Futurum
RESEARCH

# Quantum Fine-Tuning and the Energy Case for Quantum in AI

## Why Energy-to-Solution Reframes the Quantum Debate, Why Architecture is Now the Variable That Matters, and Why Ion Trap Sits at the Center of it

Analyst: Daniel Newman
Publication Date: June 25, 2026
Document #: AIODN202606

## The Thesis

The quantum conversation has been having the wrong argument for a decade. Qubit counts, gate fidelity, the perpetual debate over when quantum advantage arrives. None of that is the variable that decides whether quantum earns a line item in the AI stack. Energy is.

AI is in the Hard ROI Era. The cost of intelligence is power draw, grid contracts, and the gigawatt math now sitting inside every hyperscaler capex model. And the timing is sharp. In May 2026, OpenAI began winding down its self-serve fine-tuning platform, steering developers toward prompting and retrieval instead. The largest lab in the market is stepping back from fine-tuning as a classical economic problem at the exact moment a quantum approach to fine-tuning produces a credible energy result. When the marginal cost of training and serving frontier models is measured in datacenter power, every enterprise quietly asks whether there is a cheaper way to compute the same answer. New work out of IonQ suggests that for a narrow but expanding set of AI workloads, the answer is yes. And it runs on a quantum processor.

A May 2026 study measuring energy-to-solution for quantum fine-tuning of foundational AI models is the first credible data point that puts energy, not just accuracy, at the center of the quantum value case. This report makes the argument in three parts. The broad quantum case for energy efficiency in AI. The architecture choices that decide whether that case holds. And why ion trap, and specifically IonQ, is positioned to catch the tailwind.

## Energy-to-Solution Changes the Question

Classical AI scales by brute force. More parameters, more data, more compute, more power. The scaling laws that built the foundation model era also built the datacenter power crisis. Every efficiency lever the industry has pulled, better silicon, sparsity, quantization, and distillation, lives inside the classical paradigm. They slow the curve. They do not bend it.

Quantum is a different proposition. Workloads that scale exponentially on classical hardware can scale far more gently on a quantum processor. The interesting part of the new IonQ work is

what it leads with. Not speed. Not accuracy. Energy. The team directly instrumented power consumption on quantum hardware and compared it to the classical alternative on the same task: fine-tuning a pretrained foundation model using a parameterized quantum circuit as the fine-tuning head, then running sentiment-analysis inference. Importantly, their method applies to any foundational model, which means literally any base AI model that you already invested time, dollars, and energy into creating in the first place.

IonQ has already moved the headline numbers, a 24 percent reduction in classification error and an energy break-even at 34 qubits, onto its customer-facing materials. When a hardware company welds a research result onto the sales deck, it is telling the market which metric it intends to compete on. Here, it is not qubit count. It is energy.

Three findings carry the value case.

1. Energy scales linearly, not exponentially. QPU energy consumption scaled approximately linearly with qubit count for shallow circuits. Classical simulation of the same problem scaled exponentially. That difference in slope is the entire argument. Linear versus exponential is the difference between a curiosity and a cost structure.

2. There is a measurable break-even. The study identified an energy break-even point around 34 qubits. Below it, classical wins on energy. Above it, the quantum path pulls ahead. A defined crossover turns a philosophical debate into an engineering target. In practice, that target maps to the workloads where fine-tuning earns its keep: sparse, complex, high-dimensional datasets where classical models plateau. Think fraud and anomaly detection, drug discovery, materials science, and financial risk modeling, the domains where a few points of accuracy at lower power is the difference between a model that ships and one that does not pencil out.

3. Accuracy did not have to be sacrificed. The quantum fine-tuned models matched and in cases exceeded classical baselines such as logistic regression and support vector classifiers, reaching 91.20 percent test accuracy against an 89.56 percent classical ML baseline, a best-case classification error improvement of roughly 24 percent. This is the point that matters most. Accuracy and energy are not a tradeoff here. They are dual goals moving in the same direction. The classical world forces a choice: spend more energy to get more accuracy. This result delivers higher accuracy and lower energy at once. Which one a buyer leans on depends on the use case. A regulated risk model leans on accuracy. A high-volume inference pipeline leans on energy. The advance is having both on the table.

Read the limits before you read the headline. Sentiment classifier, not a frontier model. Measured range of 12 to 18 qubits, which means the 34-qubit break-even is extrapolation, not an operating point. The underlying comparison is QPU versus classical simulation of the same circuit, not QPU versus production GPU inference on a real fine-tuning workload. None of that is fatal. It is the line between a real result and a marketed one. The result is real. The first

end-to-end measurement of energy-to-solution on quantum hardware. The slope scales the right way. In a market that prices intelligence in watts, that is the only metric the next conversation has to organize around.

