Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Scalable and Verifiable Quantum Benchmarking with Spectrally Flat Hidden Shifts

Lior Horesh

lhoresh@us.ibm.com

IBM Research - Thomas J. Watson Research Center

Vasileios Kalantzis

IBM Research - Thomas J. Watson Research Center

Shashanka Ubaru

IBM Research - Thomas J. Watson Research Center

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11433479/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Scalable and Verifiable Quantum Benchmarking with Spectrally Flat Hidden Shifts

Lior Horesh, Vasileios Kalantzis, Shashanka Ubaru

IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

Corresponding author: Lior Horesh, lhoresh@us.ibm.com

Abstract

Scalable quantum benchmarks often lack exact reference answers that can be verified without classically simulating the device. We introduce a hidden-shift benchmark in which a secret string is planted by construction, recovered by a quantum circuit, and exactly verified at any system size. The benchmark uses spectrally flat Boolean oracles, whose Fourier structure fixes the ideal output weight of the planted answer and enables majority-vote recovery even when its single-shot probability is extremely small. Register size, circuit depth, and non-Clifford content are independently controlled through the oracle degree. On IBM superconducting processors, degree-2 instances are recovered on up to 24 qubits despite single-shot fidelities near 10^{-3}, while higher-degree oracles provide a tunable non-Clifford stress test that distinguishes processor performance. Although classically tractable, these benchmarks provide a scalable and exactly verifiable measure of how far a noisy quantum processor can reliably execute a structured computation.

Introduction

Today's noisy quantum processors perform computations at circuit depths of a few hundred gate layers. Single- and two-qubit gate fidelities largely determine that depth [1, 2], and deeper computations are generally expected to require either improved hardware or quantum error correction [3, 4]. Assessing how far a processor can reliably carry a structured computation is therefore a central challenge for quantum benchmarking.

Existing benchmarks primarily measure either computational hardness or average gate quality. Random-circuit sampling probes computational hardness but does not provide a scalable route to exact verification without classically simulating the device [1]. Quantum volume likewise relies on classical prediction and yields a pass–fail metric rather than a problem-specific answer [5]. Randomized benchmarking and mirror circuits characterize average gate performance but do not test recovery of the solution to a structured computational task [2, 6]. Consequently, there remains a need for benchmarks that combine scalable implementation with an exact reference answer that can be verified independently of classical simulation.

Here we introduce an exactly verifiable benchmark based on hidden-shift instances with a planted secret string. A hidden shift s is embedded in an oracle by construction, recovered by

a quantum circuit, and known in advance, making correctness exactly verifiable at any system size. The benchmark is built from spectrally flat Boolean functions and parameterized by a single quantity, the spectral flatness α. Register size, circuit depth, and non-Clifford content can be varied independently through the choice of oracle, enabling a family of benchmarks spanning both Clifford-dominated and non-Clifford regimes.

The key enabling property is spectral flatness. Rötteler [7] showed that spectral flatness underlies the hidden-shift algorithm for bent Boolean functions, with the Bernstein–Vazirani algorithm [8] appearing as the degree-1 special case. We show that spectral flatness also governs the robustness of hidden-shift recovery under noise. For an α-flat function, the planted answer retains a constant ideal probability share independent of system size. Under an idealized model in which lost probability is redistributed broadly across the outcome space, each incorrect outcome receives only $\mathcal{O}(2^{-n})$ of the error weight, allowing majority-vote recovery well beyond the regime in which individual shots remain reliable. More generally, the recovery guarantee depends only on the probability margin between the planted answer and its strongest competitor, a quantity that can be estimated directly from the measured output distribution. Spectral flatness therefore provides a design principle for constructing benchmark families with analytically tractable ideal outputs and scalable recovery guarantees.

On IBM hardware, the benchmark recovers degree-2 hidden shifts in 40 of 40 runs through $n = 20$ and in 10 of 10 runs at $n = 24$ (one by a tie) on two Heron processors, despite measured single-shot fidelities of only 3.0×10^{-3} and 9.2×10^{-4}. Recovery collapses at $n = 32$, where the planted answer is the unique mode in only one of ten runs. Increasing the oracle degree introduces CCZ gates and converts the same framework into a tunable non-Clifford stress test. Across five backends, the degree-3 benchmark reaches $n = 140$, and higher-degree instances produce a processor ranking that correlates strongly with measured two-qubit error ($\rho = -0.90$).

The benchmark instances studied here are intentionally classically easy. Amy and Stinchcombe [9] showed that shifted Maiorana–McFarland functions of bounded degree admit polynomial-time classical recovery, and for the public bent functions used in our experiments the hidden shift can be recovered classically in $\mathcal{O}(n)$ queries (Proposition 6). This classical tractability is a feature rather than a limitation: exact verification follows from comparison with a known planted answer rather than from a computational-hardness assumption. Accordingly, we claim neither computational advantage nor a query-complexity separation. Instead, we present a scalable benchmark that combines exact verifiability, a structured algorithmic task, and tunable non-Clifford resources. Unlike benchmarks summarized by aggregate fidelity metrics, the hidden-shift task retains an algorithmically meaningful target solution whose recovery can be tracked directly as a function of circuit scale and noise.

Results

A benchmark with a planted answer

Definition 1 (α-flat function)**.** *A Boolean function $f\colon \mathbb{Z}_2^n \to \{\pm 1\}$ is α-flat for $\alpha \in (0, 1]$ if*

$$|\hat{f}(\xi)| \geq \alpha\, 2^{n/2}$$

for all $\xi \in \mathbb{Z}_2^n$, where

$$\hat{f}(\xi) = \sum_x f(x)(-1)^{x\cdot\xi}$$

is the Walsh–Hadamard transform. Bent functions correspond to the maximally flat case $\alpha = 1$.

