

HilbertBench: Blinded, Pre-Registered Validation of a Diagnostic Instrument for Quantum Machine Learning

Md. Abdullah Al Mamun, Sadia Afrin Purba, and Joseph Picone

Neural Engineering Data Consortium, Temple University, Philadelphia, PA, USA

Variational quantum machine learning fails silently: barren plateaus, shot starvation, and hardware noise all present as a loss that stops improving. We introduce HilbertBench, a non-intrusive instrument that records circuit executions through a transparent proxy into sealed, append-only traces and attributes failures offline, and we validate it the way diagnostic instruments are validated in clinical research. In two pre-registered, blinded studies, an independent diagnostician attributed planted failure modes from traces alone, scored against cryptographically committed answer keys. On a corpus of dominant plantings she was correct on all 72 runs (36 simulated, 36 on IBM hardware). A second corpus, graded down to the detection boundary, stress-tested the method: two detection boundaries reproduced on fresh seeds, a blinded human reader and a frozen mechanical rule tracked each other closely (both near 0.6 accuracy on this harder task), and the protocol surfaced a previously unnoticed false-positive mode of the instrument itself, showing that blinded evaluation can debug an instrument as well as certify it. Applying the validated instrument to two superconducting processors, we then measure how noise displaces the expressibility–trainability frontier, the hardware-to-ideal landscape variance ratio falling from 0.93 to roughly 0.50 with depth. Instrument, traces, pre-registrations, and both corpora are openly available.

1 Introduction

Variational quantum machine learning [1, 2, 3] fails silently. A training run whose loss has stopped improving looks the same from the outside whether the cause is an exponentially flat landscape [4, 5, 6], a measurement budget too small for the signal, hardware noise that has destroyed the computation [7, 8], or nothing at all, a healthy optimization that simply needs more iterations. These four situations demand four different responses (change the ansatz, buy more shots, reduce the circuit depth, wait), yet on NISQ devices [8], where every experiment spends scarce and expensive quantum-processor time, practitioners routinely cannot tell them apart post hoc.

Classical machine learning solved the corresponding problem with observability infrastructure: experiment trackers and structured logs [9] that record what a training run did, so that failures can be attributed from evidence after the fact. Quantum machine learning largely lacks this layer. Diagnostic information is scattered across ad hoc print statements and notebook state; it is lost when the process ends, cannot be verified after the fact, and, most importantly, is produced by instrumentation entangled with the very algorithm code under study, so that adding diagnostics risks changing the experiment.

Existing quantum software tooling occupies adjacent niches, none of which is post-hoc failure attribution. Error-mitigation libraries such as Mitiq [10] *change* the computation to suppress noise rather than diagnose why an unmodified run failed. Characterization suites such as Qiskit Experiments [11] measure device properties by running *dedicated* circuits, an intrusive protocol that spends extra device time and characterizes the processor, not the training run that failed on it. The major QML frameworks [12, 13] offer in-process debugging and inspection utilities, but their outputs live and die with the process: noth-

ing is preserved that a third party could later audit. And classical experiment trackers [9] record hyperparameters and metrics but are blind to quantum evidence: shot budgets, transpiled circuit structure, and the device calibration in force at execution time. HilbertBench’s position is the complement of all four: record the evidence a diagnosis needs, at the execution boundary the run already crosses, without altering the run, and make the record verifiable after the fact.

This paper introduces **HilbertBench**, a diagnostic framework built on a separation of concerns: record everything at the circuit execution boundary, non-intrusively and verifiably; interpret the recording offline. A transparent proxy wraps the estimator or device a user already employs and records each execution into an append-only, cryptographically sealed trace; six analyzers then compute mechanism-level diagnostics: landscape flatness, shot-noise ratio, expressibility [14, 15], convergence, circuit structure, and calibration-based fidelity, from the trace alone.

An instrument of this kind is only as good as the evidence that its readings mean what they claim, and here QML faces a methodological gap: diagnostic tools are typically evaluated by their own authors, on data they generated, knowing the answers. We instead validate HilbertBench the way diagnostic instruments are validated in clinical research, adopting the pre-registration discipline of the open-science movement [16]. We planted the four failure modes above in a corpus of recorded runs, froze the analyzer thresholds on disjoint development data, *pre-registered* the protocol and analysis plan on the Open Science Framework, published a SHA-256 commitment of the answer key, and then handed the blinded traces to an independent diagnostician who had no knowledge of how they were produced. To our knowledge this is the first blinded, pre-registered validation of a QML diagnostic. The diagnostician attributed all 72 runs correctly (accuracy 1.000, 95% Wilson interval [0.949, 1.000], $p = 4.5 \times 10^{-44}$ against chance), and the pre-registered barren-noise co-occurrence on hardware appeared in her secondary labels exactly as predicted (Sec. 4.1). Because a ceiling score on a corpus designed to be diagnosable measures the workflow rather than the instrument’s limits, we complement the blinded study with an algorithmic one: a mechanical decision rule, com-

mitted before execution, replays the blinded corpus with no human in the loop and is then swept across a corpus whose planted mechanisms range from physically absent to fully expressed, mapping where the evidence stops supporting detection (Sec. 4.2). We then return the human to that boundary in a second, separately pre-registered blinded study, where accuracy falls to around 0.6 for reader and rule alike, two detection boundaries reproduce on fresh seeds, and the instrument is caught in a false positive its authors had never observed (Sec. 4.3).

With the readings validated and their limits mapped, we put the instrument to work on a question theory leaves open. On two independent superconducting processors we measure how noise displaces the expressibility–trainability frontier [15]: the hardware-to-ideal landscape variance ratio falls to roughly 0.50 in the deep regime, with bootstrap intervals separating cleanly from the matched ideal (noise-induced suppression of trainability [7] beyond the intrinsic plateau), while the measured expressibility is simultaneously inflated (Sec. 4.4). As external checks on the same readings, the instrument recovers the McClean barren-plateau scaling law [4] with fitted exponent -1.00 and predicts hardware fidelity from recorded calibration snapshots (Sec. 4.5).

The remainder of the paper is organized as follows. Section 2 describes the recorder, the sealed trace, and the analyzers. Section 3 specifies both blinded, pre-registered validation protocols and the committed detection-boundary follow-up between them. Section 4 reports the two blinded outcomes, the detection boundaries, and the hardware measurements, and Section 5 discusses the diagnostic entanglement of barren plateaus with noise on NISQ hardware, limitations, and open problems. The instrument, all traces and seeds, the pre-registrations, and both corpora with their answer keys are openly available (Data and code availability).

2 The instrument

HilbertBench is built around a single separation of concerns: the *recording* of what a quantum program did is decoupled from the *interpretation* of whether that behaviour was pathological. Recording is non-intrusive and happens online;

interpretation is a set of pure functions applied offline to the recorded evidence. Figure 1 shows the arrangement, together with the evidence one real run leaves behind and the reading the analyzers give it. This section describes the recorder, the trace format it produces, and the analyzers that read it.

The design is governed by four architectural invariants, enforced by a 319-test suite rather than by convention: observation never perturbs the observed program; a sealed trace is immutable and any tampering is detectable; the data models are generated from a versioned schema, so the documented format and the code cannot drift; and analyzers are pure functions of a trace, so every diagnostic result is reproducible from the evidence file alone. Integrations cover the two dominant QML stacks, Qiskit [12] and PennyLane [13]. These guarantees are not incidental: the blinded protocol of Sec. 3 is only meaningful because the evidence a diagnostician reads is provably the evidence the experiment produced.

2.1 Non-intrusive recording

HilbertBench records executions through a transparent proxy that wraps the estimator or backend a user already employs. The proxy forwards every call to the wrapped object unchanged and, as a side effect, writes a record of the execution; no algorithm code is modified and no results are altered. Integrations are provided for the Qiskit primitives and for PennyLane devices, so the same recording path serves both simulators and real hardware.

Non-intrusion is behavioural, and its temporal cost is small and constant per execution: recording adds approximately 1 ms (content hashing plus one append), a 4% slowdown on a shot-based simulator ($26.0 \rightarrow 27.2$ ms per evaluation at $\sim 10^3$ shots). The worst case is an idealized statevector backend, where an evaluation costs only ≈ 1.4 ms and the same absolute overhead is 72% relative; against the seconds-to-minutes latency of queued hardware it is negligible.

