WHEN WORLD MODELS MISRANK: BUDGETED VERIFICATION FOR FINITE-SHOT QUANTUM CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

A world-model planner can recommend a control that is worse than the incumbent. Detecting such errors with fresh system measurements appears attractive, but those measurements consume the same scarce interactions needed to learn the model. We study this trade-off in finite-shot quantum control using a replacement gate that compares a candidate and incumbent, charges both measurements to the interaction budget, and reuses their outcomes for learning. A frozen opportunity-cost experiment across four simulated tasks and 40 paired task–seed units compares this procedure with reallocating every verification shot to additional active-learning probes. The gate improves equal-task normalized fidelity area under the budget curve by 0.1281 (95% CI $[0.0990, 0.1564]$) and final fidelity by 0.1797 ($[0.1285, 0.2302]$); both Holm-adjusted p-values are 2×10^{-5}. Two later held-out robustness families, with rules fixed before their new evaluation seeds were run, test stronger zero-verification alternatives: Gate remains ahead of a conservative model-only replacement rule under BALD ($+0.1119$ nBAUC, $+0.2140$ final fidelity) and after replacing BALD by decision-aligned value-variance acquisition ($+0.1278$, $+0.1907$), with both global 95% intervals excluding zero in each family. A post-hoc diagnostic further finds that measured advantage separates harmful candidates better than the tested model-only scores. A separate Hoeffding rule admits a conditional family-wise replacement guarantee under its stated independent-shot assumptions, at a substantial acceptance cost. These results identify a regime in which decision-targeted verification outperforms spending the same physical interaction budget only on model learning, while weakening explanations based solely on unconditional deployment or BALD-specific acquisition.

1 INTRODUCTION

Learning a world model permits an optimizer to search over controls without physically executing every candidate (Sutton, 1991; Ha & Schmidhuber, 2018; Chua et al., 2018). Deployment nevertheless requires a decision that imagined search cannot settle by itself: should the proposed candidate replace a control already in use? The candidate may be favored by the model while performing worse on the actual system. In an interaction-limited setting, the natural response—measure before replacing—raises a second question: would those measurements be better spent improving the model instead?

We study this question in partially observed, finite-shot quantum control. The learner observes measurement counts rather than density-matrix labels or device parameters. A neural terminal-state model supports control search; a small set of targeted measurements can instead evaluate a specific candidate–incumbent comparison. These measurements compete with general learning probes under one physical-shot budget. Quantum simulations provide a controlled setting in which true performance can be audited offline without exposing it to the learner. Our scope is open-loop control of four simulated quantum tasks.

The relevant failure is ordinal. Following the model-exploitation perspective of Bhamidipaty et al. (2026), a harmful update can arise from reversing two policies' order; positive absolute prediction

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.

bias is unnecessary. In an exploratory development audit, 139 of 223 model-favored replacement proposals are worse than their incumbents. This motivates auditing actual replacement opportunities; it does not establish favorable passive accuracy or causally attribute every error to optimization.

Our central experiment evaluates the opportunity cost of verification. MEAN+GATE uses fresh candidate–incumbent measurements to decide whether to replace, whereas MEAN-NOVAL reallocates the entire verification allocation to additional BALD learning probes. A third arm, MEAN+MEASURE, acquires the verification measurements but accepts every distinct candidate. All methods use the same total system-shot budget. After different decisions, their incumbents, observations, and future models can diverge; these are comparisons of complete adaptive procedures, not identical-data interventions.

The frozen four-task opportunity-cost study finds positive mean gate–NoVal effects in every task and substantial equal-task gains on both primary endpoints. Two later held-out families target the strongest alternative explanations suggested by that result. First, Conservative-NoVal spends zero shots on replacement verification but deploys only when the ensemble-paired predicted advantage satisfies a frozen lower-confidence rule. Second, the same conservative contrast is repeated after replacing BALD with decision-aligned value-variance acquisition. Gate remains ahead in both families on new seeds, so the observed benefit is not explained solely by unconditional deployment or by the particular BALD acquisition used in the central experiment. A complementary post-hoc analysis compares direct measurements with several model-side diagnostics on actual replacement opportunities.

Contributions. We contribute a budget-accounted test of deployment verification against additional model learning; two later held-out stress tests against conservative model-only deployment and a second decision-aligned acquisition rule; an event-level comparison of direct evidence with marginal and pairwise ensemble diagnostics; and an operational finite-shot gate with a separate, explicitly conditional assurance variant. Three original disjoint held-out seed families evaluate replacement performance, alternative search baselines, and opportunity cost; the two later families are reported separately rather than pooled or retroactively treated as part of the original confirmatory design. The contribution is the decision-level evidence and resource-allocation result, not algorithmic complexity for its own sake.

Related work. Value-aware losses, probabilistic ensembles, short model rollouts, and pessimistic planning address decision-relevant model error (Farahmand et al., 2017; Chua et al., 2018; Janner et al., 2019; Yu et al., 2020). We examine verification after candidate generation rather than replace those methods. High-confidence policy improvement and baseline bootstrapping already restrict updates using evidence or baseline constraints (Thomas et al., 2015; Laroche et al., 2019). SPIN uses policy-improvement tests under non-stationarity (Chandak et al., 2020), and off-policy confidence sequences explicitly support gated deployment (Karampatziakis et al., 2021). Deep SPI studies safe improvement through world models (Delgrange et al., 2026). Our distinction is direct finite-shot measurement of the proposed controls, charged to the same budget as learning. Local-CEM is a locality ablation, not a reimplementation of Deep SPI. Quantum system learning and model-based or model-free control supply the application context (Flurin et al., 2020; Genois et al., 2021; Khalid et al., 2023; Sivak et al., 2022); we focus on whether a synthesized control should replace its incumbent.

2 REPLACEMENT DECISIONS UNDER A PHYSICAL-SHOT BUDGET

A control is an open-loop sequence $\pi = (u_0, \ldots, u_{H-1})$. An unknown context ϑ determines the latent quantum dynamics, including dissipation and, where applicable, pulse filtering. The density operator evolves under a Lindblad model (Lindblad, 1976). A reset–control–measure experiment executes π in terminal setting B and returns

$$c \sim \mathrm{Multinomial}\big(n, p_\vartheta(\cdot \mid \pi, B)\big). \tag{1}$$

Only controls, settings, and counts are available for online learning. Exact probabilities, states, and control values are reserved for offline evaluation.

Table 1: Verification requires measurements of *both* controls. K is the number of terminal settings, so $n = 16$ shots per control per setting cost $2nK$ shots per comparison. $F_\Phi = (1+\langle XX\rangle-\langle YY\rangle+\langle ZZ\rangle)/4$.

Task	Fidelity F	Verification settings	K	Shots
Prepare-1	p_1	Target Z basis	1	32
Prepare-+	p_+	Target X basis	1	32
Transmon	p_1	Level populations	1	32
Bell	F_Φ	XX, YY, ZZ	3	96

Predicted and physical orderings. Let $J_\vartheta(\pi)$ be the scalar control objective and $\bar{J}(\pi) = M^{-1}\sum_m \widehat{J}_m(\pi)$ the ensemble prediction. For proposed candidate π_c and current incumbent π_i, define

$$\widehat{d}_{\text{model}} = \bar{J}(\pi_c) - \bar{J}(\pi_i), \qquad d_\vartheta = J_\vartheta(\pi_c) - J_\vartheta(\pi_i). \tag{2}$$

A *harmful candidate* has $d_\vartheta < 0$; an *ordering inversion* additionally requires $\widehat{d}_{\text{model}} > 0$. We retain this distinction when reporting candidate harm rates. A model can underestimate both controls and still reverse their order by underestimating the incumbent more strongly.

Task values and observable quantities. The four tasks are preparation of $|1\rangle$, preparation of $|+\rangle$, leakage-aware three-level transmon control, and preparation of $|\Phi^+\rangle$ in a two-qubit system. Table 1 lists fidelity endpoints and verification costs. The scalar objective J is distinct from the fidelity endpoint F: it may penalize leakage and deterministic control cost. For a sequence with d_u control channels, the implementation uses

$$\mathcal{E}(\pi) = \frac{1}{Hd_u}\sum_{t=0}^{H-1}\sum_{a=1}^{d_u} u_{t,a}^2. \tag{3}$$

The qubit tasks use $J = F - 0.001\mathcal{E}(\pi)$, the transmon uses $J = p_1 - 0.5p_2 - 0.001\mathcal{E}(\pi)$, and Bell uses $J = F_\Phi - 0.001\mathcal{E}(\pi)$. We report scalar scores and fidelity separately; full conventions are in Appendix B.4. A guarantee in J is not a guarantee on every component.

Decision loss. For a fixed pair and acceptance indicator $g \in \{0, 1\}$, the selected value is $J_i + gd$. Regret relative to the better member of the pair is

$$\ell(g; d) = \max(d, 0) - gd = g(-d)_+ + (1 - g)d_+. \tag{4}$$

Rejecting harmful candidates prevents loss; rejecting beneficial candidates misses gain. This identity explains why a zero harmful-acceptance rate alone is not sufficient evidence of a useful controller. It is a fixed-pair identity, not a causal decomposition of two different adaptive trajectories.

Accounting. All initial, learning-probe, and verification shots enter one cumulative ledger. Imagined evaluations within CEM consume computation but no physical shots. An active verification round allocates

$$B_{\text{round}} = B_{\text{probe}} + 2nK. \tag{5}$$

Verification observations return to the training set after the decision. Their reuse does not remove their opportunity cost: under a fixed budget, they displace other probes. Readout calibration is a specified measurement-model assumption; calibration of a real device is not claimed to be free.

3 Budgeted verification after model-based search

Learning and proposing. We use a five-member neural ensemble following deep-ensemble practice (Lakshminarayanan et al., 2017). Each member is trained on a bootstrap-resampled version of the accumulated data. A complete control sequence maps to a positive semidefinite density-matrix parameterization $\widehat{\rho} = LL^\dagger/\operatorname{Tr}(LL^\dagger)$. Training uses multinomial count likelihood, without exact density-matrix supervision. The model forward likelihood evaluates POVM probabilities without

Algorithm 1 Replacement step for one adaptive controller

Require: Current model, dataset $\mathcal{D}$, incumbent π_i, reserved verification budget $2nK$
1: Propose π_c with CEM; fix both controls before obtaining validation outcomes.
2: Execute both controls with n fresh shots in each of K settings; debit $2nK$.
3: Compute readout-corrected $\widehat{d}$ and $\widehat{\mathrm{se}}(\widehat{d})$ for the configured J.
4: **if** $\widehat{d} > \delta + z\widehat{\mathrm{se}}(\widehat{d})$ **then**
5: $\quad \pi_i \leftarrow \pi_c$
6: **end if**
7: Append both batches to $\mathcal{D}$ for subsequent model fitting, regardless of acceptance.
8: **return** $\pi_i, \mathcal{D}$

an explicit readout-mixing layer, so $\widehat{\rho}$ is best interpreted as an effective parameterization fitted directly to the observed count distributions rather than as a separately readout-corrected physical state. This implementation is shared by all compared arms. CEM approximately maximizes the ensemble-mean control objective (de Boer et al., 2005); BALD selects general learning probes (Houlsby et al., 2011). The terminal-state architecture and the search and acquisition rules are held fixed across the three opportunity-cost arms.