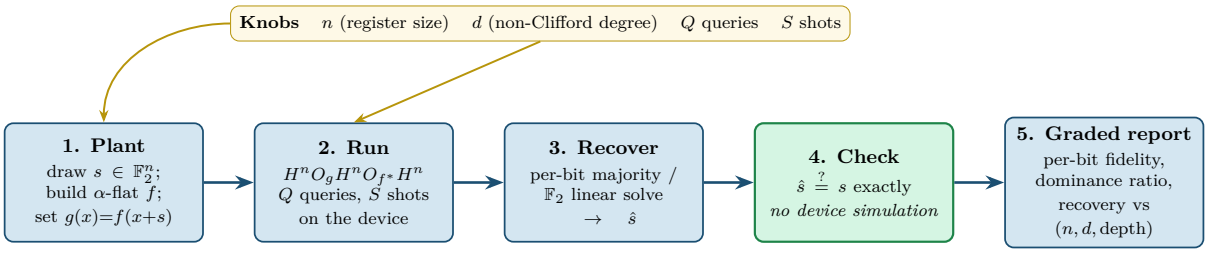


Figure 1 | Hidden-shift benchmark protocol. A hidden shift s is planted in a spectrally flat oracle. The recovery circuit runs for Q queries at S shots. The estimate $\hat{s}$ is read by full-string majority vote or per-bit majority and checked against the planted s exactly, with no classical simulation of the device. The output is a graded, depth-resolved profile. The circuit drawn is the full Rötteler dual-oracle form (Protocol C), the protocol behind every recovery claim in this paper. The combined-oracle specialization is Protocol B (a shallow compilation check) and Fourier-label sampling is Protocol A (Methods). The parameters n, d, Q, S set register size, non-Clifford content, and statistics.

Table 1 | Comparison of the α-flat hidden-shift benchmark with representative scalable quantum-processor benchmarks. "Verif. w/o sim." indicates that correctness can be established from a known reference answer without classically simulating the device. "Planted answer" indicates recovery of an exactly known target solution rather than a fidelity estimate or circuit-specific score. "Non-Cliff. control" indicates that non-Clifford content can be systematically varied within the benchmark family. The comparison focuses on the benchmarking properties emphasized in this work; other approaches, including accreditation and trap-based protocols, target complementary objectives.

	Verif. w/o sim.	Recovery beyond single-shot regime	Non-Cliff. control	Planted answer
XEB / RCS [1]	no	no	yes	no
Quantum volume [5]	no	no	limited	no
RB / mirror [2, 6]	yes	yes	partial	no
This benchmark	yes	yes	yes	yes

The benchmark plants a uniformly random hidden shift $s \in \mathbb{F}_2^n$ in the oracle

$$g(x) = f(x \oplus s)$$

and executes the Rötteler dual-oracle circuit

$$H^{\otimes n} O_g H^{\otimes n} O_{f^*} H^{\otimes n}$$

(Fig. 1). In the ideal circuit, the planted shift s is the measurement mode, returned with probability $\bar{\alpha}^2$, where $\bar{\alpha}$ is the mean normalized spectral magnitude (Proposition 7; Methods).

Recovery is extracted directly from the measured counts using two estimators: the most frequent output string (full-string majority vote) and the per-bit majority string. The benchmark exposes three independent control parameters: the register size n, the oracle degree d (degree-2 instances are Clifford, whereas $d \geq 3$ introduces $C^{(d-1)}Z$ resources), and the query and shot budgets. We refer to the IBM processors by letter: processors A, B, D, F, G and H are Heron processors and processors C, E and I are Nighthawk processors (Supplementary S21).

Because the hidden shift is planted, recovery can be verified exactly at any system size through a direct comparison of $\hat{s}$ and s. This yields a graded, depth-resolved benchmarking profile consisting

of recovery rate, per-bit fidelity, and the measured single-shot fidelity

$$F = \Pr(\text{one shot returns } s),$$

the probability that a single circuit execution produces the complete n-bit answer. This quantity plays a role analogous to the sampling fidelity used in cross-entropy benchmarking. For circuits containing on the order of 10^3 two-qubit gates with per-gate fidelities of 99.8–99.9%, single-shot fidelities near 10^{-3} are the expected scale.

Existing scalable quantum benchmarks typically sacrifice either exact verification, tunable algorithmic structure, or scalability of the verification procedure (Table 1). The present construction combines all three properties through a planted solution whose correctness can be checked exactly without classical simulation.

The following results characterize the behavior of majority-vote and per-bit recovery under noise.

Theorem 2 (Majority-vote recovery from a probability margin)**.** *Let p_y denote the probability that a single shot returns the string y, let $F = p_s$ be the single-shot fidelity, and define $\Delta := F - \max_{y \neq s} p_y > 0$ as the probability margin between the planted answer and its strongest competitor. For S independent shots, the majority-vote estimate equals s with probability at least*

$$P(\textit{correct mode}) \;\geq\; 1 - 2^n \exp\Big(-\frac{S\Delta^2}{4F + 2\Delta}\Big). \tag{1}$$

Consequently,

$$S \;\geq\; (4F + 2\Delta)\,\Delta^{-2}\big(n \ln 2 + \ln(1/\delta)\big)$$

suffices to bound the failure probability by δ.

The hypothesis depends only on the output distribution and makes no assumption about the source of the noise. Proof can be found in Supplementary S3.

Corollary 3 (Uniform redistribution)**.** *Suppose $F > 2^{-n}$ and every incorrect string carries probability at most $\frac{1-F}{2^n-1}$, corresponding to uniform redistribution of the lost probability mass. Then*

$$\Delta \geq F - \frac{1-F}{2^n - 1} > 0,$$

and Theorem 2 applies with this margin. In particular, $S = \mathcal{O}(F^{-1}n)$ shots suffice whenever $F \geq 2^{1-n}$, yielding polynomial shot complexity only when $F \geq 1/\mathrm{poly}(n)$.

