METHODOLOGY

The RQP Evidence Engine

How the studies behind qubit-lab.ch/evidence are made. Four method studies in August 2026 - quantum concentration, the noise ladder, and the two advantage-island maps - came out of the same machinery: a management question, a design locked before the first run, certified classical referees on every run, committed scripts that reproduce every published number, and revision notes that disclose corrections. This paper describes that machinery, shows its receipts, and explains how to run a question of your own through it.

August 2026

qubit-lab.ch

qubit-lab.ch
contact@qubit-lab.ch

Contents

The engine on one page

The Rapid Quantum Prototypes are tools: portfolio optimization, index replication, feature selection, QML classification, quantum Monte Carlo. This paper is about what sits around them - the machinery that turns a management question about quantum methods into a published, checkable answer. We call it the **RQP Evidence Engine**, and it has five parts:

1. **A management question, stated first.** Every study begins with the question a decision-maker would ask - not with an experiment looking for a headline. The published documents open with those questions and answer them in order.
2. **A design locked before the first run.** Hypotheses, arms and analysis plans are written down and locked before execution; from the second island study onward the design is publicly tagged in an immutable repository before the first run, so the chronology is verifiable, not asserted. Deviations are disclosed in the appendix of each study.
3. **A classical referee on every run.** No quantum method grades its own homework. The QAOA studies score every readout against a classically certified optimum (exhaustive enumeration or branch-and-bound with a proven bound); the QML studies score every arm against ten tuned classical variants, including an order-matched twin built to imitate the quantum model classically. Advantage is claimed only when the whole 95% confidence interval clears parity.
4. **Reproduction before printing.** Every published number is re-derived from the committed raw results by a committed script before it is printed. The script asserts the published value first and computes second; if the archaeology and the document disagree, the build fails and the discrepancy is investigated - twice this has caught an imprecise published sentence, and both corrections are disclosed.
5. **A closed verdict vocabulary and open revisions.** Every result is filed under one of five verdicts - Advantage demonstrated, Parity, No advantage demonstrated, Inconclusive, Method study - and the evidence page uses all of them, including the unflattering ones. Published documents are revised at the same public URL with a dated revision note saying exactly what changed.

The receipts: four studies between 15 and 30 August 2026 - 12,397 certified QAOA runs; a noise ladder ending on ibm_kingston (100,000 device shots, 44 seconds of processor time); 996 refereed QML arms mapping where a quantum kernel wins at all; 65 pre-registered training arms mapping the variational model's territory. Each with a public PDF, a machine-readable evidence record, and raw results a committed script can replay.

The invitation: the engine is not reserved for our own questions. Section 6 describes what it takes to run yours through it.

1. Management summary

Written for decision-makers; the basis for every sentence follows in sections 2-7.

Q1 - What is the Evidence Engine? The RQP suite plus a fixed discipline around it: question first, design locked before the first run, certified classical referees on every run, committed reproduction scripts, closed verdicts, disclosed revisions. None of these parts is exotic on its own; the point is that all five are applied to every study, and each leaves a verifiable trace - a tag, a referee bound, a script, a revision note - so trust never has to rest on our word.

Q2 - Why should you trust results from a vendor's own machinery? Because of what the machinery publishes against itself. Our own studies print that the products' standard warm start never escaped a wrong seed on real books (0 of 96 runs); that on today's hardware nothing past three circuit layers survives; that our own screening checklist wrongly vetoed 83 of 84 winning datasets and was replaced; that a bet we registered against the folklore lost. A machine built for marketing does not produce these sentences. Twice, conclusions that looked final were overturned by the engine's own verification stages before publication - the overturning is part of the record, not an embarrassment removed from it.

Q3 - What has the engine produced so far? Four method studies in August 2026: which QAOA method concentrates probability on the certified optimum and under what prior knowledge; how much survives realistic noise and a real device; where a quantum kernel classifier can beat a full classical referee at all; and whether the trainable quantum model owns territory of its own. Each study's findings shipped back into the products as measured defaults and printed diagnostics - the studies are not brochures about the tools, they are how the tools improve.