Empirical gate. After fixing a candidate and incumbent, collect independent fresh batches under matched terminal settings. Let

$$\widehat{d} = \widehat{J}_c - \widehat{J}_i, \qquad \widehat{\mathrm{se}}(\widehat{d}) = \sqrt{\widehat{\mathrm{Var}}(\widehat{J}_c) + \widehat{\mathrm{Var}}(\widehat{J}_i)}. \tag{6}$$

The variances are those of the reward *estimators*, including shot counts, readout correction, and covariance between populations estimated from the same multinomial batch. In the simulator, each probe batch first draws one OU process-noise trajectory and then draws its multinomial shots from the resulting terminal probabilities; the reported standard error therefore quantifies measurement-shot variability conditional on that batch and does not include between-batch process-noise variation. Deterministic control costs contribute no measurement variance. Accept the candidate only if

$$\widehat{d} > \delta + z\widehat{\mathrm{se}}(\widehat{d}). \tag{7}$$

The frozen empirical setting is $z = 0.5$, $\delta = 0$, and $n = 16$. The multiplier is a performance-oriented threshold, not a nominal confidence quantile. Algorithm 1 states the replacement subroutine; acquisition and refitting occur in the surrounding scheduled learning loop.

Two necessary controls. MEAN+MEASURE pays the same verification allocation and reuses its observations, but accepts every distinct candidate. MEAN-NOVAL performs no dedicated replacement verification and assigns all saved shots to additional BALD probes; it uses the same unconditional acceptance semantics. Thus gate–Measure tests the end-to-end effect of using a replacement rule when measurements are budgeted, whereas gate–NoVal tests its net value relative to a specific alternative use of those shots. Neither isolates a decision while holding all realized future data fixed. Ledger audits confirm subsequent trajectory divergence in 37 of 40 primary pairs and 38 of 40 opportunity-cost pairs.

3.1 A SEPARATE CONFIDENCE-VALID GATE

A more conservative rule replaces $z\widehat{\mathrm{se}}$ with a one-sided concentration radius r_t. The following standard conditional argument permits adaptive candidates and subsequent reuse of validation data when the fresh estimator targets the same replacement advantage and satisfies the stated conditional tail bound.

Proposition 1 (Family-wise replacement guarantee)**.** *At most T replacement decisions are made. Conditional on pre-validation history $\mathcal{F}_{t-1}$, each candidate and incumbent is fixed, and the fresh estimator satisfies*

$$\mathbb{P}(\widehat{d}_t - d_t \geq r_t \mid \mathcal{F}_{t-1}) \leq \alpha/T. \tag{8}$$

If acceptance requires $\widehat{d}_t > \delta + r_t$, then

$$\mathbb{P}(\exists t \leq T : \mathrm{accept}_t \text{ and } d_t \leq \delta) \leq \alpha. \tag{9}$$

Proof. On an acceptance with $d_t \le \delta$, $\widehat{d}_t - d_t > r_t$. Take expectations in Equation (8) and apply a union bound over at most T decisions. □

For bounded independent shots with known uniform readout attenuation e, direct-difference Hoeffding radii (Hoeffding, 1963) for a qubit target and Bell correlator objective are, respectively,

$$r_+(n) = \frac{1}{1-e}\sqrt{\frac{\log(T/\alpha)}{n}}, \qquad r_\Phi(n) = \frac{1}{1-e}\sqrt{\frac{3\log(T/\alpha)}{4n}}. \tag{10}$$

Appendix A derives the ranges and conditions. It protects the scalar replacement advantage only under its measurement assumptions, not the world model, every physical objective, execution of the validation candidate, or performance after device drift. In our simulator, nonzero OU process noise is sampled once per probe batch while offline J is evaluated at the fixed context without that additional draw; consequently, the development Hoeffding runs are reported as operating-point experiments rather than as an established family-wise certificate for the nominal offline objective. The empirical gate does not inherit the proposition's guarantee.

4 Experimental protocol

We separate development (seeds 0–9), three original frozen performance families, two later held-out robustness families, and retrospective mechanism analyses. Each task–seed unit defines a latent context and paired method comparison; new seeds test new contexts within the same task families. The four tasks are not independent application domains. Environment supports and model/search specifications are given in Appendix B.

Endpoints and inference. Performance is evaluated offline at system-shot checkpoints $N = (128, 256, 512, 1024)$. The primary endpoints are deployed fidelity $F(1024)$ and

$$\mathrm{nBAUC}(F) = \frac{1}{1024-128}\sum_{k=1}^{3}\frac{F(N_k)+F(N_{k+1})}{2}(N_{k+1}-N_k). \tag{11}$$

This is fidelity averaged over the recorded interval, not an integral from zero budget or the penalized reward. Per-task mean paired differences receive equal weight in global effects. FAM-1 and FAM-2 use task-stratified paired bootstrap intervals. For FAM-3, the frozen analysis uses 50,000 task-stratified paired bootstrap resamples and 100,000 one-sided task-stratified paired sign flips; the inferential units are task–seed pairs, not within-trajectory decisions. Holm correction (Holm, 1979) applies to the two prespecified FAM-3 endpoints and within the other declared performance families; taskwise exploratory analyses use Benjamini–Hochberg adjustment (Benjamini & Hochberg, 1995). FAM-4 and FAM-5 report the same equal-task paired endpoints with paired intervals and sign-flip tests; we apply a two-endpoint Holm adjustment within each later family. The Monte Carlo sign-flip p-values have the corresponding finite simulation resolution.

Original performance families. FAM-1 uses seeds 10–19 for Mean+Gate versus Mean+Measure: 80 runs and 40 paired units, with two global endpoints. FAM-2 uses seeds 20–29 for these methods, Local-CEM, and shot-matched model-free CEM: 160 comparative runs, with six global contrasts. Local-CEM restricts changes to a segmentwise L_∞ radius of 0.25 around the incumbent. FAM-3 freezes gate–NoVal as its primary contrast before running seeds 30–39: 120 runs of three adaptive arms and 40 paired units per contrast. Its two fidelity endpoints form a separate Holm family; other within-family contrasts are secondary. All 120 runs completed. These families answer distinct questions and are not pooled into one larger confirmatory sample.

In FAM-3, the post-initialization allocation is 896 shots per method. Qubit and transmon Gate/Measure runs divide it into 448 learning-probe and 448 verification shots; Bell divides it into 224 and 672. NoVal instead devotes all 896 to probes. The ledger also includes the common 128-shot initialization, reaching 1024 shots. Each task–seed triplet is run on the same machine, including triplets completed on CPU/WSL. The controls match physical resources, not necessarily realized data or all computational costs.

Table 2: FAM-3: absolute performance at 1024 shots and over $[128, 1024]$, ten seeds per task and method. Each cell is final fidelity / nBAUC. Bold indicates the largest mean within the task, not a separate significance claim. NeverReplace reports the initial-policy fidelity mean.

Task	MEAN-NOVAL	MEAN+MEASURE	MEAN+GATE	NeverReplace
Prepare-1	.8748 / .8404	.8586 / .8568	**.9362 / .8831**	.6881
Prepare-+	.7463 / .6255	.8304 / .7285	**.9305 / .8469**	.6381
Transmon	.3938 / .3806	.4943 / .4505	**.6306 / .5109**	.3133
Bell	.2541 / .2343	.4157 / .2869	**.4904 / .3524**	.2300

Later held-out robustness families. FAM-4 was designed after the earlier families were analyzed to test whether NoVal is weak simply because it replaces unconditionally. Before any seeds 40–49 were run, we froze Conservative-NoVal: for ensemble-member paired predicted advantages $d_m = \widehat{J}_m(\pi_c) - \widehat{J}_m(\pi_i)$, it replaces only if $\overline{d} - \mathrm{Std}(d_m) > 0$. The multiplier is fixed at 1.0, with no FAM-4 outcome-based tuning. Conservative-NoVal spends zero shots on candidate–incumbent verification and reallocates the saved allocation to learning. FAM-4 compares it with Gate under BALD on four tasks, two arms, and ten new seeds (80 runs). FAM-5, specified before evaluating seeds 50–59, repeats the same two-arm contrast after replacing BALD by the implemented decision-aligned value-variance acquisition rule, again for 80 runs. These are post-analysis robustness families with prospectively frozen within-family rules, not part of the original FAM-1–3 design and not pooled with them.

Audits. The original ledger checks cover all 80 FAM-1 and 120 FAM-3 runs, including chained budgets, verification charging, acceptance semantics, and same-policy queries. The later-family ledgers report 80/80 completed runs in each of FAM-4 and FAM-5 at 1024 total shots; every Conservative-NoVal run uses zero verification shots, leaves observed-validation fields unset, and records a finite model-side lower-confidence score. The objective audit reports no failures, with a readout-consistency note (Appendix F). Independent complex128 propagation checks 200 policy evaluations against the complex64 RK4 simulator. Mechanism analyses on FAM-1 and FAM-3 logs are *post-hoc analyses of held-out trajectories*, not prospectively specified tests. We report the diagnostic AUROCs descriptively, without treating correlated events as independent confirmatory replicates.

5 RESULTS

5.1 VERIFICATION OUTPERFORMS REALLOCATING ITS BUDGET TO LEARNING

The opportunity-cost comparison is the central result. Under equal total physical-shot budgets, FAM-3 yields an equal-task gate–NoVal effect of $+0.1281$ in nBAUC (95% CI $[0.09899, 0.1564]$) and $+0.1797$ in final fidelity ($[0.1285, 0.2302]$). Both Holm-adjusted p-values are 2×10^{-5}. Table 2 gives absolute outcomes; Figure 1 gives paired effects. All task means favor gating, but Prepare-1 final fidelity has an interval crossing zero and is not individually significant. We therefore distinguish consistent mean direction from universal taskwise significance.

The third arm clarifies the result. Relative to MEAN+MEASURE, Gate improves nBAUC by 0.06763 and final fidelity by 0.09718; relative to NoVal, Measure itself has higher mean performance, with NoVal–Measure differences of -0.0605 and -0.0825. These secondary contrasts suggest that both the control-pair query distribution and replacement decision can matter, but they are not a mediation decomposition because each procedure changes its later data and candidates. NeverReplace also shows that Gate exceeds the starting-policy mean in every task; NoVal's weak Bell result identifies a limitation of that tested learning allocation, not of all model-learning strategies.

