SAMPLING AND KRYLOV-TRUNCATION UNCERTAINTY IN QUANTUM FISHER INFORMATION ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating quantum Fisher information (QFI) from randomized measurements with a truncated Krylov representation introduces two errors: sampling error around the fixed-order target and truncation error relative to the full-space QFI. Bootstrap intervals quantify the former but can become narrow around a biased target, making width-based stopping unreliable. We formulate adaptive stopping as selective uncertainty quantification, with separate sampling and truncation radii and abstention when their sum does not support the requested tolerance. For independent confirmation data, we prove finite-horizon false-declaration control provided that the combined radius is a valid total-error bound. We also introduce Krylov-first positivity stabilization to mitigate finite-sample plug-in bias. With frozen within-instance truncation calibration, a six-qubit experiment using $|+\rangle^{\otimes 6}$ at $K = 48 < 64$ yields 19 declarations in 20 held-out runs, no observed false declarations, and 4.04% median relative error at a 10% target. Under matched random draws, full-space positivity produces 20 false declarations, whereas Krylov-first stabilization produces 16 correct declarations and four abstentions. The results demonstrate accurate selective QFI decisions at non-full Krylov resolution while identifying transferable truncation control as the central scalability challenge.

1 INTRODUCTION

Uncertainty quantification (UQ) is most useful when it controls the error of the quantity that a downstream decision actually uses. This requirement becomes subtle in adaptive estimators that combine stochastic observations with an iterative approximation. A confidence interval may accurately describe finite-sample variation around the current approximate target while omitting the bias between that target and the desired quantity. Increasing the sample count then narrows the interval without resolving the approximation error.

Quantum Fisher information (QFI) provides a concrete instance of this two-source UQ problem. QFI determines the local sensitivity of a parameterized quantum state and the precision limit of quantum sensing protocols (Braunstein & Caves, 1994; Giovannetti et al., 2011; Taylor & Bowen, 2016; Degen et al., 2017; Liu et al., 2020). Exact mixed-state QFI evaluation becomes expensive as the Hilbert space grows. Randomized-measurement and classical-shadow estimators replace direct access to the state by randomized observations (Aaronson, 2020; Huang et al., 2020; Elben et al., 2023; Rath et al., 2021; Beckey et al., 2022; Vitale et al., 2024). Krylov-shadow estimation additionally projects the QFI problem onto a subspace of order K (Zhang & Tong, 2025). The sample count M controls uncertainty at the selected order, whereas K controls the approximation of the full QFI. These resources need not become adequate at the same time.

This paper studies the decision made on top of the estimator: when can an adaptive routine attach a success declaration to its current QFI estimate? We call a declaration *false* when the reported estimate lies outside a target tolerance, as determined by an exact reference available only for evaluation. The central failure mode is simple but consequential. A bootstrap interval may have valid coverage for the Krylov-truncated population value $\mathcal{F}_K$ while remaining far from the desired value $\mathcal{F}$. Thus sampling convergence cannot, by itself, justify stopping.

Sequential testing and confidence sequences address data-dependent sampling and optional stopping (Wald, 1945; Howard et al., 2021). Adaptive allocation methods similarly decide where additional

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.

observations are most valuable (Even-Dar et al., 2002; Lattimore & Szepesvári, 2020). These tools control statistical uncertainty relative to a specified target. The QFI setting studied here adds a separate question: whether the computational target at the current Krylov order is itself close to the desired full-space quantity.

We organize the problem as selective UQ. A run outputs an estimate together with either a success declaration or abstention at the resource limit. This view follows the error–rejection tradeoff in selective prediction (Chow, 1970). It makes two quantities inseparable: the error risk among declarations and the fraction of instances on which the procedure declares. In particular, zero false declarations obtained with zero declarations is a safety result, not a positive accuracy result.

Our contributions are as follows.

- We isolate sampling error and Krylov approximation error in adaptive QFI estimation and show why a fixed-order bootstrap interval targets $\mathcal{F}_K$, rather than the full QFI $\mathcal{F}$.
- We introduce AKS-QFI, a component-aware interface that records separate observable scores for the two error sources. We distinguish its empirical guarded rule from a held-out rule whose finite-horizon guarantee requires calibrated component radii.
- We evaluate stopping as a selective decision using declaration rate, false-declaration rate, precision among declarations, and measurement cost. The experiments include the negative and abstention regimes rather than reporting zero false declarations in isolation.
- We introduce Krylov-first positivity stabilization for shadow reconstructions. At six qubits, a directly preparable product-state starting vector yields 19 correct declarations and one abstention at $K = 48 < 64$. A paired study at $K = 40$ changes the outcome from 20 false declarations under full-space stabilization to 16 correct declarations and four abstentions.
- We document a calibration-transfer boundary: the four-qubit guard suppresses premature declarations, but unchanged parameters fail at larger system sizes because local Krylov stability is not a transferable upper bound on approximation error.

Together, the experiments identify a sampling-width failure, a projection-order effect, and a calibrated non-full six-qubit regime. Extending that regime to unseen physical instances requires a transferable approximation bound; Section 6 discusses this remaining requirement.

2 Quantum Fisher Information with Two Sources of Uncertainty

2.1 Quantum Fisher information

For a differentiable family of density operators ρ_θ, the symmetric logarithmic derivative L_θ satisfies

$$\partial_\theta \rho_\theta = \frac{1}{2}(\rho_\theta L_\theta + L_\theta \rho_\theta), \tag{1}$$

and the QFI is $\mathcal{F}(\rho_\theta) = \mathrm{Tr}(\rho_\theta L_\theta^2)$. Equivalently, if $\rho_\theta = \sum_i \lambda_i |i\rangle\langle i|$,

$$\mathcal{F}(\rho_\theta) = 2 \sum_{i,j:\lambda_i+\lambda_j>0} \frac{|\langle i|\partial_\theta \rho_\theta|j\rangle|^2}{\lambda_i + \lambda_j}. \tag{2}$$

Equation (2) is used only to compute post hoc reference values in the controlled experiments. The spectral form follows the symmetric-logarithmic-derivative formulation (Petz & Ghinea, 2011; Meyer, 2021).

At Krylov order K, let $\mathcal{F}_K$ denote the infinite-sample population value of the projected estimator, and let $\widehat{\mathcal{F}}_{K,M}$ be its estimate from M randomized measurements. The total error obeys the exact decomposition

$$\widehat{\mathcal{F}}_{K,M} - \mathcal{F} = \underbrace{(\widehat{\mathcal{F}}_{K,M} - \mathcal{F}_K)}_{S_{K,M}\ \text{(sampling)}} + \underbrace{(\mathcal{F}_K - \mathcal{F})}_{B_K\ \text{(approximation)}}. \tag{3}$$

The sample count can reduce $S_{K,M}$ but cannot reduce the Krylov bias B_K.

2.2 WHAT A BOOTSTRAP INTERVAL COVERS

Let $[\widehat{L}_{K,M}, \widehat{U}_{K,M}]$ be an equal-tailed bootstrap interval constructed from the M shadows at fixed K, with width

$$w_{K,M} = \widehat{U}_{K,M} - \widehat{L}_{K,M}. \tag{4}$$

At fixed K, the interval targets sampling variation around $\mathcal{F}_K$ (Efron, 1979; van der Vaart, 1998; Ghysels & Morgan, 2025). Its coverage for the nonlinear plug-in estimator is not established here, and it does not include B_K unless that bias is separately bounded. Consequently, $w_{K,M} \le \varepsilon$ is not a total-error guarantee.

For independent local-Pauli snapshots X_m, the linear reconstruction $\widetilde{\rho}_M = M^{-1}\sum_m X_m$ has the exact state-level second moment

$$\mathbb{E}\|\widetilde{\rho}_M - \rho\|_{\mathrm{F}}^2 = \frac{5^n - \mathrm{Tr}(\rho^2)}{M}. \tag{5}$$

Each one-qubit snapshot has eigenvalues $2, -1$, so $\|X_m\|_{\mathrm{F}}^2 = 5^n$ and $\mathbb{E}X_m = \rho$; independence cancels the cross terms in the squared sample mean. Markov's inequality gives the state radius $\sqrt{5^n/(M\delta)}$ at failure level δ, but this is not a QFI radius without regularity of the nonlinear positivity and QFI maps. Section S3 gives the full proof and a full-rank continuity bound.

The empirical resolution-side statistic used in this work is the adjacent-order change

$$d_{K,M} = |\widehat{\mathcal{F}}_{K,M} - \widehat{\mathcal{F}}_{K-1,M}|. \tag{6}$$

This quantity detects local motion of the Krylov sequence. It is informative for allocation, but it is not generally an upper bound on $|B_K|$. A slowly changing biased sequence may have small $d_{K,M}$ and large $|B_K|$.

The failure persists with no sampling noise. Set $G = \mathrm{diag}(0, 1, 2)$, start from $v = (|0\rangle + |1\rangle)/\sqrt{2}$, and let $\rho_\eta = (1-\eta)|\psi\rangle\langle\psi| + \eta I_3/3$ with $|\psi\rangle = (|0\rangle + |2\rangle)/\sqrt{2}$ and $0 < \eta < 1$. The weighted population QFI is zero at both Krylov orders $K = 1$ and $K = 2$, while the full QFI is

$$\mathcal{F}(\rho_\eta, G) = \frac{4(1-\eta)^2}{1 - \eta/3} > 0. \tag{7}$$

Indeed, the order-two space is $\mathrm{span}\{|0\rangle, |1\rangle\}$; compression removes the $|0\rangle\langle 2|$ coherence, leaving a state diagonal in the compressed generator basis. The full state retains that coherence, and Eq. (2) gives Eq. (7). Thus no finite universal constant can bound $|B_2|$ by the adjacent-order change. The full proof appears in Supplementary Material, Section S3.

2.3 KRYLOV-FIRST POSITIVITY STABILIZATION

Let $\widetilde{\rho}_M$ be the linear shadow reconstruction and let Q_K contain the Krylov basis vectors. The original plug-in estimator first applied a full-space positivity projection Π_+ to $\widetilde{\rho}_M$ and then formed $Q_K^\dagger \Pi_+(\widetilde{\rho}_M) Q_K$. At finite M, this nonlinear projection distributes weight over the ambient Hilbert space and can strongly shrink the mass retained by a non-full Krylov subspace.

We instead reduce the unbiased reconstruction before stabilization:

$$\widetilde{\rho}_{K,M} = Q_K^\dagger \widetilde{\rho}_M Q_K, \qquad \widehat{\rho}_{K,M} = \Pi_+(\widetilde{\rho}_{K,M}). \tag{8}$$

The projected QFI is evaluated from $\widehat{\rho}_{K,M}$ and $G_K = Q_K^\dagger G Q_K$, multiplied by the clipped empirical subspace weight $\mathrm{Tr}(\widetilde{\rho}_{K,M})$. For a valid population density matrix, the two orders agree up to numerical precision. For a noisy shadow reconstruction, Eq. (8) confines the nonlinear positivity operation to K dimensions. It does not remove Krylov bias, which remains controlled separately by $R_K^{(\mathrm{approx})}$.

3 SELECTIVE, COMPONENT-AWARE STOPPING

3.1 OUTPUTS AND EMPIRICAL DECISION RULE

AKS-QFI wraps a fixed-(K, M) Krylov-shadow estimator. Given tolerance ε and resource limits $(K_{\max}, M_{\max})$, it returns a point estimate, the terminal resource pair, the scores $(d_{K,M}, w_{K,M})$, and one of two labels:

1. **declare**: the empirical gates have passed; or
2. **abstain**: the resource limit was reached without sufficient evidence for a declaration.

The observable severity score

$$\widetilde{I}_{K,M} = \max\{d_{K,M}, w_{K,M}\} \tag{9}$$

routes resource allocation. If the resolution-side score is large, the rule increases K; otherwise, if the sampling width is large, it increases M. The empirical guard additionally requires minimum resources $(K_{\min}, M_{\min})$ and P consecutive eligible passes. In the experiments, $(K_{\min}, M_{\min}, P) = (4, 128, 2)$ for the four-qubit family.

The empirical rule includes a finite-resource convention: reaching $K_{\max}$ passes its resolution-side gate because no further order is available. This convention produces a well-defined terminal decision, but it must not be confused with evidence that the approximation error is small. The distinction between this practical guard and the calibrated rule below is central to the paper.