The uniform-redistribution assumption is a sufficient condition for recovery, not a model of device noise (Discussion).

Theorem 4 (Per-bit majority recovery)**.** *Suppose each output bit is independently correct with probability $\frac{1}{2} + \gamma$ on every shot. Then per-bit majority vote over S shots recovers all n bits with probability at least $1 - n\exp(-2S\gamma^2)$. Consequently, $S = \mathcal{O}(\gamma^{-2}\log n)$ shots suffice, independent of the full-string fidelity F.*

This bound and the associated recovery statistic are established in Ref. [10]; we restate them here and claim no novelty.

The mechanism underlying Corollary 3 is the dominance ratio

$$\frac{F(2^n - 1)}{1 - F}.$$

In the noiseless circuit, the planted answer carries the constant probability mass $\bar{\alpha}^2$, whereas under uniform redistribution each incorrect string receives only $\mathcal{O}(2^{-n})$. The resulting dominance ratio is therefore

$$\Theta(\alpha^2 2^n),$$

and majority voting succeeds whenever this ratio exceeds unity, even after the single-shot fidelity has become very small.

Uniform redistribution is a sufficient condition for recovery, not a realistic model of device noise. Experimentally, error probability remains concentrated on Hamming-near competitors, and per-bit marginals become biased before full-string voting fails. Supplementary S20 analyzes these departures and shows that positively associated errors can sustain exact-string recovery well beyond the independent-bit regime. The observed overdispersion (1.4–2.1×) is consistent with such positive association, which can only increase the probability of recovering the full string relative to the independent-bit prediction. In an explicit burst-noise model, the planted answer remains the unique mode with $\mathcal{O}(n)$ shots at fidelities where per-bit majority voting fails regardless of shot count. For this reason, the benchmark reports both full-string and per-bit recovery statistics.

The prediction

$$P(s) = \bar{\alpha}^2$$

was tested in a preregistered hardware study on three processors, confirmed on one and replicated on the other two. Measured planted-string retention scaled approximately linearly with $\bar{\alpha}^2$, yielding fitted slopes between 0.87 and 0.94 across devices. Experimental design, decision rules, replication protocols, control studies, and complete numerical results are reported in Supplementary S23. Figure 2 plots the planted share against $\bar{\alpha}^2$ on the three processors.

Degree-2 recovery at the shot floor

The degree-2 record comprises two same-design dual-oracle ladders on IBM Heron processors. Registration chronology, audit records, and experimental metadata are reported in Supplementary S24. Table 2 presents the complete 132-run record, including every failure.

The record exhibits a clear recovery frontier. Through $n = 20$ ($\sim$430 two-qubit gates), every recovery metric succeeds on every run across both processors (40/40). At $n = 24$ ($\sim$ 950–970 two-qubit gates), measured single-shot fidelities have fallen to 3.0×10^{-3} on processor B and 9.2×10^{-4} on processor A. Per-bit majority begins to fail (3/5 and 1/5 recoveries, respectively), yet full-string majority still recovers the planted shift in all ten runs. The correct string wins with only 3–107 counts out of 16 384 shots because no other strings receive more counts (Fig. 3).

Multinomial resampling reveals an emerging difference between the two processors. At $n = 24$, the planted shift remains the mode with probability at least 0.96 in every processor B run, but in only two of the five processor A runs does the corresponding probability exceed 0.9; three recoveries are therefore fragile under resampling (Table 3). At $n = 32$, the frontiers separate. On processor A, the output distribution is indistinguishable from uniform at the resolution of the shot budget, and recovery fails entirely. On processor B, a residual signal survives and full-string voting recovers the correct 32-bit shift in one run of five, from 3 counts against a strongest rival count of 1, corresponding to $F = 1.8 \times 10^{-4}$, approximately three times the shot floor. Beyond this point, all recovery statistics are consistent with chance.

Two diagnostics summarize the frontier behavior directly from the measured counts. The first is the count ratio μ_c/μ_w, where μ_c and μ_w denote the counts of the planted shift and strongest rival outcome. Recovery persists while $\mu_c > \mu_w$. The second is the empirical margin

