availability. Each case presented the decision-time protocol summary, two randomized de-identified outputs, gate margins, and a separate historical-amendment reference. Two clinical reviewers and one statistician independently scored acceptability, power/safety concern, actionability, and preference before majority adjudication. The exercise evaluated de-identified system outputs, collected no patient-level data, and made no clinical decisions; reviewers completed it within their professional evaluation roles without separate compensation. Because committee precedent was available as a separate reference, this experiment measures method-blinded expert acceptability and comparative usefulness under that information condition; it is distinct from the mechanical encoded-gate recheck.

Table S4: Adjudication by archived output-availability stratum. "Non-refusal" comprises either a surfaced package or preservation; reviewers saw the actual three-action output. The action-enriched sample is not a prevalence estimate for any one cohort.

Output relationship	Cases	AmendGuard wins/ties/losses	Interpretation
Both non-refusal	60	37/12/11	Direct comparison of the displayed package or preserve decisions
AmendGuard non-refusal; baseline refuses	7	5/1/1	Additional eligible coverage is usually preferred
AmendGuard refuses; baseline non-refusal	3	0/3/0	Reviewers accept all three refusals as safety-preserving
Both refuse	2	0/2/0	Neither finite search finds an eligible package

Five sampled AmendGuard refusals are possible because the pooled source contains six total AmendGuard refusals and the design deliberately enriches rare output relationships.

A.5 Second Disjoint Later Cohort

The second later cohort is disjoint from the core audit and the first later cohort and uses the same reported decision configuration. It contains 60 historically amended trials and 36 matched no-amendment controls.

Table S5: Second disjoint later cohort. Direction and Jaccard use 60 amended trials; control surfacing uses 36 matched controls; refusal and encoded-gate violations use all 96 trials.

Policy	Direction	Jaccard	Control surface	Refuse	Violations
AmendGuard	67.8	0.59	3/36 (8.3%)	2	0
Precedent retrieval	63.5	0.53	5/36 (13.9%)	5	4
Rule enumerator	57.6	0.47	2/36 (5.6%)	11	2
Predictive utility ranker	61.9	0.50	7/36 (19.4%)	4	3

Relative to retrieval, the paired confirmation differences are 4.3 direction points (95% CI 1.1–7.8) and 0.06 Jaccard (0.02–0.10). Differences to rules are 10.2 points (5.8–14.9) and 0.12 Jaccard (0.07–0.17); differences to the predictive utility ranker are 5.9 points (2.1–9.8) and 0.09 Jaccard (0.04–0.14). AmendGuard also retains enrollment MAE 5.2% (95% CI 4.4–6.2) and screen-failure MAE 3.4% (2.6–4.4).

A.6 Hard-Case Quality and Systems Evidence

The hard-case strata overlap by design and identify settings in which calibration support, endpoint reconstruction, or power margins are most likely to bind. Table S6 reports package quality inside those strata rather than only the number filtered.

Table S6: Full-system decision quality within overlapping hard-case strata. Direction and Jaccard use amended members; violations are surfaced packages failing the encoded evaluation-reference gates.

Stratum	Trials	Direction	Jaccard	Violations
Endpoint specification	33	65.8	0.57	0
Margin-sensitive design	21	66.2	0.58	0
Fewer than 10 sites	26	62.9	0.54	0
Calibration-tail error	8	56.3	0.48	0
Closure / regulatory shock	17	59.7	0.51	0

The evidence matrix connects each encoded protection to its stressor and evaluation view. It separates a predicate's execution cost and audit trace from the method-blinded expert judgments reported in the main paper.

Table S7: Certification evidence matrix connecting protected objects, stressors, enforced properties, matched comparators, temporal tests, costs, and manuscript evidence.

Protected object	Stressor	Enforced property	Comparator	Temporal / hard-case test	Cost and evidence trace
Original power floor	Missing inputs; margin-sensitive or endpoint edits	Reconstructable power at least the original or 0.80	Retrieval, native and wrapped	Margin-sensitive and endpoint strata; both later cohorts	0.8 median minutes; Methods and Tables S6, S16
Safety rules	Unsafe exclusions; reduced monitoring	Immutable minimum monitoring and exclusion rules	Ungated policy	Safety-rule perturbations in the core candidate audit	173 first-failing rejections; Tables S1, S16
Locally supported surfacing	Thin support, closures, regulatory shocks	Local upper-MAE bound and minimum reference support	Predictive utility ranker	Tail-error and shock strata; 52- and 96-trial cohorts	3/8 tail cases surface; main hard-case table and Table S6
Protocol semantics	Malformed fields; ambiguous endpoint timing	Deterministic validation plus bounded confirmation	Unverified extraction	Endpoint and reconstruction audits before label access	64/156 confirmed; Tables S1, S16

Power, safety, calibration support, and protocol semantics are therefore not collapsed into one generic abstention score; each produces a distinct reason trace.