3.2 A CALIBRATED HELD-OUT RULE

Suppose that $R_K^{(\mathrm{approx})}$ is a valid upper bound on $|\mathcal{F}_K - \mathcal{F}|$ for the target instance or specified family, and that an independent confirmation batch provides a sampling radius $R_j^{(\mathrm{sample})}(\delta_j)$. The held-out rule declares only if

$$R_K^{(\mathrm{approx})} + R_j^{(\mathrm{sample})}(\delta_j) \le \varepsilon. \tag{10}$$

Exploration data choose a candidate pair, but the returned estimate and the sampling radius come from the independent confirmation batch.

Theorem 1 (Finite-horizon false-declaration control)**.** *Consider at most J declaration attempts. Conditional on the exploration history before attempt j, assume that the confirmation estimate $\widehat{\mathcal{F}}_j^{\mathrm{conf}}$ satisfies*

$$\Pr\Big(|\widehat{\mathcal{F}}_j^{\mathrm{conf}} - \mathcal{F}| > R_{K_j}^{(\mathrm{approx})} + R_j^{(\mathrm{sample})}(\delta_j) \,\Big|\, \mathcal{H}_j\Big) \le \delta_j. \tag{11}$$

If the procedure declares only when Eq. (10) *holds and $\sum_{j=1}^{J} \delta_j \le \delta$, then*

$$\Pr\Big(\exists j \le J :\ \textit{declare at } j \textit{ and } |\widehat{\mathcal{F}}_j^{\mathrm{conf}} - \mathcal{F}| > \varepsilon\Big) \le \delta. \tag{12}$$

The proof is a conditional union bound and is given in Supplementary Material, Section S1. The substantive requirement is the availability of a valid approximation radius. The observable adjacent-order difference $d_{K,M}$ is not substituted for this radius in the theorem. The primary and transfer studies therefore evaluate the empirical rule as an empirical selective decision. The six-qubit study instantiates the held-out architecture with an exactly calibrated truncation radius and an empirical bootstrap sampling radius. Because bootstrap validity is not proved for this plug-in estimator, that experiment provides held-out evidence rather than a formal application of Theorem 1.

4 EVALUATION PROTOCOL

4.1 CONTROLLED MIXED-STATE FAMILY

We use noisy mixed states with depolarizing probability $p_{\mathrm{dep}} = 0.03$ and dephasing probability $p_\phi \in \{0, 0.06, 0.12, 0.18, 0.24\}$. The primary study uses four qubits, absolute tolerance $\varepsilon = 0.2$, resource limits $(K_{\max}, M_{\max}) = (8, 512)$, and 50 independent runs per noise level. Exact QFI values from Eq. (2) label terminal decisions only after each run; they are not available to the stopping rule.

We compare width-only stopping with the component-aware empirical guard. A second four-qubit control uses relative tolerance 0.05, full order $K = 16$, and $M \in \{32768, 65536, 131072, 262144\}$. Finally, transfer stress tests apply the four-qubit guard without recalibration to six- and eight-qubit systems, with 10 runs per noise level. These transfer runs test robustness of the chosen calibration, not asymptotic scaling.

4.2 FROZEN SIX-QUBIT STUDIES

The six-qubit studies use $p_\phi = p_{\mathrm{dep}} = 0.03$, a fixed random-local generator, and 10% relative error, giving $\mathcal{F} = 4.3131$ and $\varepsilon = 0.4313$. The main evaluation starts the Krylov construction from the directly preparable product state $|+\rangle^{\otimes 6}$, where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$, which requires no knowledge of the target density matrix. Exact-density calculation performed before sampling selects $K = 48 < 2^6 = 64$ and gives $R_{48}^{(\mathrm{approx})} = 0.0974$. Three pilot repetitions fix $M_{\max} = 131072$; the rule is then locked and evaluated in 20 new sampling repetitions. Each attempted declaration uses an independent confirmation batch and Bonferroni allocation over at most 12 attempts.

A separate paired study isolates the order of positivity stabilization. It uses the target density matrix's dominant eigenvector, $K = 40$, and $R_{40}^{(\mathrm{approx})} = 0.0804$, then compares full-space positivity with Eq. (8) under identical random draws. Exact QFI and exact truncation errors are used only for offline calibration and evaluation. The product-state study removes the oracle starting vector from the strongest result, but both studies retain exact within-instance truncation calibration.

4.3 SELECTIVE UQ METRICS

Let D denote a success declaration and let $A = \{|\widehat{\mathcal{F}} - \mathcal{F}_{\mathrm{ref}}| \le \varepsilon\}$ denote post hoc accuracy. We report

$$\text{declaration rate} = \Pr(D), \tag{13}$$

$$\text{false-declaration rate} = \Pr(D \cap A^c), \tag{14}$$

$$\text{precision among declarations} = \Pr(A \mid D), \tag{15}$$

together with terminal measurement count. Declaration rate is the selective coverage of the rule. Reporting it prevents a procedure that always abstains from appearing preferable solely because its marginal false-declaration rate is zero.

5 RESULTS

The strongest operational result is the frozen product-start six-qubit study: 19 of 20 held-out runs declare correctly at non-full Krylov order. The four-qubit controls explain why the additional error component matters, while the transfer tests delimit the scope of a frozen calibration.

5.1 SAMPLING WIDTH CAN BE SMALL AROUND A BIASED KRYLOV TARGET

Figure 1 plots terminal bootstrap width against true absolute error in the primary four-qubit study. Width-only runs populate the region $w_{K,M} \le \varepsilon$ but $|\widehat{\mathcal{F}}_{K,M} - \mathcal{F}_{\mathrm{ref}}| > \varepsilon$. The interval is narrow at its fixed-order target, yet the target is biased. This is the empirical counterpart of Eq. (3).

5.2 THE PRIMARY GUARD IS SAFE THROUGH ABSTENTION

Width-only stopping declares in 106 of 250 primary runs, and every declaration is inaccurate. Its false-declaration rate increases from 0.16 to 0.68 across the dephasing grid. The empirical guard makes no declaration in the same 250 runs. Its false-declaration rate is therefore zero, but so is its declaration rate. Figure 2 displays both quantities. The guard demonstrates that a selective interface can refuse an unsupported claim; this experiment alone does not demonstrate useful positive coverage.

5.3 A POSITIVE CONTROL REQUIRES FULL FOUR-QUBIT RESOLUTION

In the relative-tolerance control, exact-density calibration shows that $K = 12$ retains relative approximation errors from 2.24% to 16.80% over the tested points, whereas $K = 16$ removes this error for the four-qubit Hilbert space. At $K = 16$, the guarded sample-count schedule declares in 12 of 20 runs; all 12 declarations are accurate, and all 20 terminal estimates are within the target. The matched width-only rule declares in 19 runs and makes 15 false declarations. This control shows that sampling-side selection behaves as intended once approximation error has been eliminated. Because $K = 16$ is full resolution for four qubits, it is not evidence of scalable Krylov calibration.

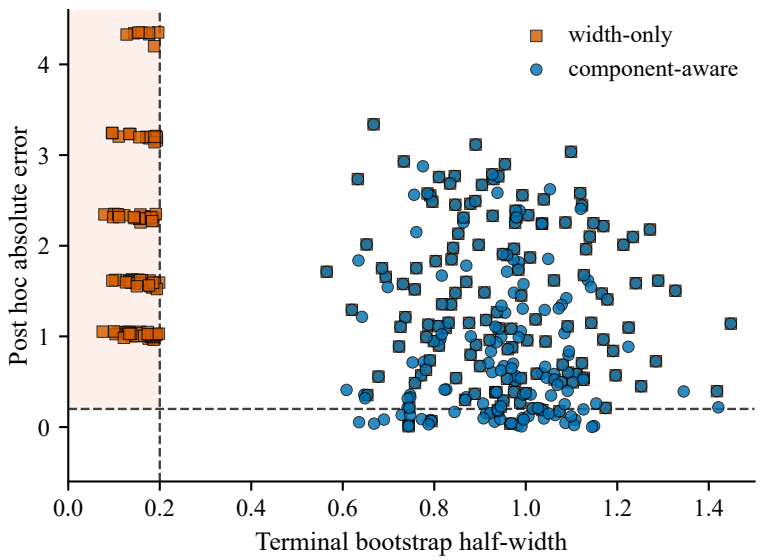

Figure 1: Narrow sampling intervals can accompany inaccurate full-QFI estimates. Terminal width is plotted against post hoc error in the four-qubit study; dashed lines mark the stopping-width and accuracy thresholds. Upper-left points are false declarations with unresolved Krylov approximation error.

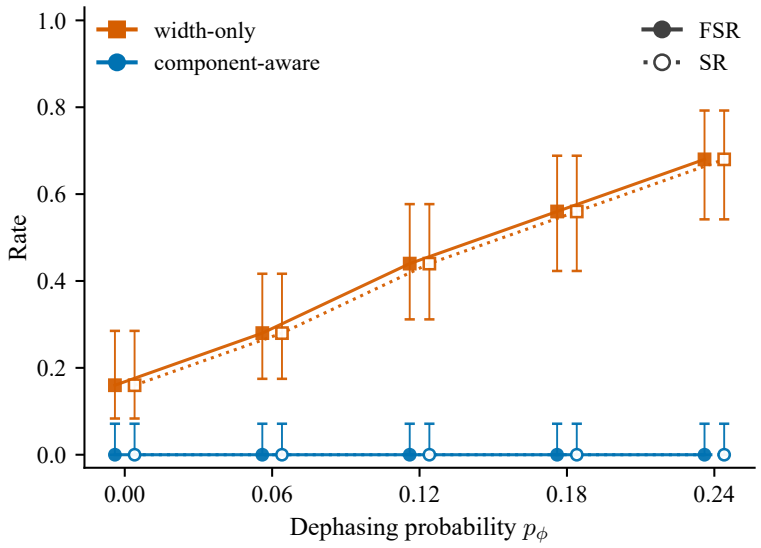

Figure 2: The primary guard has zero false declarations because it abstains at every tested point; every width-only declaration is false. Filled markers show false-declaration rate and open markers show declaration rate across the four-qubit dephasing grid. Error bars are 95% Wilson intervals.

5.4 Frozen non-full six-qubit study has positive held-out coverage

Figure 4 reports the frozen six-qubit studies. With the product-state starting vector, the Krylov-first rule declares in 19 of 20 held-out runs and abstains once, with no observed false declaration. One declaration stops at $M = 32768$, 11 at $M = 65536$, and seven at $M = 131072$. The median terminal relative error is 4.04%; the largest error among declared runs is 9.96%, below the locked 10% target. The 95% Wilson interval for declaration rate is $[0.764, 0.991]$. For the observed false-declaration rate $0/20$, its upper endpoint is 0.161. All 20 terminal estimates satisfy the target, including the single abstention.

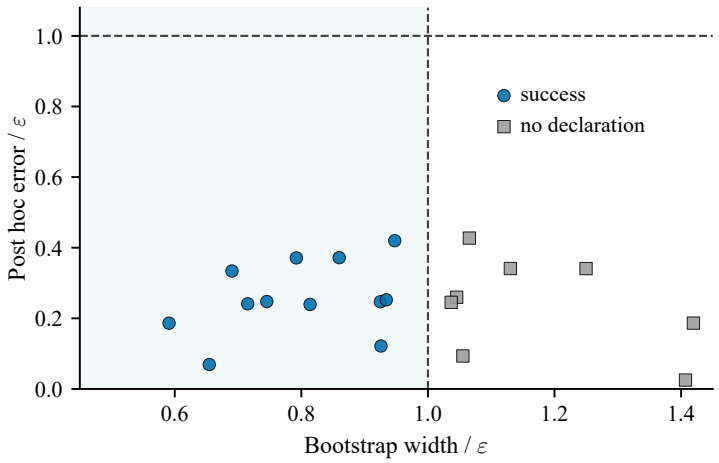

Figure 3: Relative-tolerance control at full four-qubit Krylov resolution. Blue circles are declarations and gray squares are abstentions. Every terminal estimate is accurate, while declarations occur only when the sampling-width gate also passes.

The dominant-eigenvector comparison provides the paired mechanism test. Full-space positivity declares in all 20 runs, yet every declaration is false, with median relative error 32.1%. Krylov-first stabilization on the same random draws makes 16 correct declarations and four abstentions, with median relative error 4.61%.

The state, $K = 40$, tolerance, measurement schedule, and truncation radius are matched. The contrast therefore isolates the finite-sample plug-in bias caused by enforcing positivity in the 64-dimensional ambient space before projection.

