

Supplementary information

Online Resource 1 — Discover Quantum Science

Shot saturation and evidence limits in quantum-AI hardware benchmarks

Vikram Lex

KarLex AI, Newark, DE, USA

vikram@karlexai.com

S1 Purpose and relationship to the main paper

This supplement reports the numerical record of the earlier QHM campaign [1] and the supporting analyses for the main paper. The main paper uses archived overlap moments, observed kernel entries, and reported QAOA cut scalars. Additional historical learning, simulation, and GPU records are reported here at the level the available evidence supports. Their inclusion does not reinstate interpretations excluded by the source audit.

The supplementary data contain 66 historical result summaries, source files, and logs covering 22 studies. The study registry below identifies their scope and limitations. Recorded configurations distinguish experiments with similar names; incomplete execution records and missing model checkpoints limit which comparisons can be validated.

The reanalysis includes arithmetic checks, ideal-statevector circuit checks, and exhaustive MAX-CUT enumeration for four graphs of at most 20 vertices. Earlier GPU training and quantum-hardware measurements are analyzed from stored results.

S2 Signed-overlap statistics and circuit checks

The eight full-task means each correspond to 10240 nominal shots. The integer ancilla-zero counts reconstructed from those means are, in pair order $(0, 0)$, $(0, 1)$, $(1, 2)$, $(2, 3)$:

$$(8069, 7818, 8649, 6670) \quad \text{and} \quad (7958, 7859, 8456, 6531). \quad (\text{S1})$$

This reconstruction recovers the total number of zeros, not their order, the identity of provider sub-batches, or discarded outcomes. The conditional Clopper–Pearson analysis in the main text treats each task as iid Bernoulli sampling. The saved moments do not establish that assumption.

At each shot setting, the archived standard deviation uses divisor five, not four. Consequently the exact empirical squared-error reconstruction is the squared error of the five-batch mean plus the stored variance. Replacing this variance with an unbiased sample-variance convention would change the identity. At 256, 512, 1024, and 2048 shots per batch, the squared-mean-offset shares are 84.0%, 90.4%, 94.6%, and 98.2%, respectively. They include finite-sample error in the batch mean and are not estimates of a universally stable physical bias fraction.

The available preparation routine is checked only on the relevant inputs. Its Gram–Schmidt completion is not promoted as a robust general-purpose state-preparation algorithm. The local checker reads the two pure numerical definitions from the frozen source and constructs the documented controlled unitary independently of the cloud SDK. The four ideal expectations agree with both the untranspiled and locally transpiled circuit to within 2×10^{-11} . Local CX counts are 10, 10, 8, and 10 with the supplied Qiskit version and transpiler seed; these are new local compilation results, not historical native-gate counts.

The incorrectly controlled older circuit gives approximately 0.0125, -0.0912 , -0.4707 , and 0.5443 on the same vector pairs, instead of the intended 0.7861, 0.6193, 0.6180, and 0.3637. This is a circuit-definition difference rather than a finite-shot effect. A separate antiparallel-vector check yields signed overlap -1 while its squared overlap is $+1$, ensuring the two observables are not conflated in the analysis.

S3 Kernel completion and the input audit

The IonQ success list, rather than the numerical value of a matrix cell, determines whether an entry was measured. A successful measurement of zero must remain an observed zero. An unexecuted cell is missing even if its placeholder is also zero. The analysis rejects duplicate pair identifiers, malformed indices, and nonfinite observed probabilities; it excludes the imposed diagonal from measurement counts.

The successful upper-triangular pairs are retained verbatim in the frozen input JSON. The corresponding 18 Rigetti matrix positions have mean 0.0813, but matching matrix indices do not establish matching physical inputs. The available Rigetti generator selects dataset indices

$$(48, 434, 49, 367, 247, 244, 395, 53, 114, 485) \tag{S2}$$

with NumPy seed 42. Their label sequence matches the saved IonQ labels. In contrast, the available IonQ generator takes the first ten records, all of which have label zero. Labels do not uniquely determine examples, and neither archive stores the fitted feature vectors required to resolve this inconsistency. The main paper therefore makes no provider ranking, even after restricting to common matrix positions.

The completion range $[0.0888, 0.6888]$ fixes the observed point estimates and permits every missing entry to vary independently in $[0, 1]$. It deliberately makes no statistical missingness assumption. Constraints such as positive semidefiniteness could narrow a mathematical completion set, but are not established for the noisy, partially recorded numerical matrix and were not imposed here. No sampling confidence should be attached to this range.

