

Connecting the Quantum Dots

AI pathfinders from access to discovery

joint work with Vincent Danos (CNRS, ENS)
and William Waites (Edinburgh)

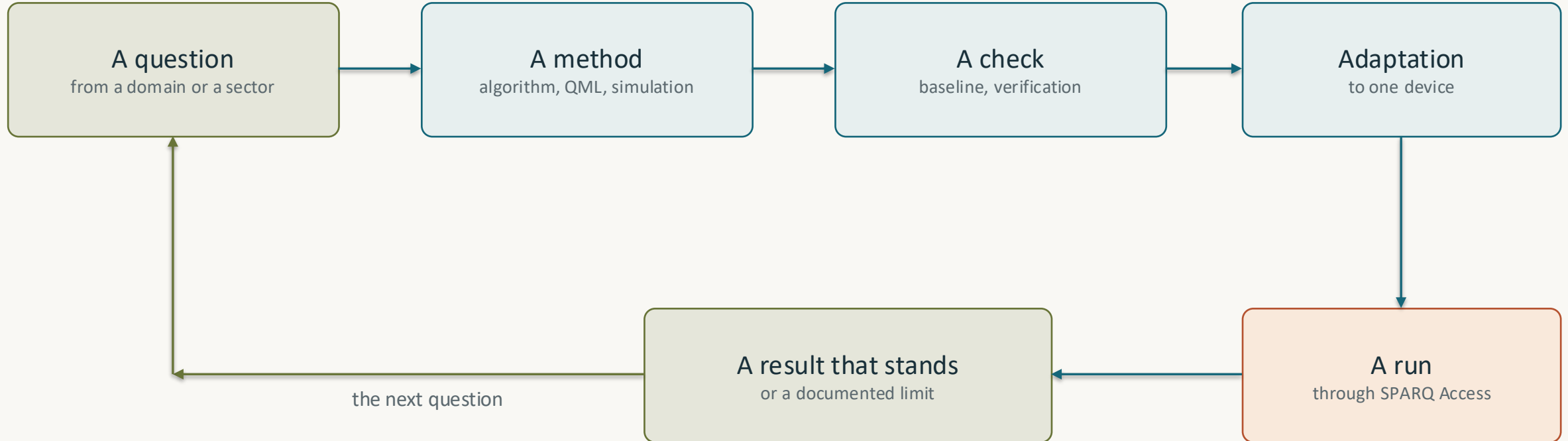
Elham Kashefi

Quantum Software Lab, University of Edinburgh
CNRS, Sorbonne University
NQCC

github.com/vd1/pathfinder



The loop every project here runs



SPARQ Access provides the device. Each team completes this loop once, within its own project.

From one run to a full paper

A QSL run, and the layers it still needs for a complete result

The run

Correlated-state preparation for an impurity model, 20 qubits, IBM Heron r3.

Result: the density-matrix error grows linearly, the energy error only quadratically.

Who covers all six layers today?

Mostly industry teams. They know their hardware and can assign a mitigation expert, a compiler engineer and a domain scientist to one paper.

Academic groups usually cover one layer per project, as funded by a grant.

SPARQ also runs industry projects, which come with end-user questions. These are another dot to connect.

End-user application: for example correlated materials or catalysis, where an impurity model is the core step.

Error mitigation: probabilistic error cancellation and symmetry post-selection, planned but not yet applied.

Device-aware compilation: native gates, connectivity and timing included in the method, rather than a default transpiler.

Verification: a certificate of correctness for the output that does not assume a noise model.

Alternative algorithms: for example sample-based Krylov methods on the same model, compared on this device.

Classical baseline: a fair comparison with the best classical method.

Combining all layers across academic groups is an ambitious goal. Recent progress in AI gives us tools to attempt it.

The same loop at national scale

QATCH: the full-stack programme QSL leads with the hubs and the NQCC, 2026 to 2030



QATCH

The hubs, the NQCC, QSL, QCi3 and IQN work on one pipeline across six sectors, with software driving hardware.