5.2 CONSERVATIVE MODEL-ONLY DEPLOYMENT AND A SECOND ACQUISITION DO NOT CLOSE THE GAP

FAM-4 targets the strongest baseline objection to FAM-3: NoVal may look weak because it deploys every distinct candidate. Conservative-NoVal instead uses no dedicated verification measurements

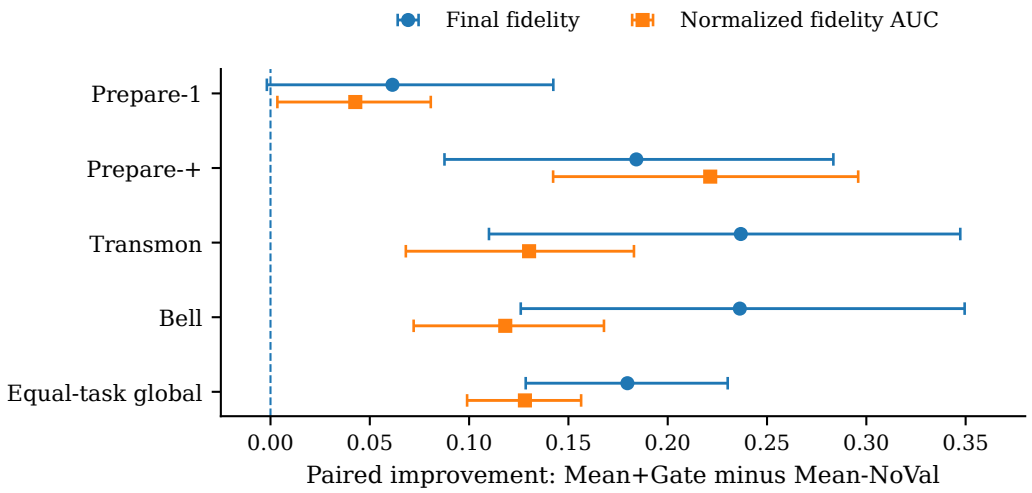

Figure 1: FAM-3 opportunity cost: paired gate–NoVal effects with reported 95% paired-effect intervals. Positive values favor verification. Global effects weight tasks equally. These are improvements on the fidelity scale, not relative percentage gains.

and replaces only when its frozen ensemble-paired model score satisfies $\overline{d} - \mathrm{Std}(d_m) > 0$. On new seeds 40–49, Gate improves equal-task nBAUC by $+0.1119$ (95% CI $[0.0710, 0.1538]$) and final fidelity by $+0.2140$ ($[0.1444, 0.2834]$). All four task means are positive on both endpoints; the smallest taskwise nBAUC effect is $+0.0689$ on Bell ($[0.0007, 0.1412]$). Conservative-NoVal still replaces 17.9–26.4% of eligible candidates across tasks, so the comparison is not a degenerate never-replace policy. Applying the same two-endpoint Holm correction gives $p = 2 \times 10^{-5}$ for both global endpoints.

FAM-5 asks whether the result is specific to BALD. Replacing BALD by decision-aligned value-variance acquisition while leaving the Gate versus Conservative-NoVal comparison otherwise unchanged, on new seeds 50–59 Gate again improves nBAUC by $+0.1278$ ($[0.0985, 0.1576]$) and final fidelity by $+0.1907$ ($[0.1302, 0.2494]$), with Holm-adjusted $p = 2 \times 10^{-5}$ for both endpoints and positive effects on all four tasks. Both arms remain active replacers (22.9% Gate, 28.6% Conservative-NoVal overall). These later families weaken explanations based only on unconditional deployment or BALD-specific acquisition; they do not establish superiority over every possible model-only rule, acquisition strategy, or model class. Appendix C summarizes the separate families.

5.3 Earlier held-out studies and alternative search baselines

FAM-1 independently establishes the advantage over unconditional replacement with measurements: nBAUC improves by 0.09544 ($[0.0610, 0.1303]$) and final fidelity by 0.1316 ($[0.07167, 0.1923]$), with Holm-adjusted $p = 2 \times 10^{-5}$ and 4×10^{-5}. FAM-2 broadens the comparison. Gate–Local-CEM gains are 0.04106 in nBAUC and 0.09196 in final fidelity; gate–model-free CEM gains are 0.1829 and 0.2733. Complete per-task results and the six-comparison family appear in Appendix C.

These averages have important exceptions. On Bell in FAM-2, model-free CEM has final fidelity 0.4954 versus 0.4934 for Gate, and nBAUC 0.4914 versus 0.4123. Local-CEM also exceeds Gate's Prepare-1 final-fidelity mean. The evidence supports a useful hybrid in the tested average, not uniform dominance or a universal preference for world-model search. Cross-family results are never merged into matched comparisons.

5.4 Direct measurements versus decision-aligned model diagnostics

The original exploratory audit contains 139 harmful candidates among 223 model-favored proposals. The FAM-3 audit compares diagnostics on 443 gated-trajectory decisions and 418 Measure-

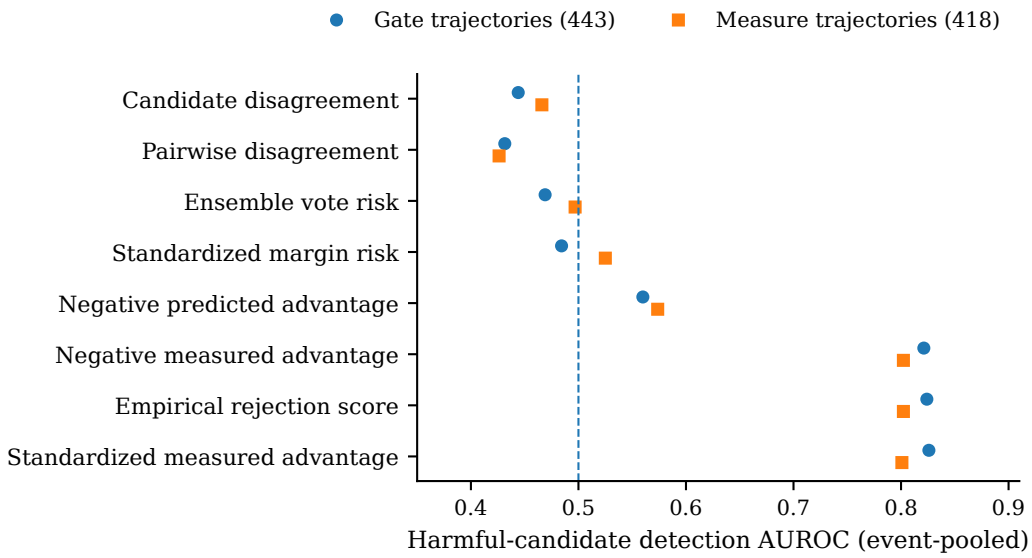

Figure 2: Post-hoc detection on FAM-3 decisions. Points are descriptive event-pooled AUROCs on different trajectory sets, not independent-event significance tests. Measurement scores use fresh system evidence; the primary experiment separately accounts for its cost.

trajectory decisions; harmful-candidate prevalence is 61.6% and 47.4%, respectively. These are different trajectory distributions, not a common set on which Gate changes proposal quality.

For member m, let $d_m = \widehat{J}_m(\pi_c) - \widehat{J}_m(\pi_i)$. We test candidate disagreement, pairwise disagreement $\mathrm{Std}_m[d_m]$, predicted advantage, member voting, and standardized margins against measurement-based scores using fresh system evidence. Figure 2 reports descriptive event-pooled AUROCs; full definitions and taskwise values are in Appendix D.

Measured advantage separates harmful and beneficial candidates better than the tested model-only scores on these trajectories. The pooled Gate AUROC is 0.8212 for measured advantage, versus 0.5599 for predicted advantage and 0.4315 for pairwise ensemble disagreement under the stated risk orientation; reversing disagreement yields 0.5685, still below the measurement score. Within-task averaging gives 0.8205 for measured advantage and 0.6175 for predicted advantage. This does not rule out other calibrated or learned diagnostics. Measurements are also imperfect: on Gate events, measured–true advantage correlation is 0.7702 and MAE is 0.1175. Their practical value therefore depends on the separate budget comparison, not diagnostic accuracy alone.

5.5 What the empirical Gate prevents, and what it does not

The retrospective FAM-1 event audit reports harmful accepted decisions per eligible proposal of 50.0–53.9% for Measure and 5.4–10.0% for Gate. Among accepted Gate proposals, the harmful fractions are approximately 18.9–35.3%, and 50–70% of gated runs contain at least one harmful acceptance. The empirical rule reduces harmful deployment frequency; it does not eliminate harm. Appendix D.1 reports acceptance, prevented degradation, and missed improvement with explicit denominators.

Scalar rewards provide a complementary check. In FAM-3 transmon control, Gate has mean logged reward 0.5687 versus 0.4309 for Measure and 0.2159 for NoVal. Mean leakage is 0.1227, 0.1255, and 0.3544, respectively. Thus the gate–NoVal fidelity gain does not hide higher mean leakage in this task. These component summaries are descriptive, not component-wise guarantees.

5.6 VERIFICATION STRENGTH AND NUMERICAL CHECKS

Development sweeps show that absolute performance need not increase with verification allocation. For Prepare-+, final Gate fidelity across $n = 4, 8, 16, 24$ is 0.8717, 0.8879, 0.9390, and 0.9077. Bell is endpoint-dependent: its nBAUC at $n = 4$ slightly exceeds that at $n = 16$, although final fidelity does not. These small-sample results characterize a trade-off rather than an optimized allocation frontier; full values are in Appendix E.

The confidence-valid comparison evaluates Prepare-+ and Bell on five development seeds each. The Hoeffding rule has no observed harmful accepts, but accepts only 3.1% of Prepare-+ proposals and none on Bell, compared with 33.8% and 26.7% for the empirical rule. Observing no violations at such low acceptance is not a substitute for Proposition 1's assumptions. Independent propagation checks over 200 evaluations give maximum fidelity discrepancy 6.188×10^{-4} and maximum trace distance 1.58×10^{-3}, supporting numerical consistency rather than real-device validity.

6 DISCUSSION AND LIMITATIONS

The evidence favors a division of labor: use a learned model for broad search and system measurements for selected replacement decisions. In the tested budget regime, directing some interactions toward the current deployment decision is more effective than directing all of them toward the tested model-learning alternatives. FAM-4 shows that this gap is not explained only by unconditional deployment, and FAM-5 shows that it persists under a second, decision-aligned acquisition rule. These are stronger than a free-validation result, but narrower than a theorem that verification is always the best use of additional data.

The gate is deliberately simple. Our central claim is therefore not that a complex new uncertainty architecture is required, but that the allocation of scarce physical interactions can matter at the decision boundary: direct evidence about the proposed replacement can retain value even when the alternative uses all saved shots for learning and deploys conservatively. The three-arm FAM-3 comparison also suggests two roles for targeted measurements—changing later model-fitting data and changing the immediate decision. Because trajectories diverge, the reported effects are end-to-end closed-loop comparisons rather than identical-data causal decompositions.

All tasks are small simulated quantum systems with reset access, one world-model family, one main planner, two tested acquisition rules, and a finite budget grid. Other model classes, acquisitions, control parameterizations, or larger budgets could change the relative benefit; model-free CEM is competitive on Bell. We have not established hardware robustness, cross-domain generality, or a universal ranking over model-only conservative rules. The event analyses are retrospective and descriptive, and near-zero advantages remain sensitive to numerical and measurement error. The world model fits observed count distributions without an explicit known-readout layer, whereas the empirical gate explicitly corrects the configured uniform readout channel; this shared modeling choice across the tested arms limits extrapolation to alternative readout-aware learners. Moreover, one process-noise trajectory is shared within each simulated probe batch, so shot-only concentration bounds do not by themselves certify the nominal offline objective under nonzero process noise. The configured simulator and scalar objective must be distinguished from experimentally calibrated hardware and component-wise safety. Appendices give the executable conventions and remaining reproduction limits.