$$\hat{\Delta} = \hat{F} - \max_{y \neq s} \hat{p}_y,$$

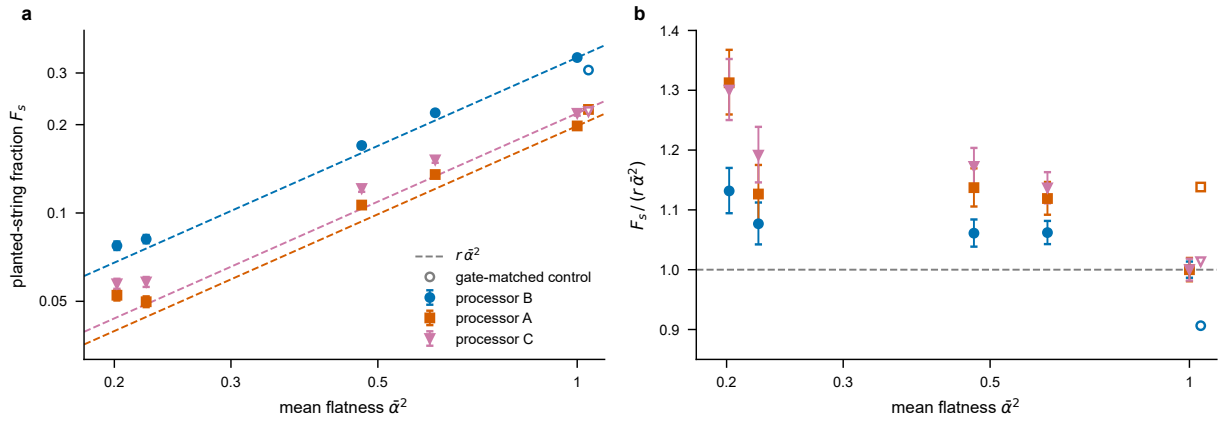


Figure 2 | Planted share against mean flatness at eight qubits. **a**, Planted-string fraction F_s (ten shifts of 4,096 shots each per family, pooled, Wilson 95% intervals) against the exact mean flatness $\bar{\alpha}^2$ of five block-product oracle families at $n = 8$ and one compiled two-qubit depth of 167, on processor B (circles), processor A (squares) and processor C (inverted triangles). Dashed lines: each device's bent-cell retention r times $\bar{\alpha}^2$, where r is the measured F_s of the bent family, so each line passes through that point by construction. Open markers, drawn just right of $\bar{\alpha}^2 = 1$ in both panels: the gate-matched control (a degree-4 bent oracle with one C^3Z per oracle). Registered log-log slopes 0.94, 0.88 and 0.87 (bootstrap 95% intervals 0.93 to 0.96, 0.85 to 0.91 and 0.83 to 0.91), each below 1. **b**, The ratio $F_s/(r\,\bar{\alpha}^2)$ with the Wilson interval of F_s at fixed r. The low-flatness cells lie 6 to 13%, 12 to 31% and 14 to 30% above the line at 1, and each stays resolved when the interval of r is added. The excess is largest in the lowest-flatness family on every device. The controls sit at 0.91, 1.14 and 1.01 of the bent family, so gate content alone moves retention by up to 14% on two devices. The planted shift was the unique mode in 50 of 50 circuits on each device. The sweep measures the flatness factor at $n = 8$ only, and every frontier cell of Table 2 is bent.

which estimates the separation between the planted answer and its closest competitor. Table 4 reports both quantities for every frontier cell, and Supplementary S25 lists every run. Theorem 2 certifies all ten runs at $n = 20$ at the observed shot budget and none beyond (Supplementary S20). At $n = 32$, a wrong string out-counts the planted shift in eight of ten runs.

The Hamming distance of the strongest rival provides an additional diagnostic. At $n = 20$, the strongest competitor typically differs from the planted shift by a single bit, indicating that remaining errors are highly localized. By $n = 24$, most rivals appear at substantially larger distances, and by $n = 32$ the strongest competitors lie 5–20 bit flips away from the answer. The detailed per-run analysis is reported in Supplementary S25.

The per-bit statistics reveal a different frontier. At $n = 20$, every bit marginal exceeds one half and per-bit majority recovers all ten runs. By $n = 24$, several bit marginals have crossed below one half, and per-bit recovery drops to 3/5 on processor B and 1/5 on processor A. Notably, all three processor A recoveries that are fragile under resampling also fail under per-bit majority. Proposition 8 certifies one of the successful processor B runs; in the other three per-bit recoveries, the weakest bit lies within one binomial standard deviation of random guessing.

The observed frontier arrived one to two ladder rungs earlier than the preregistered forecasts, consistent with the stated expectation that the underlying two-qubit-gate depolarizing model is optimistic (Supplementary S16). Nevertheless, the predicted qualitative signature was observed on both processors: single-shot fidelity reached the shot floor before full-string recovery collapsed, and processor B consistently matched or exceeded processor A in all overlapping cells.

The record also distinguishes between competing noise pictures. Assuming independent per-bit flips, the observed per-bit fidelities at the processor A frontier would predict approximately 10^{-3}

Table 2 | Complete degree-2 dual-oracle record from two same-design hardware ladders (132 runs). "2Q" denotes transpiled two-qubit gate count. "Mode" and "Bit" denote recovery by full-string majority vote and per-bit majority vote, respectively. $\bar{F}$ is the mean measured single-shot fidelity (fraction of shots equal to the planted shift among 16 384 shots). One processor A mode success at $n = 24$ arose from a 3-count tie; scored conservatively as a failure, the pooled record through $n = 24$ becomes 49/50. ♯The processor B $n = 32$ success recovers the 32-bit shift from 3 counts against a best wrong outcome of 1, at $F = 1.8 \times 10^{-4}$. Registration, resampling, routing-audit, entanglement, and exploratory-frontier details are reported in Supplementary S8, S24 and S25.

		B (primary)			A			
n	2Q	Mode	Bit	$\bar{F}$	Mode	Bit	$\bar{F}$	Ent.
8	52	5/5	5/5	0.64	5/5	5/5	0.69	yes
12	106	5/5	5/5	0.40	5/5	5/5	0.42	yes
16	380	5/5	5/5	0.10	5/5	5/5	0.085	yes
20	422	5/5	5/5	0.073	5/5	5/5	0.049	yes
24	948	5/5	3/5	3.0×10^{-3}	5/5*	1/5	9.2×10^{-4}	yes
32	1590	1/5♯	0/5	3.7×10^{-5}	0/5	0/5	1.2×10^{-5}	yes
40	2440	0/5	0/5	0	0/5	0/5	0	yes
48	3823	0/5	0/5	0	0/5	0/5	0	yes
56	5305	0/5	0/5	0	0/5	0/5	0	yes
64	6929	0/5	0/5	0	0/5	0/5	0	yes
72	8720	0/5	0/5	0	0/5	0/5	0	yes
80	10996	0/5	0/5	0	0/5	0/5	0	yes
100	17782	0/2	0/2	0	0/2	0/2	0	yes
128	29494	0/2	0/2	0	0/2	0/2	0	yes
154	43962	0/2	0/2	0	0/2	0/2	0	yes
Total		26/66	23/66		25/66	21/66		

correct 24-bit strings in 16 384 shots, whereas the data contain 3–45 occurrences. Recovery therefore relies on residual correlated structure in the output distribution. At the frontier, the measured count ratio between the planted shift and its strongest rival is only of order unity, far smaller than the 10^4–10^5 dominance expected under uniform redistribution of the error mass. The pass/fail boundary instead tracks the shot floor itself: every run containing at least three occurrences of the planted answer recovers it, whereas no run with fewer does. Vote-carried recovery down to the shot floor, followed by a sharp collapse, is precisely the qualitative behavior anticipated by Theorem 2. The depth at which this transition occurs provides a natural graded benchmark of processor performance.