Q4 - What does a study cost? The studies print their own resource accounting: the entire hardware rung cost 44 seconds of processor time; the QML island grids ran on standard product workers; the variational study discloses its training-hours against a registered ceiling. The receipts table in section 4 shows the calendar span of each study from its first run to its published PDF. We let those dates speak for themselves.

Q5 - What question of yours could it answer? Any question of the form the four studies answer: should we use method X on problem Y, and what is the honest evidence? - on your data, your books, your constraints, under the same referees, the same verdict vocabulary and the same publication standard (or private delivery). Section 6 describes the commissioning path; the natural first candidate is placing a real dataset on the two advantage-island maps.

2. The machinery, part by part

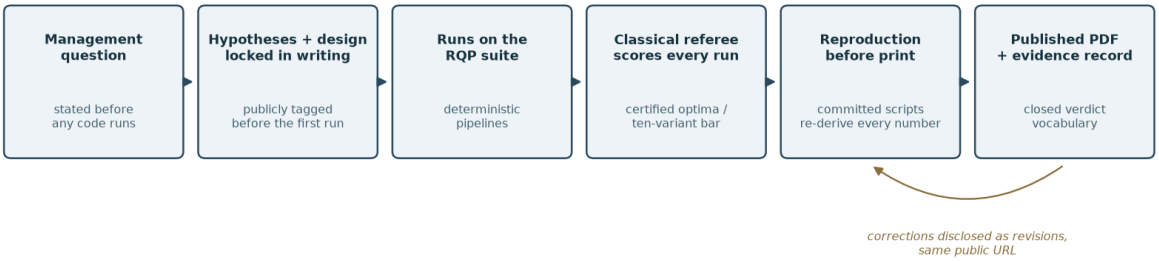

Figure 1 - The Evidence Engine's loop. Every stage leaves a verifiable trace; corrections flow back as disclosed revisions at the same public URL.

2.1 The question comes first

Each study opens with the questions a decision-maker would commission - "which method should we run by default?", "does trained depth survive a real device?", "when is the trainable model reasonable at all?" - and the management summary answers exactly those questions, in order, with the measured basis referenced section by section. This is a constraint on us, not a formatting choice: a study that cannot state its management questions up front has not decided what it is for.

2.2 The design is locked - and, now, provably so

Every study's design - hypotheses, arms, analysis plan, stopping rules - is written and locked before execution, with deviations disclosed in an appendix. The series is honest about the strength of that claim over time: the first three studies were pre-specified - locked in writing before execution, but with registration-grade public chronology unproven; the fourth was publicly pushed and tagged in an immutable repository before its first run, making the chronology verifiable by anyone; the next will deposit its design with an independent registry before running. Each study states plainly which of these grades it carries.

2.3 The referee is classical, and it is not ours to influence

The engine's central rule: **no quantum method grades its own homework.**

- In the QAOA studies, every one of 12,397 runs is scored against a certified optimum - proven by exhaustive enumeration or by branch-and-bound with a certified bound - so "the circuit found the answer" is a checkable fact, not a claim.
- In the QML studies, every arm is scored against ten tuned classical variants: logistic regression, random forest, gradient boosting, an order-matched periodic twin built to imitate the quantum model classically, a tuned RBF kernel SVM and a neural network - each threshold-tuned, so the bar is the strongest classical attempt we can mount, not a strawman. Advantage is claimed only when the whole 95% confidence interval of the paired comparison clears parity, with a selection-aware robustness check on top.

2.4 Every number is reproduced before it is printed

Each study folder contains its committed raw results and a committed analysis script whose job is adversarial: it asserts the published numbers first and computes intervals second, failing loudly on any mismatch. In August this discipline caught two imprecise published sentences - an attribution in one management summary, a range computed from rounded table entries in another. Both were corrected in dated revision notes at the same public URLs, with the convention stated: what changed, what did not, and that no verdict moved.

2.5 Closed verdicts, open revisions

Every evidence record carries one of five verdicts: **Advantage demonstrated, Parity, No advantage demonstrated, Inconclusive, Method study.** The vocabulary is closed so that it cannot inflate: there is no "promising", no "near-quantum-advantage". The evidence page files records under all five - including a product benchmark filed as No advantage demonstrated (0.92x against the classical bar, interval entirely below parity) sitting directly above an Advantage demonstrated record (1.45x, interval entirely above parity). Both link the raw runs. The page is also machine-readable (**/evidence-runs.json**), because a claim that cannot be scraped and checked is a weaker claim.