A.7 Boundary Sensitivity and Direction–Surfacing

The three sweeps answer different questions and are interpreted separately: utility weights change ranking, calibration bounds change candidate eligibility, and the low-benefit cutoff changes intervention.

Table S8: Representative settings from the utility, local-support, and low-benefit sweeps. Direction, control surfacing (CS), non-refusal coverage, and encoded-gate violations are percentages; Jaccard is unitless. The low-benefit row additionally reports overall surface rate because that threshold cannot change searched-set eligibility.

Setting	Direction	Jaccard	CS (%)	Coverage (%)	Viol. (%)
Nominal equal weights	74.2	0.62	7.7	99.4	0
Enrollment-heavy weights	75.6	0.60	10.3	98.7	0
Screen-failure-heavy weights	73.4	0.63	5.1	97.4	0
Dropout-heavy weights	72.8	0.61	7.7	98.1	0
Strict calibration	69.7	0.58	5.1	91.7	0
Permissive calibration	76.0	0.63	12.8	99.4	4/156 (2.6%)
Low-benefit cutoff 0.20	70.8	0.60	2.6	94.2 surface; 99.4 coverage	0

Across the intervention-threshold sweep, direction–surface-rate area integrates trial-level direction concordance over the fraction of trials receiving a surfaced package, with larger values preferred. AmendGuard reaches 0.694 (95% CI 0.651–0.731), compared with 0.635 (0.589–0.677) for precedent retrieval. Non-refusal coverage is invariant to this threshold because eligible below-threshold packages map to PRESERVE; only the local-support sweep can change searched-set eligibility.

A.8 MECHANISM DECOMPOSITION

Table S9 probes the power gate and the performance consequence of restricting the action space to one edit family on the same 156 trials. Removing the power gate trades a 1.9-point direction increase for 19 surfaced power violations, while restricting the action vocabulary to any single lever family loses more than 21 direction points. The single-lever rows are policy restrictions; together they show that each one-family action space misses substantial structure present in historical multi-family packages.

Table S9: Paired internal probes. Power status distinguishes refusal from a surfaced encoded-gate violation. Control-surfacing cells report observed counts and proportions before labeled bootstrap-mean diagnostics.

Configuration	Dir. agree.	Power / refusal status	Enroll. Δ	Control surface
Full AmendGuard	74.2%	1/156 refuse; 0 viol.	+16.8%	3/39 = 7.7% obs.; 7.4% boot.
− Power gate	76.1%	19/156 surfaced viol.	+22.4%	3/39 = 7.7% obs.; 8.1% boot.
Single-lever: eligibility only	52.8%	2/156 refuse; 0 viol.	+9.1%	4/39 = 10.3% obs.; 11.2% boot.
Single-lever: visit burden only	45.6%	1/156 refuse; 0 viol.	+7.3%	5/39 = 12.8% obs.; 12.6% boot.
Single-lever: endpoint specification only	41.2%	3/156 refuse; 0 viol.	+5.8%	5/39 = 12.8% obs.; 13.9% boot.
Feasibility only	63.1%	1/156 refuse; 0 viol.	+14.2%	6/39 = 15.4% obs.; 14.8% boot.

Feasibility-only scoring further shows that an intervention threshold matters: it loses 11.1 direction points and roughly doubles the bootstrap-mean control-surfacing rate.

A.9 Chronology and Evaluation Tracks

Table S10 separates the policy-specification date from the later extraction implementation and makes cohort overlap and target-label cutoffs explicit. The target trial is always isolated from its own eventual committee decision; retrieval can access only chronologically eligible source-trial amendments inside the training fold.

Table S10: Evaluation chronology, cohort overlap, and information availability.