Recording operates in one of two modes. In *passive* mode the recorder observes a real optimization trajectory. In *active* mode it performs a controlled intervention, evaluating the cost at a batch of uniformly random parameter points, to characterize a landscape without the bias an optimizer introduces by seeking descent directions.

The distinction is recorded, so an analyzer never confuses a guided trajectory with an unbiased landscape sample.

2.2 The sealed, append-only trace

A trace is an append-only log of *spans*, one per recorded circuit execution, each carrying its bound parameters, the measured outcome, a reference to the circuit, and, on hardware, a snapshot of the device calibration at the moment of execution. Large artifacts (circuit definitions in OpenQASM, measurement outcomes) are stored once in a content-addressed file store and referenced by their hash, so identical circuits across thousands of optimizer steps are de-duplicated automatically.

The trace format is schema-first: a versioned JSON Schema defines every record, and the in-memory data models are generated from that schema rather than written by hand, eliminating drift between the documented format and the code. Once written, a trace is sealed with a cryptographic hash so that any later modification is detectable; this is what allows a diagnostic result to be tied to immutable evidence, and what makes the blinded protocol of Sec. 3 possible.

2.3 The six analyzers

Each analyzer is a pure function from a sealed trace to a structured result that separates *evidence* from *interpretation*: it reports the raw quantity it computed (with a bootstrap confidence interval where applicable) alongside a human-readable verdict obtained by comparing that quantity to a fixed threshold. A reader who disagrees with a threshold can re-interpret the evidence without re-running the experiment. Four of the six analyzers are quantitative detectors; we state each one’s estimator as implemented.

Barren plateau (trainability). For a recorded cost $C(\theta) = \langle \psi(\theta) | O | \psi(\theta) \rangle$ evaluated at parameter vectors $\theta_1, \dots, \theta_M$ drawn uniformly from $[0, 2\pi]^m$ (active mode), the detector reports the sample variance

$$\text{Var}[C] = \frac{1}{M-1} \sum_{i=1}^M (C(\theta_i) - \bar{C})^2, \quad (1)$$

with a bootstrap 95% interval; a landscape is flagged barren when $\text{Var}[C] < \tau_{\text{var}}$, the flat-

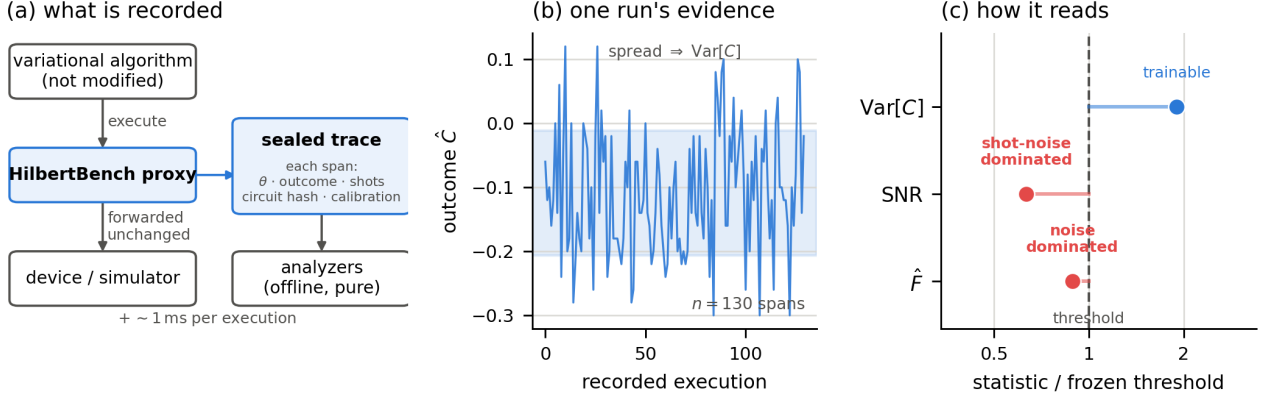


Figure 1: The instrument, and what it produces. (a) The recording boundary. A transparent proxy sits between the unmodified variational algorithm and the device, forwards every call unchanged, and appends one span per execution to a sealed trace; analyzers read that trace offline as pure functions. Nothing on the algorithm’s path is altered, at a cost of ~ 1 ms per execution. (b) The evidence for one real corpus run: the 130 recorded outcomes of a noise-dominated training trajectory. The shaded band is ± 1 standard deviation, the spread that becomes $\text{Var}[C]$ in Eq. (1). (c) The same run read by three frozen thresholds, each statistic plotted as a ratio to its own threshold so that the dashed line at 1 is the decision boundary. This run reads as trainable ($\text{Var}[C]$ above threshold) but both shot-noise dominated and noise dominated: the two mechanisms that were in fact planted in it.

ness signature of an exponentially concentrating cost [4, 6].

Shot noise (measurement). An expectation estimated from S shots carries an estimator variance bounded below by the shot-noise floor $\sigma_{\text{shot}}^2 \approx 1/S = p^2$, where p is the recorded target precision. The detector compares the variance of the recorded outcome trajectory to that floor,

$$\text{SNR} = \text{Var}[\hat{C}] / \sigma_{\text{shot}}^2, \quad (2)$$

flagging shot starvation when $\text{SNR} < \tau_{\text{SNR}}$: the recorded signal is then statistically indistinguishable from sampling noise.

Expressibility. Following [14], the recorded states of an active-mode probe yield pairwise fidelities F whose histogram $\hat{P}_b$ (over bins b) is compared to the Haar-random fidelity distribution $P_{\text{Haar}}(F) = (N-1)(1-F)^{N-2}$ with $N = 2^n$; the exact Haar mass q_b per bin is obtained by integrating its distribution function $1 - (1-F)^{N-1}$ over the bin, and the analyzer reports

$$D_{\text{KL}}(\hat{P} \parallel P_{\text{Haar}}) = \sum_b \hat{P}_b \ln \frac{\hat{P}_b}{q_b}, \quad (3)$$

with a bootstrap interval; small divergence means near-Haar coverage (high expressibility) [15].

Hardware noise (fidelity). From the calibration snapshot recorded at execution time, scoped

to the qubits the circuit actually used, the analyzer estimates the circuit fidelity by the standard NISQ product formula over the transpiled circuit’s operation counts,

$$\hat{F} = (1 - \bar{e}_1)^{n_1} (1 - \bar{e}_2)^{n_2} (1 - \bar{e}_{\text{ro}})^{n_{\text{ro}}}, \quad (4)$$

where $\bar{e}_1, \bar{e}_2, \bar{e}_{\text{ro}}$ are the mean single-qubit-gate, two-qubit-gate, and readout error rates on the active qubits and n_1, n_2, n_{ro} the corresponding operation counts; a run is noise-dominated when $\hat{F} < \tau_F$. The remaining two analyzers are descriptive: **optimization convergence** (summary statistics of the recorded trajectory) and **circuit structure** (depth, gate, and entanglement counts parsed from the recorded circuit).

The default thresholds used throughout ($\tau_{\text{var}} = 0.005$, $\tau_{\text{SNR}} = 1.5$, $\tau_F = 0.50$) are order-of-magnitude statements rather than fitted constants. For the local two-qubit cost used throughout, trainable landscapes at small width show $\text{Var}[C] \gtrsim 5 \times 10^{-2}$ while intrinsic plateaus sit below 10^{-3} (measured in Sec. 4.5); τ_{var} marks the geometric middle of that separation. $\tau_{\text{SNR}} = 1.5$ demands the recorded signal variance exceed the shot floor by half again; at $\text{SNR} \leq 1$ the trajectory carries no resolvable signal by construction. $\tau_F = 0.5$ marks majority signal loss: less than half the ideal expectation survives the device. All three are frozen before any validation

data are generated, and Appendix A shows each verdict is stable across a wide band around them.

3 Validation methodology

An instrument is only as trustworthy as the evidence that its readings mean what they claim. We validate HilbertBench with a protocol borrowed from clinical trials: we plant known failure modes, hide the labels, have an independent analyst diagnose each run from the trace alone, and score the result against an answer key that was cryptographically committed before the analyst saw any data.

3.1 Planted failure modes and the corpus

We consider the four failure modes that dominate practical variational QML and that are indistinguishable from the outside; the loss simply stops improving: a *healthy* run that trains normally, a *barren plateau* (an exponentially flat landscape), *shot starvation* (too few measurements, so the signal is lost in sampling noise), and *noise domination* (device noise destroys the computation). Each class is produced by a fixed generation recipe that plants exactly one mechanism by construction, so the ground-truth label is known: a wide, deep random landscape for the barren plateau; a precision-starved shallow variational circuit for shot starvation; a deep transpiled circuit on a noisy device for noise domination; and a shallow, well-resourced circuit on a clean simulator for the healthy case.