3. What the engine publishes against itself

The strongest evidence for the machinery is the set of findings it published that a marketing instrument would have suppressed.

- **The concentration study** prints that the products' classical warm start, given a slightly wrong seed, never escaped on the real books - 0 of 96 runs, below 3.1% at 95% confidence - and that a wrong seed comes back as a confident wrong answer.
- **The noise-ladder study** prints that noise-free circuit depth is worthless on today's hardware for this problem class: from three layers upward the device readout is statistically uniform, and the adaptive circuit's entire simulated advantage is erased (0 hits in 20,000 shots). It also prints that a prediction built from the device's own calibration data is optimistic by a factor of about three - a result that makes one of the suite's own noise sources an optimistic bound, and the product now labels it so.
- **The island study** refuted its own instruments twice: the screening checklist the product shipped with vetoed 83 of 84 datasets on which the quantum kernel then won (the checklist was replaced by calibrated measurements), and the study's own working conjecture about why shot noise was survivable failed its dedicated ablation and is printed as refuted.
- **The variational study** registered a bet against the folklore - that initial-gradient variance would not predict training outcomes - and lost it; the loss is printed with its AUC and interval.

Two further reversals happened before publication, and the record keeps them: the first island study's variational arm initially looked like a flat null, and the engine's verification stage - the levers-and-checks pass that every surprising result must survive - overturned it, which is why the variational model got its own study with an adapted fair methodology rather than a dismissive paragraph. The published documents disclose this pilot-and-overturning history rather than presenting the final design as the first attempt.

This is the sense in which the verdicts are trustworthy: the machinery demonstrably prefers being right to being flattering.

4. The four studies and their receipts

Study	First run - published	Scale (as printed)	Referee	Verdict class	Public PDF, revision
QAOA Concentration	15-16 Aug 2026	12,397 certified runs, 24 instances	exhaustive enumeration	Method study	18 pp, Rev. 1.2
QAOA Noise Ladder	21-22 Aug 2026	5 circuits x (3 noise rungs + calibrated twin + ibm_kingston); 100,000 device shots	certified optimum (branch-and-bound)	Method study	16 pp, Rev. 1.3
Advantage Island (quantum kernels, QSVM)	27-28 Aug 2026	996 pre-specified referee arms	ten classical variants	Method study	14 pp, Rev. 1.1
Advantage Island (variational, VQC)	29-30 Aug 2026	65 fair-recipe training arms (registered ceiling 150)	ten variants + kernel head-to-head	Method study	10 pp, first edition; design publicly tagged before the first run

One paragraph each:

QAOA Concentration asked which quantum optimization method concentrates probability on the certified optimum, and under what prior knowledge. Its decision tree - trust a seed you can certify, insure a seed you cannot, tune the problem's penalty ratio before touching the circuit - now ships as the products' guidance, with the study's escape rates and correlations (uncertainty intervals attached in Revision 1.2) as the measured basis.

QAOA Noise Ladder asked how much of that concentration survives noise and a real device. Its answers - the routing overhead is the first number that matters, a calibrated prediction is an optimistic bound with a measured margin, measuring the device is cheaper than predicting it - are now printed in the products' hardware guidance.

Advantage Island (QSVM) asked where a quantum kernel classifier beats a full classical referee at all, between the classical-learnability wall and the kernel-concentration wall. Its island map, kernel-concentration indicator and shots-to-resolve estimate now ship in the QML product as printed diagnostics.

Advantage Island (VQC) asked whether the trainable quantum model owns territory the kernel method does not. Its size-gated crossover, fair-run recipe (depth plus proportionate stopping, near-identity warm start) and pre-run screening instruments now ship as the product's training defaults and controls.

The pattern across all four: the studies are how the products improve. Findings become defaults, diagnostics and printed warnings - which is also why the engine's incentives point toward honest verdicts: we have to live with them in our own tools.