One SPARQ run

The same pipeline, run once by one team on one device.

The gap

The pieces exist across the UK. There is no systematic way to connect them.

For three years our scientists connected the pieces by hand. We are now building a tool to do much of this work.

Before the engine: QSL on real hardware

Runs and results on NQCC testbeds, SparQ Access and vendor access. Status from the QSL hardware access programme

Piece of the loop	Project and lead	Device	Status
Verification	On-chip verification of MBQC (Gustiani, Leichtle, Mills, Polacchi)	Quantinuum H1-1	Published, PRL 2025: 52-vertex graph state
Benchmarking	Entropy accumulation (Demarty, Hammam, Umathe)	Rigetti, NQCC testbed	Completed, April 2026
Benchmarking	Entropy benchmarking (Hammam, Demarty, Besserve, García-Patrón)	IBM Kingston	Reported, June 2026
Physics	Quench spectroscopy of magnetic excitations (Thomson, with IBM)	IBM	Posted, July 2026
State preparation	Correlated states for an impurity model (Besserve)	IBM Heron r3, 20 qubits	Running
Mitigation and verification	Accredited error cancellation (Mills, Yang)	IBM, via SparQ	In progress
Verification	Accreditation of an Ising quench on an error-detecting code (Coyle)	IonQ Forte, via SparQ	Simulator results so far
Error correction	Verified error correction on colour-code qubits (Mills, Roffe)	Quantinuum Helios	In progress
Tooling	Device-agnostic backend and resource estimation (Umathe)	All devices	Usage within 1.5% on Heron and Nighthawk

Each piece runs in a separate project. Mitigation, verification, compilation and algorithms rarely appear in the same paper.

And more in the pipeline

Ready, waiting for a device, or proposed to the NQCC testbed programme

Ready or waiting

Photonic state injection and energy model

Monbroussou. Code ready for ORCA. Deployment is being arranged.

Generative models, trained classically

Herrero Gonzalez. Code ready for Infleqtion. Runs planned for December 2026, when the device returns.

Analogue accreditation

Jackson. Neutral atoms.

Proposed to the testbed programme, 2025

- Coarse-grained quantum-enhanced MCMC for optimisation
- Distributed computing on neutral atoms
- Adiabatic-inspired preimage search
- Building blocks of error correction across platforms
- Quantum pseudorandomness proof of concept
- On-chip MBQC verification across platforms
- Natural-orbital state preparation
- Noise harvesting with a dissipative bath
- Adiabatic-variational chemistry ground states

About twenty projects on nine vendors' devices. Each covers one piece of the loop, in a separate project.

Asking AI to connect the dots

The task

Two bodies of work: papers on methods, such as verification, mitigation, compilation and algorithms, and papers on hardware runs and applications. For each pair, decide whether the method could test, correct or extend the result. If so, propose a next step.

Why agents, and why now

Useful pairs are often not similar papers, so keyword search is not enough. The task requires reading both papers in full, checking a derivation or a number, and writing down the argument. Agents can now do this for thousands of pairs.

