INSTITUTE FOR PHYSICAL AI @ BMI
The Charlot Lab

Technical Report TR-2026-26
Criteria / Method · v0.1
4 August 2026

CRITERIA FOR A NOVEL COMPUTING PARADIGM

Building the Energy Compute Future

Why performance claims for new machines keep collapsing, which claims survive the attack, and how to state the two of them together.

Institute for Physical AI @ Bailey Military Institute

Charlot Lab · Energy First Architecture (EFA) research program

Drafted with AI assistance from public sources. Quantitative figures are labelled measured, reported, claimed or modelled. A modelled figure is never presented as a measurement.

Abstract. A new computing machine is announced with a large speedup. Within a few years an ordinary computer running a cleverer algorithm reaches the same place, and the advantage is gone. This has happened often enough to have a name, dequantization, and it is the central obstacle to arguing that any novel paradigm is worth building. This report separates two claims that are usually made in one breath. The first is that the new machine runs an existing workload better; that claim is attackable by anyone with an idea and no laboratory, and it should be expected to fall. The second is that the machine does work the incumbent cannot do; stated as computability that claim is simply false, because a classical computer can compute anything computable, and the version that survives is about resources: the work is not reachable within the energy and latency budget the application actually has. We show why an advantage located in an algorithm is unstable while an advantage located in the arrangement of matter is not, give a runnable miniature of the attack, and set out a two-sided test a paradigm must pass on both sides at once. The test is offered as teaching material: the reasoning transfers to any efficiency claim a student will meet.

1. The pattern worth studying before the physics

In 2018 a nineteen-year-old undergraduate removed the best-known speedup in quantum machine learning. Ewin Tang, asked to show that the quantum recommendation-systems algorithm of Kerenidis and Prakash could not be matched classically, found instead that it could.[1] The quantum algorithm was exponentially faster than the best classical algorithm known at the time. Tang's classical algorithm was exponentially faster than that classical algorithm, and the quantum advantage

evaporated, not because the quantum computer had been mismeasured but because the comparison had been against the wrong opponent.

What made this possible is worth stating precisely, because it is the whole mechanism. The quantum algorithm was not given a matrix in the ordinary sense. It was given a particular kind of access to the matrix: the ability to draw rows with probability proportional to their weight. Tang's contribution was to notice that the access, not the hardware, was doing the work, and to hand the same access to a classical routine. The results generalised quickly: low-rank matrix inversion, principal component analysis, supervised clustering and a family of related routines were all dequantized within about two years by the same technique.[2]

The pattern did not stop at machine learning. When a 127-qubit processor was used in 2023 to argue for utility on a physics simulation beyond brute-force classical reach, several groups reproduced the results on ordinary computers within months, using tensor-network methods that exploited the structure of the specific problem.[3] Through 2026 the same fate has met short-path algorithms proposed for constraint problems and the simulation of local observables at short times.[4] In each case the classical attack cost a fraction of what the hardware cost.

The temptation is to read this as a story about quantum computing. It is not. It is a story about a kind of claim, and the claim is available to any novel paradigm: thermodynamic samplers, Ising machines, photonic processors, analog in-memory arrays, neuromorphic fabrics. Each of them will, at some point, report a large speedup on a workload that classical hardware already performs. The lesson from twenty years of dequantization is that this specific claim is the fragile one, and that a field organised around it will spend its credibility defending results that were never load-bearing.

2. Two claims that arrive wearing the same clothes

When an advocate says a new machine is better, the sentence usually contains two propositions that have almost nothing in common.

The first is a *complexity claim*: for this problem, the new machine reaches the answer in fewer steps, or with better scaling, than the incumbent. This is a statement about algorithms. Its truth depends on the state of human cleverness about classical algorithms, which is not a physical constant. It can be falsified by a graduate student with an idea, retroactively, without touching the hardware. Every dequantization result in Section 1 is a complexity claim being falsified.

The second is a *thermodynamic claim*: for this problem, the new machine reaches the answer having spent fewer joules, because of where the physics is happening. This is a statement about the arrangement of matter. It cannot be falsified by a better algorithm, because a better algorithm changes what operations are performed, not where the operands live. It can only be falsified by measurement, or by someone building a better arrangement.