The corpus has two strata, blinded together. The *simulated* stratum provides clean, isolated ground truth (in simulation a single mechanism can be induced without the others) and forms the primary analysis: 36 runs, nine per class. A *real-hardware* subset, generated on a single pinned superconducting device in balanced rounds, provides real-device credibility. Because on a real device the circuit depth required to induce an intrinsic barren plateau also collapses fidelity into the noise-dominated regime, the barren and noise classes are physically entangled on hardware; we therefore treat the hardware barren runs as expected co-occurrence cases (Sec. 4.4) rather than clean ground truth, and diagnose every hardware run with a primary and an optional secondary label.

3.2 The blinded, pre-registered protocol

The full protocol, analysis plan, and thresholds were pre-registered on the Open Science Framework (DOI [10.17605/OSF.IO/2PWHU](https://doi.org/10.17605/OSF.IO/2PWHU)) *before* any test-corpus run was generated; the instrument was frozen at version 1.0.0 (Zenodo DOI [10.5281/zenodo.21009245](https://doi.org/10.5281/zenodo.21009245)). The custodian of the corpus (Researcher A) generated the runs, blinded each one by copying it verbatim under a random identifier, audited every run for label leakage in its metadata, and published the SHA-256 hash of the answer key (the commitment) to two independent, timestamped locations before the diagnostician saw anything. An independent diagnostician (Researcher B), who did not build the analyzers and had no knowledge of the corpus design, then assigned to each run a primary label, an optional secondary label, a confidence, and free-text notes, using only the instrument’s outputs. The diagnosis was a single pass and was itself hash-committed before the answer key was unsealed, so neither party could adjust their record afterward. Generation order was shuffled so that timestamps carry no label information, and the class balance was not disclosed to the diagnostician.

Researcher B is a graduate researcher in the authors’ group. Before the first study she had [XX: prior experience with QML / with this instrument], and her preparation consisted only of the public documentation; she completed each corpus in a single pass over [XX: duration]. This matters for interpretation: the ceiling result and the null of Sec. 4.3 both describe this reader, at this level of expertise, on these corpora.

Two design limits are disclosed up front. The study used a single diagnostician, so inter-rater agreement is out of scope; the binomial test below rejects random guessing, not variation across readers, and the result should be read as an existence proof: a trained reader, from traces alone, can attribute every run, rather than a claim about diagnostician populations. The machine-only replay of Sec. 4.2 removes the human from the loop entirely and is reproducible by any referee. Second, both roles were held within one research group (the corpus custodian is the lead author; the diagnostician is a co-author). The commitments are what make this auditable: the answer key and the diagnosis sheet were each hashed and published before unsealing, the pre-

registration is timestamped, and the unsealing session was recorded, so any tampering after the fact is cryptographically detectable, a stronger guarantee than most single-lab validations offer, though a cross-laboratory study remains the natural strengthening.

3.3 Frozen thresholds and analysis plan

The analyzer thresholds were fixed at the library defaults on a disjoint twelve-run development corpus and frozen before the test corpus existed, so the detector could not be tuned to its own exam; their robustness is reported as a registered supporting analysis (Appendix A). Freezing is a procedural commitment, not a claim of optimality: the constants are order-of-magnitude statements (Sec. 2.3), and what the freeze guarantees is that they were set *before* any data they would be scored on existed. The two properties are orthogonal: a threshold can be approximate and still be untuned, and the sensitivity analysis shows the results are stable under substantial perturbation of the constants. The registered primary outcome is the diagnostician’s primary-label accuracy on the simulated stratum, with a 95% Wilson score interval [17] and a one-sided exact binomial test [18] against the four-class chance level of 25%. Secondary outcomes are the per-class precision, recall, and F_1 ; the full confusion matrix; the top-two accuracy (whether the true class is the primary or secondary label); and, on the hardware stratum, the rate at which planted barren runs are diagnosed as noise-dominated, a pre-registered, falsifiable prediction of the barren-noise entanglement.

3.4 Detection-boundary mapping

The blinded study grades the workflow on a corpus each of whose runs expresses one dominant mechanism; it cannot say where detection *fails*. A follow-up study (non-preregistered; its complete design of grids, seeds, decision rule, and metrics was committed to the public repository before any run was generated or scored) maps that boundary algorithmically. Each mechanism is swept along a physical knob using the byte-identical generators of Sec. 3.1: circuit width with depth for the barren mechanism (four to twelve qubits at the depth- n rule), measurement budget for shot starvation (target precision 0.02 to 0.45,

i.e. $\sim 2,500$ down to ~ 5 shots), and circuit depth on a fixed calibrated noise model for noise domination (two to sixteen layers), with five replicate runs per level and a healthy-control arm for false positives. The knob is never the diagnostic statistic itself; planting is always physical, and the achieved statistics are measured from the traces afterward. A frozen mechanical decision rule (the naive reading of the analyzer verdicts under the same frozen thresholds, with a fixed precedence for the primary label) classifies every run with no human in the loop. The rule is also replayed, unmodified, over all 72 runs of the original blinded corpus, separating the instrument’s discriminative power from the diagnostician’s competence on identical evidence. Finally, a mixture arm plants two mechanisms simultaneously at strong settings, with the expected outcome of each mixture stated in the committed protocol in advance, including one mixture (barren+shot-starved) predicted to be *undetectable* in principle, because a five-shot measurement budget cannot resolve a 10^{-3} landscape variance.

3.5 Design of the second blinded study

The ceiling result of Sec. 4.1 and the mechanical boundaries of Sec. 4.2 leave one question open: at the strengths where the evidence stops being conclusive, what does a human reader do? We therefore ran a second blinded study, this one pre-registered on the Open Science Framework (DOI [10.17605/OSF.IO/UYN46](https://doi.org/10.17605/OSF.IO/UYN46)) before the diagnostician received any data.

The corpus is fresh: 63 runs on new random seeds, generated in shuffled order by the same generators, because the boundary corpus of Sec. 4.2 is not blindable: its grid and seeds are public, so a human reader could decode it. Mechanism strengths bracket each measured boundary (barren plateau at widths $n = 5, 7, 8, 9, 11$; shot starvation at target precisions 0.05 to 0.45; noise domination at depths 6 to 20 on a fixed device-noise model), with three replicates per level, mechanism-free controls including the depth-one converged-trajectory traps, and the three two-mechanism mixtures. The class balance and grid were withheld from the diagnostician and committed by SHA-256; the registration was embargoed for the same reason, and the diagnostician was deliberately not given access to it. The same diagnostician served, blind to this corpus. The

comparator is the frozen rule of Sec. 3.4, applied unchanged, with its outputs hash-committed *before* she received the corpus so the comparison could not react to her answers. Her completed sheet was hash-committed before the answer key was unsealed.

The registered primary outcome is a paired, one-sided exact McNemar test [19] of her primary labels against the rule’s over the discordant runs. We chose a paired test because in the transition region the knob-defined label and the trace evidence genuinely diverge, and pairing cancels an ambiguity both graders face identically. Registered secondary outcomes are her recovery of the two rule-failure classes identified in Sec. 4.2, her confidence inside versus outside the transition regions, per-level detection curves for both graders, mixture recall, and false positives. The registration also fixed a power contingency: below six discordant pairs the exact test is underpowered and is to be reported descriptively, and a manipulation check: the achieved evidence statistics must be verified to move monotonically with each planted knob.

4 Results

4.1 Blinded diagnosis

The diagnostician attributed all 72 blinded runs to their planted failure mode correctly (Table 1): overall primary-label accuracy 1.000, with a 95% Wilson interval of [0.949, 1.000] and a one-sided exact binomial $p = 4.5 \times 10^{-44}$ against the 25% four-class chance level, the registered primary outcome. Accuracy was 1.000 on each stratum separately (simulated and hardware, 36 runs each; Wilson interval [0.904, 1.000], $p = 2.1 \times 10^{-22}$), so per-class precision, recall, and F_1 are unity throughout and the confusion matrices are diagonal. The result was scored in a recorded unsealing after the answer key was verified against its published SHA-256 commitment and the diagnosis sheet against its own.