7 CONCLUSION

A world-model proposal is a hypothesis about an improvement, not an improvement itself. The original held-out studies show that budgeted verification can improve deployment relative to unconditional replacement, alternative search baselines, and reallocating every verification shot to additional BALD learning. Two later held-out families preserve the advantage against a conservative zero-verification deployment rule and after switching to value-variance acquisition. Decision-aligned ensemble scores also do not match direct measurement's descriptive discrimination on the audited events. The supported lesson is conditional but sharper: under scarce physical interaction, checking the proposed replacement can be a better investment than spending the same interactions only on learning, even when the model-only alternative is conservative and the acquisition rule is changed.

AI USE STATEMENT

Generative AI tools assisted with methodological critique and experimental planning, refinement of hypotheses and the conceptual framing, interpretation of experimental summaries, mathematical exposition and proof writing, literature discovery, manuscript drafting and editing, reference formatting, data-table reformatting, figure-generation code, and LaTeX preparation. The empirical numbers presented here originate from the experiment records and summaries; they were not generated as hypothetical results by a language model. The paper separates reported experimental results from derived quantities and from conditional mathematical statements. Responsibility for the final text, implementations, analyses, and submission rests with the authors.

REPRODUCIBILITY STATEMENT

The main text specifies the replacement rules, budgets, endpoints, and separation of original and later held-out families. Appendix A gives the measurement-range derivations and proof; Appendix B records the executed environment, objective, training, readout, and process-noise conventions; Appendices C–F provide the numerical tables, diagnostics, robustness analyses, and accounting scope. The anonymized supplementary code archive contains the executable implementation, frozen configurations, saved raw counts and trajectory logs, tests, and analysis scripts for the released experiment package; the manuscript source contains machine-readable aggregate tables and figure-generation scripts. Model checkpoints were not saved, and bitwise equality across software or hardware environments is not claimed. The released records support direct recomputation of reported endpoints and event-level estimators without retraining where the corresponding run records are included.

ETHICS STATEMENT

The experiments involve simulated quantum systems, not human participants or personal data. A statistical non-degradation statement about a scalar control objective does not guarantee safe physical execution. Measuring a candidate can itself incur risk, and neither reset access nor an accurate readout model should be assumed in a new application without validation. The work does not establish a hardware safety certificate or a quantum computational advantage.

REFERENCES

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

L. M. Bhamidipaty, E. S. Whitammer, D. Abel, M. J. Kochenderfer, and S. Ramamoorthy. Imperfect World Models Are Exploitable. arXiv preprint arXiv:2605.15960, 2026. URL https://arxiv.org/abs/2605.15960.

Yash Chandak, Scott M. Jordan, Georgios Theocharous, Martha White, and Philip S. Thomas. Towards Safe Policy Improvement for Non-Stationary MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9156–9168, 2020. URL https://arxiv.org/abs/2010.12645.

K. Chua, R. Calandra, R. McAllister, and S. Levine. Deep Reinforcement Learning in a Handful of Trials Using Probabilistic Dynamics Models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134(1):19–67, 2005. doi: 10.1007/s10479-005-5724-z.

F. Delgrange, R. Avalos, and W. Röpke. Deep SPI: Safe Policy Improvement via World Models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=24C3bSaH3F.

A.-M. Farahmand, A. Barreto, and D. Nikovski. Value-Aware Loss Function for Model-Based Reinforcement Learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1486–1494, 2017.

E. Flurin, L. S. Martin, S. Hacohen-Gourgy, and I. Siddiqi. Using a Recurrent Neural Network to Reconstruct Quantum Dynamics of a Superconducting Qubit from Physical Observations. *Physical Review X*, 10(1):011006, 2020. doi: 10.1103/PhysRevX.10.011006.

É. Genois, J. A. Gross, A. Di Paolo, N. J. Stevenson, G. Koolstra, A. Hashim, I. Siddiqi, and A. Blais. Quantum-Tailored Machine-Learning Characterization of a Superconducting Qubit. *PRX Quantum*, 2(4):040355, 2021. doi: 10.1103/PRXQuantum.2.040355.

D. Ha and J. Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning. arXiv preprint arXiv:1112.5745, 2011. URL https://arxiv.org/abs/1112.5745.

M. Janner, J. Fu, M. Zhang, and S. Levine. When to Trust Your Model: Model-Based Policy Optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-Policy Confidence Sequences. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5301–5310, 2021. URL https://proceedings.mlr.press/v139/karampatziakis21a.html.

I. Khalid, C. A. Weidner, E. A. Jonckheere, S. G. Schirmer, and F. C. Langbein. Sample-Efficient Model-Based Reinforcement Learning for Quantum Control. *Physical Review Research*, 5(4):043002, 2023. doi: 10.1103/PhysRevResearch.5.043002.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

R. Laroche, P. Trichelair, and R. Tachet Des Combes. Safe Policy Improvement with Baseline Bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3652–3661, 2019.

G. Lindblad. On the Generators of Quantum Dynamical Semigroups. *Communications in Mathematical Physics*, 48(2):119–130, 1976. doi: 10.1007/BF01608499.

V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. H. Devoret. Model-Free Quantum Control with Reinforcement Learning. *Physical Review X*, 12(1):011059, 2022. doi: 10.1103/PhysRevX.12.011059.

R. S. Sutton. Dyna, an Integrated Architecture for Learning, Planning, and Reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High Confidence Policy Improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2380–2388, 2015.

T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma. MOPO: Model-Based Offline Policy Optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A MEASUREMENT ESTIMATORS AND FULL REPLACEMENT PROOF

This appendix makes explicit the estimator ranges behind Proposition 1. The guarantee is about the task objective J under the stated measurement model, not about every component of a scalarized reward.

A.1 DIRECT-DIFFERENCE CONCENTRATION

Let $\mathcal{F}_{t-1}$ include all information used to choose the candidate, incumbent, and validation allocation before the round's measurements. Conditional on this history and on any variables defining the common target distribution, write the unbiased difference estimator as

$$\widehat{d}_t - d_t = \sum_{k=1}^{q_t} \left(X_{t,k} - \mathbb{E}[X_{t,k} \mid \mathcal{F}_{t-1}] \right), \tag{12}$$

where the $X_{t,k}$ include their sample-mean and reward weights, are conditionally independent, and lie in known intervals $[a_{t,k}, b_{t,k}]$. Hoeffding's inequality implies

$$\mathbb{P}(\widehat{d}_t - d_t \geq r \mid \mathcal{F}_{t-1}) \leq \exp\left(-\frac{2r^2}{V_t} \right), \qquad V_t = \sum_k (b_{t,k} - a_{t,k})^2. \tag{13}$$

Thus $r_t = \sqrt{V_t \log(T/\alpha)/2}$ suffices. The deterministic difference of control-energy penalties changes d_t and $\widehat{d}_t$ by the same amount, so it contributes no random range. The corresponding two-sided or sequential construction would need its own error allocation; repeatedly peeking at a fixed-sample interval is not licensed by this result.

A.2 QUBIT TARGET FIDELITY

Under the configured known uniform readout channel, correcting a target-basis Bernoulli outcome gives range width $w = (1-e)^{-1}$. The direct candidate–incumbent difference uses $2n$ independent weighted shots, each of range width w/n. Consequently

$$V_t = 2n(w/n)^2 = 2w^2/n, \qquad r_+(n) = w\sqrt{\log(T/\alpha)/n}. \tag{14}$$

This e parameter describes uniform readout attenuation. Replacing it by the error probability of a different readout model, such as a symmetric bit-flip channel, without changing the range calculation would be incorrect.

A.3 BELL FIDELITY AND CONTROL COST

For the Bell target,

$$F_\Phi = \frac{1 + \langle XX \rangle - \langle YY \rangle + \langle ZZ \rangle}{4}. \tag{15}$$

The constant cancels in a candidate–incumbent difference. Each corrected correlator shot has range width $2w$. The two controls and three independently measured settings give $6n$ weighted observations, with absolute weight $1/(4n)$ each. Hence

$$V_t = 6n \left(\frac{2w}{4n} \right)^2 = \frac{3w^2}{2n}, \qquad r_\Phi(n) = w\sqrt{\frac{3\log(T/\alpha)}{4n}}. \tag{16}$$

The Bell task uses $J = F_\Phi - 0.001\mathcal{E}(\pi)$. The same concentration radius applies to J because energy is deterministic for a fixed sequence. However, certifying an improvement in J does not certify an improvement in F_Φ separately.

A.4 GENERAL READOUT CORRECTION AND THE TRANSMON COVARIANCE

If a known invertible readout matrix C maps true outcome probabilities p to observed probabilities $q = Cp$, and the stochastic reward is $v^\top p$, then the corrected score for an observed outcome a is $s(a) = v^\top C^{-1} e_a$. The sample average is unbiased and its range is determined by $\min_a s(a)$

and $\max_a s(a)$. This gives a direct way to calculate a valid range without assuming independent population estimates.

For transmon reward $p_1 - 0.5p_2$, set $v = (0, 1, -0.5)^\top$. Without readout correction, a single observation has score $\mathbf{1}\{a = 1\} - 0.5\mathbf{1}\{a = 2\}$ and

$$\mathrm{Var}(s) = p_1 + 0.25p_2 - (p_1 - 0.5p_2)^2. \tag{17}$$

The variance of its n-shot mean is this quantity divided by n. With correction, use the variance of the corrected categorical score. The shared multinomial observations induce covariance between population estimates; treating $\widehat{p}_1$ and $\widehat{p}_2$ as independent would not reproduce this expression. This calculation clarifies an estimator requirement; it is not an additional transmon assurance experiment.

A.5 Adaptive proposals, fresh validation, and the margin

Let $A_t = \{\widehat{d}_t > \delta + r_t,\ d_t \le \delta\}$. On A_t,

$$\widehat{d}_t - d_t > \delta + r_t - d_t \ge r_t. \tag{18}$$

Therefore $\mathbb{P}(A_t \mid \mathcal{F}_{t-1}) \le \alpha/T$, and the tower property and union bound give

$$\mathbb{P}\left(\bigcup_{t=1}^{T} A_t\right) \le \sum_{t=1}^{T} \mathbb{E}[\mathbb{P}(A_t \mid \mathcal{F}_{t-1})] \le \alpha. \tag{19}$$

The margin is a *required improvement*: with probability at least $1 - \alpha$, every accepted replacement has $d_t > \delta$. At $\delta = 0$, harmful or non-improving acceptance is controlled. No independence between successive model updates is required, but a candidate must not be selected by optimizing over the same validation noise used to certify it. Verification observations may be reused after that decision to train future models; those future decisions require new validation data.

The result controls an instantaneous comparison for a common target system. It does not guarantee monotone realized returns across changing environments, accuracy under an unknown readout channel, or absence of harm while experimentally executing a rejected candidate. If there is nonzero unaccounted estimator bias, the stated radius alone is not a valid certificate.

A.6 Scope under batch-level process noise

The executed simulator samples one Ornstein–Uhlenbeck detuning-noise trajectory for each probe batch, propagates the control under that trajectory, and then draws all n multinomial shots from the resulting terminal probabilities. Conditional on the sampled trajectory, the shot contributions are independent and the linear readout-corrected estimator targets that batch-specific terminal score. Marginalizing over the trajectory introduces shared batch-level randomness that is not included in the shot-only Hoeffding radius.