These are graded differently by evidence and they decay differently over time. A complexity claim is provisional by construction: it is true until someone is cleverer. A thermodynamic claim, once measured under a stated workload with stated conditions, stays true. Reporting both in one figure, as almost every announcement does, means the durable part of the result inherits the fragility of the provisional part. When the algorithmic advantage falls, the energy result falls with it in the public account, even though nothing about the energy result was disturbed.

3. The two-sided test

A paradigm that intends to exist in ten years has to answer two questions at once, and the honest form of each is narrower than the form usually offered.

Side A. On work the incumbent already does, what does this machine deliver per joule? The comparison must be against a competently optimised classical implementation, not a naive one, and the figure must state whether it was measured or modelled. Side A is where credibility is earned, because it is the side the audience can check. It is also the side that will be attacked, and a paradigm should expect to lose ground here repeatedly as classical methods improve. Losing ground on Side A is normal and is not disqualifying. Claiming Side A as permanent is the error.

Side B. What work does this machine make reachable that the incumbent does not? Here the usual phrasing, work that is impossible on traditional hardware, is false as literally stated and should not be used. A classical computer can compute anything that is computable; this is not a matter of opinion but of the definition of computation, and any claim of the form *only our machine can compute X* is either wrong or is smuggling in a resource bound without saying so.

The version of Side B that is both true and checkable is a resource claim:

> **The honest form of Side B.** There exists a task, a stated energy budget, and a stated latency bound, such that the incumbent cannot complete the task within both bounds and this machine can. The task is named in advance. The bounds come from an application that actually has them, not from the machine's convenience.

This reformulation is not a weakening. It is the version that has teeth, for three reasons. It is falsifiable, because someone can attempt the task on classical hardware within the same budget and report the outcome. It is not vulnerable to dequantization in the way Side A is, because a cleverer algorithm has to beat the budget rather than beat a scaling exponent, and budgets are set by the physical situation rather than by the state of the art. And it forces a real application into the argument, which quietly rules out the largest category of unfalsifiable claims: those where the machine is fast at a task nobody needed done.

The two sides must be presented together. Side A alone produces the announcement-then-collapse cycle. Side B alone is unfalsifiable in practice, because a paradigm with no measured parity result has no way to demonstrate that its Side B numbers mean anything. A machine that is credible on Side A and specific on Side B is making a claim that can survive an adversary.

4. Why an advantage in an algorithm is unstable

The mechanism is small enough to build in an afternoon, and the exercise is the fastest way to understand why the field keeps being surprised. A version of it runs in the browser as part of the Institute's course on this material.[5]

Take a quantity that requires reading every row of a large table to compute exactly. Now grant a machine the ability to sample rows in proportion to a cheaply known weight, and to correct for the sampling bias afterwards. With a few hundred samples out of several thousand rows the machine returns an estimate within a percent or two of the exact answer, having read a small fraction of the data. Reported as a ratio, this is a twentyfold reduction in work, and it looks like the machine is doing something the ordinary computer cannot.

Then perform the attack. Give an ordinary routine, running on the same ordinary hardware, exactly the same sampling access. It returns an estimate of the same quality having read the same number of rows. The advantage is gone. It was never in the hardware; it was in the access model, and the access model was granted by assumption rather than earned by the machine.

The general lesson is that a complexity advantage is a claim about a gap between two algorithms, and gaps of that kind are closed by ideas. Ideas are cheap relative to fabrication. A field whose central claim can be attacked for the price of a graduate student's attention, while defending it costs a fabrication run, is structurally on the wrong side of an asymmetry.

5. Why an advantage in an arrangement of matter is not

Now apply the same attack to an energy claim, and watch it fail to land.

Where computing spends its energy is not where most people assume. Arithmetic is nearly free; moving the operands to the arithmetic is the bill. On a canonical accounting at 45 nm, a 32-bit floating-point multiply costs a few picojoules while fetching its operands from off-chip memory costs several hundred, a ratio of roughly two orders of magnitude.[6] Process improvements since then have made the arithmetic cheaper faster than they have made the movement cheaper, so the ratio has widened rather than closed. The consequence is that for the workloads that dominate current practice, the majority of the energy is spent on transport, and a very large majority of it is spent in ways that have nothing to do with the computation's logical content.