S4 QAOA normalization and simulator checks

Table S1 converts the same stored IonQ mean-cut scalars to edge fractions. It is an arithmetic audit. The original outcome histograms and provider bit mappings remain unavailable, so these numbers are not a corrected physical-device comparison. Their random-bit references in the original ratio units are 0.6000, 0.5625, 0.5455, 0.5769, and 0.5921, respectively. These references are graph-specific.

Table S1: Denominator audit of the IonQ size-sweep scalars. A stored optimum denominator is a field in the historical file; this table does not independently re-score the physical outcomes. Historical data: Ref. [1]

n	$ E $	Stored denominator	Stored ratio	Same cut / $ E $
8	12	10	0.6154	0.5128
12	18	16	0.5857	0.5207
16	24	22	0.5580	0.5115
20	30	26	0.5962	0.5167
30	45	38	0.6077	0.5132

The four smaller simulator graphs have complete edge lists. We independently enumerate all 2^n assignments for $n = 8, 12, 16, 20$, confirming the archived optima 10, 16, 22, and 26. Table S2 preserves selected expected QAOA cuts and the recorded cuts returned by simulated annealing (SA) and the Goemans–Williamson (GW) implementation. This compares different output statistics in common cost units; it is not an equal-budget runtime comparison. The optimized QAOA simulations were not rerun in this reanalysis. Their saved parameters, costs, and source remain available.

Table S2: Earlier statevector QAOA results in cut units, with newly verified graph optima. QAOA entries are expectations; SA and GW entries are stored returned cuts. No best-case QAOA sample is substituted for its expectation. Historical data: Ref. [1]

n	Verified OPT	QAOA $p = 3$	QAOA $p = 5$	SA cut	GW cut
8	10	9.6582	9.8322	10	10
12	16	14.2294	14.5308	16	16
16	22	19.6039	20.1507	22	21
20	26	23.7550	24.5709	26	26

The larger GPU log reports a 24-vertex, 36-edge graph and a simulated-annealing cut of 31. It records four completed depths before stopping during $p = 5$ (Table S3). The JSON does not contain those completed QAOA records. We retain the rounded expected cuts as log evidence and explicitly label the original cut/31 column as a heuristic-baseline ratio. It is not an approximation ratio to a certified optimum. The previous audit reported an optimum of 32, but this supplement’s fresh exhaustive checks are restricted to the four fully specified graphs of at most 20 vertices. No missing depth or additional precision is inferred.

Table S3: Rounded 24-vertex QAOA log entries. The third column reproduces the recorded heuristic normalization; it is not labelled cut/OPT. Historical data: Ref. [1]

Depth	Rounded expected cut	Stored cut/31	Stored time (s)
1	24.76	0.7988	35.3
2	27.08	0.8735	352.6
3	28.06	0.9053	865.2
4	28.40	0.9161	1118.1

S5 Earlier GPU learning runs

The recorded configurations for the larger Fashion-MNIST and AG-News experiments specify 10000 training examples, 2000 reported test examples, 80 epochs, ten qubits, and six circuit layers. Separate preliminary runs used 100/50 or 200/50 training/test examples and only two or three epochs. Tables S4 and S5 report the larger 80-epoch experiments.

The model names are historical labels for compact implementations. They do not identify reproductions of the full published DeltaNet, GLA, Mamba, or RWKV architectures. The stored label **Dequantized** is rendered as *Fourier-feature surrogate*; it denotes a classical control, not an efficient simulation theorem. The basic hybrid designation follows the historical record. Execution hardware cannot be established from the model label alone.

Table S4: Recorded Fashion-MNIST learning metrics from the larger archive. These are exploratory single-run scores, not fresh training results or independent-test-set estimates. Historical data: Ref. [1]

Stored model label	Accuracy	F1	Time (s)
Classical	0.9940	0.9940	31.78
Basic VQC hybrid	0.9820	0.9821	862.99
Fourier-feature surrogate	0.9935	0.9935	38.81
Transformer hybrid	0.9880	0.9880	933.23
DeltaNet hybrid	0.9815	0.9814	962.87
Gated DeltaNet hybrid	0.9805	0.9806	1013.60
GLA hybrid	0.9780	0.9779	956.21
Mamba-3 hybrid	0.9870	0.9869	979.73
RWKV-7 hybrid	0.9840	0.9841	1437.10

Table S5: Recorded AG-News learning metrics from the larger archive, with the same limitations. The text preprocessing cannot be uniquely authenticated from the available alternative scripts. Historical data: Ref. [1]