A non-Clifford frontier that ranks processors

Increasing the oracle degree transforms the same planted-instance benchmark into a non-stabilizer stress test. Degree-d oracles introduce $C^{(d-1)}Z$ gates, taking the circuit beyond the efficiently simulable stabilizer (Gottesman–Knill) regime while preserving both the planted solution and the exact verification procedure.

Table 5 summarizes the complete degree-3 record: 156 recoveries in 259 trials across five backends. Recovery remains near ceiling for 2–4 CCZ gates, degrades smoothly for 6–10 CCZ gates, and approaches the noise floor for 10–12 CCZ gates under the baseline compilation recipe. Because the reported tiers aggregate multiple backends and mitigation strategies, the table should be viewed primarily as a frontier map. The principal fixed-protocol measurements are the preregistered

Table 3 | Frontier diagnostics for the degree-2 record. Runs on processor B and processor A, $S = 16\,384$ shots, five runs per cell. Reported quantities include single-shot fidelity $\bar{F}$, planted-shift and strongest-rival counts (μ_c, μ_w), count ratios, predicted uniform-noise dominance, Hamming-distance diagnostics, and mode stability under multinomial resampling. S_{95} denotes the shot count required for mode stability 0.95. * denotes the $n = 24$ tie; ♯ denotes the $n = 32$ 3-against-1 recovery. Details are given in Supplementary S20.

Device	n	$\bar{F}$	μ_c	μ_w	count ratio	R_{unif}	d=1 (%)	d=2 (%)	$d\geq$3 (%)	rival d
B	20	0.073	744–1814	145–229	5.1–9.9	5.0×10^4–1.3×10^5	3.19–5.54	2.58–3.75	79.8–89.7	1–5
A	20	0.049	564–1016	68–196	4.8–11.3	3.7×10^4–6.9×10^4	2.73–4.91	2.47–3.64	85.2–91.4	1–4
B	24	3.0×10^{-3}	25–107	9–18	1.9–5.9	2.6×10^4–1.1×10^5	0.12–0.73	0.20–0.59	98.0–99.5	5–7
A	24	9.2×10^{-4}	3–45	3–15	1.0–4.3	3.1×10^3–4.6×10^4	0.03–0.30	0.04–0.22	99.2–99.9	1–7
B	32	3.7×10^{-5}	0–3	1–2	0.0–3.0	0–7.9×10^5	0.00–0.01	0.00–0.02	100.0	1–11
A	32	1.2×10^{-5}	0–1	1	0.0–1.0	0–2.6×10^5	0.00–0.01	0.00–0.01	100.0	1–6

Device	n	run	μ_c	μ_w	P naive	P tail	S_{95} naive	S_{95} tail
B	24	1–5	25–107	9–18	0.97–1.00	0.96–1.00	2,048–32,768	2,048–16,384
A	24	1	13	3	0.94	0.97	32,768	16,384
A	24	2	6	4	0.24	0.41	262,144	262,144
A	24	3	45	15	1.00	1.00	4,096	4,096
A	24	4*	3	3	0.02	0.14	n/a	n/a
A	24	5	8	3	0.53	0.78	65,536	32,768
B	32	5♯	3	1	0.02	0.79	262,144	32,768
A	32	1–5	0–1	1	0.00	0.00–0.26	n/a	n/a

replication cells executed under a single recipe on a single backend: **10/10** recoveries at ($n = 80, 8$ CCZ), **10/10** at $(100, 6)$, and **6/10** at $(140, 10)$. In the final cell, every non-recovery differs from the planted answer by exactly one bit out of 140 (per-bit fidelity 99.7%), illustrating a gradual rather than catastrophic loss of performance.

The largest-n demonstrations use an interleaved layout for which the prepared state factorizes into the monomial-support qubits together with two-qubit blocks. Accordingly, these experiments primarily probe depth scaling and large-scale recovery, while entanglement remains localized to the monomial support (Supplementary S14). Full-register-entangled recoveries are demonstrated up to $n = 24$ in the degree-2 benchmark and up to $n = 22$ for degree-10 instances ($C^{(9)}Z$, 8/8 recoveries; Supplementary S14).

The benchmark becomes most discriminating near the recovery frontier, where success rates begin to depart from unity. In this regime, a preregistered comparison across five Heron processors was performed using the frontier cells

$$(n, d) \in \{(18, 8), (20, 9), (22, 10)\},$$

and scored against each processor's measured median two-qubit error at submission. The resulting ranking achieved a Spearman correlation of $\rho = -0.90$, exceeding the preregistered acceptance threshold of -0.6, and was monotonic except for one pair of processors whose reported two-qubit error rates agreed to three significant figures (Supplementary S16).

First, this was the second ranking study; an earlier version employing a different scoring rule did not meet its acceptance criterion and is documented in Supplementary S16. Second, with only five processors, the corresponding one-sided permutation p-value is $5/120 \approx 0.04$, providing suggestive rather than definitive statistical evidence. Nevertheless, the observed sensitivity is not restricted to a single hardware family: a vetted-map replication on the Nighthawk R2 processor C reproduced the same frontier collapse pattern (Supplementary S21 and S24).