The offline fidelity, scalar reward, and harmful-candidate labels used in the main performance and mechanism tables evaluate the fixed latent context without an additional OU draw. Therefore, with nonzero process noise, the batch-specific validation target need not equal this nominal offline d_t. We consequently treat the development Hoeffding runs in Appendix E as a conservative threshold operating-point study, not as an empirical demonstration that Equation (9) certifies the nominal offline objective. A nominal-objective certificate would require, for example, conditioning the target accordingly, disabling this shared process noise, or deriving a radius that also controls its contribution.

B Task, model, and implementation specifications

This appendix records the executed specifications used by the frozen runs. All online learners receive controls, measurement settings, and finite-shot counts. Simulator states, exact probabilities, and true rewards used in evaluation are privileged offline quantities, not training labels or inputs to a deployable gate. Unless stated otherwise, each context coordinate is sampled independently and uniformly over its listed interval; all systems start in the computational-basis ground state ($|0\rangle$, or $|00\rangle$ for Bell).

B.1 DRIVEN-QUBIT TASKS

The quantum component follows

$$\dot{\rho} = -i[H_t, \rho] + \gamma_1 \mathcal{D}[\sigma_-](\rho) + \frac{\gamma_\phi}{2}\mathcal{D}[Z](\rho), \tag{20}$$

$$H_t = \frac{1}{2}(\Delta Z + g_x u_{x,t} X + g_y u_{y,t} Y), \tag{21}$$

$$\mathcal{D}[L](\rho) = L\rho L^\dagger - \tfrac{1}{2}\{L^\dagger L, \rho\}. \tag{22}$$

There are 20 time steps with $dt = 0.15$, controls in $[-1, 1]^2$, and uniform readout mixing $e = 0.02$. Context supports are $\Delta \in [-0.3, 0.3]$, $g_x, g_y \in [0.8, 1.2]$, $\gamma_1 \in [0.002, 0.02]$, and $\gamma_\phi \in [0.002, 0.03]$. The two targets are $|1\rangle$ and $|+\rangle$. The simulated detuning process noise is Ornstein–Uhlenbeck with $\sigma = 0.01$ and $\tau = 4.0$.

B.2 THREE-LEVEL TRANSMON

The truncated model includes detuning, anharmonicity, two drive quadratures, dissipation, and first-order pulse filtering. The horizon is 20 with $dt = 0.12$, controls in $[-1, 1]^2$, and uniform readout mixing $e = 0.02$. The filter is initialized at $v_{-1} = 0$ and updated componentwise by

$$v_t = \beta v_{t-1} + (1 - \beta)u_t. \tag{23}$$

Context supports are detuning $[-0.2, 0.2]$, anharmonicity $[-0.35, -0.2]$, gain $[0.85, 1.15]$, filter coefficient $\beta \in [0, 0.5]$, and damping and dephasing parameters in $[0.001, 0.02]$. The simulated detuning process noise uses $\sigma = 0.008$ and $\tau = 4.0$.

B.3 TWO-QUBIT BELL CONTROL

With g_{ex} denoting exchange coupling, the coherent generator is

$$\begin{aligned} H_t = &\sum_{i=1}^{2} \tfrac{1}{2}(\Delta_i Z_i + u_{x,i,t} X_i + u_{y,i,t} Y_i) \\ &+ \tfrac{g_{\mathrm{ex}}}{2}(X_1 X_2 + Y_1 Y_2) + \tfrac{\zeta}{2} Z_1 Z_2. \end{aligned} \tag{24}$$

The horizon is 24 with $dt = 0.16$ and four controls in $[-1, 1]$. The frozen context supports are $\Delta_1, \Delta_2 \in [-0.2, 0.2]$, $g_{\mathrm{ex}} \in [0.35, 0.55]$, $\zeta \in [-0.15, 0.15]$, $\gamma_{1,1}, \gamma_{1,2} \in [0.001, 0.008]$, and $\gamma_{\phi,1}, \gamma_{\phi,2} \in [0.001, 0.01]$. Uniform readout mixing is $e = 0.02$. The shared antisymmetric detuning-noise draw uses OU parameters $\sigma = 0.006$, $\tau = 5.0$, entering the two qubits with opposite signs.

B.4 FIDELITY, SCALAR REWARD, AND DETERMINISTIC PENALTIES

The fidelity endpoints are p_1 for Prepare-1, p_+ for Prepare-+, p_1 for transmon, and F_Φ for Bell. For d_u control channels the implementation defines

$$\mathcal{E}(\pi) = \mathrm{mean}_{t,a}\, u_{t,a}^2 = \frac{1}{H d_u} \sum_{t=0}^{H-1} \sum_{a=1}^{d_u} u_{t,a}^2. \tag{25}$$

The executed scalar objectives are

$$J_{\mathrm{qubit}} = F - 0.001\mathcal{E}(\pi), \tag{26}$$

$$J_{\mathrm{transmon}} = p_1 - 0.5p_2 - 0.001\mathcal{E}(\pi), \tag{27}$$

$$J_{\mathrm{Bell}} = F_\Phi - 0.001\mathcal{E}(\pi). \tag{28}$$

Thus the small qubit differences $F - J$ in Table 9 are the deterministic energy penalty rather than an unidentified residual. The main confirmatory endpoints remain fidelity; harm labels use the corresponding logged scalar-reward difference. A certificate on scalar J is not automatically a certificate on each component or on fidelity alone.

B.5 Readout model and observed-score estimators

The simulator maps ideal terminal probabilities p to observed probabilities $q = (1-e)p+e/K$, with K outcomes and $e = 0.02$. The empirical replacement estimator applies the inverse uniform-mixing correction to finite-shot frequencies, clips negative corrected probabilities to zero, and renormalizes before forming the task score; its plug-in standard error uses the corresponding linear corrected categorical score. The separate Hoeffding branch uses an unclipped linear corrected estimator so that the bounded-score concentration calculation is not applied to the clipped empirical estimator.

The world model does *not* insert the known readout-mixing channel in its forward likelihood. Its predicted density-matrix parameterization is converted directly to POVM probabilities and fitted to the observed counts. It should therefore be interpreted as an effective model of the observed terminal distributions, not as an explicitly readout-corrected reconstruction of the physical state. All opportunity-cost arms use the same world-model implementation, so this choice is controlled within their comparisons but limits extrapolation to a readout-aware learner.

B.6 World model, acquisition, and search

The five-member bootstrap ensemble uses two-layer GRUs of width 64. Each predicts a complex triangular factor L, giving $\widehat{\rho} = LL^\dagger / \operatorname{Tr}(LL^\dagger)$. Up to terms independent of the parameters, the fitting objective is

$$\mathcal{L}(\theta) = -\sum_{e,a} c_{e,a} \log \widehat{p}_{\theta,a}(\pi_e, B_e). \tag{29}$$

Each fit uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-5}, mini-batch size at most 32, and gradient-norm clipping at 5. Qubit fits allow 100 initial and 12 update epochs with patience 20; transmon uses 110/14 with patience 20; Bell uses 120/16 with patience 22. Each ensemble member draws a fresh bootstrap resample at each fit, the optimizer is re-created at each fit, and early stopping monitors that bootstrap training loss with improvement threshold 10^{-5}, restoring the best state from that fit. These runs did not save model checkpoints.

CEM uses 256 samples, elite fraction 0.1, six iterations, two restarts, and minimum standard deviation 0.05 for qubits/Bell and 0.04 for transmon. Controls are parameterized by ten segments for qubits and transmon and twelve for Bell. General acquisition uses BALD. FAM-3 Probe/verification totals are specified in Table 18. They determine the total physical resource without assuming that a model evaluation or a CEM restart consumes a shot. Local-CEM restricts changes to a segmentwise L_∞ radius of 0.25. The model-free CEM baseline queries the simulator through finite-shot observations, not exact simulator rewards.

A simulator-informed search reference is configured with 1024 samples, ten iterations, four restarts, elite fraction 0.1, and minimum standard deviation 0.02. This is a finite optimization reference, not a demonstrated global optimum. Our principal results do not use reference regret as a substitute for the fidelity endpoints.

B.7 Reproduction scope

The anonymized experiment supplement contains the executable simulator, learning and deployment implementation, frozen configurations, tests, saved raw counts and per-event trajectory logs, and analysis scripts used for the reported experiment families. The manuscript archive separately contains machine-readable aggregate tables and plotting scripts. Saved records permit recomputation of the reported FAM-3 endpoints and deployment estimators without retraining; model checkpoints were not saved, and software/hardware differences can prevent bitwise-identical retraining. The release therefore supports executable reruns and raw-record reanalysis without claiming checkpoint-level or cross-environment bitwise replication.

C Complete held-out performance results

FAM-1, FAM-2, and FAM-3 use disjoint original seed sets 10–19, 20–29, and 30–39. FAM-4 and FAM-5 are later robustness families on new seeds 40–49 and 50–59, respectively. Statistical units are task–seed pairs. The later families were designed after analysis of the earlier results, but their

Table 3: FAM-1 absolute means, ten seeds per task and arm. Scheduled-round acceptance differs from the eligible-event denominator used in the mechanism audit.

Task	Arm	Final F	nBAUC	a_{round}
Prepare-1	Measure	0.8352	0.8311	0.6571
Prepare-1	Gate	0.9341	0.8945	0.2500
Prepare-+	Measure	0.8300	0.7469	0.9214
Prepare-+	Gate	0.9411	0.8770	0.2429
Transmon	Measure	0.4922	0.4368	0.9143
Transmon	Gate	0.7033	0.5592	0.2643
Bell	Measure	0.3497	0.3336	0.8571
Bell	Gate	0.4551	0.3995	0.2429

Table 4: FAM-1 Gate–Measure paired effects, reported 95% intervals, raw one-sided taskwise sign-flip p, and BH q over eight secondary tests.

Task	Endpoint	Δ	95% interval	p	q
Prepare-1	Final F	0.0989	$[0.0367, 0.1696]$	0.0068	0.0176
Prepare-1	nBAUC	0.0634	$[0.0298, 0.0990]$	0.0059	0.0176
Prepare-+	Final F	0.1111	$[0.0169, 0.2435]$	0.0273	0.0332
Prepare-+	nBAUC	0.1301	$[0.0659, 0.2046]$	0.0029	0.0176
Transmon	Final F	0.2111	$[0.0202, 0.3894]$	0.0332	0.0332
Transmon	nBAUC	0.1224	$[0.0152, 0.2302]$	0.0322	0.0332
Bell	Final F	0.1053	$[0.0276, 0.1859]$	0.0205	0.0328
Bell	nBAUC	0.0659	$[0.0260, 0.1046]$	0.0088	0.0176

rules and evaluation seed sets were fixed before the corresponding runs; they are not pooled with FAM-1–3 or relabeled as part of the original confirmatory design.