Track	Configuration	Separation guarantee
Core decision audit	156 trials: 117 amended and 39 controls; earliest 125 use grouped five-fold cross-fitted predictions, latest 31 are excluded from model fitting	Target-trial amendment labels and all post-index fields are unavailable to every policy
Temporal calibration	Latest 31 core trials; 20% temporal split	Evaluates ingredient-model calibration with no target-trial fitting
Policy specification	30 June 2022	Fixes the action taxonomy, normalization, candidate grammar, intervention and veto thresholds, search budget, seeds, and metrics
Extraction implementation	Temperature-zero LLaMA-3 70B Instruct, released in 2024, followed by deterministic validators and bounded confirmation	Instantiates the graph schema fixed by the 2022 policy specification
Fold-local fitting and output lock	Predictive models and baseline mappings use only their permitted development folds	The held-out target amendment remains unavailable to extraction, fitting, search, and baselines until the target output is fixed
Matched comparisons	Full system, native baselines, wrapped baselines, and ablations on identical held-out outputs	Paired trial resampling preserves multi-lever dependence
Blinded adjudication	72 fixed cases pooled from all three evaluation cohorts; 216 ratings	Action-enriched sampling precedes judgments; system identity and left/right order are concealed
Later-cohort evaluation	A 52-trial cohort followed by a second cohort of 96 trials	Cohorts are mutually disjoint and disjoint from the core audit; both use the reported decision configuration
Replay diagnostic	89 records from 119; 12 unstable pre-trends and 18 incomplete windows removed	Record-level historical-consistency analysis separated from trial-level decision metrics

The two later cohorts remain mutually disjoint and disjoint from the core audit, while replay alone uses record-level windows linked to core trials.

A.10 Metric Mechanics and Bootstrap

Table S11 maps every recurring metric to its scoring unit and action handling. This prevents family-level descriptive counts, trial-level macro-averages, and record-level replay quantities from being combined as if they shared a denominator.

Table S11: Metric definitions, denominators, and abstention handling.

Metric	Scoring unit / denominator	Mechanical rule
Direction concordance	Trial macro-average over 117 historically amended trials	Trial i receives the fraction of its historically amended signed lever families matched by the output; preserve and refuse receive zero. The headline is $100(1/117)\sum_i s_i = 74.2\%$. Descriptive family ratios are 57/73 eligibility, 57/80 visit burden, and 26/38 endpoint specification, so their pooled ratio is not the macro-average.
Control surfacing	Matched no-amendment trial, n=39	Any surfaced package is an event; preserve and refuse are non-events. The point estimate is 3/39 = 7.7%; 7.4% is the bootstrap mean. The label records observed amendment status, not counterfactual necessity.
Searched-set eligibility	Trial, n=156	A trial has at least one evaluated package satisfying reconstruction, safety, power, and local-calibration checks; 155 do and one does not. Eligibility does not imply that the final action surfaces a package.
Final action and violation	Trial, n=156	Surface requires benefit above threshold; otherwise preserve. Refuse requires no eligible package in the searched set. A violation is a surfaced package that fails an encoded evaluation-reference gate; none occur for AmendGuard.
Package agreement	Amended trial, n=117	Exact match compares canonical edit sets; Jaccard overlap is intersection over union. Preserve and refuse receive zero.
Replay eligibility	Replay record, starting n=119	Exclude 12 unstable six-month pre-trends and 18 incomplete pre/post windows, leaving 89 records; multiple windows may map to one amended trial.
Uncertainty	Trial bootstrap	1000 paired replicates, stratified by amended/control status and area; all levers from a sampled trial stay together; locked models are not refit; percentile 95% intervals.

All bootstrap intervals preserve the trial as the dependency unit even when one sampled trial contributes multiple lever-family decisions.

A.11 Software and Data Environment

Table S12 records the exact software and data environment used by the study. The extraction model is restricted to protocol structuring; all package ranking is performed by the explicitly described feasibility and utility models.

Table S12: External software, compute, and data assets used in the study.

Asset	Version	Role in study	Upstream terms / status
LLaMA-3 (Grattafiori et al., 2024)	70B Instruct	Structured protocol-field extraction only; no amendment ranking	Llama 3 Community License
vLLM (Kwon et al., 2023)	0.4.3	Deterministic extraction serving at temperature zero	Apache-2.0
Python	3.11	Experiment orchestration and figure generation	Python Software Foundation License
PyTorch	2.2.1	Modeling and GPU execution runtime	BSD-style upstream license
CUDA	12.1	GPU compute runtime	NVIDIA CUDA Toolkit EULA / proprietary upstream terms
Matplotlib	3.10.8	Manuscript figure rendering	Matplotlib License
Clinical trial registries and sponsor repositories	Study inputs	Protocol documents, amendment histories, and de-identified operational aggregates	Registry terms of use and governing data-use agreements

Protocol and de-identified aggregate operational inputs were used under the registry terms and governing data-use agreements listed in the final row.