The registered hardware prediction (H2) was confirmed exactly: all nine planted **barren_plateau** hardware runs received **noise_dominated** as the diagnostician’s *secondary* label (co-occurrence rate 1.0, 95% CI [0.701, 1.000]), while the primary label remained **barren_plateau** in every case; the diagnostician saw the barren-noise entanglement of Sec. 4.4

in the evidence and resolved it, evidently by the variance signature (a barren hardware run shows collapsed landscape variance *and* low fidelity; a noise-dominated one shows low fidelity with an unremarkable landscape).

Three exploratory observations, reported in Appendix B and not part of the registered analysis, add texture: every one of her 23 secondary labels fell on a mechanism pair the registration had identified in advance as confusable; her reported confidence was lowest exactly where mechanisms physically co-occur; and the sim-versus-hardware provenance visible in a trace’s backend metadata (Sec. 5) did not manifest as any asymmetry in accuracy between strata.

A perfect score deserves a candid reading. The corpus was constructed so that each planted mechanism is the dominant signature of its run, a ground-truth validity requirement, verified on the disjoint development corpus and, for hardware, in a pilot before the recipes were frozen, so the examination is passable by a careful reader *by design*. What the blinded protocol establishes is that passing required no tuning, leakage, or author involvement: thresholds were frozen before the test data existed, the answer key was committed before diagnosis, and the diagnostician saw only traces. The ceiling result therefore validates the evidence pipeline end to end, but it cannot, by itself, locate the instrument’s limits. Section 4.2 does exactly that: a committed mechanical rule replays this corpus without a human and is then pushed across mechanism strengths from absent to fully expressed, until detection fails. Section 4.3 then returns a blinded human reader to that boundary, on a second pre-registered corpus built to be hard, where the ceiling does not recur.

4.2 Detection boundaries and machine-only replay

The strength sweeps locate each mechanism’s detection boundary sharply, and in physically interpretable places. (These are the machine-only measurements; the same two boundaries, re-measured blind on fresh seeds in Sec. 4.3, are the ones plotted in Fig. 2.) Barren-plateau detection is a step function of circuit width: zero detections through seven qubits, five of five from eight qubits on. The step is exactly where the McClean decay puts the evidence across the frozen

Table 1: Blinded-diagnosis confusion matrices (planted truth versus the diagnostician’s primary label), reported separately per stratum as registered. Both matrices are diagonal: 72/72 correct.

planted	simulated: diagnosed as				hardware: diagnosed as			
	hea.	bar.	shot	noi.	hea.	bar.	shot	noi.
healthy	9	0	0	0	9	0	0	0
barren_plateau	0	9	0	0	0	9	0	0
shot_starved	0	0	9	0	0	0	9	0
noise_dominated	0	0	0	9	0	0	0	9

threshold: the median landscape variance moves from 0.0087 at $n = 7$ to 0.0043 at $n = 8$, straddling 0.005, and because the variance falls exponentially in n , the transition is intrinsically narrower than one qubit. Shot-starvation detection is instead graded, a textbook psychometric curve: 0/5 at target precision 0.12, 1/5, 3/5, then 5/5 at 0.28, tracking the estimated SNR through the threshold. Noise domination steps between depth twelve and sixteen on the fixed device model, precisely where the predicted fidelity crosses 0.50 ($0.52 \rightarrow 0.44$). The healthy-control arm produced 3/15 false positives, all of one kind and all informative: depth-one runs whose optimizer converged quickly and then idled, leaving a low-variance *trajectory* that the variance detector reads as a plateau. The instrument’s own convergence analyzer reports these runs as converged; a reader who integrates that evidence does not make the error, but the naive rule, which reads one verdict per mechanism, does.

The mixture arm behaved as committed in advance. Noise+shot starvation was fully recovered (both mechanisms flagged in 5/5 runs). Barren+shot starvation confirmed the pre-registered masking prediction exactly: the barren flag was raised in 0/5 runs, because a five-shot budget puts a sampling-noise floor of ~ 0.2 on the measured variance, four orders of magnitude above the planted 10^{-3} landscape signal. The instrument does not fail silently here; it reports, correctly, that the measurement budget cannot support conclusions, but no trace-based method could recover the plateau from this evidence, and we report it as the information-theoretic limit it is. Barren+noise, the algorithmic form of the H2 entanglement, raised both planted flags in 5/5 runs, plus the pre-registered SNR-depression co-flag.

Replaying the frozen rule over the original

blinded corpus separates the instrument from its human reader on identical evidence. On the simulated stratum the rule scores 35/36: the evidence is machine-separable, and the blinded result cannot be attributed to diagnostician skill alone. Overall, however, the rule scores 56/72 (flag-set top-two analogue 66/72), against the diagnostician’s 72/72, and the sixteen-point deficit concentrates exactly where mechanisms entangle. Nine misses are the hardware barren runs: the rule raises both the barren and noise flags but its fixed precedence picks noise as primary, where the diagnostician used the collapsed landscape variance to order the two (Sec. 4.1). Six are hardware shot-starved runs that device noise pushed just across the SNR flag line (estimated SNR 1.5–1.7, graded “marginal” rather than “dominated”): a borderline evidence regime encountered in the wild rather than by design. The diagnostician resolved these from the recorded shot counts and the marginal verdict; the one-verdict rule defaulted to healthy. The remaining miss is a single simulated precedence flip. On this corpus, then, the instrument’s evidence does the separating and the human’s contribution, worth sixteen runs, is the resolution of entangled and borderline cases that a fixed precedence cannot supply. That margin is corpus-dependent, and Sec. 4.3 measures it again on a corpus built to be hard, where it largely disappears.

4.3 A second blinded diagnosis, at the boundary

On the graded corpus (Sec. 3.5) the same diagnostician scored 40/63 (0.635, 95% CI [0.511, 0.743]) and the frozen rule 38/63 (0.603, [0.480, 0.715]). Neither is near the ceiling of Sec. 4.1, and unlike that study the confusion matrix is far from diagonal (Table 2): making the plantings graded