5.5 Declarations occur before the sample cap

For the product-state study, the approximation radius is fixed at 0.0974 and the median held-out sampling radius among declarations is 0.2540, giving a median total radius of 0.3513 against the locked threshold 0.4313. The declarations are distributed over three terminal sample counts: one at 32768 measurements, 11 at 65536, and seven at 131072. The sole abstention reaches the cap because its total radius remains above threshold; its terminal estimate is nevertheless accurate under post hoc evaluation. Thus positive selective coverage does not arise by treating the resource cap as automatic success, and the stopping label remains distinct from terminal accuracy.

The paired dominant-vector experiment isolates a different point. Its full-space baseline has a small sampling radius around a badly biased plug-in target and therefore passes the numerical gate in all 20 runs. Krylov-first stabilization changes the finite-sample target before the same held-out gate is applied. The resulting contrast is evidence about estimator construction, not about a looser acceptance threshold. Complete per-run radii and measurement counts are reported in Supplementary Material, Section S2.

5.6 Local stability does not transfer as an approximation bound

Table 1 summarizes the selective outcomes. At six qubits, the unchanged guard makes seven declarations across 50 runs and all seven are false. At eight qubits, it declares in all 50 runs and every declaration is false. The Krylov bias is nearly flat over $K \leq 8$, even when adjacent-order changes are small. Figure 5 directly compares local changes d_K with the true bias $|B_K|$. This transfer failure rules out treating d_K or a fixed minimum order as a system-size-independent approximation radius.

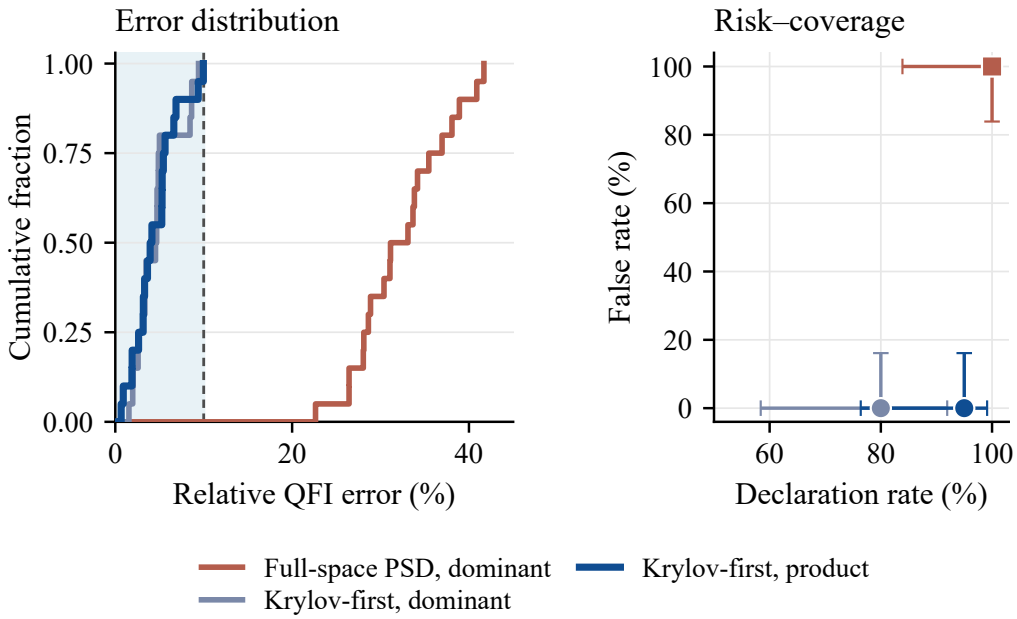

Figure 4: A directly preparable product start yields 19 correct declarations and one abstention at non-full Krylov order. Left: empirical cumulative distributions of terminal relative QFI error at 10% tolerance; the shaded band is the target region. The full-space and Krylov-first dominant-vector curves use paired random draws at $K = 40$, while the product-state curve uses a separate held-out set at $K = 48$. Right: selective risk–coverage plane with 95% Wilson intervals.

Table 1: Selective stopping outcomes. The false rate is the marginal false-declaration probability $\Pr(D \cap A^c)$. Precision is undefined when there are no declarations.

Setting	Rule	Runs	Decl. rate	False rate / precision
$n = 4$, primary	width-only	250	0.424	0.424 / 0.000
$n = 4$, primary	guarded	250	0.000	0.000 / –
$n = 4$, control	width-only	20	0.950	0.750 / 0.211
$n = 4$, control	guarded	20	0.600	0.000 / 1.000
$n = 6$, frozen	full-space, dominant	20	1.000	1.000 / 0.000
$n = 6$, frozen	Krylov-first, dominant	20	0.800	0.000 / 1.000
$n = 6$, frozen	Krylov-first, $\|+\rangle^{\otimes 6}$	20	0.950	0.000 / 1.000
$n = 6$, transfer	guarded	50	0.140	0.140 / 0.000
$n = 8$, transfer	guarded	50	1.000	1.000 / 0.000

6 Discussion and Limitations

The studies separate three effects. The four-qubit sweep shows that a narrow sampling interval can track the wrong Krylov target. The paired six-qubit study isolates a finite-sample bias from full-space positivity, while the product start demonstrates selective decisions at non-full order on a frozen physical instance. Transfer failures show that small adjacent-order change cannot replace an approximation bound.

The product start uses neither full Krylov resolution nor an oracle vector, but its truncation radius is calibrated from the exact density matrix of the same instance. Independent confirmation avoids reuse of exploration data; bootstrap coverage for this nonlinear QFI estimator remains unproved. Thus the decisions show within-instance feasibility, not a formal 95% guarantee for unseen states.

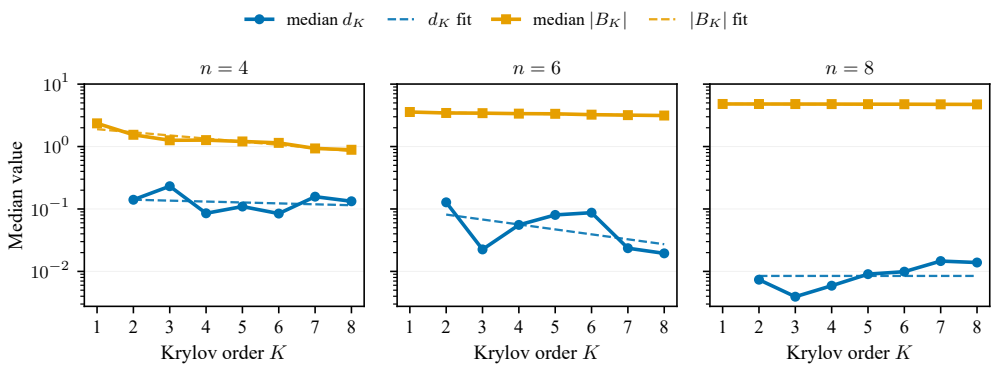

Figure 5: Small adjacent-order changes coexist with large approximation bias at six and eight qubits. The plot compares local change d_K with post hoc $|B_K|$ over $K \leq 8$; the local score can guide allocation but does not provide a transferable UQ radius.

The fixed-resource paired ladder tests projection order beyond the frozen low-noise setting. At $M = 32768$, median relative errors at $p_\phi = 0.03$ are 5.1% for Krylov-first and 26.7% for full-space positivity; at $p_\phi = 0.12$ they reverse to 33.5% and 6.1% across three paired seeds. The latter population truncation error is only 6.56%, so truncation alone cannot explain the finite-sample reversal.

6.1 Decision implications

When sampling width is small but Krylov error is unresolved, the rule should increase resolution or abstain. On the frozen low-noise instance, Krylov-first stabilization addresses the full-space projection bias; the higher-noise reversal calls for comparing the two orders under the target conditions. Decisions on unseen physical instances require a new approximation bound or state-level calibration and testing, rather than more samples from the same instance.

Scalability requires a computable upper bound on $|B_K|$ rather than a local surrogate. Candidate routes include residual-based Krylov bounds under spectral assumptions or calibration separated at the state or Hamiltonian level from evaluation instances.

On the sampling side, independent confirmation data provide the cleanest finite horizon analysis. Time-uniform confidence sequences (Howard et al., 2021) could reduce repeated hold-out costs while retaining validity under adaptive stopping. Either approach should be evaluated through risk–coverage–cost curves rather than a marginal false-declaration rate alone.

7 Conclusion

AKS-QFI separates sampling uncertainty around a fixed Krylov target from the target's approximation error and treats success as a selective decision. Krylov-first positivity stabilization reduces the finite-sample plug-in bias on the frozen low-noise instance. With the directly preparable $|+\rangle^{\otimes 6}$ start, the six-qubit study makes 19 correct declarations in 20 held-out runs at $K = 48 < 64$; a paired dominant-vector comparison changes 20 false full-space declarations into 16 correct Krylov-first declarations and four abstentions. Extending this within-instance result requires a computable approximation radius and validated sampling coverage for unseen physical states.

AI use statement

OpenAI Codex was used for coding assistance and language editing. The authors reviewed all AI-assisted content and take full responsibility for the manuscript, implementation, and reported results.

REFERENCES

Scott Aaronson. Shadow tomography of quantum states. *SIAM Journal on Computing*, 49(5): STOC18–368–STOC18–394, 2020. doi: 10.1137/18M120275X.

Jacob L Beckey, M Cerezo, Akira Sone, and Patrick J Coles. Variational quantum algorithm for estimating the quantum fisher information. *Physical Review Research*, 4(1):013083, 2022. doi: 10.1103/PhysRevResearch.4.013083.

Samuel L. Braunstein and Carlton M. Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439–3443, 1994. doi: 10.1103/PhysRevLett.72.3439.

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.

Christian L Degen, Friedemann Reinhard, and Paola Cappellaro. Quantum sensing. *Reviews of modern physics*, 89(3):035002, 2017. doi: 10.1103/RevModPhys.89.035002.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Andreas Elben, Steven T. Flammia, Hsin-Yuan Huang, Richard Kueng, John Preskill, Benoît Vermersch, and Peter Zoller. The randomized measurement toolbox. *Nature Reviews Physics*, 5(1):9–24, 2023. doi: 10.1038/s42254-022-00535-2.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT)*, volume 2375 of *Lecture Notes in Computer Science*, pp. 255–270. Springer, 2002. doi: 10.1007/3-540-45435-7_18.

Eric Ghysels and Jack Morgan. On the classical shadow nonparametric bootstrap. *arXiv e-prints*, 2025.

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Advances in quantum metrology. *Nature photonics*, 5(4):222–229, 2011. doi: 10.1038/nphoton.2011.35.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann. Statist.*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020. doi: 10.1038/s41567-020-0932-7.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Jing Liu, Haidong Yuan, Xiao-Ming Lu, and Xiaoguang Wang. Quantum fisher information matrix and multiparameter estimation. *Journal of Physics A: Mathematical and Theoretical*, 53(2):023001, 2020. doi: 10.1088/1751-8121/ab5d4d.

Johannes Jakob Meyer. Fisher information in noisy intermediate-scale quantum applications. *Quantum*, 5:539, 2021. doi: 10.22331/q-2021-09-09-539.

Dénes Petz and Catalin Ghinea. Introduction to quantum fisher information. In *Quantum probability and related topics*, pp. 261–281. World Scientific, 2011. doi: 10.1142/9789814338745_0015.

Aniket Rath, Cyril Branciard, Anna Minguzzi, and Benoît Vermersch. Quantum fisher information from randomized measurements. *Physical Review Letters*, 127(26):260501, 2021. doi: 10.1103/PhysRevLett.127.260501.

Michael A Taylor and Warwick P Bowen. Quantum metrology and its application in biology. *Physics Reports*, 615:1–59, 2016. doi: 10.1016/j.physrep.2015.12.002.

A. W. van der Vaart. *Asymptotic Statistics*. Number 3 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998. ISBN 9780521784504. doi: 10.1017/CBO9780511802256.

Vittorio Vitale, Aniket Rath, Petar Jurcevic, Andreas Elben, Cyril Branciard, and Benoît Vermersch. Robust estimation of the quantum fisher information on a quantum processor. *PRX Quantum*, 5(3): 030338, 2024. doi: 10.1103/PRXQuantum.5.030338.

Abraham Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 1945. doi: 10.1214/aoms/1177731118.