A.12 Matched Native Baselines

Table S13 evaluates native comparator policies under the shared package/preserve/refuse action space. Table S3 then applies the common veto to every ranking to isolate ranking quality from gate access. Low-impact counts use the enrollment-only $< 2\%$ diagnostic; a package can still clear the multi-outcome utility threshold through screen-failure or dropout benefit.

Table S13: Native decision-matched baselines. No-amendment control cells lead with observed package-surfacing counts and rates; bootstrap means and intervals follow.

Configuration	Direction (117 amended)	Control surfacing (39 controls)	Refusal, coverage, violations, low-impact (156 trials)
Full AmendGuard	74.2% (95% CI 67.8–80.1)	3/39 = 7.7% observed; 7.4% bootstrap mean (95% CI 3.8–12.6)	1 refuse; 155 coverage; 0 violations; 14 low-impact surfaced
Precedent retrieval	70.6% (95% CI 63.1–77.8)	4/39 = 10.3% observed; 10.6% bootstrap mean (95% CI 5.6–17.3)	151 coverage: 146 valid + 5 violations; 5 refuse; 21 low-impact
Rule-based governed enumerator	62.4% (95% CI 54.1–70.8)	2/39 = 5.1% observed; 5.9% bootstrap mean (95% CI 2.4–11.3)	146 coverage: 142 valid + 4 violations; 10 refuse; 17 low-impact
Predictive utility ranker	68.3% (95% CI 60.4–75.7)	6/39 = 15.4% observed; 15.1% bootstrap mean (95% CI 9.0–22.8)	153 coverage: 149 valid + 4 violations; 3 refuse; 26 low-impact

AmendGuard improves all three package-concordance measures over the implemented baselines while retaining 155/156 non-refusal coverage and no encoded-gate violations.

A.13 WHOLE-PACKAGE AGREEMENT

Whole-package scoring tests coupled construction directly. AmendGuard leads retrieval by 6.7 points in exact match and 0.06 Jaccard, with paired intervals excluding zero.

Table S14: Whole-package agreement on the amended-trial cohort.

Configuration	Exact package match (117 amended trials)	Mean package overlap (Jaccard over in-scope surfaced edits)
Full AmendGuard	31.4% (95% CI 23.5–39.6)	0.62 (95% CI 0.55–0.69)
Precedent retrieval	24.7% (95% CI 17.8–32.2)	0.56 (95% CI 0.48–0.63)
Rule-based governed enumerator	19.6% (95% CI 13.3–26.8)	0.49 (95% CI 0.42–0.56)
Predictive utility ranker	22.1% (95% CI 15.6–29.4)	0.53 (95% CI 0.46–0.60)

The exact-match endpoint is deliberately strict; Jaccard shows that the advantage also holds when historically defensible packages differ by a subset of edits.

A.14 FIRST DISJOINT LATER COHORT

Table S15 applies the reported policies to 52 later trials using only decision-time information. AmendGuard retains the leading direction score, high coverage, and zero encoded-gate violations.

Table S15: First disjoint later-cohort evaluation. Control-surfacing cells state observed counts before bootstrap means; status partitions final actions into non-refusal valid/refuse/surfaced encoded-gate violation and is distinct from searched-set eligibility.

Configuration	Later-time split	Decision behavior	Later-time calibration
Full AmendGuard	52 total; 37 amended; 15 control	Direction 69.1% (95% CI 58.8–78.4); FP 2/15 = 13.3% observed, 14.0% bootstrap mean (95% CI 4.9–24.8); 49 non-refusal valid / 3 refuse / 0 violations	Enrollment MAE 5.0% (95% CI 4.1–6.3); screen-failure MAE 3.3% (95% CI 2.4–4.6)
Precedent retrieval	52 total; 37 amended; 15 control	Direction 65.4% (95% CI 55.1–75.0); FP 3/15 = 20.0% observed, 19.8% bootstrap mean (95% CI 8.1–33.4); 46 non-refusal valid / 3 refuse / 3 violations	Enrollment MAE 5.4% (95% CI 4.4–6.8); screen-failure MAE 3.7% (95% CI 2.8–5.1)
Rule-based governed enumerator	52 total; 37 amended; 15 control	Direction 59.2% (95% CI 48.7–69.1); FP 1/15 = 6.7% observed, 8.9% bootstrap mean (95% CI 1.7–20.5); 43 non-refusal valid / 8 refuse / 1 violation	Enrollment MAE 6.1% (95% CI 4.9–7.6); screen-failure MAE 4.0% (95% CI 3.0–5.3)
Predictive utility ranker	52 total; 37 amended; 15 control	Direction 63.8% (95% CI 53.0–73.6); FP 4/15 = 26.7% observed, 25.1% bootstrap mean (95% CI 11.7–39.8); 47 non-refusal valid / 2 refuse / 3 violations	Enrollment MAE 5.6% (95% CI 4.6–7.0); screen-failure MAE 3.8% (95% CI 2.9–5.2)