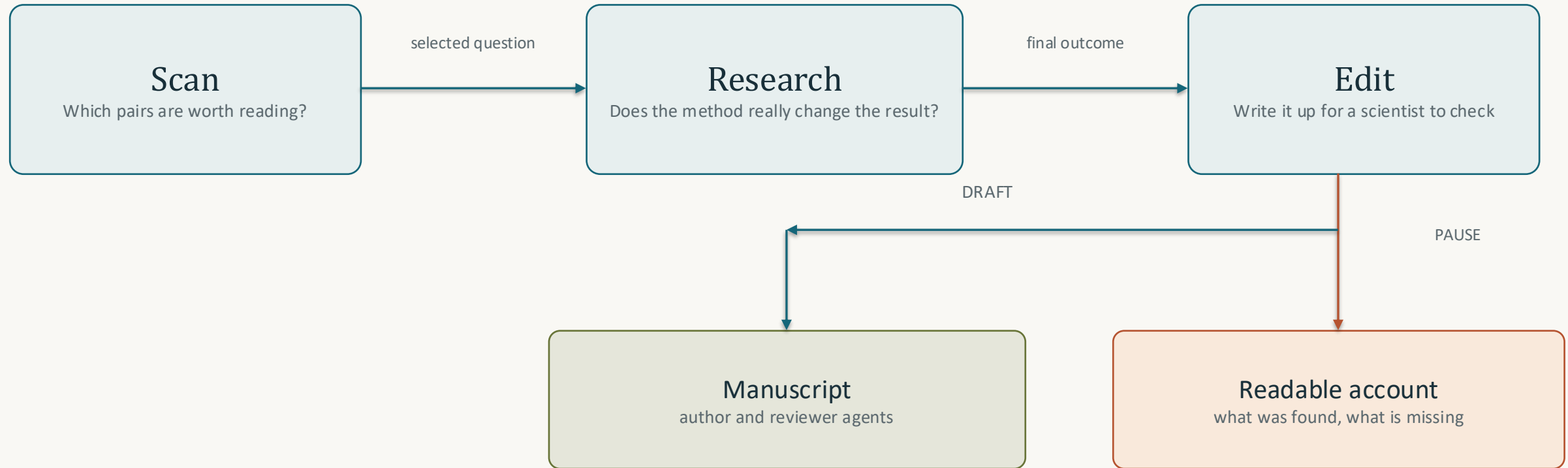
The output

A proposal with its evidence, not a result. A scientist can check it, reject it or develop it. We built Pathfinder for this task and tested it first on our own papers.

The agent searches, reads and proposes. The scientist decides what is worth exploring.

The Pathfinder pipeline: three stages

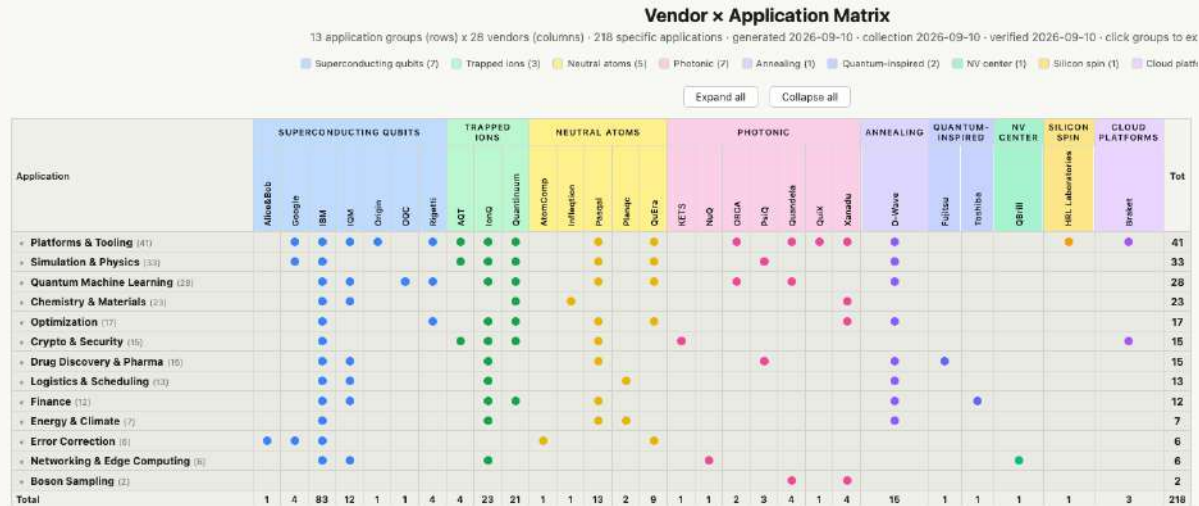
The task of slide 7, built with Vincent Danos and William Waites. The design changes as the agents improve.