Da-Jian Zhang and D. M. Tong. Krylov shadow tomography: Efficient estimation of quantum fisher information. *Phys. Rev. Lett.*, 134:110802, 2025. doi: 10.1103/PhysRevLett.134.110802.

SUPPLEMENTARY MATERIAL

This supplement groups the decision rule and its conditional guarantees in Section S1, the frozen experimental protocols and complete results in Section S2, additional theory in Section S3, and the exact-density and paired-sampling stress tests in Section S4. The later diagnostics do not enlarge the held-out guarantee claimed for the frozen six-qubit instance.

The tables below give complete per-run results; the anonymous supplementary archive contains full-precision records and experimental settings.

S1 DECISION RULE AND CONDITIONAL GUARANTEES

For a full-space QFI value $\mathcal{F}$, a Krylov-truncated population value $\mathcal{F}_K$, and a randomized-measurement estimate $\widehat{\mathcal{F}}_{K,M}$, write

$$\widehat{\mathcal{F}}_{K,M} - \mathcal{F} = \underbrace{(\widehat{\mathcal{F}}_{K,M} - \mathcal{F}_K)}_{S_{K,M}} + \underbrace{(\mathcal{F}_K - \mathcal{F})}_{B_K}. \tag{S1.1}$$

The empirical sampling score is the fixed-order bootstrap width $w_{K,M}$; the empirical resolution score is the adjacent-order change $d_{K,M} = |\widehat{\mathcal{F}}_{K,M} - \widehat{\mathcal{F}}_{K-1,M}|$. These scores are kept separate because $w_{K,M}$ targets sampling variation around $\mathcal{F}_K$, whereas $d_{K,M}$ is only a local diagnostic and need not upper-bound $|B_K|$.

The empirical interface starts from (K_0, M_0) and a zero persistence counter. At each iteration it performs the following steps.

1. Compute $\widehat{\mathcal{F}}_{K,M}$ and, for $K > 1$, $\widehat{\mathcal{F}}_{K-1,M}$ using the same sample count. Record $d_{K,M}$.
2. Construct the fixed-K bootstrap interval and record $w_{K,M}$.
3. Mark the iteration empirically eligible if $K \geq K_{\min}$, $M \geq M_{\min}$, ($d_{K,M} \leq \varepsilon$ or $K = K_{\max}$), and $w_{K,M} \leq \varepsilon$.
4. Increment the persistence counter after an eligible iteration and reset it otherwise. When the counter reaches P, either return an empirical declaration or nominate the pair for held-out confirmation.
5. If $d_{K,M} > \varepsilon$ and $K < K_{\max}$, increase K. Otherwise, if $w_{K,M} > \varepsilon$ and $M < M_{\max}$, double M. If neither update is available, return the estimate with an abstention label.

In held-out mode, exploration data only nominate (K_j, M_j). An independent confirmation batch produces $\widehat{\mathcal{F}}_j^{\mathrm{conf}}$ and a sampling radius $R_j^{(\mathrm{sample})}(\delta_j)$. Given a separately calibrated truncation radius $R_{K_j}^{(\mathrm{approx})}$, the rule declares only if

$$R_{K_j}^{(\mathrm{approx})} + R_j^{(\mathrm{sample})}(\delta_j) \leq \varepsilon. \tag{S1.2}$$

The returned point estimate comes from the independent confirmation batch. Reaching $K_{\max}$ can make the empirical exploration gate eligible, but it does not set $R_K^{(\mathrm{approx})}$ to zero and does not force a held-out declaration.

The following guarantees require separately justified bounds on Krylov approximation and sampling error.

Assumption S1 (Krylov approximation)**.** *Over a specified calibrated family, the population approximation bias satisfies $|B_K| \le C_{\mathrm{approx}}\mu^K$ for constants $C_{\mathrm{approx}} > 0$ and $\mu \in (0,1)$.*

Assumption S2 (Sampling concentration)**.** *For fixed K, the sampling error $S_{K,M}$ has bias bounded by $\beta_{K,M}$, and $S_{K,M} - \mathbb{E}S_{K,M}$ is sub-Gaussian with proxy variance σ_K^2/M.*

Proposition S3 (Fixed-resource error bound)**.** *Under Assumptions S1–S2, for any $\delta \in (0,1)$,*

$$\Pr\left(|\widehat{\mathcal{F}}_{K,M} - \mathcal{F}| > C_{\mathrm{approx}}\mu^K + \beta_{K,M} + \sigma_K\sqrt{\frac{2\log(2/\delta)}{M}}\right) \le \delta. \tag{S1.3}$$

Proof. Equation (S1.1) and the triangle inequality give $|\widehat{\mathcal{F}}_{K,M} - \mathcal{F}| \le |B_K| + |S_{K,M}|$. Apply the deterministic approximation bound and the two-sided sub-Gaussian tail inequality to the centered sampling error, then add the sampling-bias bound. □

Theorem S4 (Finite-horizon held-out rule)**.** *Consider at most J declaration attempts. Conditional on the exploration history $\mathcal{H}_j$ before attempt j, suppose*

$$\Pr\Big(|\widehat{\mathcal{F}}_j^{\mathrm{conf}} - \mathcal{F}| > R_{K_j}^{(\mathrm{approx})} + R_j^{(\mathrm{sample})}(\delta_j) \,\Big|\, \mathcal{H}_j\Big) \le \delta_j. \tag{S1.4}$$

If declarations obey Eq. (S1.2) *and $\sum_{j=1}^J \delta_j \le \delta$, then*

$$\Pr\Big(\exists j \le J:\ \text{declare at } j \text{ and } |\widehat{\mathcal{F}}_j^{\mathrm{conf}} - \mathcal{F}| > \varepsilon\Big) \le \delta. \tag{S1.5}$$

Proof. Let A_j be the event that attempt j declares and its confirmation estimate has error exceeding ε. On a declaration, Eq. (S1.2) holds, so A_j is contained in the conditional failure event controlled at level δ_j. Thus $\Pr(A_j \mid \mathcal{H}_j) \le \delta_j$ and $\Pr(A_j) \le \delta_j$. A union bound gives $\Pr(\bigcup_{j=1}^J A_j) \le \sum_j \delta_j \le \delta$. □

The theorem is deliberately conditional. In the reported positive six-qubit experiment, $R_K^{(\mathrm{approx})}$ is obtained by exact within-instance calculation and the sampling radius is an empirical held-out bootstrap radius. Independence prevents adaptive reuse of exploration data, but bootstrap validity for the nonlinear plug-in QFI estimator is not proved. Consequently, the experiment tests the architecture and is not presented as a formal instantiation of the theorem's 95% guarantee.

S2 Frozen Protocols and Complete Results

The controlled mixed-state family applies depolarizing noise $p_{\mathrm{dep}} = 0.03$ and dephasing noise p_ϕ to a fixed circuit family. Exact post hoc references use the spectral QFI formula

$$\mathcal{F}(\rho_\theta) = 2\sum_{i,j:\lambda_i+\lambda_j>0} \frac{|\langle i|\partial_\theta\rho_\theta|j\rangle|^2}{\lambda_i+\lambda_j}. \tag{S2.1}$$

These references label terminal accuracy and calibrate the explicitly identified control studies; they are not available to the online stopping decision.

The primary four-qubit study uses $p_\phi \in \{0, 0.06, 0.12, 0.18, 0.24\}$, $\varepsilon = 0.2$, $K_0 = 1$, $M_0 = 16$, $K_{\max} = 8$, $M_{\max} = 512$, and 50 independent runs per noise level and rule. The guard uses $K_{\min} = 4$, $M_{\min} = 128$, and persistence $P = 2$. Width-only and component-aware variants use the same resource limits.

For the full-resolution four-qubit control, at $p_\phi = 0.03$, $\mathcal{F}_{\mathrm{ref}} = 3.7696$ and the 5% target is $\varepsilon = 0.1885$. Exact-density evaluation gives relative approximation errors of 2.24%, 4.65%,

9.21%, and 16.80% at $K = 12$ for $p_\phi = 0.03, 0.06, 0.12, 0.24$, respectively. The control therefore fixes $K = 16$, the full four-qubit resolution, and varies only the sample count over $M \in \{32768, 65536, 131072, 262144\}$. It contains 20 independent runs.

Both frozen six-qubit studies use $p_\phi = p_{\mathrm{dep}} = 0.03$, the same random-local generator with family seed 17, $\mathcal{F} = 4.313116$, and a 10% absolute tolerance $\varepsilon = 0.431312$. The schedule starts at $M = 2048$, doubles to at most $M = 131072$, uses 32 bootstrap replicates, and allocates $\delta = 0.05$ by Bonferroni correction over at most 12 confirmation attempts. Each candidate uses an independent confirmation batch of equal size.

For the paired projection-order study, the Krylov start is the exact dominant eigenvector. Exact calibration gives $K = 40$ and $R_{40}^{(\mathrm{approx})} = 0.080417$. Pilot seeds 0–2 freeze the resource cap; evaluation seeds 100–119 are then used without further tuning. Full-space and Krylov-first estimators receive identical random draws.

For the operational study, the start is the directly preparable product state $|+\rangle^{\otimes 6}$. Exact calibration gives $K = 48$ and $R_{48}^{(\mathrm{approx})} = 0.097358$. Pilot seeds 0–2 again freeze the cap, and evaluation seeds 200–219 are accessed only after the configuration is locked. This removes the oracle starting vector but retains exact within-instance truncation calibration.

The transfer study applies the four-qubit guard unchanged to six- and eight-qubit problems. It uses relative tolerance 5%, $(K_{\min}, M_{\min}, P) = (4, 128, 2)$, and $(K_{\max}, M_{\max}) = (8, 512)$, with 10 runs per system size and noise level. This is an intentionally adverse transfer test rather than a scaling study.

S2.1 SIX-QUBIT POSITIVE EVIDENCE

Table S1 separates the mechanistic paired study from the operational product-start result. The product-start rule has 95% declaration rate with no observed false declaration. Its median error is 4.04%, and its maximum error among declarations is 9.96%.

Table S1: Aggregate six-qubit results at 10% relative tolerance. The two dominant-start rows use paired random draws. The product-start row uses a distinct held-out set.

Stabilization	Start / K	Decl.	False decl.	Precision	Med. error
full-space PSD	dominant / 40	20/20	20/20	0.00	32.11%
Krylov-first	dominant / 40	16/20	0/20	1.00	4.61%
Krylov-first	$\|+\rangle^{\otimes 6}$ / 48	19/20	0/20	1.00	4.04%

Table S2: Complete held-out results for the six-qubit product-start experiment. The exact reference is $\mathcal{F}_{\mathrm{ref}} = 4.3131$, the order is $K = 48$, the absolute target is $\varepsilon = 0.4313$, and $R^{(\mathrm{approx})} = 0.0974$ in every run. Runs are ordered by the frozen simulation seed.

Run	Seed	Decision	M	$\widehat{\mathcal{F}}$	Rel. error (%)	$R^{(\mathrm{sample})}$	$R^{(\mathrm{total})}$
1	200	declare	131,072	4.079	5.44	0.230	0.328
2	201	declare	131,072	4.176	3.17	0.225	0.322
3	202	declare	32,768	3.909	9.38	0.311	0.409
4	203	declare	131,072	4.086	5.26	0.242	0.339
5	204	declare	65,536	4.084	5.30	0.222	0.320
6	205	declare	65,536	4.232	1.88	0.293	0.390
7	206	declare	131,072	4.200	2.63	0.244	0.341
8	207	declare	65,536	4.177	3.15	0.306	0.403
9	208	declare	65,536	4.156	3.64	0.243	0.341
10	209	declare	131,072	4.170	3.31	0.213	0.310
11	210	declare	65,536	4.231	1.91	0.295	0.392
12	211	declare	131,072	4.070	5.63	0.264	0.361
13	212	declare	65,536	4.143	3.94	0.298	0.395
14	213	declare	65,536	3.883	9.96	0.283	0.380
15	214	declare	65,536	4.285	0.66	0.289	0.387

Continued on next page

Table S2 (continued)

Run	Seed	Decision	M	$\widehat{\mathcal{F}}$	Rel. error (%)	$R^{(\mathrm{sample})}$	$R^{(\mathrm{total})}$
16	215	declare	65,536	4.017	6.87	0.301	0.398
17	216	declare	65,536	4.134	4.14	0.207	0.305
18	217	abstain	131,072	4.274	0.90	0.335	0.433
19	218	declare	65,536	4.085	5.29	0.243	0.341
20	219	declare	131,072	4.028	6.62	0.244	0.341

The measurement-count distribution in Table S2 is not an all-cap artifact: one declaration occurs at 32768 measurements, 11 at 65536, and seven at 131072. Across the 19 declarations, the median sampling radius is 0.2540 and the median total radius is 0.3513. All 20 terminal estimates, including the abstention, satisfy the post hoc error target.