Under the later temporal slice, the package-concordance ordering remains and only AmendGuard has no encoded-gate violation.

A.15 Replay and Calibration Diagnostics

Figure S1 separates the record-level replay denominator, mechanism ablations, and site-count calibration strata. Replay concerns observed historical amendments, while the other panels describe the policy's candidate checks; the panels therefore provide distinct diagnostics for historical consistency and boundary behavior.

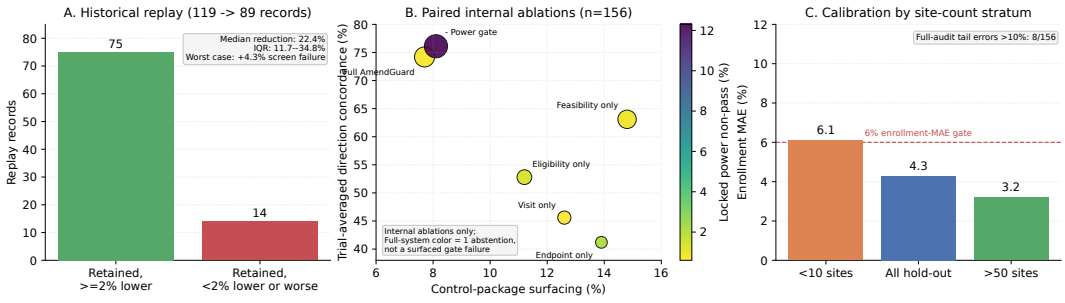

Figure S1: Replay, mechanism, and certification diagnostics. Panel A uses records: 89 of 119 survive filtering, and its 14 low/negative records are distinct from the 14/156 low-impact surfaced decisions. Panel B shows the safety–direction trade-off among package and gate ablations. Panel C reports enrollment calibration by site-count stratum against the 6% candidate boundary.

The calibration panel makes the thin-site regime visible: its 6.1% enrollment MAE lies just above the 6% candidate boundary, motivating local support checks instead of relying on the 4.3% aggregate hold-out value.

A.16 Extraction Confirmation and Gate Audit

Table S16 quantifies manual confirmation and candidate-level filtering before recommendations are surfaced. Bounded confirmation corrects protocol semantics, and rejection rows record the first failing condition, yielding an attributable trace for each filtered package.

Table S16: Extraction-confirmation and gate-audit summary on the 156-trial corpus.

Audit trace	Scoring unit	Value
Protocols with any manual confirmation	Trial in the 156-trial corpus	64/156 protocols
Corrected extracted fields	Field correction across the full corpus	187 corrected fields
Confirmed fields per manually reviewed protocol	Manual-review protocol subset	Median 2 fields (IQR 1–3)
Extraction-triggered abstentions	Trial in the 156-trial corpus	0 surfaced abstentions attributed to unresolved extraction ambiguity
Unique candidate packages entering first-failing gate audit	Deduplicated candidate package across the full corpus	10842 candidate packages
Rejected for power-reconstruction failure	Candidate package	614 first-failing rejections
Rejected for calibration failure	Candidate package	321 first-failing rejections
Rejected for safety-rule violation	Candidate package	173 first-failing rejections
Rejected for unsupported endpoint-specification edits	Candidate package	96 first-failing rejections
Retained below the intervention threshold	Candidate package	1287 candidate packages
Pass listed gates and intervention threshold	Candidate package	8,351 packages (derived remainder before top-package selection)

The rows now account for all 10842 packages: 1,204 first-failing gate rejections, 1287 below-threshold packages, and 8,351 packages that remain eligible before one top output is chosen per trial. Zero final violations verifies compliance with these encoded predicates; the blinded concern endpoint in the main-paper adjudication table evaluates the broader scientific judgment separately.