Danos, Kashefi, Waites. *Pathfinder: hypothesis generation across the quantum software-hardware boundary*. AIQxQIA 2026. Preprint: hal.science/view/index/docid/5766056

A research prototype, not a product: about ten designs in four months, each changed after feedback from QSL researchers. The newest generation is open: github.com/vd1/pathfinder

The Pathfinder pipeline: inside the scan stage



The scan is useful on its own: a current map of hardware results, and a ranked list of where a lab's methods could apply. We propose this corpus for your use.

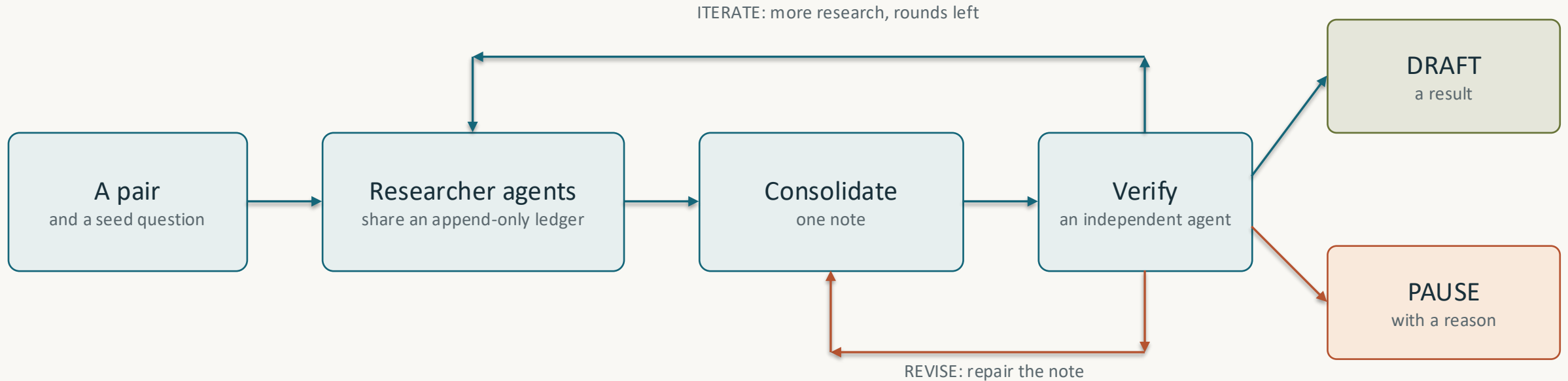
Building the corpus

Ranking the pairs

Ranking from abstracts costs pennies per pair. Reading both papers in full costs pounds per pair. A cheaper model, tested as a pre-filter, kept 21 of the 22 pairs that the strong judge marked positive.

The Pathfinder pipeline: inside the research stage

Agents share one record, an independent agent verifies, and every loop has a limit