The floor is far below either number. Landauer's principle sets the minimum energy to irreversibly erase one bit at $kT \ln 2$, which is about 2.9 zeptojoules at room temperature.[7] A single picojoule-scale operation is therefore some eight orders of magnitude above the thermodynamic floor. The gap between what physics requires and what current practice spends is not a few percent. It is a factor of roughly a hundred million, and almost all of it is in the plumbing.

Consider what a better algorithm can and cannot do to this. It can reduce the number of operations, which reduces both the arithmetic and the transport proportionally. It cannot change the cost of a fetch, because the fetch cost is set by the distance the charge moves and the capacitance it moves through. If a machine computes where the data already sits, it does not perform the expensive fetch at all. There is no algorithmic move that takes this back, because nothing is being optimised away. The operation is not happening.

> **The asymmetry in one line.** An algorithm can remove operations. It cannot relocate your memory. A claim that rests on the first is provisional. A claim that rests on the second is a statement about the arrangement of matter, and survives contact with a cleverer opponent.

This is the reason to be more interested in a modest measured energy result than in a large reported speedup. The speedup is a bet on the current state of classical algorithms. The energy result is a measurement of the world. Both can be wrong, but they are wrong in different ways and on different timescales, and only one of them can be undone by someone else having an idea.

Two cautions keep this from being overstated. First, the argument establishes that a thermodynamic advantage is *durable*, not that it is *large*: durability is a property of the claim type, and the magnitude still has to be measured for the specific machine and workload. Second, the advantage is only real if it survives composition. A substrate that settles cheaply but must be fed by a conventional host, converted at the boundary and read back, may spend more at the interface than it saves in the core. Composed cost across the whole path, not core cost, is the figure that counts, and it is the figure most often omitted.[8]

6. What Side B looks like when it is stated properly

Because the impossibility framing is unavailable, Side B requires more care than Side A, and it is where most of the intellectual work remains. Three forms are worth distinguishing.

The budget form. A task must complete within an energy and latency envelope fixed by the physical situation: a device on a battery, a control loop that must close in milliseconds, a sensor that must run for a year on a coin cell. The claim is that the incumbent cannot fit and the new machine can. This is the most checkable form and

the most useful, because the envelope is imposed from outside and cannot be negotiated by either party.

The native-operation form. Some machines perform in one physical step an operation the incumbent must emulate with many. Drawing a sample from a distribution is the clearest example: a thermodynamic device can be arranged so that its equilibrium fluctuations are the sample, while a conventional processor must generate pseudorandomness and reject. The claim here is not that the incumbent cannot do it but that it must pay a conversion cost that the substrate does not. Stating this form honestly requires reporting the conversion cost at the boundary, which is where such claims usually fail.

The scale form. A task is reachable in principle but the incumbent's version would require more energy than the operator can obtain, or more time than the answer remains useful for. This is the weakest of the three because both quantities move, and a claim of this kind should be dated and revisited rather than treated as settled.

What none of these permit is the sentence *only this machine can do it*. A paradigm that cannot state its Side B in one of these three forms does not yet have a Side B, and saying so plainly is more useful than a claim that will not survive its first serious reader.

7. Stating a claim so that it can be checked

The two-sided test only functions if claims arrive in a form that permits checking, which most do not. Three conventions do most of the work, and each exists because of a specific failure that is easy to commit and hard to detect afterwards.

Record how a figure was obtained, alongside what it was. A measured joule, a modelled joule and a joule projected from a component datasheet are three different kinds of statement. Stored as a bare number they become interchangeable, which is the mechanism by which a simulation becomes a headline. Provenance has to be a field in the record rather than a footnote in the prose.

When receipts are composed, sum the energy but take the weakest provenance. If a measured stage is combined with a modelled stage, the result is modelled. Averaging provenance is what allows a simulated component to inherit a measured component's credibility, and it is the single most consequential error available in this area.