On the same device, an out-of-sample validation exceeded the prediction of a fit fixed before the validation run, achieving a pooled per-bit fidelity of 0.609 against a predicted band of 0.557 ± 0.026. Across architectures, the degree-4 lattice realization compiles the same benchmark cells using 35–38%

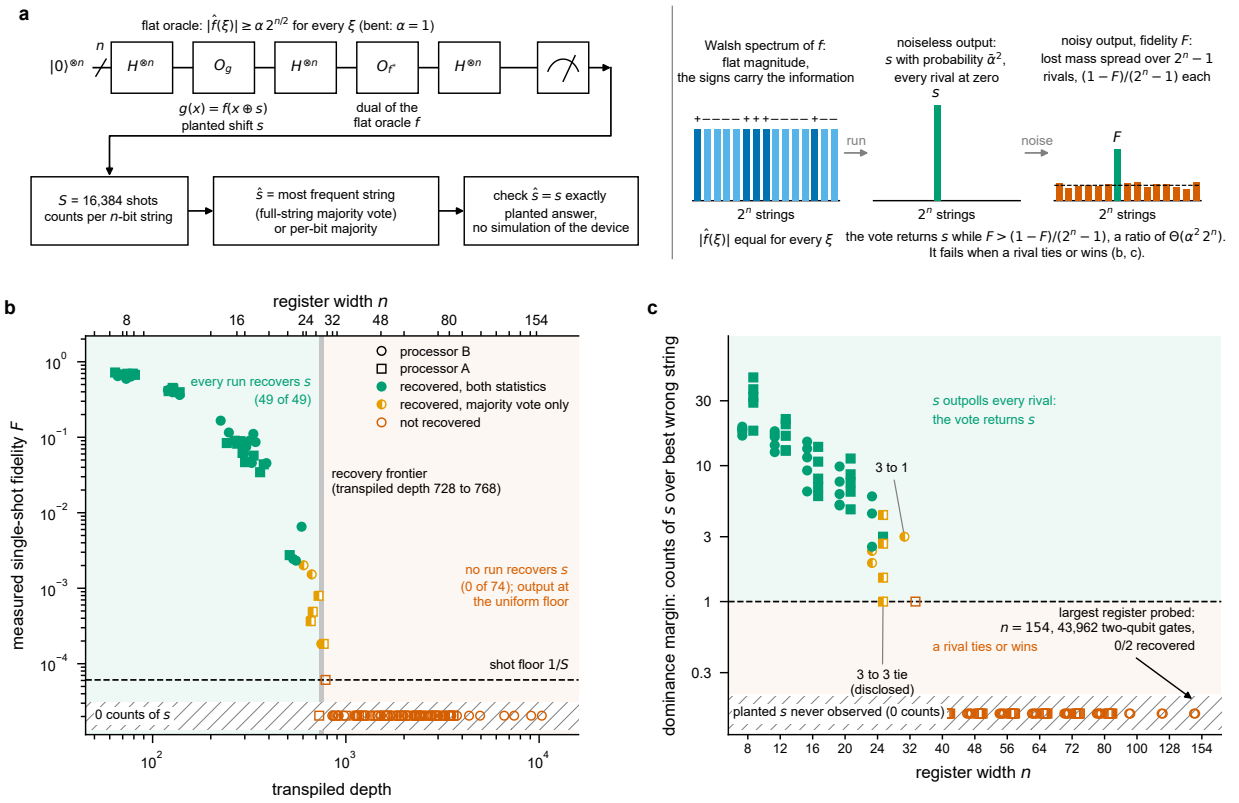


Figure 3 | Noise resilience of the flat-state hidden-shift benchmark. **a**, A planted shift s is read back from $S = 16{,}384$ shots by full-string or per-bit majority and checked exactly. Under uniform redistribution, lost probability spreads over 2^n outcomes. The vote wins while s outpolls every rival. **b**, Single-shot fidelity F versus transpiled depth, processor B (circles) and processor A (squares). F falls four orders of magnitude to the shot floor $1/S$, yet every run left of the frontier band recovers s (40/40 through $n = 20$, 10/10 at $n = 24$ by vote alone, three processor A runs unstable under resampling), the band holds the last recovery ($n = 32$, 3 counts against 1), and none beyond it recovers (0 of 74). **c**, Count ratio of s over its best rival. The vote wins above 1. Void processor A runs at $n \geq 100$ omitted. Largest register probed: $n = 154$, 43,962 two-qubit gates, 0/2.

fewer two-qubit gates. The graded benchmark therefore tracks delivered circuit quality rather than a single calibration scalar and correctly distinguishes devices whose reported calibration metrics alone would misrank (Supplementary S21).

Discussion

Gate fidelities ultimately limit how deeply a noisy quantum processor can carry a computation. The benchmark reported here recovers planted hidden shifts in all tested degree-2 instances through $n = 24$, despite measured single-shot fidelities near 10^{-3}, and retains a final vote-carried recovery at $n = 32$ near the shot floor. Spectral flatness fixes the planted answer's ideal output weight and provides a mechanism by which majority-vote recovery can persist even when individual shots are rarely correct. Because the oracle degree continuously introduces non-Clifford operations, the same framework also provides a depth-graded benchmark of processor performance.

Theorem 2 identifies the probability margin between the planted answer and its strongest competitor as the key quantity governing recovery. Experimentally, the frontier is reached when

Table 4 | Margin diagnostics for all degree-2 frontier cells ($S = 16{,}384$ shots; both processors; $n = 20$, 24, and 32; one row per cell, ranges over its five runs). "Rec." counts full-string recoveries. $\hat{F}$ is the planted-shift fraction, d the strongest rival's Hamming distance from the planted shift, and $\hat{\Delta} = \hat{F} - \hat{p}$ the margin over the strongest rival fraction $\hat{p}$. "$\hat{\Delta} > 0$" counts runs whose 95% percentile bootstrap interval from 2 000 multinomial resamples excludes zero. The ratio μ_c/μ_w compares planted-shift and rival counts. "Stable" denotes mode stability under resampling, and "Cert." indicates that Theorem 2 certifies recovery at the observed shot budget. *The 3-count tie of Table 2. Supplementary S20 provides methodological details, and Supplementary S25 lists every run with its bootstrap.