C.1 FAM-1: REPLACEMENT WITH MEASUREMENTS

Table 3 gives absolute primary-study outcomes and Table 4 gives taskwise effects. Global gate–Measure effects are 0.09544 ($[0.0610, 0.1303]$) in nBAUC and 0.1316 ($[0.07167, 0.1923]$) in final fidelity. The corresponding raw one-sided sign-flip p-values are 10^{-5} and 4×10^{-5}; the two-endpoint Holm values are 2×10^{-5} and 4×10^{-5}. The displayed intervals are those of the original primary analysis, not a newly generated bootstrap of rounded means.

C.2 FAM-2: ALTERNATIVE SEARCH BASELINES

Table 5 reports all method means and standard deviations. Across tasks, Gate has larger mean performance than the three comparators, but does not dominate every task. In particular, the Bell model-free baseline matches final fidelity and exceeds nBAUC. The six predeclared global contrasts are listed in Table 6. Their displayed raw p-values imply joint Holm rejection at 0.05; rounded-input adjusted values are provided only as arithmetic reconstruction, not as a new empirical analysis. The largest reconstructed adjusted value is 0.0214.

C.3 FAM-3: OPPORTUNITY COST

The central contrast is Gate–NoVal, not Gate–Measure. Both global primary endpoints have reported raw $p = 10^{-5}$ and Holm-adjusted $p = 2 \times 10^{-5}$. Table 7 reports eight taskwise secondary comparisons. Prepare-1 final fidelity has $p = 0.0869$ and an interval spanning zero. All four task means have positive direction, which must not be described as all eight taskwise tests being significant.

Secondary contrasts. Gate–Measure is 0.06763 ($[0.04086, 0.09668]$, raw $p = 2 \times 10^{-5}$) in nBAUC and 0.09718 ($[0.04806, 0.1441]$, raw $p = 2.6 \times 10^{-4}$) in final fidelity. NoVal–Measure is -0.0605 ($[-0.09329, -0.0264]$) and -0.0825 ($[-0.1382, -0.02871]$), respectively. Its one-sided

Table 5: FAM-2 means (standard deviations) across ten seeds per task and arm. The scalar reward column is a recorded value, not substituted fidelity.

Task	Arm	Final F (SD)	nBAUC (SD)	Mean J
Prepare-1	Measure	0.8380 (0.119)	0.8089 (0.101)	0.8373
Prepare-1	Gate	0.8777 (0.081)	0.8294 (0.063)	0.8770
Prepare-1	Local-CEM	0.8991 (0.095)	0.8026 (0.107)	0.8984
Prepare-1	Model-free	0.2673 (0.134)	0.2440 (0.112)	0.2671
Prepare-+	Measure	0.8029 (0.140)	0.7030 (0.122)	0.8024
Prepare-+	Gate	0.9268 (0.057)	0.7784 (0.120)	0.9262
Prepare-+	Local-CEM	0.7853 (0.181)	0.7576 (0.124)	0.7848
Prepare-+	Model-free	0.7395 (0.152)	0.7373 (0.161)	0.7393
Transmon	Measure	0.4247 (0.238)	0.4157 (0.114)	0.3312
Transmon	Gate	0.6288 (0.161)	0.5030 (0.098)	0.5672
Transmon	Local-CEM	0.5554 (0.157)	0.4631 (0.144)	0.5057
Transmon	Model-free	0.3315 (0.036)	0.3186 (0.045)	0.2621
Bell	Measure	0.4040 (0.147)	0.3298 (0.091)	0.4036
Bell	Gate	0.4934 (0.160)	0.4123 (0.105)	0.4930
Bell	Local-CEM	0.3190 (0.114)	0.3355 (0.074)	0.3187
Bell	Model-free	0.4954 (0.034)	0.4914 (0.022)	0.4952

Table 6: FAM-2 global Gate-minus-comparator contrasts. Holm values are arithmetic reconstructions from the six reported rounded raw p-values, not recomputed tests.

Endpoint	Comparator	Δ	95% interval	Raw p	Holm*
Final F	Measure	0.1143	$[0.0574, 0.1720]$	3.2×10^{-4}	9.6×10^{-4}
Final F	Local-CEM	0.0920	$[0.0436, 0.1396]$	0.0010	0.0020
Final F	Model-free	0.2733	$[0.2254, 0.3193]$	1.0×10^{-5}	6.0×10^{-5}
nBAUC	Measure	0.0664	$[0.0349, 0.0991]$	1.3×10^{-4}	6.5×10^{-4}
nBAUC	Local-CEM	0.0411	$[0.0038, 0.0782]$	0.0214	0.0214
nBAUC	Model-free	0.1829	$[0.1413, 0.2251]$	1.4×10^{-4}	6.5×10^{-4}

tests were oriented toward NoVal superiority, giving $p = 0.999$ and 0.996; they are not evidence of a significant effect in that direction, nor automatically the correct tests for reverse-direction superiority. The difference-of-means identity between the three contrasts is an arithmetic consistency check, not a causal attribution to two independent mechanisms.

C.4 FAM-4 AND FAM-5: STRONGER ZERO-VERIFICATION CONTROLS

FAM-4 tests a conservative model-only deployment alternative under BALD. Conservative-NoVal uses no dedicated candidate–incumbent verification and accepts only when $\overline{d} - \mathrm{Std}(d_m) > 0$ for the five-member paired predicted advantage, with multiplier 1.0 fixed before seeds 40–49 were evaluated. Its raw sign-flip p-values are 2×10^{-5} for nBAUC and 10^{-5} for final fidelity; applying Holm across the two endpoints yields 2×10^{-5} for both. Every task has a positive mean effect on both endpoints. The smallest taskwise nBAUC effect is Bell at $+0.0689$ ($[0.0007, 0.1412]$). Conservative-NoVal accepts 17.9–26.4% of eligible candidates across tasks, excluding a never-replace explanation.

FAM-5 changes the acquisition rule to the implemented decision-aligned value-variance score while otherwise retaining the Gate versus Conservative-NoVal contrast. On seeds 50–59, the raw sign-flip p-value is 10^{-5} for both endpoints and the two-endpoint Holm value is 2×10^{-5}. All four task means are positive for both endpoints; taskwise effects range from approximately $+0.099$ to $+0.287$ across the two reported metrics. Overall replacement rates are 22.9% for Gate and 28.6% for Conservative-NoVal. These later families are evidence against two specific alternative explanations—unconditional deployment and BALD specificity—not evidence of invariance to arbitrary uncertainty estimators, acquisition rules, or world-model classes.

Table 7: FAM-3 taskwise Gate–NoVal effects and exploratory BH adjustment over eight tests. These taskwise comparisons are distinct from the two Holm-adjusted global endpoints.

Task	Endpoint	Δ	95% interval	p	q
Prepare-1	Final F	0.0614	$[-0.0018, 0.1424]$	0.0869	0.0869
Prepare-1	nBAUC	0.0427	$[0.0035, 0.0807]$	0.0342	0.0391
Prepare-+	Final F	0.1842	$[0.0876, 0.2834]$	0.0039	0.0063
Prepare-+	nBAUC	0.2214	$[0.1423, 0.2959]$	0.0019	0.0059
Transmon	Final F	0.2368	$[0.1100, 0.3473]$	0.0049	0.0065
Transmon	nBAUC	0.1302	$[0.0681, 0.1830]$	0.0029	0.0059
Bell	Final F	0.2363	$[0.1260, 0.3495]$	0.0029	0.0059
Bell	nBAUC	0.1182	$[0.0721, 0.1679]$	9.8×10^{-4}	0.0059

Table 8: Later held-out robustness families. Effects are equal-task Gate-minus-comparator differences; positive values favor Gate. FAM-4/5 were designed after analysis of the original families but frozen before their respective new evaluation seeds were run.

Family	Acquisition	Comparator	Δ nBAUC [95% CI]	Δ final F [95% CI]
FAM-4	BALD	Cons-NoVal	$+.1119$ $[.0710, .1538]$	$+.2140$ $[.1444, .2834]$
FAM-5	Value variance	Cons-NoVal	$+.1278$ $[.0985, .1576]$	$+.1907$ $[.1302, .2494]$

C.5 Scalar objective and physical components

Table 9 reports terminal values at 1024 shots. For the qubit and Bell tasks, $F - J$ is the implemented deterministic energy penalty; for transmon it combines $0.5p_2$ leakage and the energy penalty and therefore cannot be equated with energy alone. Mean transmon leakage for Gate is 0.1227, compared with 0.3544 for NoVal and 0.1255 for Measure. These component means and standard deviations are descriptive, not independently multiplicity-corrected component tests. No claim that all energy terms improve is made.

Initial-controller reference. NeverReplace initial-fidelity means (standard deviations) in FAM-3 are: Prepare-1, 0.6881 (0.1150); Prepare-+, 0.6381 (0.3009); transmon, 0.3133 (0.2090); and Bell, 0.2300 (0.1467). A constant-deployment reference assumes the context is unchanged. The values are not a simulator-optimized performance ceiling.

D Replacement-event diagnostics

D.1 Event definitions and denominators

Let N be the number of eligible distinct candidate–incumbent decisions on one arm's trajectory, g acceptance, and d the true scalar advantage. We distinguish

$$a = \frac{\#\{g = 1\}}{N}, \qquad h = \frac{\#\{g = 1, d < 0\}}{N}, \tag{30}$$

$$h_{\text{accepted}} = \frac{\#\{g = 1, d < 0\}}{\#\{g = 1\}} = h/a, \tag{31}$$

$$p_{\text{prevented}} = \frac{\#\{g = 0, d < 0\}}{N}, \qquad p_{\text{missed}} = \frac{\#\{g = 0, d > 0\}}{N}. \tag{32}$$

Thus "prevented" and "missed" in the event table use *all eligible decisions* as denominator, not all harmful or all beneficial candidates. The run-level statistic records whether at least one harmful acceptance occurred anywhere in a run. Exact ties form a separate possible category. These rates exclude ineligible comparisons and hence need not equal the deployment acceptance rate computed over every scheduled round. The FAM-1 audit is retrospective and uses the original held-out trajectories. It is not an additional set of independent learning runs. Gate and Measure generate different candidate distributions after divergent decisions; cross-arm frequencies describe the complete procedures rather than a fixed-event causal intervention. On the Gate trajectories the derived

Table 9: FAM-3 terminal scalar reward, fidelity, leakage, and $F - J$. Parentheses contain standard deviations. For qubit/Bell, $F - J$ is the implemented deterministic energy penalty; for transmon it also includes the leakage penalty. Values are taken from the saved report rather than recalculated from rounded displayed means.

Task	Arm	J (SD)	F	Leakage (SD)	$F - J$
Prepare-1	NoVal	0.8741 (0.108)	0.8748	—	0.000769
Prepare-1	Measure	0.8579 (0.090)	0.8586	—	0.000754
Prepare-1	Gate	0.9355 (0.030)	0.9362	—	0.000747
Prepare-+	NoVal	0.7457 (0.154)	0.7463	—	0.000621
Prepare-+	Measure	0.8298 (0.150)	0.8304	—	0.000629
Prepare-+	Gate	0.9299 (0.043)	0.9305	—	0.000597
Transmon	NoVal	0.2159 (0.213)	0.3938	0.3544 (0.241)	0.177900
Transmon	Measure	0.4309 (0.169)	0.4943	0.1255 (0.105)	0.063390
Transmon	Gate	0.5687 (0.205)	0.6306	0.1227 (0.070)	0.061910
Bell	NoVal	0.2538 (0.145)	0.2541	—	0.000336
Bell	Measure	0.4153 (0.162)	0.4157	—	0.000416
Bell	Gate	0.4900 (0.114)	0.4904	—	0.000356

Table 10: FAM-1 post-hoc eligible-event audit. a: accepted/all; h: harmful accepted/all; p: rejected harmful/all; m: rejected beneficial/all. Last column is the fraction of runs with any harmful acceptance, not an event rate.