Figure S1 exposes both ingredients that are compressed into the main-paper summary. The exact resolution curves identify non-full orders satisfying the pre-specified 2.5% truncation allocation on the frozen instance. The held-out run plot shows the fixed approximation radius and run-specific sampling radius separately. Nineteen totals lie below the locked threshold; the final run is an abstention rather than an automatic cap-level declaration.

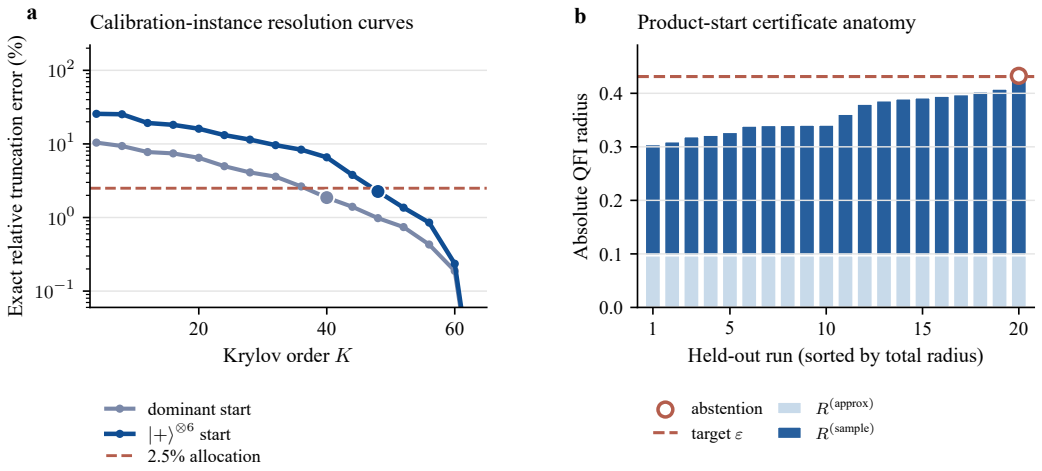

Figure S1: Six-qubit calibration and certificate anatomy. (a) Exact relative truncation error on the frozen calibration instance for the dominant and product starts. Highlighted points are the locked orders used in evaluation. (b) The approximation and sampling components of the held-out product-start radius, sorted only for display. The open circle marks the single abstention.

S2.2 Four-qubit and transfer evidence

Table S3: Complete four-qubit primary sweep. Declaration and false-declaration counts use 50 runs per noise setting.

p_ϕ	Rule	Decl.	False decl.	Med. K	Med. M	Med. abs. error
0.00	width-only	8/50	8/50	8	512	2.533
0.00	component-aware	0/50	0/50	8	512	2.458
0.06	width-only	14/50	14/50	8	512	1.744
0.06	component-aware	0/50	0/50	8	512	1.590
0.12	width-only	22/50	22/50	8	512	1.240
0.12	component-aware	0/50	0/50	8	512	0.994
0.18	width-only	28/50	28/50	2	64	1.553
0.18	component-aware	0/50	0/50	8	512	0.393
0.24	width-only	34/50	34/50	2	32	1.006
0.24	component-aware	0/50	0/50	8	512	0.211

The width-only rule has false-declaration rates $0.16, 0.28, 0.44, 0.56, 0.68$ across increasing four-qubit dephasing, including the $p_\phi = 0$ row reported in the main-paper aggregate. Its precision among declarations is zero at every point. The guarded rule abstains in every primary run. At $p_\phi = 0.18$ and 0.24, its median terminal absolute errors are 0.393 and 0.211, compared with 1.553 and 1.006 for width-only stopping. These terminal errors belong to abstentions and are not declared accuracies.

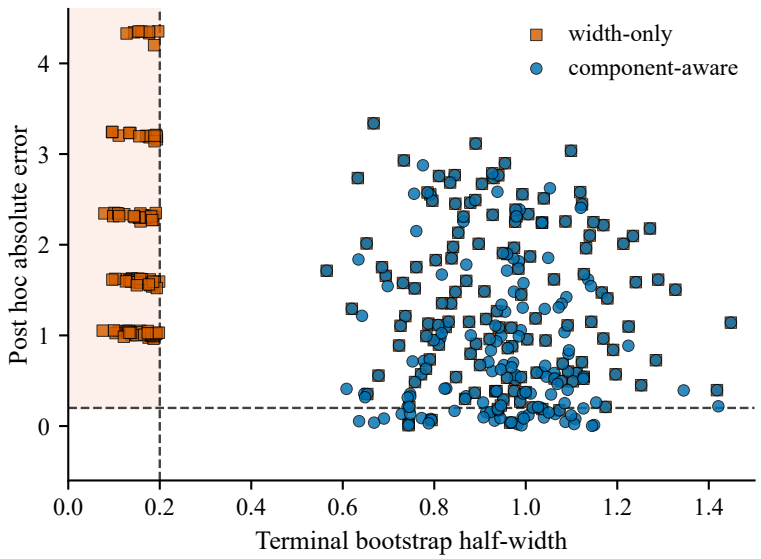

Figure S2: Terminal sampling width versus post hoc error in the primary four-qubit experiment. The upper-left quadrant contains false declarations: small sampling width coexists with large total error.

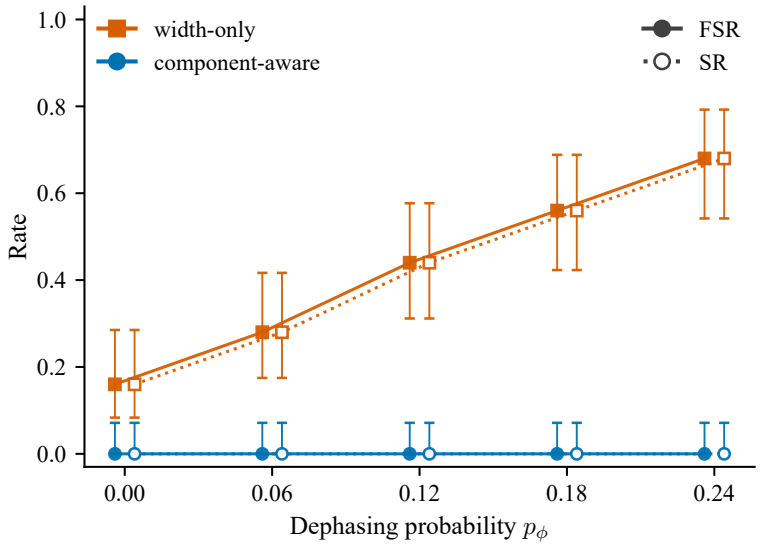

Figure S3: Declaration and false-declaration rates over the primary dephasing grid. Reporting both makes explicit that the guard's zero false-declaration rate is obtained through complete abstention.

In the full-resolution control, the component-aware schedule declares in 12 of 20 runs, all correctly, while all 20 terminal estimates meet the error target. The matched width-only schedule declares in 19 runs and makes 15 false declarations. This validates the sampling-side selection once approximation error is removed, but it is not evidence of scalable Krylov truncation.

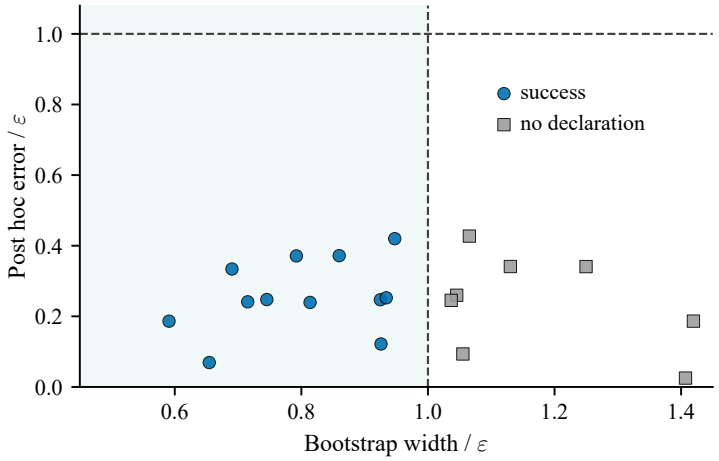

Figure S4: Full-resolution four-qubit control. Blue circles are declarations and gray squares are abstentions. The experiment isolates sampling-side selection by setting $K = 16$.

Under unchanged transfer, the six-qubit guard makes seven declarations over 50 runs and all seven are false; the eight-qubit guard declares in all 50 runs and every declaration is false. Figure S5 shows the mechanism: local adjacent-order changes can be small while the true approximation bias is large and nearly flat. Thus d_K is useful for routing resources but cannot be treated as a transferable error radius.

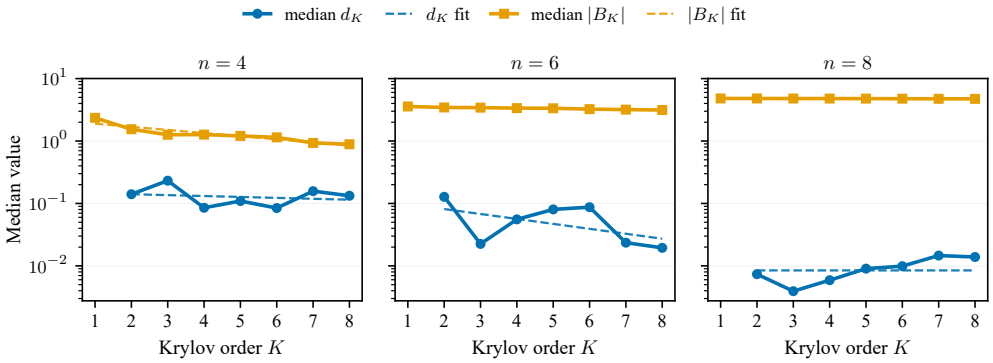

Figure S5: Adjacent-order change and post hoc approximation bias across the four-, six-, and eight-qubit transfer study. Small local motion does not imply small remaining bias.

Complete per-run and calibration records accompany the anonymous submission. Evaluation seeds are disjoint from the three pilot seeds. Declaration labels use Eq. (S1.2), whereas accuracy is computed only afterward from the exact reference. The product start is $|+\rangle^{\otimes 6}$ rather than an eigenvector of the target density matrix, and paired dominant-start runs use identical random draws across projection modes.

S3 Additional Theory and Diagnostic Limits

The failure of an adjacent-order change to bound Krylov bias persists even without sampling noise. The following construction is full rank when $\eta \in (0, 1)$ and has an exactly flat non-full Krylov sequence.

Proposition S5 (A flat Krylov sequence with nonzero QFI bias)**.** *Let* $\{|0\rangle, |1\rangle, |2\rangle\}$ *be an orthonormal basis, set* $G = \mathrm{diag}(0, 1, 2)$*, and start the Krylov construction from* $v = (|0\rangle + |1\rangle)/\sqrt{2}$*. For* $|\psi\rangle = (|0\rangle + |2\rangle)/\sqrt{2}$ *and* $\rho_\eta = (1 - \eta)|\psi\rangle\langle\psi| + \eta I_3/3$ *with* $0 \le \eta < 1$*, the weighted projected*

QFI is zero at both $K = 1$ and $K = 2$. Nevertheless, the full unitary-encoding QFI is

$$\mathcal{F}(\rho_\eta, G) = \frac{4(1-\eta)^2}{1-\eta/3} > 0. \tag{S3.1}$$

Thus the adjacent-order change at $K = 2$ is zero while the approximation error equals Eq. (S3.1)*. No universal inequality $|B_K| \le C|\mathcal{F}_K - \mathcal{F}_{K-1}|$ can hold with finite C for this estimator class.*

Proof. The first two Krylov vectors span $\mathcal{K}_2(G, v) = \mathrm{span}\{|0\rangle, |1\rangle\}$ because $Gv = |1\rangle/\sqrt{2}$. The one-dimensional projection at $K = 1$ has a scalar generator and hence zero QFI. At $K = 2$, compression removes the $|0\rangle\langle 2|$ coherence in ρ_η; the resulting state and compressed generator are both diagonal in $\{|0\rangle, |1\rangle\}$, so its QFI is again zero, regardless of the retained weight. For the full state, $|\psi\rangle$ and $(|0\rangle - |2\rangle)/\sqrt{2}$ have eigenvalues $1 - 2\eta/3$ and $\eta/3$, respectively, and the generator matrix element between them has magnitude one. Their eigenvalue difference is $1 - \eta$ and their sum is $1 - \eta/3$. Substitution into the spectral QFI formula gives Eq. (S3.1). The $|1\rangle$ sector contributes nothing. □

Krylov powers of G cannot recover a component absent from the starting vector's spectral support. Small local motion is therefore a useful resource-allocation diagnostic, but not a certificate of distance from full-space QFI. This construction separates an exact within-instance truncation radius from a transferable one.