Device	n	Rec.	$\hat{F}$	d	$\hat{\Delta}$	$\hat{\Delta} > 0$	μ_c/μ_w	Stable	Cert.
B	20	5/5	0.045–0.11	1–5	0.037–0.10	5/5	5.1–9.9	5/5	5/5
A	20	5/5	0.034–0.062	1–4	0.030–0.052	5/5	4.8–11.3	5/5	5/5
B	24	5/5	1.5–6.5×10^{-3}	5–7	0.73–5.4×10^{-3}	4/5	1.9–5.9	5/5	0/5
A	24	5/5*	0.18–2.8×10^{-3}	1–7	0–1.8×10^{-3}	1/5	1.0–4.3	2/5	0/5
B	32	1/5	0–1.8×10^{-4}	5–20	-1.2 to 1.2×10^{-4}	0/5	0–3	0/5	0/5
A	32	0/5	0–6.1×10^{-5}	15–19	-6.1×10^{-5} to 0	0/5	0–1	0/5	0/5

that margin collapses, even though residual signal remains visible in the output distribution. Error probability accumulates preferentially on Hamming-near competitors rather than spreading uniformly across the output space. Nevertheless, the planted answer remains the mode through the recovery frontier, indicating that modal recovery persists well beyond the theorem's idealized regime. The theory therefore identifies a recovery mechanism, while the measurements quantify its robustness under realistic noise.

The benchmark also provides diagnostic information beyond a binary success metric. The Hamming distance of the strongest competitor reveals how error probability accumulates around the correct answer, distinguishing localized failure modes from broadly distributed errors. A simple two-component model of near-clean and scrambled shots does not fully explain the observed frontier behavior, suggesting that the structure of the near-answer probability mass remains an open question.

A central feature of the benchmark is the trade-off between exact verification and classical hardness. Because the hidden shift is planted and known in advance, correctness can be verified exactly at any system size without classical simulation of the quantum device. For the public oracle families used here, that same structure also permits efficient classical recovery of the hidden shift. Accordingly, the benchmark is not intended as a demonstration of computational advantage or classical intractability. Its purpose is instead to provide a scalable, structured computational task with an exactly known target solution whose recovery can be tracked as hardware noise increases.

Whether comparable noise resilience can be combined with computational hardness remains an open question. One possible direction is to replace the public oracle structure with trapdoor-based constructions while preserving the favorable output statistics associated with spectrally flat states.

Taken together, these results establish a benchmark that combines exact verification, a structured algorithmic task, tunable non-Clifford content, and experimentally demonstrated recovery far below the regime in which individual shots remain reliable. Rather than measuring computational hardness, the benchmark measures how far a noisy quantum processor can carry a known computation while retaining a verifiable answer. More broadly, the results show that structured computations with exactly known solutions can provide scalable diagnostics of quantum-processor performance, complementing existing fidelity-based benchmarking methodologies.

Table 5 | Complete higher-degree record (degree-3 oracles with CCZ gates). Entries report pooled successes/trials for each (n, CCZ) cell across all backends and recipes. †Includes the preregistered replication batch (10 fresh seeds). ‡Executed only under the chunked XY4-plus-twirling recipe. "Ent." indicates whether the implementation prepares full-register entanglement. Per-run records, replicate analyses, and entanglement audits are provided in Supplementary S11 and S14.

n	CCZ count k						
	2	4	6	8	10	12	Ent.
8	23/25	–	–	–	–	–	yes
12	15/15	–	–	–	–	–	yes
20	5/5	–	–	–	–	–	no
32	5/5	–	–	–	–	–	no
48	10/10	–	–	–	1/3	0/2	no
64	5/10	–	–	–	0/3	–	no
80	4/5	5/5	5/8	**11/13**†	0/6	0/6	no
90	–	–	–	2/3	–	–	no
100	15/24	7/11	**15/23**†	0/13	1/1‡	–	no
108	–	–	1/1‡	–	–	–	no
112	4/5	–	–	–	–	–	no
116	5/5	–	1/3	–	–	–	no
120	0/8	–	0/3	–	–	–	no
124	–	–	–	0/3	–	–	no
126	0/12	–	–	–	–	–	no
130	3/3	–	–	–	–	–	no
136	–	–	–	0/3	–	–	no
140	3/3	–	1/1‡	2/2‡	**7/11**†	–	no
Total: 156/259 (per-tier subtotals in Supplementary S11)							

Methods

Function construction. The degree-2 benchmark uses quadratic bent functions

$$f(x, y) = (-1)^{x^T B y},$$

where $x, y \in \mathbb{F}_2^{n/2}$ and B is an invertible matrix over $\mathbb{F}_2$ (fixed seed 42 throughout the degree-2 study). Spectral flatness follows from the bent property (Supplementary S1). Higher-degree benchmarks are generated by adding independently sampled cubic monomials (CCZ gates; Supplementary S11), while the unbounded-degree family uses a single monomial whose degree increases with system size (Supplementary S14). The exact dual oracle

$$f^*(u, v) = v \cdot B^{-1} u$$

is compiled directly from B^{-1}. No explicit Walsh–Hadamard transform is constructed.

Theory.