Stored model label	Accuracy	F1	Time (s)
Classical	0.6570	0.6525	28.19
Basic VQC hybrid	0.6445	0.6352	355.00
Fourier-feature surrogate	0.6605	0.6689	31.49
Transformer hybrid	0.6865	0.6786	371.33
DeltaNet hybrid	0.5975	0.6295	380.45
Gated DeltaNet hybrid	0.5710	0.5484	430.69
GLA hybrid	0.6185	0.6217	372.16
Mamba-3 hybrid	0.5545	0.5862	395.74
RWKV-7 hybrid	0.6200	0.6021	431.20

The available pipeline monitors the reported test data during training, and its text path can re-split combined official train and test examples after preprocessing. Another script uses cached scores and a different text representation. The full per-model configurations, exact checkpoints, prediction vectors, and run-bound source revisions needed to disambiguate these records are not supplied. The tables therefore cannot establish a generalization

improvement, statistical significance, or a fair comparison of full-scale architectures. The stored training times are descriptive measurements from the earlier environment and were not reproduced here.

Two extra entries labelled IonQ and H2-1E in the Fashion-MNIST summary have no corresponding per-example provider outputs in the supplied archive. They are retained in the source file but excluded from these local-training tables and from claims of successful hardware deployment. Likewise, sequential MNIST uses a nearly saturated binary task and combines cached and newly trained records; its output remains in the archive but is not used for an architecture-ranking claim. Ablation figures without matching numerical records are not transcribed into invented tables.

S6 Larger Grover simulations and synthetic noise

Table S6 retains the stored up-to-30-qubit Grover simulation. For a single marked item in $N = 2^n$ possibilities and t iterations, the ideal supplied-oracle success probability is

$$P_t = \sin^2((2t + 1) \arcsin(N^{-1/2})) . \quad (\text{S3})$$

The recorded probabilities agree with this scalar expression to within 10^{-7} . This is a new local formula check, not a rerun of the large statevectors. The algorithm is established; these records preserve computational work without presenting the ideal query ratio as a measured quantum speedup. The stored routine applies an oracle sign vector and a mean-based diffuser, omitting reversible predicate construction. The times are earlier classical simulation times.

Table S6: Archived larger Grover simulation outputs. Probabilities are rounded here; the JSON preserves full precision. Time is classical simulation time, not QPU latency. Historical data: Ref. [1]

Qubits	Oracle iterations	Stored success probability	Stored time (s)
10	25	0.999461245	0.44
15	142	0.999986830	0.05
20	804	0.999999757	0.36
22	1608	1.000000000	0.80
25	4549	1.000000000	11.71
28	12867	0.999999997	254.42
30	25735	0.999999999	2032.19

Other retained files include SV1 Grover measurements and local density-matrix noise sweeps. The former are cloud-simulator outputs. The latter use an explicitly synthetic noise model and must not be relabelled as QPU measurements, current device calibrations, or evidence for a universal unit-constant noise law. Neither category is pooled with the 114 core hardware submissions counted in the main paper.

S7 Registry of the 22 earlier studies

Table S7 summarizes the scope and evidential role of each study. *Registry context* denotes observations retained for methodological context that do not support a new quantitative claim. Studies excluded from the analysis are listed with the limitations that prevent their quantitative use. These assessments follow the source audit and the checks described in the main text and above.

Table S7: Retained scope and limitations of the 22 original study identifiers. Historical data: Ref. [1]

ID	Study and role	Supported use and limits
1	Quantum-kernel classification. <i>Supplementary descriptive record</i>	Small simulator classification results and classical comparator scores are preserved. Selected small datasets and splits; no claim of a general quantum advantage. Stored Dequantized labels do not establish a complexity-theoretic dequantization theorem.
2	Simulated QAOA, 8–20 vertices. <i>Supplementary descriptive record</i>	Stored cuts are compared in common units; exact graph optima are independently enumerated in the reanalysis. One graph per size and selected optimized parameters. Crossing 0.878 is not outperforming the actual Goemans–Williamson solver.
3	Overlap state preparation and the original controlled circuit. <i>Partial exclusion</i>	Separate state preparations can check vectors; the current study independently verifies the corrected full controlled unitary. Original Method 2 controls only the inverse query preparation and measures a different observable. Its attention-fidelity curves are excluded.
4	Ideal Grover search. <i>Supplementary descriptive record</i>	Success probabilities can be compared with the supplied-oracle amplitude-amplification formula. Statevector sign-vector oracles omit reversible predicate construction, data loading, and real execution costs. Query ratios are not measured speedups.
5	Cloud core circuits. <i>Registry context</i>	Retain backend-specific simulator and syntax-check records. Cloud service names do not imply QPU execution. The old inner-product implementation requires a separate circuit check.
6	Cloud Grover circuits. <i>Registry context</i>	Retain simulator success probabilities and recorded circuit sizes. Not physical QPU validation or an end-to-end verification task.
7	Cloud scaled QAOA and kernel. <i>Registry context</i>	Retain simulator outputs in their original observable units. Different backends, shot counts, and configurations cannot be pooled as a physical-device ranking.
8	Quantinuum H2-1E. <i>Registry context</i>	Retain explicitly labelled emulator outputs. H2-1E is an emulator; low-shot circuit checks do not identify real-device causes or a universal noise hierarchy.