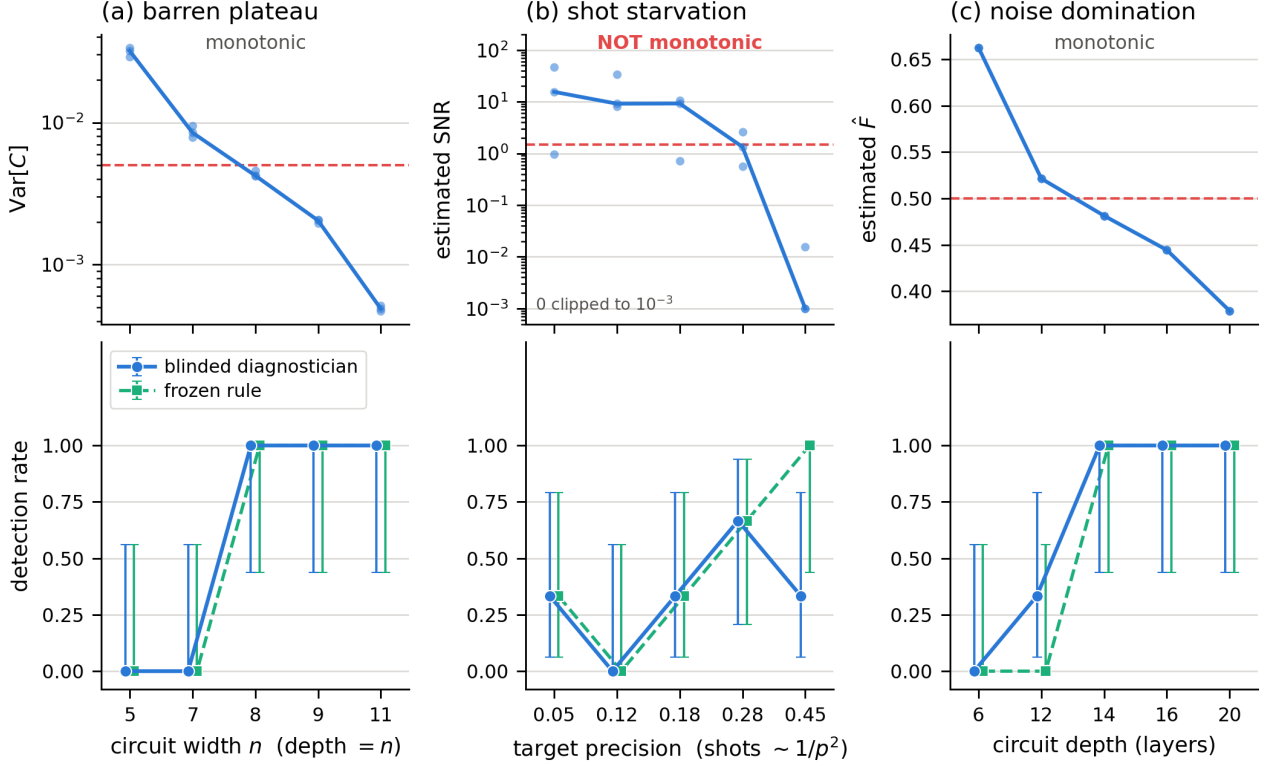


Figure 2: Detection boundaries in the second blinded study (three runs per level). *Top row, the registered manipulation check:* the evidence statistic each planted knob is supposed to move, with individual runs (dots), the per-level median (line), and the frozen decision threshold (dashed). *Bottom row:* the resulting detection rate for the blinded diagnostician and for the frozen mechanical rule, with 95% Wilson intervals. (a) Barren plateau: the landscape variance falls monotonically through the threshold and both graders switch from 0/3 to 3/3 between $n = 7$ and $n = 8$. (c) Noise domination: the estimated fidelity falls monotonically through 0.50 and both graders switch between depths 12 and 14. Both boundaries reproduce, on fresh seeds, the ones measured algorithmically in Sec. 4.2. (b) Shot starvation fails its manipulation check: the achieved SNR is not monotonic in the target precision and varies over more than an order of magnitude *within* a single level, so the planted label and the actual evidence disagree at three of five levels and neither grader’s curve is interpretable as skill (Sec. 4.3).

rather than dominant moves trace-based diagnosis from a solved problem to a hard one, which is the first thing this study establishes.

The manipulation check partitions the corpus. As registered, we verified that each planted knob moved the evidence it was supposed to move (Fig. 2, top row). For two arms it did. The barren-plateau arm’s median landscape variance falls monotonically ($0.0319 \rightarrow 0.0085 \rightarrow 0.0042 \rightarrow 0.0020 \rightarrow 0.0005$ for $n = 5, 7, 8, 9, 11$), and the noise arm’s median estimated fidelity falls monotonically ($0.663 \rightarrow 0.521 \rightarrow 0.481 \rightarrow 0.444 \rightarrow 0.379$ for depths 6 to 20). The shot-starvation arm *failed*: median SNR runs $15.50 \rightarrow 9.15 \rightarrow 9.22 \rightarrow 1.34 \rightarrow 0.00$ as the budget tightens, non-monotonic, and within the weakest-starvation level alone the achieved SNR spans

$[0.95, 46.14]$, more than an order of magnitude across three nominally identical runs. The cause is structural: SNR is an empirical variance divided by a shot floor, and along an optimizer trajectory that variance depends on the path taken, not only on the shot budget, so the precision knob does not control it. We report the arm rather than dropping it: a knob that fails to plant its mechanism is a finding about the experimental design, and concealing it would inflate both graders’ apparent difficulty.

This failure also qualifies a result reported earlier. The machine-only sweep of Sec. 4.2 drove the same precision knob and produced an orderly graded curve, but with five replicates per level rather than three, and its per-level medians happened to fall monotonically. Given the within-level spread measured here, that orderliness is

better read as the benefit of more replicates than as evidence that the knob is well behaved, and the shot-starvation curve of Sec. 4.2 should be treated with the same caution as the one here. The barren-plateau and noise knobs move their evidence statistics monotonically with tight replicate spread in *both* studies; the boundary claims of this paper rest on those two.

Two boundaries reproduce on fresh seeds.

Where the manipulation held, the boundaries are sharp and land exactly where the frozen thresholds sit. Barren-plateau detection steps from 0/3 to 3/3 between $n = 7$ and $n = 8$ for both graders, the same step measured algorithmically in Sec. 4.2 on an independently generated corpus. Noise detection steps between depths 12 and 14, where the estimated fidelity crosses 0.50; the machine-only study bracketed this between 12 and 16, so the blinded corpus tightens it. That a detection boundary replicates on new seeds, read by a diagnostician who did not know the grid, is the most direct reproducibility evidence we can offer for these thresholds.

On the failed arm, the diagnostician was right and the label was wrong. Of her fifteen shot-arm runs she called eight **healthy**. All eight are runs the instrument itself graded “Signal Clear” (seven) or “Marginal” (one); that is, she read the recorded evidence correctly and was scored incorrect against a label the run did not exhibit. Excluding this arm, accuracy rises to 36/48 (0.750) for her and 35/48 (0.729) for the rule; we report this as an exploratory sensitivity, not as the registered outcome.

Extreme shot starvation mimics a barren plateau. At the strongest starvation (~ 5 shots per estimate) the recorded expectation values collapse toward a constant, so the landscape variance falls below the plateau threshold and the trainability analyzer reports “Barren Plateau Detected”, while the measurement analyzer simultaneously reports an SNR of essentially zero. The diagnostician called two of these three runs **barren_plateau** with confidence 1.0: she read the instrument correctly, and the instrument pointed the wrong way. This is a genuine false-positive mechanism, discovered only because a blinded reader was pushed into a regime we had

not probed. It is also repairable: the disambiguating evidence is already in the trace, since an intrinsic plateau does not come with a vanishing SNR, but we report it rather than patch it, so that the instrument evaluated here remains the frozen v1.0.0 that every other result in this paper was produced with.

The registered primary outcome is null.

The paired comparison found only two discordant runs, both favouring the diagnostician ($b = 2$, $c = 0$), which is below the registered underpower threshold of six; per the analysis plan we report the discordance descriptively and headline no p -value. The hypothesis that a human reader outperforms the frozen rule is *not* supported by this study. The reason is visible in the raw agreement: the two graders assigned an identical primary label on 57 of 63 runs (90.5%). Both were reading the same analyzer verdicts, and the diagnostician’s own description of her procedure, submitted with her sheet and archived, is close in structure to the mechanical rule, so with a single rater the discordance had nowhere to accumulate. The two runs that did discriminate are exactly the failure classes predicted in Sec. 4.2: a depth-one converged trajectory the rule called barren, and a strongly starved run the rule also called barren.

What the human did add. Three registered secondary outcomes came out in her favour. She made no false positives on the mechanism-free runs (0/9 versus the rule’s 1/9), including 3/3 on the converged-trajectory traps built to fool a variance-only reading. And her reported confidence was markedly lower inside the transition regions than on the strongly expressed anchors (0.728, $n = 24$, versus 0.904, $n = 9$; one-sided Mann–Whitney [20] $U = 42.0$, $p = 0.0037$, surviving Holm correction [21] across the secondary tests). Nobody told her which runs sat near a boundary; she located them from the evidence. On the mixtures her flag sets matched the rule’s exactly, including the barren+shot case where the plateau is masked beyond recovery, as predicted in advance.

4.4 The noise-displaced frontier

The expressibility–trainability frontier is a provable tension: deeper circuits are more expres-

Table 2: Second blinded study: planted mechanism versus the diagnostician’s primary label (63 runs). Unlike the ceiling corpus, this matrix is not diagonal. Nearly all misses fall in the healthy column and concentrate at the weak end of each sweep, where the planted mechanism was deliberately dialled toward absent, or, in the shot arm, never took at all.

planted	diagnosed as			
	hea.	bar.	shot	noi.
healthy (9)	9	0	0	0
barren plateau (15)	6	9	0	0
shot starvation (15)	8	3	4	0
noise domination (15)	0	0	6	9
barren + noise (3)	0	3	0	0
barren + shot (3)	0	0	3	0
noise + shot (3)	0	0	0	3

sive but have flatter landscapes [15]. What that theory does not say is how a real device *moves* the frontier. Using the instrument’s active mode, we measured both axes on the superconducting device `ibm_marrakesh` and against a matched noiseless simulator driven by the same parameter draws, as a function of circuit depth (Fig. 3).