5. What a study costs

The studies print their own resource accounting, and it is deliberately unglamorous:

- The entire hardware rung of the noise study - five circuits, 100,000 shots on ibm_kingston - cost **44 seconds of processor time**, within a standard monthly open-plan allowance; the calibrated predictions of the same five circuits cost 17-26 minutes of simulation each. On this class of circuit, measuring the device is both cheaper and closer to the truth than predicting it.
- The QML island grids ran on the products' standard workers; the variational study discloses its training-hours (about 24) against a registered ceiling, and found that with a sane stopping rule a quarter of the naive training budget already holds the verdicts.

- The calendar spans are in the receipts table above. We make no claim about them; the dates are public and the reader can do the arithmetic.

The honest summary: for questions of this shape - method comparisons at prototype scale, with certified referees, on problems up to about 16 qubits - the binding cost is no longer the computing. It is the discipline: locking the design, building the referee, and letting the reproduction scripts veto the prose.

6. Running your question through the engine

The engine is not reserved for our own questions. A commissioned study follows the identical path, and the deliverable is the same standard you can inspect in the four published PDFs:

1. **Your question, in management form.** "Should we use method X on problem Y?" - scoped together until it is answerable and worth answering.
2. **A locked design you approve.** Arms, referee set, verdict rules and stopping conditions, written before the first run - registered publicly, or privately with a verifiable timestamp, at your choice.
3. **Your data, the same referees.** Real books and datasets replace the synthetic instances; the classical bar is tuned as hard as we can make it, because an advantage verdict that dies on contact with a competent classical baseline is worse than no study.
4. **The same publication standard - your distribution choice.** A PDF with a management summary answering your questions in order, every number reproduced from raw results by a committed script, verdicts from the closed vocabulary, and revisions disclosed - published on the evidence page, or delivered privately.

The natural first commission is the one our series points at next: placing a real dataset on the two advantage-island maps and reading off, with intervals, which method - classical included - the data actually deserves.

7. How the engine is operated - scope and limits

The RQP suite is built to be operated programmatically: every product exposes a machine interface through which datasets are prepared, runs are launched, licensed and logged, and results and reports are retrieved - the same interface a licensed client can script. The studies were produced through that interface, including by AI-assisted engineering; the safeguards of section 2 exist precisely so that trust never depends on who or what operates the tools. The referee does not care who launched the run it certifies, and neither should the reader: the design tags, the certified bounds, the committed scripts and the revision notes are checkable by anyone, which is the point of having them.

Limits, stated plainly:

- **Prototype scale.** Everything so far lives at 4-16 qubits, mostly in exact or shot-sampled simulation, with one hardware rung. That is the scale where certified referees are possible and where honest method comparisons are therefore cheapest; claims do not extend beyond it.
- **Benchmark data.** Three of the four studies run on synthetic instances and constructed label families - best-case envelopes by design, and labeled as such. The engine has not yet run a client's real dataset through the island maps; that is the next study, not an accomplishment of this paper.
- **One vendor's discipline, not a standard.** The engine is our practice, documented so it can be inspected and criticized. Where an element has an external anchor - registration chronology, referee certificates, machine-readable records - we say which; where it is self-imposed, the reader should treat it as exactly that.

References

1. qubit-lab.ch, QAOA Concentration Study (August 2026, Rev. 1.2), qubit-lab.ch/evidence.
2. qubit-lab.ch, QAOA Noise-Ladder Study (August 2026, Rev. 1.3), qubit-lab.ch/evidence.
3. qubit-lab.ch, The Advantage Island of Quantum Kernels (August 2026, Rev. 1.1), qubit-lab.ch/evidence.
4. qubit-lab.ch, The Advantage Island of Variational Classifiers (August 2026), qubit-lab.ch/evidence.
5. qubit-lab.ch, evidence page and machine-readable records: qubit-lab.ch/evidence and qubit-lab.ch/evidence-runs.json.

Disclaimer

This document describes the methodology behind the method studies published by qubit-lab.ch and summarizes their published findings. It is intended for technical review, education and structured discussion. It makes no performance claim for real-world datasets, no financial advice, and no hardware claim beyond what the cited studies print. The individual studies remain the canonical source for every number quoted here.