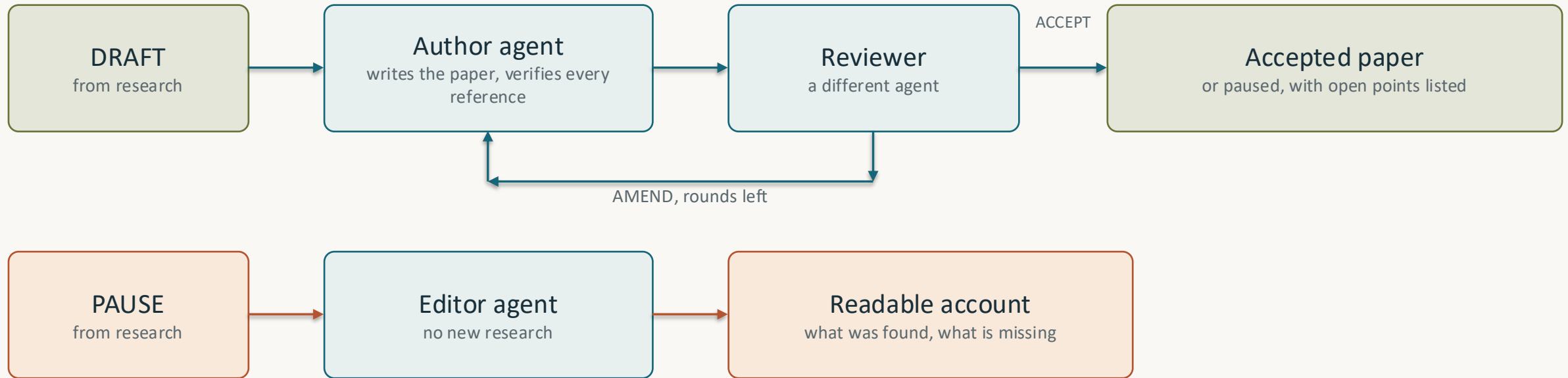


Researcher agents may disagree, and a new finding reopens the discussion. The verifier has not seen the earlier work. Every loop has a round limit, and the cost of every call is recorded.

The seed question may change during the research. The record states the reason for each change.

The Pathfinder pipeline: inside the edit stage

From agent research to a paper that a researcher wants to develop



This stage has changed most. Three edit designs are in use: William Waites's, Vincent Danos's and mine. William's is an adversarial review loop, with an editor, a fact-checker and domain-specialist critics, packaged as a skill: github.com/wwaites/pce. In mine, a referee from a different model family checks every number, and a scientist decides each point, until a named researcher will sign the paper.

Editing determines whether a researcher will take the work further. This is where human judgement enters.

Similar papers are not the useful pairs

Against text similarity

Cosine similarity of abstract embeddings, same 10,912 pairs.

Top K	10	20	50	100
Shared pairs	0	1	3	8
Random	0.01	0.04	0.23	0.92

Rank correlation with the judges: 0.394.

Repeatability

The same judge scores a fixed set of 100 pairs again.

Kept on rescoring	Opus 5	GPT-5.6 Sol
Top 10	90%	40–60%
Top 20	95%	50%
Top 75		85.3%

Random expectation at top 75 of 100 is 75%.

Over all 10,912 pairs the two judges agree well: rank correlation 0.740. But their two top-100 lists share only 12 to 36 pairs, against 0.92 by chance. They agree on which region is promising, but not on the best pairs.

The judges detect links between papers, not shared topics. Their shortlist is a reading list, not a ranking of quality.

Did the agents connect the dots?

Latest completed campaign: 400 pairs from 66 QSL and 148 vendor papers, 397 concluded by 5 September 2026

56 documented result

301 on hold:
passed a first check on invented data, not yet tested on real data

40 paused or dropped

Agents read both papers in full, ran declared checks, and recorded a verdict. Most results check or correct a published paper. None is a new hardware run.
In an earlier cohort of 100 pairs, the scan rank did not predict success: the top 20 gave 6 results, the bottom 20 gave 18, partly because the method improved as it ran.

Earlier cohort of 100: what the 58 results answered

Answered the question asked



8

Answered a narrower version



19

Found a different result



31

Work in progress: the agents find and check connections at scale. The bottleneck now is human validation, one pair at a time.

Which dots were connected: the missing layers

Research notes that reached a documented result in the August cohort of 100, grouped by us into the layers of slide 3

Error mitigation

A closed-form repair and a missing control for bounded zero-noise extrapolation · Interval containment and held-out accuracy choose different ZNE controls

Verification and metrics

Average infidelity is not an error-event rate · Purity does not certify a quantum kernel · Equal entropy does not identify which GHZ observable failed

Algorithms

A symmetry obstruction to zero-angle adiabatic QAOA continuation · A paired-state control collapses the UCC gate hierarchy for H_2 · A classical-reducibility test for verified order finding

Baselines and claims

A trend test for device drift · Order finding has a precision-register entropy baseline · A reproducibility check for a linked chemistry workflow

Device and decoding

Strict single-window decoding breaks in a finite radial-code memory · A printed pulse-optimisation basis is orthogonal to the stated square pulse

Each connection could supply a missing layer for another group's run: a check, a baseline, a correction or an algorithmic limit. Not yet reached: an end user, or a new run on hardware.