Name the baseline and state its optimisation level. A hundredfold gain against an unoptimised reference is a statement about the reference. Where the baseline is not named, the claim is not yet a claim. The Institute's open receipt draft formalises these conventions, and is published under CC0 for the ordinary reason that a measurement standard which can be captured by any party is worth nothing.[9]

8. What is open

Several things this report depends on are not settled, and are recorded here so that the argument can be attacked at its weakest points rather than at its strongest.

The energy figures in common circulation for edge and browser workloads are overwhelmingly modelled rather than measured, including the per-operation constants used in the miniature of Section 4. Converting them into measured tables for specific device classes is unglamorous, entirely tractable and currently rate-limiting for everyone working in this area.

There is no accepted way to score a system that adapts while it is running, which is the capability several of these substrates are actually for. Without such a benchmark, the field's central claim has no scoreboard, and progress on it is not distinguishable from selection of favourable examples.

Whether composed cost across a heterogeneous path can be kept favourable in practice is not known. The individual core results exist; the end-to-end results, measured across the conversion boundaries, largely do not.

Finally, the durability argument of Section 5 is an argument about claim types, and it does not by itself establish that any particular machine has a large advantage. It says only that if such an advantage exists and is measured, a cleverer algorithm will not remove it. Treating the argument as a substitute for measurement would repeat the error it describes.

9. Why this belongs in a curriculum rather than a position paper

The reasoning above is not specialised. A student who can separate a complexity claim from a thermodynamic claim can read almost any efficiency announcement correctly, in this field or another, and the separation is learnable in a single sitting if the attack is performed rather than described. That is why the Institute teaches it as an exercise in which the student builds an advantage, destroys it, and then finds the one that will not break.[5] Reading about dequantization produces agreement. Performing it produces the habit, and the habit is the transferable part.

The wider point is that the discipline this area most needs is not new physics but ordinary carefulness applied consistently: naming baselines, recording provenance, dating claims, and publishing failed reproductions. None of that requires an invitation or a budget. It requires that somebody do it and write down what happened.

Notes and sources

1. Tang, E. A quantum-inspired classical algorithm for recommendation systems. *STOC* (2019); arXiv:1807.04271 (2018). Dequantizes Kerenidis, I., Prakash, A., *Quantum recommendation systems*, ITCS (2017).
   `verified: peer-reviewed, widely replicated`
2. Chia, N.-H., Gilyén, A., Li, T., Lin, H.-H., Tang, E., Wang, C. Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing quantum machine learning. *STOC* (2020).

verified: peer-reviewed

3. Kim, Y., et al. Evidence for the utility of quantum computing before fault tolerance. *Nature* 618 (2023). Classical reproductions: Tindall, J., Fishman, M., Stoudenmire, E. M., Sels, D., *PRX Quantum* 5 (2024); Begušić, T., Chan, G. K.-L., *Science Advances* 10 (2024); and related tensor-network results.
   verified: multiple independent reproductions
4. Continuing dequantization of short-path constraint-problem algorithms and of short-time local-observable simulation, reported through 2026.
   reported: individual results vary in independent confirmation; treat as a trend, not a settled ledger
5. Institute for Physical AI @ BMI. PAI-270, *The Energy-First Turn*, Module 4: the bar a new machine must clear. Runnable exercise, no installation required.
   primary: Institute teaching material
6. Horowitz, M. Computing's energy problem (and what we can do about it). *ISSCC* (2014). The canonical per-operation and per-fetch accounting at 45 nm.
   verified: peer-reviewed, standard reference
7. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* 5 (1961). Experimental confirmation: Bérut, A., et al. *Nature* 483 (2012).
   verified: theory confirmed experimentally
8. Per-operation constants used in this report's worked examples are modelled, not measured. See Section 8.
   modelled
9. Institute for Physical AI @ BMI. EFA-RFC-001, OER/1 energy receipts and DRIFT/1 benchmark protocol; EFA-RFC-002, RELAX/1 dialect. Draft, released CC0, not a ratified standard.
   primary: Institute draft specification

Technical Report TR-2026-26 · Institute for Physical AI @ Bailey Military Institute · Charlot Lab. Companion material: the topic page *Building the energy compute future*, the Energy Lab, the global database of eighty-five organisations and nineteen national programmes with per-row verification status, and PAI-270. Corrections are welcome and will be recorded rather than quietly applied.