Theorem 5 (Quantum query upper bound)**.** *There is a quantum algorithm recovering the hidden shift with $Q = \mathcal{O}(n/\alpha^2)$ queries to g and $\mathcal{O}(n^3)$ classical post-processing.*

Proposition 6 (Classical recovery for public f)**.** *When f is public, the shift is recoverable from $\mathcal{O}(n + \log(1/\delta))$ classical queries with probability at least $1 - \delta$.*

Proposition 7 (Exact recovery probability)**.** *For any α-flat function, the dual-oracle circuit returns the planted shift with probability*

$$P(s) = \bar{\alpha}^2,$$

where

$$\bar{\alpha} = 2^{-n/2}\,\mathrm{mean}_\xi\,|\hat{f}(\xi)| \geq \alpha.$$

Proposition 8 (Per-bit majority under biased marginals)**.** *Let q_j denote the probability that output bit j of a single shot equals s_j. If $q_j > \frac{1}{2}$ for every j, then per-bit majority over S independent shots recovers s with probability at least*

$$1 - \sum_{j=1}^{n} \exp\!\left(-2S(q_j - \tfrac{1}{2})^2\right).$$

No independence between output bits is required.

Proofs appear in Supplementary S1-S3. Experimental validations of Proposition 7 are reported in Supplementary S23.

Protocols. Three circuit protocols were implemented (Fig. 1). Protocol A (Fourier-label sampling) applies

$$H^{\otimes n} O_g H^{\otimes n}$$

and interprets each shot as a linear constraint $b = s \cdot \xi$, recovering the shift through Gaussian elimination over $\mathbb{F}_2$ once sufficiently many constraints span. For bent f the sign bits are physically obtainable only through the dual construction. Protocol B combines the degree-2 oracles

$$H^{\otimes n} O_f O_g H^{\otimes n},$$

which simplifies substantially through gate cancellation and serves primarily as a compilation check. Protocol C is the full Rötteler dual-oracle construction. All hidden-shift recovery results reported in the manuscript use Protocol C.

Execution. Experiments were implemented in Qiskit 2.x using Runtime SamplerV2 on IBM Heron and Nighthawk processors. Degree-2 experiments employed optimization level 3 with XY4 dynamical decoupling, whereas degree-3 experiments used noise-aware routing at optimization level 2. Backend characteristics and compiled-circuit statistics are reported in Supplementary S21.

Recovery and scoring. Recovery uses only measured counts. Two estimators are reported:
1. the most frequent measured bit string (mode), and
2. the per-bit majority string.

The planted shift enters only after recovery for scoring. Verification is therefore an exact string comparison between the recovered and planted shifts.

Statistics. Reported confidence intervals are Wilson 95% intervals. Backend ranking experiments use permutation tests, and bootstrap intervals are reported where stated. Detailed statistical methodology appears in Supplementary S23 and S24.

Reproducibility. Hardware experiments were conducted under documented registrations and an auditable execution workflow. Registration records, submission manifests, code, analysis scripts, and machine-readable result files accompany the manuscript. Details of the verification pipeline, registration chronology, and formal verification efforts are described in Supplementary S12 and S24.

Data availability

The result files that support the findings of this study (job identifiers, per-circuit transpiled metrics, full count tables, recovered shifts and the machine-generated accounting tables), the registration files, the submission manifests, the registration chronology, the two unbounded-degree ladder summaries and the per-coupler routing records of the audited jobs are deposited with the code at Zenodo under the DOI 10.5281/zenodo.23086962, the data under the Creative Commons Attribution 4.0 license. Editors and reviewers can access the deposit through a private link during review, and the deposit becomes public on publication. Source data for Figs. 3 and 2 are provided with this paper.

Code availability

The circuit generation, submission and analysis code, the Lean 4 development, the verification scripts, the figure script, the noise-mixture fit of Supplementary S25 with its report and a tutorial notebook are deposited with the data at Zenodo under the DOI 10.5281/zenodo.23086962, with the code licensed under the Apache License 2.0 and on the same access terms. The code will also be released in a public repository on acceptance.

Acknowledgements

The authors used Claude (Anthropic) and co-pilot (Microsoft) as coding and writing assistants for analysis scripts, literature checks and manuscript editing, under the authors' direction.

Author contributions

L.H. conceived the study, with contributions to the conceptualization from S.U. L.H. developed the theory, designed and performed the experiments, developed the software and supervised the project. S.U. analyzed the data and validated the results. L.H., S.U. and V.K. wrote the manuscript. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Funding

This work received no external funding.

References

[1] Arute, F. *et al.* Quantum supremacy using a programmable superconducting processor. *Nature* **574**, 505–510 (2019).

[2] Proctor, T., Rudinger, K., Young, K., Nielsen, E. & Blume-Kohout, R. Measuring the capabilities of quantum computers. *Nature Physics* **18**, 75–79 (2022). ArXiv:2008.11294.

[3] Aharonov, D. & Ben-Or, M. Fault-tolerant quantum computation with constant error rate. *SIAM Journal on Computing* **38**, 1207–1282 (2008).

[4] Takagi, R., Endo, S., Minagawa, S. & Gu, M. Fundamental limits of quantum error mitigation. *npj Quantum Information* **8**, 114 (2022).

[5] Cross, A. W., Bishop, L. S., Sheldon, S., Nation, P. D. & Gambetta, J. M. Validating quantum computers using randomized model circuits. *Physical Review A* **100**, 032328 (2019). ArXiv:1811.12926.

[6] Proctor, T. *et al.* Scalable randomized benchmarking of quantum computers using mirror circuits. *Physical Review Letters* **129**, 150502 (2022). ArXiv:2112.09853.

[7] Rötteler, M. Quantum algorithms for highly non-linear Boolean functions. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 448–457 (2010).

[8] Bernstein, E. & Vazirani, U. Quantum complexity theory. *SIAM Journal on Computing* **26**, 1411–1473 (1997).

[9] Amy, M. & Stinchcombe, L. S. Polynomial-time classical simulation of hidden shift circuits via confluent rewriting of symbolic sums. *Quantum* (2025). ArXiv:2408.02778.

[10] Baron, D., Patil, O. & Zhou, H. Maximum likelihood quantum error mitigation for algorithms with a single correct output. In *2024 IEEE International Conference on Quantum Computing and Engineering (QCE)*, 124–127 (2024). ArXiv:2402.11830.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HSRsuppl.pdf