On the trainability axis, the recorded landscape variance on hardware lies below the ideal at every depth, and the gap grows with depth [Fig. 3(a,b)]: on `ibm_marrakesh` the hardware-to-ideal variance ratio falls from 0.93 at depth one, through 0.73 at depth twelve, to roughly 0.50 across depths fourteen to eighteen (100 random parameter points per depth). In this deep regime the hardware bootstrap intervals lie *entirely* below the matched ideal (at depth sixteen, [0.025, 0.038] against 0.063), turning the marginal suppression seen at depth twelve into an unambiguous one. An independent run on a second 156-qubit device, `ibm_fez`, reproduces the effect over depths one to twelve, its ratio falling from 0.88 to 0.50. Noise is therefore suppressing trainability *beyond* the intrinsic plateau, consistently across two devices, the signature of a noise-induced barren plateau [7]. On the expressibility axis [Fig. 3(c)], the measured Kullback–Leibler divergence from Haar is shifted under noise: at depth six the hardware value (0.066) exceeds the matched ideal (0.029) by +0.037, with bootstrap intervals separating, so a noisy circuit is mis-measured as *less* expressive than it truly is. Taken together, hardware noise displaces the frontier

point, increasingly with depth, a direct measurement of an effect that had been predicted but not quantified on real devices.

4.5 Recovery of known theory

Two checks confirm that the instrument’s readings track quantitative theory. First, analysing 500-point simulated landscapes blind from their traces, the recorded cost variance decays exponentially with qubit count exactly as the McClean scaling law predicts [4] (Fig. 4): a least-squares fit of $\log_2(\text{Var})$ against n recovers a scaling exponent of -1.00 for circuits of depth n and -0.97 for depth $n/2$, while single-layer circuits show no decay (slope -0.01): shallow circuits do not plateau, as theory requires. The detector flags the plateau precisely where it is planted (at $n \geq 8$ for the depth- n rule), with bootstrap intervals that do not straddle the threshold.

Second, the hardware-noise analyzer predicts, from nothing but a trace’s recorded calibration snapshot, how much of a computation should survive the device that produced it. Across 25 runs on a calibration-snapshotted noisy device model spanning depths one to fourteen, the predicted fidelity tracks the observed attenuation along the identity line with linear correlation $r = 0.72$ ($R^2 = 0.45$ against $y = x$, RMSE = 0.135; five runs whose ideal expectation was near zero are excluded because the attenuation ratio is undefined there). On real hardware, we repeated the prediction at seven parameter points: six single-parameter runs on `ibm_marrakesh` spanning depths one to ten, plus a deeper sixteen-layer point, together covering predicted fidelities from 0.52 to 0.94. These track the observed attenuation along the same identity line, with RMSE = 0.123, indistinguishable from the simulator arm’s 0.135. Because each hardware point is a single parameter draw rather than a variance over many, its observation noise is large (one depth-ten run overshoots its prediction by 0.29), so for the hardware arm we report agreement magnitude rather than a seven-point correlation. The deepest point is representative: the estimator predicted a circuit fidelity of 0.518 where the device delivered 0.559, within 8%, unmitigated (Fig. 5). The prediction is a first-order product formula over the errors the device reports, so the residual scatter is itself informative: it measures the error mechanisms the calibration omits

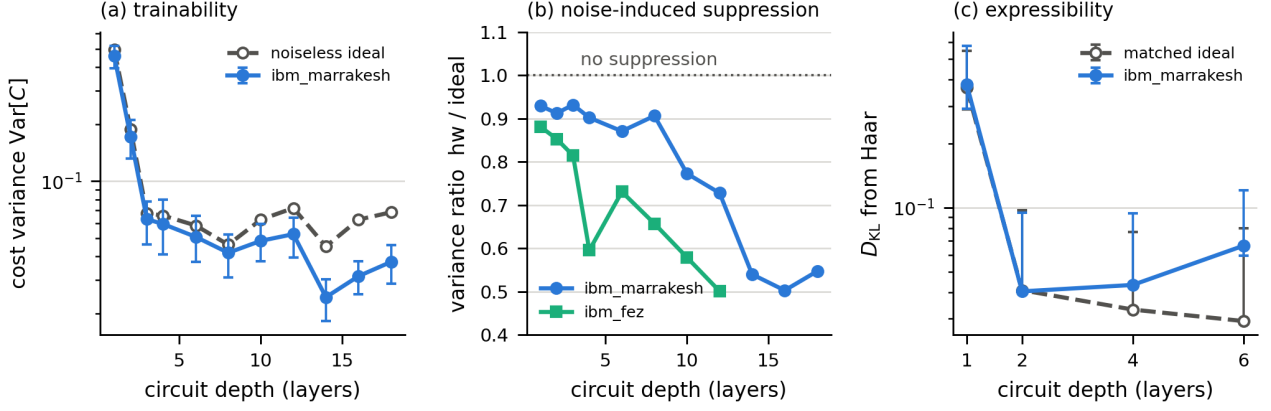


Figure 3: The noise-displaced frontier. (a) Cost-landscape variance versus circuit depth on `ibm_marrakesh` (100 random parameter points per depth, bootstrap 95% intervals) against a noiseless simulator driven by the same parameter draws. (b) Hardware-to-ideal variance ratio on two independent devices: `ibm_marrakesh` and `ibm_fez`. The ratio falls with depth on both: noise-induced suppression of trainability beyond the intrinsic plateau. (c) Measured expressibility (Kullback–Leibler divergence from the Haar distribution, 500 compute–uncompute pairs per depth) on hardware versus the matched ideal; under noise the divergence is inflated at depth, mismeasuring the circuit as less expressive.

(Sec. 5).

5 Discussion

Barren plateaus and noise are diagnostically entangled on NISQ hardware. A consequence of Sec. 4.4 is that, on a real device, an intrinsic barren plateau cannot be cleanly separated from noise domination: the circuit depth needed to make a landscape barren is precisely the depth at which fidelity collapses. This is not a shortcoming of the instrument but a property of the regime, and it has a concrete implication for diagnosis: the two failure modes overlap in the evidence a hardware trace can provide. It is why we establish barren-plateau ground truth only in simulation and treat the hardware barren runs as a registered co-occurrence case. The blinded study then confirmed the prediction quantitatively: a diagnostician who knew nothing of the corpus design flagged noise as co-occurring on all nine planted hardware barren runs, while still attributing each primarily to the barren mechanism from its collapsed landscape variance (Sec. 4.1). The entanglement is real, measurable, and, with the right evidence, resolvable.

Where a human reader actually helps.

Across the two blinded studies the division of labour proved narrower, and more specific, than we expected: an advantage of sixteen runs on the

ceiling corpus collapsed to two on the graded one, as a careful reader and a mechanical rule converged on the same evidence. What did *not* collapse is her calibration and her refusal to false-positive on healthy runs. The useful reading is therefore not that human judgement beats automation on trace-based diagnosis, but that it supplies something automation did not: a reliable signal about *when the evidence is thin*. For a practitioner deciding whether to spend more quantum time, that signal is arguably worth more than a marginal accuracy gain.

A design lesson we would apply differently.

The second study’s stated purpose was to locate the boundary at which the instrument’s readings stop being conclusive, an estimation goal. Its registration nonetheless made a human-versus-rule comparison the primary outcome and recorded boundary mapping as a descriptive secondary. Given that both graders read the same verdicts, low discordance was foreseeable; indeed we pre-specified the underpower contingency that the result then triggered, which is itself evidence that it was foreseeable. A better registration would have made the boundary location the primary estimand and the grader comparison secondary. We report the outcome as registered rather than promoting the descriptive analysis after the fact, because retrospective outcome-switching is precisely what pre-registration exists to prevent, but

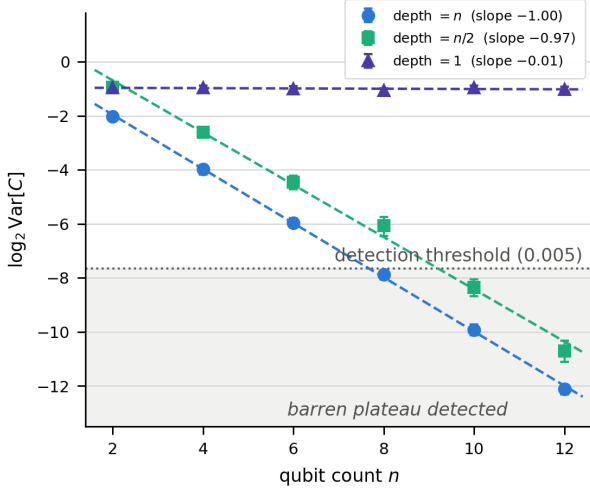


Figure 4: Recovery of the McClean scaling law from trace evidence alone. Cost-landscape variance versus qubit count for the hardware-efficient ansatz at three depth rules (500 random points per landscape; bootstrap 95% intervals smaller than the markers where not visible). Depth-scaled circuits decay exponentially with fitted exponents -1.00 and -0.97 ; single-layer circuits do not decay (-0.01). The detector’s verdict flips to “barren plateau detected” exactly where the planted decay crosses the frozen threshold (shaded region).

we flag the mismatch plainly, since the study’s scientific value lies mostly in the boundaries it reproduced and in the false-positive mechanism it exposed, not in its nominal primary test.