One connection in full: verification on a shared processor

Q Blind verification on a noisy device

A client hides its computation among secret test runs.
A failed test aborts; an accepted answer is correct.

P Sharing one processor

Several users' circuits run side by side on one chip.
A score ranks which neighbours disturb each circuit.

Seed question

Can a client on a shared processor still get a verified answer, whatever the neighbours do?

→ Obstacle

A score that compares outputs cannot distinguish crosstalk from a neighbour and drift of the device.

→ Result

The verification guarantee carries over to a shared processor under two stated conditions: a stronger guarantee than output comparison can give.

This case applies the verification layer of slide 3 to a real multi-user device, by pairing a theory paper with a hardware paper.

What the connection found: output similarity is not enough

Q Blind verification on a noisy device

A client hides its computation among secret test runs.
A failed test aborts; an accepted answer is correct.

P Sharing one processor

Several users' circuits run side by side on one chip.
A score ranks which neighbours disturb each circuit.

A one-qubit counterexample

Identity and a Z error give the same computational-basis output, so the fidelity is 1 in both cases. A trap qubit on the equator fails with probability 0 in one case and 1 in the other.

What the score measures

Sensitivity = $-(6/5) \cdot H(i, \text{post})$: a rescaled control column. It measures how much the output changes, not whether a test fails.

Cost of a direct test

A trap test that compares solo and shared runs. Under the stated noise assumptions, an error below 10^{-3} requires

6,250 full protocol runs

The ten-qubit case takes microseconds to simulate classically. This is the cost of a certificate, not a quantum advantage.

Only a test with known answers separates crosstalk from drift. Output comparison alone cannot.

What the human added

Seven rounds of refereeing on this case

The loop

A writer agent turns the research note into a paper. A referee from a different model family, which never sees the note, checks every number against both source papers. The expert decides every point.

21

referee points, all accepted

4

defects missed by the agents and by me

They reappeared when the note was regenerated from scratch, so they came from the research note, not the writing.

What the expert added

Security needs no noise model. Crosstalk, drift or a neighbour's gates either pass the tests or trigger an abort.

An abort does not identify its cause. A client cannot distinguish device noise from a neighbour. Only completion rates, solo against shared, separate them.

State every assumption about the other paper. The sign convention of the other group's score was taken from its text, and the paper says so.

The agents find and check. The expert interprets the result and decides whether to sign it.

From one link to a stack

Four QSL links on one study of several users sharing an IBM processor, by Woods and Shyu

Verification, refereed

Trap-based tests replace output similarity: identical outputs can hide opposite test results. Accepted by a cross-family referee, with every point decided by a person.

Scheduler design

A cleaner scheduling score: without self-comparisons and with a fixed orientation, the scheduler prefers the least disturbed pairings.

Verification headroom

A bound, per user, on how much disturbance verification can absorb on a shared device. An analytic result on a pair held on synthetic data.

Compilation

A pinned compiler and a seeded layout, so that equivalent circuits map to reproducible physical controls. Also on a held pair.