Table S7 continued

ID	Study and role	Supported use and limits
9	Rigetti kernel overlaps. <i>Main reanalysis</i>	All 45 stored off-diagonal probabilities are observed entries. The diagonal was assigned to one. No matching saved feature vectors, classifier evaluation, or raw result bundle.
10	Rigetti QAOA submissions. <i>Main descriptive reanalysis</i>	All 43 submission scores are expressed as mean cut divided by edge count. Heterogeneous parameters; absent raw counts, per-task angles, and compiled circuits. Reported scores cannot establish a noise mechanism or hardware scaling law.
11	Rigetti signed-overlap batches. <i>Main quantitative reanalysis</i>	Reconstruct empirical squared error from five-batch moments and compare eight full-task means with recorded ideal overlaps. Only four vector pairs and two submissions per pair. Shot levels reuse data; ordered shots and authenticated per-job circuits are unavailable.
12	Braket SV1 Grover. <i>Supplementary descriptive record</i>	Retain SV1 success probabilities as simulator outputs. The service was a cloud simulator, not the Rigetti QPU. Oracle costs are assumed.
13	Density-matrix noise sweep. <i>Supplementary descriptive record</i>	Retain QAOA responses under the specified synthetic noise rates. Local simulator data do not diagnose historical device noise or validate a unit-constant big-O bound.
14	IonQ QAOA. <i>Denominator audit only</i>	Demonstrate the mismatch between cut/OPT and cut/edge-count using the stored scalars. Raw counts and bit mapping must be recovered before treating the converted means as validated device-performance results; provider defaults differ.
15	IonQ kernel overlaps. <i>Main completion audit</i>	Preserve the 18/45 completion mask and bound zero-filled-matrix sensitivity. Available input generator contradicts recorded labels; identical feature vectors across providers are not established. Attention jobs failed for lack of quota.
16	Hybrid learning and sequential MNIST. <i>Supplementary exploratory record</i>	Preserve the larger 80-epoch files separately from smoke tests and show reported local training metrics. One recorded run; test data were monitored during training in the available routines. Alternate text encodings, cached runs, and missing per-model checkpoints prevent a fully reproduced accuracy comparison. Ablation figures lack matching numeric records.
17	Larger GPU QAOA and Grover. <i>Supplementary limited record</i>	Preserve 30-qubit Grover oracle-model outputs and rounded p=1–4 QAOA costs from the console log. The JSON lacks the completed 24-vertex QAOA outputs; the log used a heuristic denominator of 31 and stops during p=5. No invented p=5 result or quantum speedup.

Table S7 continued

ID	Study and role	Supported use and limits
18	Image-retrieval kernel. <i>Excluded pending matching rerun</i>	Retain source and outputs to recover and repair the experiment. Log and JSON use different dimensions and sample counts. Prior source audit found a nonmatching multi-repetition ZZ feature map; no single performance table is adopted.
19	Emulator QAOA extension. <i>Excluded pending matching rerun</i>	Retain code for sign and bit-order diagnosis. Prior source audit identified a Hamiltonian-sign and bit-order mismatch. Recorded ratios are not evidence of intrinsic graph hardness or a physical noise floor.
20	qRAM-free attention. <i>Excluded pending matching rerun</i>	Retain an example of why target definitions must match before comparing shot curves. Raw dot-product softmax was compared with normalized-overlap softmax. More shots cannot remove that deterministic target mismatch.
21	Closed-loop neurosymbolic verification. <i>Excluded pending matching rerun</i>	Retain the prototype and logs as implementation history. The archived loop does not establish a complete reversible predicate oracle with comparable costs; neither speedup nor operational AI-safety validation is retained.
22	Post-hoc mitigation. <i>Excluded as hardware evidence</i>	Keep simulated noise diagnostics distinct from a claimed analysis of hardware counts. The real-count source script and folding records were not located; the available synthetic procedure is not evidence of successful device-level zero-noise extrapolation.

References

- [1] Lex, V.: Quantum Hybrid Modules for AI: Attention, Optimization, and Verification on Near-Term Quantum Hardware. Research Square. Preprint, version 1 (2026). <https://doi.org/10.21203/rs.3.rs-9825475/v1>