Task	Arm	N	a	h	p	m	Any/run
Prepare-1	Measure	92	1.000	0.511	0.000	0.000	1.00
Prepare-1	Gate	116	0.302	0.095	0.629	0.069	0.70
Prepare-+	Measure	129	1.000	0.519	0.000	0.000	1.00
Prepare-+	Gate	130	0.262	0.069	0.631	0.108	0.70
Transmon	Measure	128	1.000	0.539	0.000	0.000	1.00
Transmon	Gate	130	0.285	0.054	0.600	0.115	0.50
Bell	Measure	60	1.000	0.500	0.000	0.000	1.00
Bell	Gate	60	0.283	0.100	0.517	0.200	0.50

conditional harmful fractions among accepted proposals are about 26.5%, 31.4%, 18.9%, and 35.3% for Prepare-+, Prepare-1, transmon, and Bell. Half or more of the Gate runs still contain a harmful acceptance. This limits any safety interpretation of the empirical rule.

Measured versus true advantage on the 436 FAM-1 Gate events has Pearson correlation 0.7816, Spearman correlation 0.7291, bias 0.006483, and MAE 0.1120. There are 66 measured-positive/true-negative events and 31 measured-negative/true-positive events. These correlations are descriptive properties of the visited event distribution, not a calibrated failure-probability guarantee.

D.2 E02-B SCORE DEFINITIONS

Let $\mu = M^{-1}\sum_m d_m$, $s_d = \mathrm{Std}_m(d_m)$, $s_c = \mathrm{Std}_m(\widehat{J}_m(\pi_c))$, and $s_i = \mathrm{Std}_m(\widehat{J}_m(\pi_i))$. The evaluated risk scores are

$$U_c = s_c, \qquad U_i = s_i, \qquad U_\Delta = s_d, \tag{33}$$

$$R_{\mathrm{vote}} = 1 - M^{-1}\sum_m \mathbf{1}\{d_m > 0\}, \qquad R_{\mathrm{margin}} = -\mu/s_d, \qquad R_{\mathrm{pred}} = -\mu, \tag{34}$$

$$R_{\mathrm{obs}} = -\widehat{d}, \qquad R_{\mathrm{reject}} = \theta - \widehat{d}, \qquad R_z = -\widehat{d}/\widehat{\mathrm{se}}. \tag{35}$$

Here θ is the threshold stored by the decision rule. On Gate it is $\delta + z\widehat{\mathrm{se}}$; Measure's logged threshold convention makes its rejection score match negative observed advantage in the aggregate AUROC. Ratio scores are defined when their denominator is nonzero; the aggregate report does not specify the implementation's zero-denominator convention. Reproduction of the exact score tables requires that convention from the analysis script. In particular, the generic disagreement measures do not themselves estimate a signed failure probability.

Table 11: E02-B descriptive pooled AUROC and average precision (AP). Gate has 443 events and 61.6% harmful prevalence; Measure has 418 and 47.4%. AP must be interpreted with these different prevalences.

Score	Gate AUC	Measure AUC	Gate AP	Measure AP
U_c	0.4440	0.4660	0.5687	0.4529
U_i	0.4439	0.4611	0.5839	0.4597
U_Δ	0.4315	0.4262	0.5719	0.4266
R_{vote}	0.4690	0.4969	0.5849	0.4839
R_{margin}	0.4843	0.5250	0.5942	0.4848
R_{pred}	0.5599	0.5737	0.6478	0.5270
R_{obs}	0.8212	0.8022	0.8798	0.7979
R_{reject}	0.8240	0.8022	0.8795	0.7979
R_z	0.8259	0.8008	0.8871	0.7705

Table 12: Taskwise E02-B AUROC on Gate trajectories. The final row is a derived arithmetic mean of the four reported task-specific values, not a pooled-event AUROC.

Task	N	U_Δ	U_c	R_{pred}	R_{obs}	R_{reject}	R_z
Prepare-1	125	0.3793	0.2719	0.6380	0.8057	0.7972	0.7959
Prepare-+	129	0.3710	0.5657	0.5489	0.8434	0.8496	0.8591
Transmon	129	0.3564	0.4008	0.6401	0.8304	0.8253	0.8313
Bell	60	0.3906	0.4747	0.6431	0.8025	0.8025	0.8081
Task mean	—	0.3743	0.4283	0.6175	0.8205	0.8186	0.8236

There are 443 Gate and 418 Measure decisions, for a combined 861. NoVal has no verification measurements, so it is excluded from this matched model-versus-measurement diagnostic; that exclusion does not mean simulator truth could never be computed for its stored controls. The analysis is post-hoc and does not enter the FAM-3 confirmatory family.

The pooled AUROC is based on all events within an arm. The arithmetic average of four within-task AUROCs is a separate equal-task estimand; pooled event count weighting also involves cross-task comparisons and is not generally equivalent to either a simple or an event-weighted average of within-task AUROCs. Table 12 permits this distinction to be checked. We use descriptive point estimates here rather than treat the 443 or 418 within-trajectory decisions as independent replicates. Aggregate interval endpoints alone do not identify the resampling unit or justify a paired test of the difference between scores.

Orientation sensitivity. For a fixed labelled set, reversing a score yields $\mathrm{AUC}(-s) = 1 - \mathrm{AUC}(s)$ with the usual half-credit treatment of ties. Pairwise disagreement is therefore anti-ranked under the original high-uncertainty-is-risky convention. Its reversed pooled values are 0.5685 on Gate and 0.5738 on Measure. This remains below the observed-advantage values in these data, but is not evidence that all uncertainty information is absent. No held-out-tuned sign reversal is presented as a trained, independently evaluated model. Similarly, selecting the best of several measured scores after observing the table is not a new confirmatory method selection result; the main text emphasizes the direct observed advantage.

D.3 Calibration summaries

The regression of true on measured advantage has slope 0.6083 on Gate and 0.5795 on Measure, rather than one. Correlation and detection AUROC consequently do not imply calibrated effect magnitude. Finite-shot noise can induce such shrinkage in a regression of true values on noisy measurements; the slope alone is not a proof of estimator bias. The reported average biases are small relative to RMSE, but this is not a finite-sample coverage test of the empirical gate.

Table 13: FAM-3 observed-versus-true advantage summaries. Slope is OLS of true advantage on measured advantage; bias is mean measured minus true. These do not constitute a coverage guarantee.

Arm	N	Pearson	Spearman	Slope	Bias	MAE	RMSE
Gate	443	0.7702	0.6992	0.6083	-0.00386	0.1175	0.1515
Measure	418	0.7361	0.6683	0.5795	-0.00478	0.1315	0.1637

D.4 Exploratory source and optimism analyses

The development audit of 223 model-favored pairs reports 139 harmful candidates. The original equal-task summary gave inversion rate 0.620 and diagnostic AUROC 0.412 for ensemble disagreement versus 0.795 for a finite-shot gate score. The later event-pooled table on those 223 events reports 0.4275 and 0.8020, respectively. These values correspond to different aggregation conventions; they are not substituted for one another or for the new FAM-3 held-out diagnostic.

Development CEM-minus-random mean optimism is 0.06045 for Prepare-+, -0.04550 for Prepare-1, 0.03777 for transmon, and 0.13930 for Bell. This nonuniform pattern does not establish that optimization always amplifies optimism. The controls are not a documented margin-matched causal comparison, and favorable passive predictive accuracy has not been established. Accordingly, these exploratory observations motivate the decision audit without supporting a stronger average-accuracy-versus-optimization claim.

E DEVELOPMENT ROBUSTNESS, ASSURANCE, AND NUMERICAL VALIDATION

E.1 ABSOLUTE VERIFICATION-ALLOCATION CURVES

Table 14: Development allocation sweep, absolute means. Measure pays for validation but always accepts distinct candidates. Fraction refers to the active-round budget, not necessarily the complete budget including initialization.

Task	n	Fraction	Measure F	Gate F	Measure AUC	Gate AUC
Prepare-+	4	0.1250	0.8765	0.8717	0.7488	0.8066
Prepare-+	8	0.2500	0.9022	0.8879	0.7840	0.8154
Prepare-+	16	0.5000	0.8832	0.9390	0.7652	0.8397
Prepare-+	24	0.7500	0.8815	0.9077	0.7821	0.8241
Bell	2	0.0938	0.3060	0.3363	0.2716	0.3356
Bell	4	0.1875	0.5208	0.3763	0.4153	0.3803
Bell	8	0.3750	0.3548	0.4572	0.2960	0.3407
Bell	16	0.7500	0.3213	0.5008	0.2450	0.3792

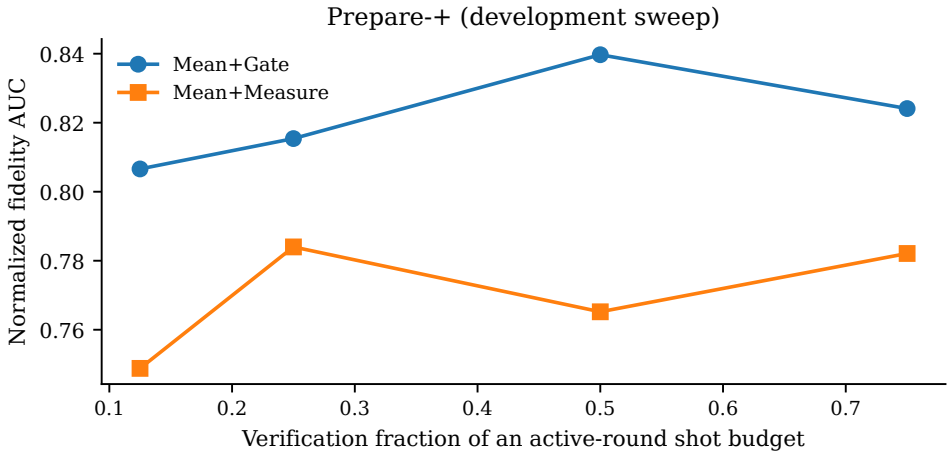

Figure 3: Prepare-+ development allocation sweep: absolute nBAUC for Gate and Measure. Each point is a reported configuration mean; lines guide the eye, not a fitted response model. No zero-verification point from another seed family is appended to this curve.

Table 14 retains both absolute arm values. A maximum of Gate–Measure is not necessarily a maximum of Gate itself. For Prepare-+, the tested $n = 16$ has the highest Gate final mean, but these small development comparisons do not establish a general optimum. For Bell, the nBAUC mean at $n = 4$ is 0.3803 and at $n = 16$ is 0.3792, despite the larger final-fidelity mean at $n = 16$. Budget allocation must specify the endpoint as well as the cost model. FAM-3 separately confirms the frozen allocation against a zero-verification alternative on new contexts.