Blinding finds what authors do not think to look for. The shot-starvation false positive of Sec. 4.3 (a five-shot budget collapsing the landscape variance until the trainability analyzer reports a barren plateau) survived a 319-test suite, a development corpus, and a first blinded study. The test suite is not at fault, and that is the point: it enforces architectural invariants (Sec. 2): that recording does not perturb, that traces are tamper-evident, that analyzers are pure functions of the evidence. Every one of those held here. No amount of such testing can establish that a threshold applied to faithfully recorded evidence yields the right *interpretation* in a regime the authors never probed. That is a semantic question about the instrument’s readings, and it surfaced only when an independent reader was pushed into such a regime and reported the wrong mechanism with full confidence, because the instrument told her so. The failure is repairable from evidence already in the trace (an intrinsic plateau does

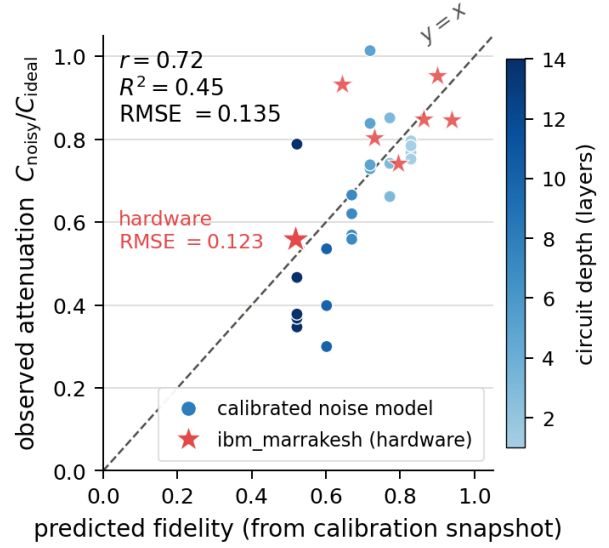


Figure 5: Predicted versus observed fidelity. Each point is one run: fidelity predicted by the noise analyzer from the trace’s calibration snapshot (abscissa) against the attenuation the run actually suffered (ordinate). Circles are the 25-run simulator arm, coloured by circuit depth. Red stars are seven real-hardware runs on *ibm_marrakesh*, which scatter around the identity line with $\text{RMSE} = 0.123$; the single-parameter estimates carry observation noise (visible at the two points furthest from the diagonal). The large star is the deep point: predicted 0.518, observed 0.559.

not arrive with a vanishing SNR), and we leave it unrepaired here so that the version evaluated throughout this paper stays frozen. We record it as the clearest practical argument for the methodology: an adversarial, blinded protocol is not only a way to certify an instrument, it is a way to *debug* one.

Measured expressibility can be misleading under noise. Section 4.4 shows that the Haar divergence measured on hardware is shifted from its ideal value, so a circuit’s apparent expressibility on a real device is a noise-contaminated quantity. Reading expressibility off hardware without accounting for this shift can misplace a circuit on the frontier.

Limitations. The hardware-noise analyzer uses a product-formula fidelity estimate built from the errors a device *reports*; systematic residuals against observation therefore measure error mechanisms the calibration omits (for example coherent errors), and the estimate

should be read as a first-order predictor rather than an exact fidelity. The analyzer thresholds are practical heuristics rather than universal constants; we freeze and pre-register them and report their sensitivity (Appendix A), but they are calibrated for small circuits; the detection boundaries of Sec. 4.2 are measured for this circuit family and should be re-mapped before the thresholds are trusted at other scales. Both blinded studies used the same single diagnostician, who is a co-author, with the corpus custodian also the lead author; the commitments, registration timestamps, and unsealing records make the protocol auditable rather than trust-based (Sec. 3.2), but a cross-laboratory replication with multiple independent raters is the natural strengthening, and both corpora are public precisely to enable it. A single rater also bounds what the human-versus-rule comparison can show: the null result of Sec. 4.3 means this reader did not beat the frozen rule on this corpus, not that human judgement generally cannot. Finally, the shot-starvation arm of the second corpus failed its registered manipulation check, so that arm measures neither grader’s skill; the boundary claims rest on the two arms that passed. The hardware results use a small number of devices, so device-to-device generalization beyond the two tested is out of scope. Finally, because a sealed trace records the backend it ran on, a diagnostician can in principle infer whether a run is simulated or executed on hardware; this does not reveal the failure class, and we report simulated and hardware results separately, but we disclose it as a residual channel that a fully device-blind protocol would need to close.

Open problems. Three follow from this work directly. *Adversarial examinations:* we have now graded diagnosis on dominant plantings, on graded and mixed ones, and at the detection boundary itself. The remaining escalation is adversarial: failure modes planted specifically to *mimic* one another’s trace signatures, of which the shot-starvation-as-plateau confusion found in Sec. 4.3 is a naturally occurring instance. Because the same reader served both blinded studies, the complementary question is a multi-rater, cross-laboratory replication: our null human-versus-rule result bounds one reader on one corpus, not human judgement in general. *The noise–*

expressibility interaction deserves its own focused study: the shift we measure at depth is a bias on a widely used metric, and characterizing it across devices and ansatz families would make expressibility readings noise-correctable. *A community trace standard:* the sealed, schema-first trace format is instrument-independent by construction, and adopting a shared format across QML software would make blinded, cross-tool diagnostic studies routine rather than exceptional.

6 Conclusion

An instrument is trustworthy when a stranger can rely on its readings. We built a diagnostic framework whose evidence is recorded non-intrusively and sealed cryptographically, validated it the way clinical instruments are validated (pre-registered protocol, committed answer key, independent blinded diagnostician) and it passed at ceiling: 72 of 72 planted failure modes correctly attributed across simulation and IBM hardware, with the one physically entangled pair (barren plateaus and noise on NISQ devices) predicted in advance, observed in the diagnostician’s secondary labels, and resolved by the recorded evidence. We then measured where that reliability ends. A mechanical rule committed before execution reproduces the blinded separation without a human; a strength-swept corpus locates the boundary at which each mechanism’s trace evidence stops supporting detection; and a second pre-registered blinded study, run at that boundary, reproduced two of those boundaries on fresh seeds, dropped both graders to around 0.6 accuracy, and turned up a false-positive mechanism (extreme shot starvation masquerading as a barren plateau) that a large test suite and a ceiling-scoring blinded study had both missed. Its registered human-versus-rule comparison came out null, which we report as registered. Inside the boundaries so measured, the instrument earned the trust to report physics: it reproduced the McClean scaling law from traces alone, forecast hardware fidelity from recorded calibration snapshots, and measured, on two devices, how noise displaces the expressibility–trainability frontier. The instrument, every trace, the pre-registrations, and both corpora with their answer keys are public; the next examinations, harder ones, can be set by anyone.

Data and code availability

The instrument is open source (<https://github.com/mamuncseru/hilbertbench>; MIT license), with version 1.0.0, the frozen version used throughout this work, archived at Zenodo, DOI [10.5281/zenodo.21009245](https://doi.org/10.5281/zenodo.21009245). The two pre-registrations are on the Open Science Framework: the ceiling study at DOI [10.17605/OSF.IO/2PWHU](https://doi.org/10.17605/OSF.IO/2PWHU), and the boundary study of Sec. 4.3 at DOI [10.17605/OSF.IO/UYN46](https://doi.org/10.17605/OSF.IO/UYN46) (filed, embargoed, before the diagnostician received that corpus; the embargo kept the withheld corpus composition out of her reach and is lifted on publication). The complete study dataset (the blinded corpus of 72 sealed traces, the answer key with its pre-published commitment, the diagnostician’s frozen sheet with hers, the scoring output, the ground-truth manifest, and the unsealing session log) is archived at Zenodo (CC BY 4.0), DOI [10.5281/zenodo.21269900](https://doi.org/10.5281/zenodo.21269900), with a step-by-step recipe for independently re-verifying every commitment and re-scoring the study. The detection-boundary study (Sec. 4.2) is fully specified by a protocol document whose SHA-256 commitment was published to the repository before execution; the corpus is deterministically regenerable from the committed grids and seeds, and the frozen decision rule is fifteen lines of code any referee can rerun. The second blinded study (Sec. 4.3) is archived separately at Zenodo (CC BY 4.0), DOI [10.5281/zenodo.21482518](https://doi.org/10.5281/zenodo.21482518), containing its 63 blinded traces, the unsealed assignment, answer key, frozen-rule outputs and diagnosis sheet, the four SHA-256 commitments, the unsealing record, and a verification recipe; re-scoring from that archive reproduces the reported numbers exactly. All experiment generators and figure scripts are in the instrument repository; every figure in this paper is regenerable from the archived result files.