The implemented local-Pauli shadow snapshot is $X = \bigotimes_{q=1}^{n}(3|b_q\rangle\langle b_q| - I_2)$, with an independently selected Pauli basis on each qubit. Let $\widetilde{\rho}_M = M^{-1}\sum_{m=1}^{M} X_m$ before positivity stabilization. The next identity quantifies the state-reconstruction scale without invoking bootstrap validity.

Proposition S6 (Second moment of the linear reconstruction)**.** *For independent standard local-Pauli snapshots of a density matrix ρ,*

$$\mathbb{E}\|\widetilde{\rho}_M - \rho\|_{\mathrm{F}}^2 = \frac{5^n - \mathrm{Tr}(\rho^2)}{M}. \tag{S3.2}$$

Consequently, for any $r > 0$,

$$\Pr(\|\widetilde{\rho}_M - \rho\|_{\mathrm{F}} > r) \le \frac{5^n - \mathrm{Tr}(\rho^2)}{Mr^2} \le \frac{5^n}{Mr^2}. \tag{S3.3}$$

Proof. Each one-qubit factor has eigenvalues 2 and -1, so its squared Frobenius norm is 5 and $\|X_m\|_{\mathrm{F}}^2 = 5^n$ for every snapshot. The standard inverse Pauli-measurement channel makes $\mathbb{E}X_m = \rho$. Expanding the squared norm of the mean of independent centered snapshots removes all cross terms and yields $M^{-1}(\mathbb{E}\|X_m\|_{\mathrm{F}}^2 - \|\rho\|_{\mathrm{F}}^2)$. Since $\|\rho\|_{\mathrm{F}}^2 = \mathrm{Tr}(\rho^2)$, this is Eq. (S3.2). Markov's inequality applied to the nonnegative squared norm gives Eq. (S3.3). □

At nominal failure level δ, Eq. (S3.3) gives the conservative radius $\sqrt{5^n/(M\delta)}$. This is a state-level Frobenius bound, not a QFI confidence radius: converting it requires regularity of the nonlinear positivity and QFI maps. Its exponential factor also makes it too loose to replace the empirical bootstrap radii in the reported six-qubit experiment.

The gap between a state-reconstruction bound and a QFI bound can be made explicit when both compared states are full rank. The price is an inverse spectral-gap factor, which is severe for the noisy states in this work. For a density matrix τ, define the symmetric linear map $\mathcal{L}_\tau(X) = (\tau X + X\tau)/2$ and the Hermitian tangent $A_\tau = -i[G, \tau]$. For unitary encoding by G, $\mathcal{F}(\tau, G) = \langle A_\tau, \mathcal{L}_\tau^{-1} A_\tau\rangle_{\mathrm{F}}$.

Proposition S7 (Full-rank QFI continuity)**.** *Let ρ and σ be density matrices of the same dimension with $\lambda_{\min}(\rho), \lambda_{\min}(\sigma) \ge m > 0$, and let G be a fixed Hermitian generator with $\|G\|_{\mathrm{op}} \le g$. Then*

$$|\mathcal{F}(\rho, G) - \mathcal{F}(\sigma, G)| \le \left(\frac{8g^2}{m} + \frac{4g^2}{m^2}\right)\|\rho - \sigma\|_{\mathrm{F}}. \tag{S3.4}$$

Proof. Write $R_\tau = \mathcal{L}_\tau^{-1}$ and $e = \|\rho - \sigma\|_{\mathrm{F}}$. Since every eigenvalue of $\mathcal{L}_\tau$ is an average of two eigenvalues of τ, $\|R_\rho\|_{\mathrm{F}\to\mathrm{F}}, \|R_\sigma\|_{\mathrm{F}\to\mathrm{F}} \le m^{-1}$. The commutator inequality gives $\|A_\rho - A_\sigma\|_{\mathrm{F}} \le$

$2ge$ and $\|A_\rho\|_{\mathrm{F}}, \|A_\sigma\|_{\mathrm{F}} \le 2g$ because density matrices have Frobenius norm at most one. Moreover, $\|\mathcal{L}_\rho - \mathcal{L}_\sigma\|_{\mathrm{F}\to\mathrm{F}} \le \|\rho - \sigma\|_{\mathrm{op}} \le e$. The resolvent identity $R_\rho - R_\sigma = R_\rho(\mathcal{L}_\sigma - \mathcal{L}_\rho)R_\sigma$ therefore implies $\|R_\rho - R_\sigma\|_{\mathrm{F}\to\mathrm{F}} \le e/m^2$. Add and subtract $\langle A_\sigma, R_\rho A_\sigma\rangle_{\mathrm{F}}$ in the two QFI quadratic forms. The difference from changing the tangent is at most $(\|A_\rho\|_{\mathrm{F}} + \|A_\sigma\|_{\mathrm{F}})\|R_\rho\|_{\mathrm{F}\to\mathrm{F}}\|A_\rho - A_\sigma\|_{\mathrm{F}} \le 8g^2e/m$; the difference from changing the inverse map is at most $\|A_\sigma\|_{\mathrm{F}}^2\|R_\rho - R_\sigma\|_{\mathrm{F}\to\mathrm{F}} \le 4g^2e/m^2$. Summing proves Eq. (S3.4). □

The proposition is a sufficient bound for two physical full-space states; it is not a bound on the weighted non-full Krylov estimator. It also requires an independently justified lower eigenvalue bound for any estimated state. For the depolarized n-qubit population family before additional channels, $m \ge p_{\mathrm{dep}}/2^n$ is available, but substituting this small value into Eq. (S3.4) produces a very loose radius. Thus a formal sampling certificate for useful QFI decisions needs a sharper, problem-specific argument rather than a direct combination of Eq. (S3.3) with global full-rank continuity.

At a population state $\rho \succeq 0$, full-space positivity leaves ρ unchanged, and $Q_K^\dagger \rho Q_K \succeq 0$; the two projection orders therefore agree for a fixed Krylov basis. They need not agree on an indefinite shadow reconstruction. The following calculation shows a retained-weight effect without relying on a particular random draw.

Let $X = \mathrm{diag}(0.3, 0.4, 0.5, -0.2)$, which has trace one, and let Q select the first two coordinates. Write $\mathcal{P}(X) = X_+/\operatorname{Tr}(X_+)$ for eigenvalue clipping followed by trace normalization. The normalized two-dimensional compressed state is $\sigma = \mathrm{diag}(3/7, 4/7)$ in both orders, but the weights differ:

$$w_{\mathrm{full}} = \operatorname{Tr}(Q^\dagger \mathcal{P}(X) Q) = \frac{7}{12}, \qquad w_{\mathrm{Krylov}} = \operatorname{Tr}(Q^\dagger X Q) = \frac{7}{10}. \tag{S3.5}$$

With compressed generator $G_K = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\mathcal{F}(\sigma, G_K) = 4/49$. The weighted estimates are $1/21$ and $2/35$, respectively. This proves that the maps do not commute; it does not imply that either order is uniformly closer to true QFI. The paired sampling ladder in Section S4.2 tests that empirical question in two noise regimes.

The finite-horizon theorem also need not spend equal error probability on each attempt. For countably many independent-confirmation attempts, the deterministic allocation $\delta_j = \delta 2^{-j}$ sums to δ. The conditional union-bound proof then controls the probability of any false declaration, provided each attempted sampling radius has the stated conditional coverage and the approximation radius is valid. This observation does not supply a coverage proof for the bootstrap radius used here.

S4 Additional Experimental Stress Tests

S4.1 Exact-density truncation grids

To test transfer of the favorable frozen six-qubit truncation radius, we ran a pre-specified population calculation on the server. The grid crosses $n \in \{4, 6, 8\}$, dephasing $p_\phi \in \{0.03, 0.12, 0.24\}$, random-local generator seeds $s_G \in \{17, 23, 41\}$, and two starts: the directly preparable $|+\rangle^{\otimes n}$ and the exact dominant eigenvector. Depolarizing probability is fixed at 0.03 and the entangling angle at 0.25. These 54 configurations contain no sampled measurements. Changing s_G changes the generator, while the physical state remains the same at fixed (n, p_ϕ); the three seeds are settings, not independent state draws.

For $d = 2^n$, we evaluated the implemented population estimator at $K \in \{d/4, d/2, 3d/4, d\}$ and at $K = 3d/4 - 1$. Define

$$T_K = \frac{|\mathcal{F}_K - \mathcal{F}|}{\mathcal{F}}, \qquad D_K = \frac{|\mathcal{F}_K - \mathcal{F}_{K-1}|}{\mathcal{F}}. \tag{S4.1}$$

The 2.5% threshold matches the exact truncation allocation used for the frozen six-qubit result. We call a row *locally false stable* when $D_{3d/4} \le 0.025$ but $T_{3d/4} > 0.025$. This is an oracle diagnostic, because both relative quantities use exact QFI. The full-order computation recovers the reference in every row to numerical precision.

The single frozen product-start configuration used in the main positive study appears in Table S6: at $n = 6$, $p_\phi = 0.03$, and $s_G = 17$, $T_{48} = 2.26\%$, consistent with its stated absolute radius 0.0974. At the same noise, changing only the generator to seed 23 or 41 raises the plus-start error to 5.39% or 4.98%. At higher dephasing, all plus-start six-qubit rows exceed the 2.5% truncation allocation. The favorable calibrated order is specific to the frozen instance.

The local-motion diagnostic is especially weak at six and eight qubits: 14 of 18 rows at each size are locally false stable. Table S4 separates the directly preparable start from the oracle dominant start. Neither start provides a transferable bound merely through small adjacent-order change. The complete results, including larger errors and exact-reference values, follow; percentages are rounded for display while the CSV preserves full precision.

Table S4: Summary of the 54 exact-density configurations at $K = 3d/4$. Counts are over nine fixed generator/noise combinations per row. False stable means $D_K \leq 2.5\%$ but $T_K > 2.5\%$. Generator seeds are descriptive settings, not independent state draws.

n	Start	median T_K	max T_K	$T_K \leq 2.5\%$	false stable
4	$\|+\rangle^{\otimes n}$	7.02%	20.44%	1/9	2/9
4	dominant	5.17%	16.66%	3/9	2/9
6	$\|+\rangle^{\otimes n}$	9.49%	18.42%	1/9	8/9
6	dominant	5.81%	16.75%	3/9	6/9
8	$\|+\rangle^{\otimes n}$	9.34%	20.70%	1/9	8/9
8	dominant	6.87%	17.66%	3/9	6/9

Table S5: Exact-density truncation stress test at $n = 4$ ($d = 16$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	17	$\|+\rangle^{\otimes n}$	4.180	11.75	5.53	1.96	0.86
0.03	17	dominant	4.180	4.86	2.08	1.07	0.36
0.03	23	$\|+\rangle^{\otimes n}$	1.752	28.65	25.06	6.62	1.97
0.03	23	dominant	1.752	5.64	3.51	1.22	0.74
0.03	41	$\|+\rangle^{\otimes n}$	2.009	24.90	8.55	4.11	1.13
0.03	41	dominant	2.009	4.62	2.80	0.73	0.45
0.12	17	$\|+\rangle^{\otimes n}$	2.576	26.57	12.80	5.74	2.51
0.12	17	dominant	2.576	20.30	9.45	5.17	1.85
0.12	23	$\|+\rangle^{\otimes n}$	1.179	42.69	35.11	12.63	4.76
0.12	23	dominant	1.179	23.34	14.95	6.76	2.89
0.12	41	$\|+\rangle^{\otimes n}$	1.293	39.61	18.03	7.02	2.56
0.12	41	dominant	1.293	20.03	13.00	4.82	2.08
0.24	17	$\|+\rangle^{\otimes n}$	1.129	44.23	22.94	11.41	4.81
0.24	17	dominant	1.129	38.93	20.34	11.85	4.40
0.24	23	$\|+\rangle^{\otimes n}$	0.587	58.36	46.70	20.44	8.66
0.24	23	dominant	0.587	44.13	31.06	16.66	5.48
0.24	41	$\|+\rangle^{\otimes n}$	0.607	56.88	31.23	12.63	4.73
0.24	41	dominant	0.607	39.98	28.71	14.05	4.26

At four qubits and $p_\phi = 0.24$, all six $K = 3d/4$ errors in Table S5 exceed 11%, despite exact full-order recovery. The corresponding six-qubit low-noise rows show a different start dependence: all three dominant-start errors are below 1.3%, while two of three plus-start errors exceed the 2.5% allocation. The same truncation fraction therefore has no uniform accuracy meaning across the tested dimension, generator, and start settings.