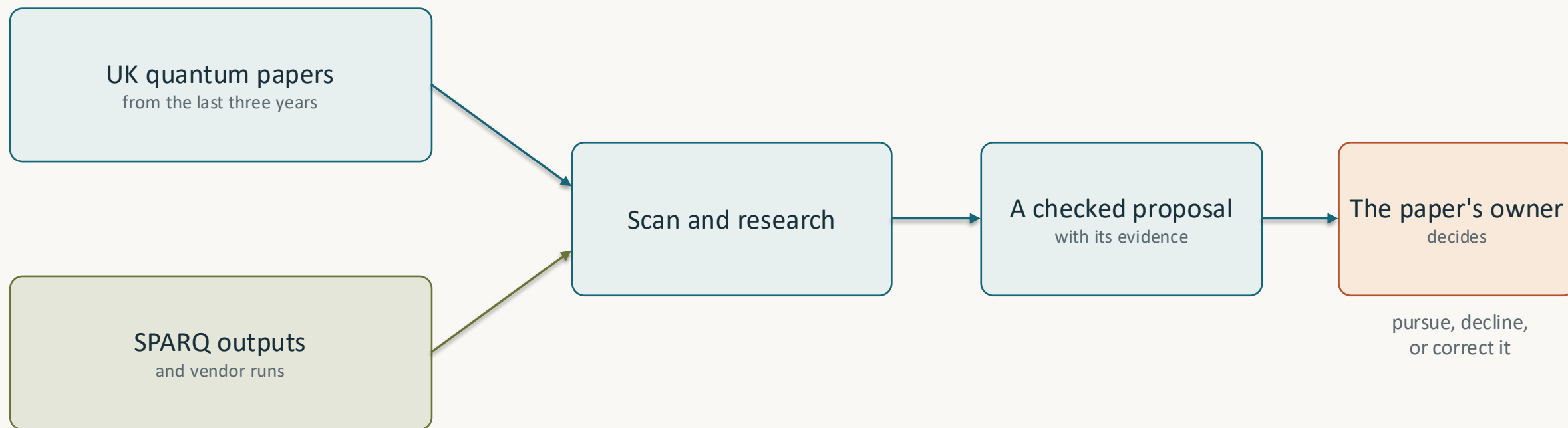
Still missing

A classical baseline, such as running each user alone. An end user. A run: one IBM Heron device, a verified circuit alone and beside a second user, testing whether an output-based score predicts test failures.

Together the links sketch a verified multi-user service on one device. The agents found the pieces. Composing them into one experiment is the next step.

Connecting the pieces across the UK

The QATCH loop of slide 4, with SPARQ outputs as one of the two bodies of work



Of the seventeen other speakers at this meeting, **ten** have an arXiv paper since 2023 that the engine could read today. One is already in our vendor-application map.

The engine proposes. The owner decides. Nothing is claimed until a named researcher checks and signs it.

Four likely objections

“You read my abstract and told me what to do.”

The message comes from a colleague: this result appears to test your work. Would you like to see the argument?

“Your own ranking does not predict outcomes.”

Correct, and we measured it. So the owner decides, not the score.

“Whose paper is it?”

The researcher's. Agents are named as tools. Nothing is claimed unsigned.

“My SPARQ report is not public.”

Then it is included only if you opt in. Opting in is the first step to connecting projects.

The engine is useful only if researchers trust how it selects their work.

Next steps proposed with the UK hubs

With QCi3 and IQN, two of the QATCH partners on slide 4. Submitted in August 2026 to the EPSRC AI in Quantum fund. Neither is yet decided.

QCi3 Hub: an AI-enabled quantum scientist

- Extend the method from one lab's papers to the whole hub.
- Each proposed connection is judged by an expert in the relevant platform: superconducting, semiconductor, photonic, modelling and verification.
- Publish the expert verdicts and the measured miss rate of the cheap filter.

IQN Hub: the Quantum Protocol Zoo

- Agents read new quantum-network papers and draft entries for the Zoo, with their sources.
- Every draft goes to a human editor. Nothing is published without approval.
- Measure how often the editor rejects a draft.

Both projects record the expert verdicts. A validation study needs this record.

A question I am asking, not answering

What QSL can say

Here is what we built, and what it did on our own papers.

Here is what it would look like on yours.

The method is open, and so are its failures.

What this community decides

Who should run this for the UK?

On which corpus, with whose consent?

Who validates, and how is that credited?

I speak here for the Quantum Software Lab. I also hold a role at the NQCC, which is why I ask rather than propose.