Author contributions

M.A.A.M. conceived the instrument and the study, developed the software, ran the experiments, served as Researcher A (corpus generation and answer-key custody), analysed the results, and wrote the manuscript. S.A.P. served as Researcher B in both blinded studies: her

sole contribution is the blinded diagnosis: she did not develop the instrument, design either corpus, or access either registration before submitting her frozen diagnosis sheets. For the second study she was deliberately excluded from the registration’s contributor list so that its embargoed contents, which specify the corpus composition, could not reach her. J.P. supervised the research and reviewed the protocol and manuscript. AI/LLM-use disclosure: AI assistance (Anthropic’s Claude) was used for software development support and manuscript drafting; all experiments, scientific claims, and results were designed, executed, and verified by the authors.

Acknowledgments

[XX: funding, IBM Quantum access, and colleague acknowledgments.]

References

- [1] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, “Quantum machine learning,” *Nature* **549**, 195–202 (2017).
- [2] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles, “Variational quantum algorithms,” *Nature Reviews Physics* **3**, 625–644 (2021).
- [3] K. Bharti, A. Cervera-Liarta, T. H. Kyaw, T. Haug, S. Alperin-Lea, A. Anand, M. Degroote, H. Heimonen, J. S. Kottmann, T. Menke, W.-K. Mok, S. Sim, L.-C. Kwek, and A. Aspuru-Guzik, “Noisy intermediate-scale quantum algorithms,” *Reviews of Modern Physics* **94**, 015004 (2022).
- [4] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, “Barren plateaus in quantum neural network training landscapes,” *Nature Communications* **9**, 4812 (2018).
- [5] M. Cerezo, A. Sone, T. Volkoff, L. Cincio, and P. J. Coles, “Cost function dependent barren plateaus in shallow parametrized quantum circuits,” *Nature Communications* **12**, 1791 (2021).

- [6] M. Larocca, S. Thanasilp, S. Wang, K. Sharma, J. Biamonte, P. J. Coles, L. Cincio, J. R. McClean, Z. Holmes, and M. Cerezo, “Barren plateaus in variational quantum computing,” *Nature Reviews Physics* **7**, 174–189 (2025).
- [7] S. Wang, E. Fontana, M. Cerezo, K. Sharma, A. Sone, L. Cincio, and P. J. Coles, “Noise-induced barren plateaus in variational quantum algorithms,” *Nature Communications* **12**, 6961 (2021).
- [8] J. Preskill, “Quantum computing in the NISQ era and beyond,” *Quantum* **2**, 79 (2018).
- [9] M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, “Accelerating the machine learning lifecycle with MLflow,” *IEEE Data Engineering Bulletin* **41**(4), 39–45 (2018).
- [10] R. LaRose, A. Mari, S. Kaiser, P. J. Karalekas, A. A. Alves, P. Czarnik, M. El Mandouh, M. H. Gordon, Y. Hindy, A. Robertson, P. Thakre, M. Wahl, D. Samuel, R. Mistry, M. Tremblay, N. Gardner, N. T. Steffen, N. Shammah, and W. J. Zeng, “Mitiq: A software package for error mitigation on noisy quantum computers,” *Quantum* **6**, 774 (2022).
- [11] N. Kanazawa, D. J. Egger, Y. Ben-Haim, H. Zhang, W. E. Shanks, G. Aleksandrowicz, and C. J. Wood, “Qiskit Experiments: A Python package to characterize and calibrate quantum computers,” *Journal of Open Source Software* **8**(84), 5329 (2023).
- [12] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, B. R. Johnson, and J. M. Gambetta, “Quantum computing with Qiskit,” arXiv:2405.08810 (2024).
- [13] V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, S. Ahmed, V. Ajith, M. S. Alam, G. Alonso-Linaje, B. AkashNarayanan, A. Asadi, *et al.*, “PennyLane: automatic differentiation of hybrid quantum-classical computations,” arXiv:1811.04968 (2018).
- [14] S. Sim, P. D. Johnson, and A. Aspuru-Guzik, “Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms,” *Advanced Quantum Technologies* **2**, 1900070 (2019).
- [15] Z. Holmes, K. Sharma, M. Cerezo, and P. J. Coles, “Connecting ansatz expressibility to gradient magnitudes and barren plateaus,” *PRX Quantum* **3**, 010313 (2022).
- [16] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, “The preregistration revolution,” *Proceedings of the National Academy of Sciences* **115**, 2600–2606 (2018).
- [17] E. B. Wilson, “Probable inference, the law of succession, and statistical inference,” *Journal of the American Statistical Association* **22**, 209–212 (1927).
- [18] C. J. Clopper and E. S. Pearson, “The use of confidence or fiducial limits illustrated in the case of the binomial,” *Biometrika* **26**, 404–413 (1934).
- [19] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika* **12**, 153–157 (1947).
- [20] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *Annals of Mathematical Statistics* **18**, 50–60 (1947).
- [21] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics* **6**, 65–70 (1979).

A Threshold sensitivity

Before freezing the analyzer thresholds we verified, on the disjoint twelve-run development corpus, that each class verdict is stable around its default; that is, that the defaults are not finely tuned to a knife edge. For the barren-plateau threshold, the planted barren runs cluster near a cost variance of ≈ 0.001 while every other run sits above ≈ 0.0095 , leaving a clean separating gap; the default of 0.005 falls near the centre of that gap, and any threshold in the band $[0.002, 0.009]$ classifies the development corpus without error. For the shot-noise threshold, the shot-starved runs sit at signal-to-noise ratios below ≈ 0.21 while clean runs exceed ≈ 0.95 ; the default of 1.5 lies above this gap and so is conservative,

additionally flagging genuinely low-SNR noise-dominated runs, the shot-noise overlap noted in Sec. 5, which the fidelity estimate rather than the SNR is designed to resolve. The fidelity threshold cannot be exercised on the development corpus, whose non-hardware runs carry no calibration snapshot and hence no fidelity value; its validation is the hardware predicted-versus-observed agreement of Sec. 4.5. This analysis uses development data only, never the test corpus.

B Exploratory observations from the first blinded study

Figure 6 shows two exploratory readings of the first blinded corpus, referred to in Sec. 4.1. Neither was part of the registered analysis. Panel (a) records which mechanism pairs the diagnostician’s secondary labels connected: all 23 fell on pairs the registration had named in advance as confusable, and none connected a non-confusable pair. Panel (b) shows her reported confidence per class: lowest on the entangled hardware barren runs (mean 0.72) and the overlapping simulated noise runs (0.71), highest on the clean simulated classes (0.95; corpus-wide mean 0.81).

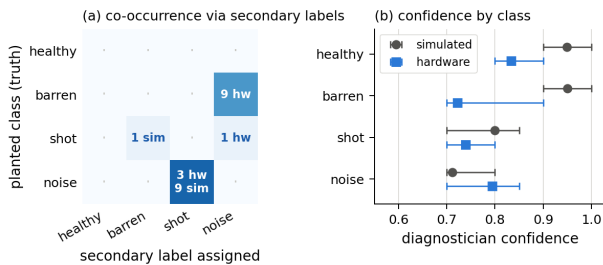


Figure 6: What the blinded diagnostician saw (exploratory). (a) Secondary labels by planted class: every one of the 23 secondary labels falls on a mechanism pair the registration identified in advance as confusable: all nine hardware barren runs flagged as noise-co-occurring (H2), and the simulated noise runs flagged for their shot-noise overlap; dots mark empty cells. (b) The diagnostician’s reported confidence per class (mean and range): lowest exactly where mechanisms physically co-occur, highest on the clean simulated classes.