Table S6: Exact-density truncation stress test at $n = 6$ ($d = 64$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	17	$\|+\rangle^{\otimes n}$	4.313	18.17	9.62	2.26	0.39
0.03	17	dominant	4.313	7.43	3.59	0.98	0.05
0.03	23	$\|+\rangle^{\otimes n}$	1.995	35.38	17.68	5.39	0.11
0.03	23	dominant	1.995	8.91	4.37	1.05	0.09
0.03	41	$\|+\rangle^{\otimes n}$	3.880	30.85	15.58	4.98	0.08
0.03	41	dominant	3.880	8.12	3.96	1.21	0.10
0.12	17	$\|+\rangle^{\otimes n}$	2.676	36.46	20.99	6.56	0.73
0.12	17	dominant	2.676	29.25	16.46	5.71	0.22
0.12	23	$\|+\rangle^{\otimes n}$	1.334	53.04	29.97	10.41	0.12
0.12	23	dominant	1.334	33.74	18.74	5.81	0.36
0.12	41	$\|+\rangle^{\otimes n}$	2.617	48.27	25.74	9.49	0.35
0.12	41	dominant	2.617	31.92	18.30	6.56	0.46
0.24	17	$\|+\rangle^{\otimes n}$	1.192	57.03	37.20	15.01	1.37
0.24	17	dominant	1.192	53.97	35.92	15.61	0.57
0.24	23	$\|+\rangle^{\otimes n}$	0.659	70.08	45.41	18.42	0.48
0.24	23	dominant	0.659	59.34	37.72	14.69	0.78
0.24	41	$\|+\rangle^{\otimes n}$	1.308	67.28	42.07	18.40	0.90
0.24	41	dominant	1.308	57.13	38.72	16.75	1.01

Table S7: Exact-density truncation stress test at $n = 8$ ($d = 256$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	17	$\|+\rangle^{\otimes n}$	5.628	21.02	9.20	2.10	0.20
0.03	17	dominant	5.628	8.60	3.98	0.96	0.03
0.03	23	$\|+\rangle^{\otimes n}$	2.337	28.98	13.48	3.71	0.13
0.03	23	dominant	2.337	9.61	4.62	1.20	0.03
0.03	41	$\|+\rangle^{\otimes n}$	4.046	33.17	16.38	5.91	0.25
0.03	41	dominant	4.046	8.31	3.93	1.49	0.04
0.12	17	$\|+\rangle^{\otimes n}$	3.636	41.02	21.14	6.29	0.24
0.12	17	dominant	3.636	33.31	17.89	5.64	0.15
0.12	23	$\|+\rangle^{\otimes n}$	1.569	49.53	27.49	9.78	0.27
0.12	23	dominant	1.569	36.50	20.58	6.87	0.14
0.12	41	$\|+\rangle^{\otimes n}$	2.743	48.53	24.97	9.34	0.26
0.12	41	dominant	2.743	33.39	17.74	7.11	0.19
0.24	17	$\|+\rangle^{\otimes n}$	1.716	63.06	39.19	15.36	0.38
0.24	17	dominant	1.716	60.00	38.18	15.53	0.37
0.24	23	$\|+\rangle^{\otimes n}$	0.781	69.77	46.49	20.70	0.41
0.24	23	dominant	0.781	63.23	42.17	17.61	0.31
0.24	41	$\|+\rangle^{\otimes n}$	1.376	67.72	41.57	17.74	0.44
0.24	41	dominant	1.376	61.19	38.64	17.66	0.40

Figure S6 makes the population solutions visible at each computed order for the six-qubit initial grid. Each curve uses one generator setting and one Krylov start; values are divided by that setting's own exact QFI. The green band indicates the fixed 2.5% truncation allocation. The curves display how the final $K = 48$ value depends on the generator and the noise setting, even at a fixed dimension.

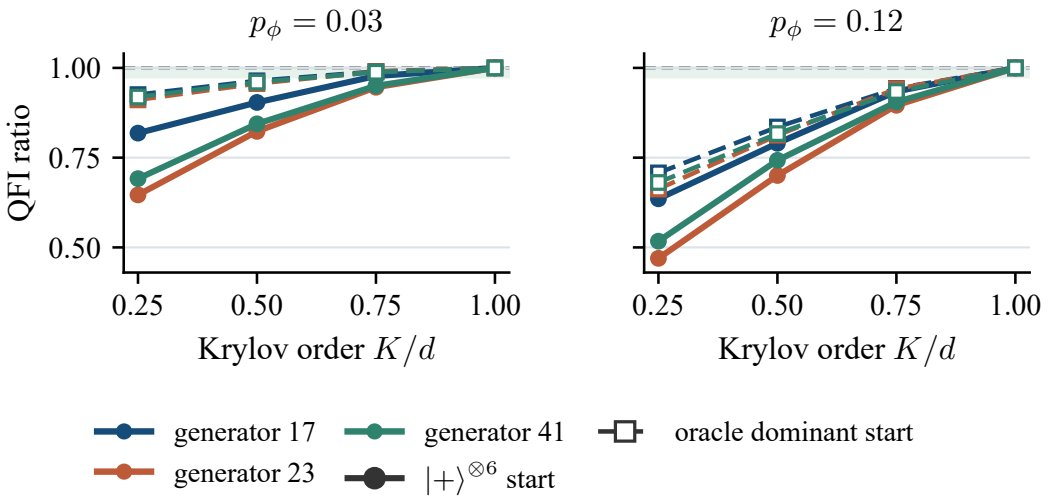

Figure S6: Representative exact-density QFI solution profiles from all six $n = 6$ rows at each shown dephasing setting in the initial grid. Solid circles use the operational product start; dashed open squares use the oracle dominant eigenvector. Colors distinguish generator seeds 17, 23, and 41. The horizontal reference is exact QFI; the green band is the 2.5% relative-error allocation. These are population calculations, without measurement noise.

After the first grid had been inspected, we fixed a follow-up grid with random-local generator seeds $53, 59, 61$ and no other changes. This is a transparent replication check, not a prospectively held-out calibration test. All 54 new configurations are reported separately. At six qubits, all nine plus-start rows are locally false stable; the same is true at eight qubits. None of the nine plus-start rows at either size meets the 2.5% truncation allocation at $K = 3d/4$. The oracle dominant start meets it in three of nine rows at each size. These outcomes reinforce the non-transferability of a local adjacent-order threshold while preserving the distinction between generator variation and unseen physical states.

Table S8: Summary of the 54 exact-density configurations in the disjoint-seed replication at $K = 3d/4$. Counts are over nine fixed generator/noise combinations per row. False stable means $D_K \leq 2.5\%$ but $T_K > 2.5\%$. Generator seeds are descriptive settings, not independent state draws.

n	Start	median T_K	max T_K	$T_K \leq 2.5\%$	false stable
4	$\|+\rangle^{\otimes n}$	12.29%	25.04%	0/9	4/9
4	dominant	9.00%	22.30%	3/9	2/9
6	$\|+\rangle^{\otimes n}$	8.80%	20.10%	0/9	9/9
6	dominant	7.77%	18.59%	3/9	6/9
8	$\|+\rangle^{\otimes n}$	9.22%	24.42%	0/9	9/9
8	dominant	8.77%	22.37%	3/9	6/9

The four-qubit replication also rules out treating this as only a large-dimension effect. None of its nine plus-start rows reaches the 2.5% allocation at $K = 3d/4$, and four have a local change below that threshold. The oracle dominant start reaches the allocation in three rows, yet two other rows are locally false stable. These are configuration counts within one state family; they do not estimate coverage over new physical states.

The low-noise four-qubit rows give a paired view of the start effect. At $p_\phi = 0.03$, the three product-start errors at $K = 3d/4$ are 3.38%, 6.30%, and 3.48%, whereas the corresponding dominant-start errors are 1.94%, 2.28%, and 1.94%. At $p_\phi = 0.24$, all six errors exceed 17%. An oracle start that helps at mild noise therefore does not provide a calibration that survives this noise shift.

Table S9: Exact-density truncation stress test (disjoint-seed replication) at $n = 4$ ($d = 16$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	53	$\|+\rangle^{\otimes n}$	0.394	20.30	4.95	3.38	0.20
0.03	53	dominant	0.394	5.77	3.51	1.94	0.17
0.03	59	$\|+\rangle^{\otimes n}$	1.227	16.53	10.71	6.30	0.68
0.03	59	dominant	1.227	6.97	5.40	2.28	0.35
0.03	61	$\|+\rangle^{\otimes n}$	2.099	24.58	14.71	3.48	3.14
0.03	61	dominant	2.099	7.28	5.88	1.94	2.04
0.12	53	$\|+\rangle^{\otimes n}$	0.276	35.27	17.67	9.45	1.63
0.12	53	dominant	0.276	23.62	15.53	8.40	0.94
0.12	59	$\|+\rangle^{\otimes n}$	0.883	35.33	22.93	12.29	1.57
0.12	59	dominant	0.883	28.47	22.72	10.30	1.72
0.12	61	$\|+\rangle^{\otimes n}$	1.462	42.29	30.46	13.27	2.77
0.12	61	dominant	1.462	28.80	24.45	9.00	6.66
0.24	53	$\|+\rangle^{\otimes n}$	0.145	52.64	34.59	18.83	3.30
0.24	53	dominant	0.145	44.28	31.59	17.22	2.55
0.24	59	$\|+\rangle^{\otimes n}$	0.484	55.06	37.57	21.53	2.92
0.24	59	dominant	0.484	52.19	42.10	22.30	3.51
0.24	61	$\|+\rangle^{\otimes n}$	0.755	59.89	47.92	25.04	2.81
0.24	61	dominant	0.755	50.62	45.72	19.99	8.74

At six qubits, the replicated low-noise product-start errors are 5.36%, 2.88%, and 2.61%, all above the 2.5% allocation, although their adjacent-order changes are only 0.28%, 0.16%, and 0.41%. Each paired dominant-start error is below 1.7% at this noise setting. At $p_\phi = 0.12$, even the dominant-start errors rise above 7%. These rows show why local stability and a favorable start need an independent approximation check before either supports a decision.