Tomorrow's feedback session is a good place to start.

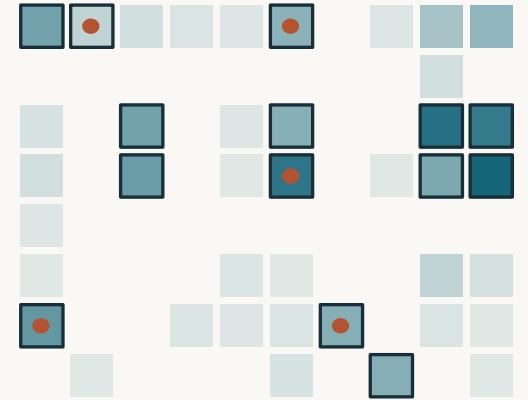
Listen to every talk that follows with one question

Which result presented here could test,
constrain or extend another SPARQ project?

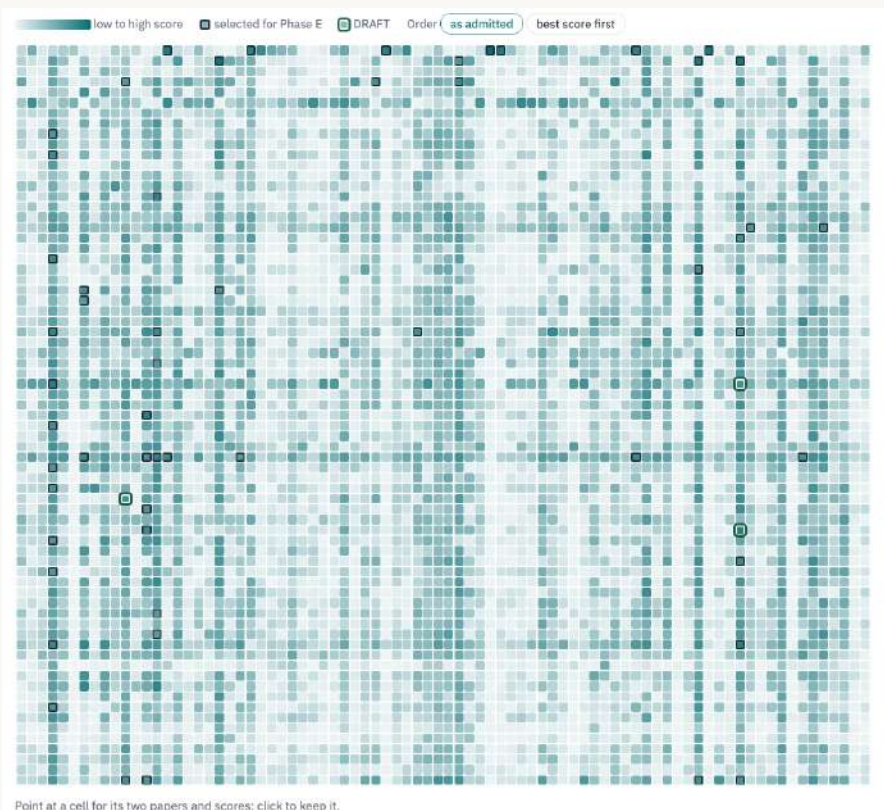
One request: add your SPARQ output to the corpus, if you agree.

Paper: hal.science/view/index/docid/5766056

Data: github.com/quantumsoftwarelab/aiqxqia-pathfinder-2026



Backup: other corpora, same method



Quantum methods for computational chemistry: an expert's own papers against the field.

Chemistry corpus, one expert

Pairs scored	5,822
Pairs taken to research	59
Draft-ready outcomes	3
QSL against vendor runs, later campaign	
Final research notes	147
Paused with a stated reason	144
Draft-ready outcomes	3

Counts at 14 September 2026. DRAFT is an internal outcome, not external validation.

Most research threads end in a documented pause. A human validator needs this record.