E.2 THRESHOLD SENSITIVITY AND DEVELOPMENT ABLATIONS

The tested z grid is $\{0, 0.25, 0.5, 0.75, 1, 1.64\}$. The development summary marks all grid points within one standard error of the best taskwise mean. This indicates imprecise separation on that sample, not statistical equivalence of thresholds or proof that tuning is irrelevant. Bell's nBAUC at $z = 1.64$ is lower than at intermediate values. The confirmatory studies retain $z = 0.5$ rather than reselect it using held-out outcomes.

Earlier development-only planner ablations show gate-alone fidelity effects of 0.1666 on transmon and 0.1469 on Bell, whereas calibrated-planning-alone effects are -0.09336 and 0.01443. Adding calibration to an already gated planner gives -0.1014 and -0.0258. These comparisons motivate

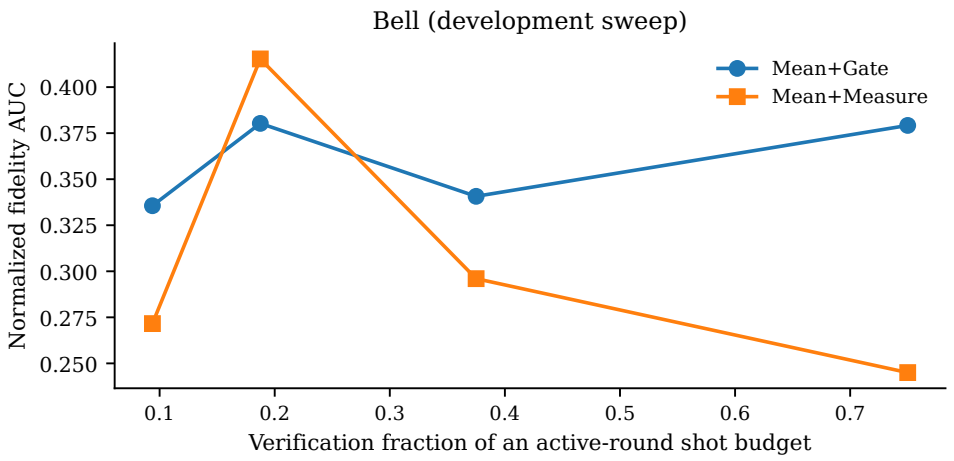

Figure 4: Bell development allocation sweep. Different numbers of measurement settings make an equal shots-per-setting value have a different physical cost from the qubit tasks.

Table 15: Development nBAUC under different empirical thresholds. The frozen confirmatory setting remains $z = 0.5$. No equivalence between grid points is inferred.

z	Prepare-1	Prepare-+	Transmon	Bell
0	0.8604	0.8569	0.5109	0.3725
0.25	0.8601	0.8527	0.5700	0.3771
0.5	0.8599	0.8397	0.5306	0.3792
0.75	0.8674	0.8385	0.5429	0.3861
1	0.8728	0.8397	0.5541	0.3451
1.64	0.8463	0.8279	0.5567	0.3133

focusing on verification rather than making calibration a headline contribution. Early acquisition explorations are not an independent confirmatory evaluation of a new acquisition algorithm.

E.3 Empirical versus confidence-valid verification

The assurance table covers two development tasks with five seeds per arm, not an empirical validation of the theorem on all four tasks. The declared family-wise level is $\alpha = 0.05$. Hoeffding accepts no Bell candidates and about 3.1% of Prepare-+ candidates in these runs. Zero observed harmful accepts under such conservative behavior is not a test proving the family-wise guarantee. In addition, the executed nonzero-OU simulator shares one process-noise trajectory within each probe batch while the nominal offline objective omits that draw, so the shot-only radius is not interpreted as an established certificate for the nominal objective in these runs; Appendix A gives the exact scope.

Empirical-minus-Hoeffding final-fidelity differences are 0.07216 ($[0.03447, 0.1111]$) on Prepare-+ and 0.2159 ($[0.09344, 0.3410]$) on Bell. Corresponding nBAUC differences are 0.01808 ($[-0.02886, 0.06110]$) and 0.09436 ($[0.03958, 0.1491]$). These are small-sample development characterizations. No adaptive-sequential gate is evaluated in this work.

E.4 Independent propagation checks

The primary complex64 RK4 simulator is checked against an independently implemented complex128 Liouvillian matrix exponential. For each task, the audit includes ten final deployed controls, twenty gate-boundary candidates, and twenty random controls, totalling 200 evaluations. The largest absolute fidelity discrepancy is 6.188×10^{-4} on transmon final controls; the largest absolute reward discrepancy is 8.448×10^{-4}, and the largest trace distance is 1.58×10^{-3}. These are different metrics and should not share a single unlabeled "error" figure.

Table 16: Two-task development assurance comparison, five seeds per arm. Harmful frequency uses all eligible decisions; the final column records runs with any harmful acceptance. No observed violations at very low acceptance do not empirically establish a nominal error rate.

Task	Gate	Final F	nBAUC	a	h	Any/run
Prepare-+	Empirical	0.9390	0.8397	0.3385	0.1231	1.00
Prepare-+	Hoeffding	0.8669	0.8216	0.0308	0.0000	0.00
Bell	Empirical	0.5008	0.3792	0.2667	0.0333	0.20
Bell	Hoeffding	0.2848	0.2848	0.0000	0.0000	0.00

Table 17: Independent simulator audit: maximum absolute reward error, fidelity error, and trace distance within each policy group. Fifty controls per task, 200 total.

Task	Controls	N	Max $\lvert\Delta J\rvert$	Max $\lvert\Delta F\rvert$	Max trace dist.
Prepare-1	Deployed	10	1.29×10^{-5}	1.29×10^{-5}	3.78×10^{-5}
Prepare-1	Boundary	20	1.44×10^{-5}	1.44×10^{-5}	3.85×10^{-5}
Prepare-1	Random	20	8.58×10^{-7}	8.57×10^{-7}	1.53×10^{-6}
Prepare-+	Deployed	10	6.17×10^{-6}	6.15×10^{-6}	1.71×10^{-5}
Prepare-+	Boundary	20	4.93×10^{-6}	4.95×10^{-6}	2.27×10^{-5}
Prepare-+	Random	20	1.08×10^{-6}	1.07×10^{-6}	1.94×10^{-6}
Transmon	Deployed	10	0.000845	0.000619	0.00132
Transmon	Boundary	20	0.000574	0.000409	0.00158
Transmon	Random	20	3.17×10^{-5}	2.49×10^{-5}	2.82×10^{-5}
Bell	Deployed	10	7.14×10^{-5}	7.15×10^{-5}	0.000259
Bell	Boundary	20	7.14×10^{-5}	7.15×10^{-5}	0.000328
Bell	Random	20	2.22×10^{-5}	2.22×10^{-5}	4.73×10^{-5}

Agreement between two numerical propagators checks numerical implementation, not the physical accuracy of the assumed Lindblad system. Small absolute error is also insufficient to guarantee the sign of every nearly tied pair. No claim that all tiny negative advantages are numerically resolved is made.

F Accounting, statistical scope, and data provenance

Ledger checks. The original experiment report records 80/80 FAM-1 and 120/120 FAM-3 runs passing six checks: chained shot increments, terminal totals, verification-cost fields, acceptance semantics, nested record consistency, and charging same-policy measurements. All 40 pairs in each family have a first acceptance disagreement; 37 and 38 pairs subsequently have diverging trajectories. These numbers are evidence against an identical-data interpretation of the closed-loop comparisons, not a defect in equal-budget pairing. The report does not document the same full ledger audit for all FAM-2 comparison runs, so the original 200-run audit count is not enlarged to include them.

The later FAM-4 and FAM-5 ledgers each record 80/80 completed runs and all declared C1–C7 checks passing. In each family, all 40 Conservative-NoVal runs use zero verification shots, have no observed candidate/incumbent validation fields, and retain finite model-side lower-confidence scores. All 80 runs per family terminate at 1024 physical shots. These checks establish the intended accounting and branch semantics; they do not convert the later, post-analysis family design into part of the original FAM-1–3 confirmatory plan.

The FAM-3 per-run totals give 896 post-initialization shots and a 128-shot remainder to the terminal budget. Qubit and transmon verification uses 448 shots, consistent with fourteen 32-shot comparisons; Bell uses 672, consistent with seven 96-shot comparisons. The corresponding allocation implied by the totals is 32 learning plus 32 verification shots per active qubit/transmon round, and 32 plus 96 for Bell. Exact query identities and update timing remain properties of the run logs.

Table 18: FAM-3 per-run physical-shot totals. Initial allocation is the 128-shot remainder between reported probe/validation totals and the terminal budget. Values are not multiplied by the ten seeds.

Tasks	Arm	Initial	Probes	Verify	Total
Qubits/transmon	Gate	128	448	448	1024
Qubits/transmon	Measure	128	448	448	1024
Qubits/transmon	NoVal	128	896	0	1024
Bell	Gate	128	224	672	1024
Bell	Measure	128	224	672	1024
Bell	NoVal	128	896	0	1024

Objective checks. The source-level reconciliation identifies two manuscript-description issues corrected in Appendix B: the executed energy penalty averages over both time and control channels, and the qubit scalar objective includes that deterministic energy term. The code also makes explicit that the world model fits observed counts without an explicit readout-mixing layer, while the empirical gate applies inverse readout correction. These are implementation facts, not claims of physical calibration. Raw-count replay of the saved Gate/Measure records reproduces 1,960 policy estimates over 980 deployment events to numerical precision under the released estimator functions.

Uncertainty and multiplicity. The primary performance procedures resample or sign-flip paired task–seed effects. Under the paired sign-flip null, signs are exchangeable within the relevant strata; the resulting one-sided tests have that inferential scope. Taskwise tests on ten pairs enumerate 2^{10} sign patterns. The frozen FAM-3 global analysis uses 50,000 task-stratified paired bootstrap resamples and 100,000 task-stratified sign-flip draws, with deterministic analysis seeds; the reported raw value 9.9999×10^{-6} is the minimum nonzero Monte Carlo value under the $(1 + k)/(R + 1)$ convention at $R = 100{,}000$. FAM-4 and FAM-5 remain separate two-endpoint families; their global effects are not pooled with FAM-3. Diagnostic decisions within a trajectory are correlated; saved FAM-3 diagnostic intervals therefore resample task–arm–seed clusters rather than treating events as independent.

Holm controls each stated family of endpoint claims. Defining several families does not imply control of an arbitrary post-hoc claim selected across all studies. The abstract emphasizes the original FAM-3 opportunity-cost result and identifies FAM-4/5 as later held-out robustness families. Effect estimates, units, sample counts, temporal ordering, and multiplicity families must travel together when moving results into the abstract or conclusion.

What can be reproduced from the released materials. The manuscript archive bundles the numerical tables and figure scripts. The separate anonymized experiment supplement provides the executable implementation, frozen configurations, saved datasets/counts, deployment logs, tests, and analysis scripts for the released experiment package. Recomputing FAM-3 endpoints from saved per-run metrics reproduces the headline effects and intervals to floating-point precision; replaying saved counts through the released estimators reproduces logged deployment estimates. These checks are raw-record reanalysis using the released implementation, not an independent reimplementation or a claim of bitwise-identical retraining. Model checkpoints were not saved.