Table S10: Exact-density truncation stress test (disjoint-seed replication) at $n = 6$ ($d = 64$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	53	$\|+\rangle^{\otimes n}$	0.550	32.43	18.05	5.36	0.28
0.03	53	dominant	0.550	9.44	5.41	1.69	0.18
0.03	59	$\|+\rangle^{\otimes n}$	2.405	28.55	8.40	2.88	0.16
0.03	59	dominant	2.405	9.32	4.36	1.65	0.02
0.03	61	$\|+\rangle^{\otimes n}$	3.264	23.06	12.14	2.61	0.41
0.03	61	dominant	3.264	9.09	4.73	1.31	0.08
0.12	53	$\|+\rangle^{\otimes n}$	0.395	49.91	29.82	10.04	0.79
0.12	53	dominant	0.395	34.83	21.12	7.77	0.73
0.12	59	$\|+\rangle^{\otimes n}$	1.767	46.33	20.70	7.71	0.69
0.12	59	dominant	1.767	34.29	18.67	7.93	0.38
0.12	61	$\|+\rangle^{\otimes n}$	2.196	44.62	26.03	8.80	0.80
0.12	61	dominant	2.196	33.83	21.02	7.34	0.57
0.24	53	$\|+\rangle^{\otimes n}$	0.215	67.98	44.72	19.17	1.50
0.24	53	dominant	0.215	60.21	39.86	17.94	1.45
0.24	59	$\|+\rangle^{\otimes n}$	0.982	64.24	37.19	16.73	1.47
0.24	59	dominant	0.982	59.05	37.69	17.94	1.26
0.24	61	$\|+\rangle^{\otimes n}$	1.086	64.63	43.80	20.10	1.12
0.24	61	dominant	1.086	57.90	41.75	18.59	1.39

Table S11: Exact-density truncation stress test (disjoint-seed replication) at $n = 8$ ($d = 256$). Each row uses the specified random-local generator seed and start. T_K is relative population QFI error and D_K is the relative adjacent-order change, both in percent. The full-order error is below numerical precision in every row.

p_ϕ	s_G	Start	$\mathcal{F}$	$T_{d/4}$	$T_{d/2}$	$T_{3d/4}$	$D_{3d/4}$
0.03	53	$\|+\rangle^{\otimes n}$	1.578	23.78	8.75	2.71	0.01
0.03	53	dominant	1.578	8.41	3.87	1.03	0.02
0.03	59	$\|+\rangle^{\otimes n}$	2.625	30.97	11.97	3.69	0.03
0.03	59	dominant	2.625	10.10	5.07	1.99	0.13
0.03	61	$\|+\rangle^{\otimes n}$	3.906	27.65	12.86	4.12	0.13
0.03	61	dominant	3.906	9.70	4.99	1.77	0.04
0.12	53	$\|+\rangle^{\otimes n}$	1.052	42.86	20.31	6.93	0.12
0.12	53	dominant	1.052	33.47	17.50	5.76	0.10
0.12	59	$\|+\rangle^{\otimes n}$	1.937	48.76	24.16	9.22	0.12
0.12	59	dominant	1.937	36.77	20.69	8.77	0.43
0.12	61	$\|+\rangle^{\otimes n}$	2.626	49.51	28.86	12.21	0.32
0.12	61	dominant	2.626	37.13	22.22	9.62	0.22
0.24	53	$\|+\rangle^{\otimes n}$	0.517	64.51	38.10	15.84	0.34
0.24	53	dominant	0.517	60.73	37.47	15.34	0.28
0.24	59	$\|+\rangle^{\otimes n}$	1.083	67.55	41.55	18.95	0.23
0.24	59	dominant	1.083	62.99	41.01	19.99	0.72
0.24	61	$\|+\rangle^{\otimes n}$	1.296	69.99	48.21	24.42	0.38
0.24	61	dominant	1.296	63.39	43.77	22.37	0.40

Figure S7 displays every row from both grids. The dashed lines mark the fixed 2.5% threshold. At $n = 6$ and $n = 8$, 29 of 36 rows in each panel fall in the upper-left false-stability region: the local change is small even though the remaining population error is above the truncation allocation. The points are paired through the same state and generator within each configuration, but the plotted rows should not be treated as independent physical-state replicates.

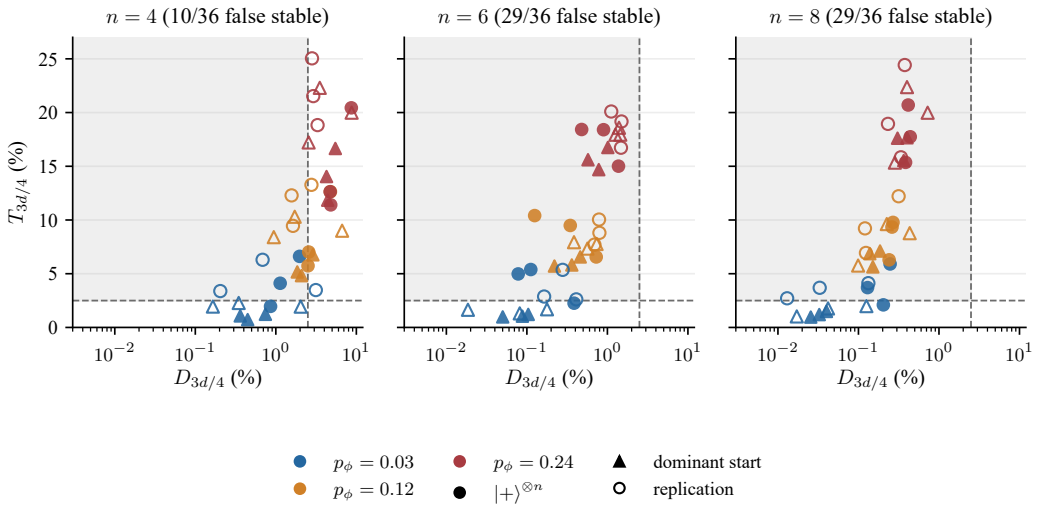

Figure S7: Complete exact-density diagnostic across the initial and disjoint-generator grids (108 configurations). $D_{3d/4}$ is the relative adjacent-order change and $T_{3d/4}$ is relative full-QFI error. The shaded upper-left region is locally false stable at the fixed 2.5% threshold. Filled markers are from the first grid; open markers are from the follow-up replication. Color gives dephasing, and shape gives the Krylov start.

This exact-density study tests approximation only. It does not turn the empirical stopping score into a certificate, establish bootstrap coverage, or estimate declaration risk under new sampled states. The

dominant-start rows use an oracle eigenvector and serve as a comparison to the operational plus start, not as a deployable calibration.

S4.2 Paired shadow-sampling ladder at fixed Krylov order

We next tested projection order with new measurement draws. The pre-specified ladder fixes $n = 6$, random-local generator seed 17, $|+\rangle^{\otimes 6}$ start, and $K = 48$, while crossing $p_\phi \in \{0.03, 0.12\}$, seeds 300–302, and nested prefixes $M \in \{2048, 8192, 32768\}$. A maximum-length Pauli-shadow data set is drawn once per noise/seed pair; both projection orders receive identical prefixes. There are 18 paired outcomes. No bootstrap, adaptive decision, or declaration is applied in this fixed-resource diagnostic.

Table S12: Descriptive summary of the paired ladder. Each row has three new seeds; medians are over these paired draws. T_{48} is exact population truncation error and does not depend on M.

p_ϕ	M	T_{48}	median err. full	median err. Krylov	Krylov better
0.03	2,048	2.26%	31.1%	1.2%	3/3
0.03	8,192	2.26%	30.0%	4.2%	3/3
0.03	32,768	2.26%	26.7%	5.1%	3/3
0.12	2,048	6.56%	9.4%	38.5%	0/3
0.12	8,192	6.56%	3.8%	42.4%	0/3
0.12	32,768	6.56%	6.1%	33.5%	0/3

At $p_\phi = 0.03$, Krylov-first has lower absolute error in all nine paired rows; the median errors across three seeds at $M = 32768$ are 5.1% versus 26.7% for full-space positivity. Full-space positivity keeps roughly 78–82% of the reconstructed mass in these rows, whereas Krylov-first keeps roughly 91–100%. The new data are compatible with the original paired result, though they use smaller measurement budgets.

At $p_\phi = 0.12$, the ordering reverses in all nine paired rows. At $M = 32768$, median relative error is 33.5% for Krylov-first and 6.1% for full-space positivity. The exact population truncation error is only 6.56% at this order, so the large Krylov-first error here is a finite-sample effect rather than the population truncation term alone. A larger retained weight is not itself a QFI accuracy certificate. This negative regime rules out presenting Krylov-first stabilization as uniformly superior across the noise family.

The sign of the paired difference is consistent within each noise setting, not only in the medians. At the largest budget, Krylov-first reduces relative error by 17.5–23.7 percentage points across the three low-noise seeds, but increases it by 26.5–27.3 points across the three higher-noise seeds. Nested prefixes and shared draws remove one source of between-method variation; three seeds and one physical family do not establish a population-level risk bound.

Figure S8 shows the corresponding sample-level solutions. It preserves all three seed trajectories at each budget, alongside their median and range; the shaded ranges are descriptive, not confidence intervals. The reference and population $K = 48$ value are different targets, which makes the low-noise and higher-noise behavior visually comparable.

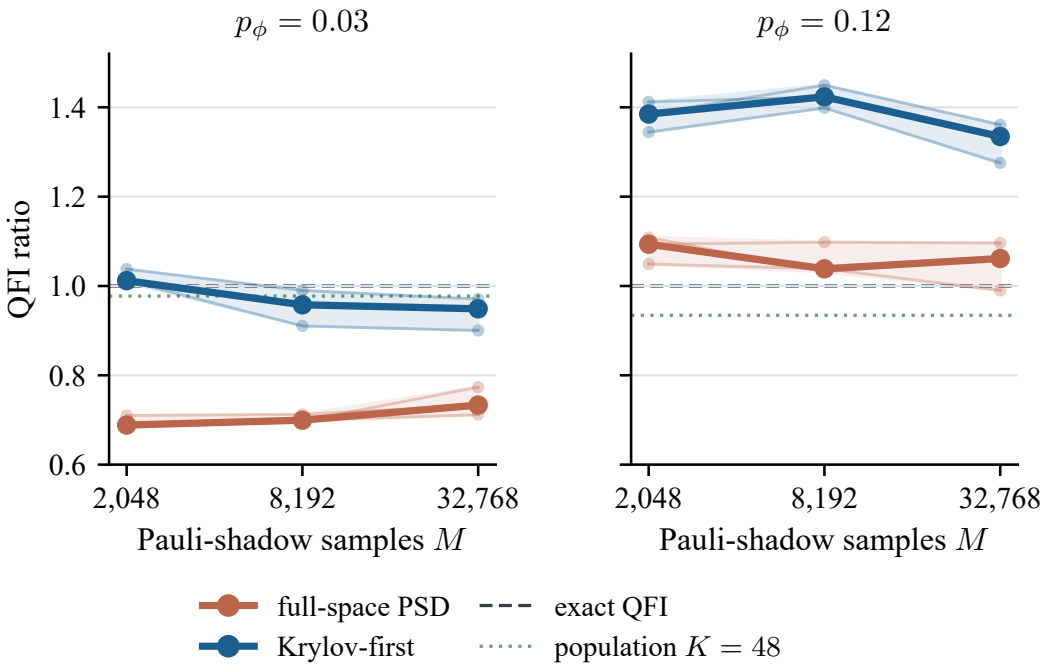

Figure S8: QFI estimates from the complete paired shadow-sampling ladder, normalized by the exact QFI for each dephasing setting. Thin lines and small markers show all three measurement seeds (300–302); thick lines show their medians and pale bands their observed ranges. The black dashed line is exact QFI; the green dotted line is the exact $K = 48$ population estimator. Nested budgets reuse the same measurement prefixes within a seed, and the two projection orders use identical prefixes.

Table S13: New paired six-qubit shadow samples at fixed $K = 48$. Full-space and Krylov-first stabilization receive the same data; each seed uses nested measurement prefixes. Errors and retained subspace weights are percentages. This diagnostic has no stopping decision or bootstrap certificate.

p_ϕ	seed	M	$\widehat{\mathcal{F}}_{\text{full}}$	$\widehat{\mathcal{F}}_{\text{Krylov}}$	err. full	err. Krylov	weight full	weight Krylov
0.03	300	2,048	2.959	4.347	31.4	0.8	79.0	100.0
0.03	300	8,192	3.007	4.131	30.3	4.2	79.1	93.5
0.03	300	32,768	3.337	4.093	22.6	5.1	81.0	90.8
0.03	301	2,048	2.970	4.364	31.1	1.2	79.3	100.0
0.03	301	8,192	3.017	3.928	30.0	8.9	79.6	90.9
0.03	301	32,768	3.070	3.884	28.8	10.0	80.9	91.2
0.03	302	2,048	3.062	4.476	29.0	3.8	78.1	100.0
0.03	302	8,192	3.073	4.270	28.8	1.0	79.6	98.0
0.03	302	32,768	3.163	4.186	26.7	3.0	82.4	97.0
0.12	300	2,048	2.927	3.707	9.4	38.5	76.3	83.4
0.12	300	8,192	2.939	3.879	9.8	44.9	78.3	89.1
0.12	300	32,768	2.934	3.643	9.6	36.1	79.7	87.7
0.12	301	2,048	2.967	3.779	10.9	41.2	76.8	83.2
0.12	301	8,192	2.762	3.810	3.2	42.4	77.9	91.5
0.12	301	32,768	2.841	3.572	6.1	33.5	80.0	87.6
0.12	302	2,048	2.808	3.599	4.9	34.5	77.7	85.6
0.12	302	8,192	2.779	3.744	3.8	39.9	77.6	90.9
0.12	302	32,768	2.650	3.414	1.0	27.5	79.3	89.2