## Architecture Is the Variable That Decides Who Wins

Here is what the qubit-count headlines obscure. Not all qubits are equal, and the energy result is not architecture-neutral. The break-even, the fidelity, the circuit depth you can actually run before noise swamps the signal, all of it depends on which physical platform you build on. The quantum industry has not converged on a single qubit the way classical computing converged on the transistor. That divergence is the investment thesis. There can be more than one winner here. IBM, Google, IonQ, Quantinuum, QuEra, PsiQuantum, they will all win different layers of the stack and different workloads. The question is which architecture wins the energy layer, specifically. That is a narrower question with a narrower answer.

The field splits across several major modalities, each with a distinct profile of strengths and tradeoffs:

| Modality | Leading players | Strength | Key Tradeoff |
|---|---|---|---|
| Superconducting | IBM, Google, Rigetti | Fast gate speeds, highest physical qubit counts, mature fabrication | Limited connectivity, cryogenic cooling, heavy error-correction overhead |
| Trapped ion | IonQ, Quantinuum | Highest gate fidelity, longest coherence, all-to-all connectivity | Slower gate speeds, laser-control complexity at scale |
| Neutral atom | QuEra, Pasqal, Atom Computing | Fastest scaling on raw atom count, flexible geometry | Earlier on fidelity and control maturity |
| Photonic | PsiQuantum, Xanadu | Room-temperature operation, CMOS-compatible silicon path | Probabilistic gates, large component counts to reach scale |
| Silicon spin | Diraq, Quantum Motion | Strongest manufacturing scalability path on existing fabs | Lags on qubit count today |

*Source: Futurum Research, 2026*

Two axes decide the energy argument specifically. Connectivity and qubit quality. Superconducting qubits are fixed in place and can only interact with near neighbors, which means more physical operations to do the same logical work, more depth, more error, and more energy spent compensating for the topology. Lower gate error means fewer physical qubits burned on error correction to produce one usable logical qubit, which is the largest hidden energy tax in any quantum system. Whichever architecture compounds both wins the energy layer.

This is offense versus defense. Modalities chasing the qubit-count scoreboard are playing offense on the metric the market currently rewards. That metric may not survive contact with the energy question. Modalities compounding quality, connectivity, and circuit efficiency are playing the longer game. The energy-to-solution result is a signal that the longer game is the one that pays. And it is not an accident that the result was produced on the modality it was produced on.

## Why This Is an Ion Trap Story, and an IonQ Tailwind

The energy-to-solution result was produced on an IonQ Forte Enterprise trapped-ion system. That is not incidental. Ion trap was the right platform to ask the energy question on in the first place.

### The Architectural Fit

Trapped-ion qubits are individual atoms held in electromagnetic fields and manipulated with lasers. Because they are natural rather than synthetic, they are identical to one another and exceptionally stable, which yields the lowest gate error rates and longest coherence times of any quantum technology. Critically, they offer all-to-all connectivity. Any qubit can interact directly with any other, with no routing penalty.

Now connect those properties back to the energy argument. The shallow, high-fidelity circuits that produced the favorable energy scaling are exactly the circuits ion trap runs best. All-to-all connectivity means fewer operations to express the same algorithm, which means shallower circuits, which means less energy and less accumulated error. High fidelity means less of the system burned on error-correction overhead. Ion trap is good at the energy question by design, not by accident.

### Energy is One Leg of a Trifecta

Energy is not the whole story, and IonQ is not treating it that way. In April 2026, the company published an application-level benchmarking framework, modeled on MLPerf, that reframes how quantum systems get measured. It moves the scoreboard off qubit count and onto three metrics that map to commercial value: solution quality, Time-to-Solution, and Energy-to-Solution. That is the trifecta. High accuracy, low time, low energy. Whoever delivers all three wins the commercial workload, not just the benchmark.

The fine-tuning result is the energy leg. The framework already shows ion trap leading on the time leg. On a 36-qubit MaxCut optimization at a 0.90 approximation-ratio threshold, IonQ Forte reached a qualifying result in roughly 34 seconds against roughly 512 seconds for a leading superconducting system, which produced no qualifying samples at all above that quality bar. That is a 15x time advantage on a real workload, not a spec-sheet claim. Put the legs together and the argument compounds. Quality and time are demonstrated. Energy is the

newest data point. Reading them as a single framework, not three separate press releases, is how IonQ intends the market to evaluate quantum, and it is an evaluation built around the metrics enterprise buyers actually price.

## The Bottom Line

The quantum value debate is migrating from speed to energy. That migration favors a different set of winners than the qubit-count scoreboard did. Energy-to-solution gives the industry a measurable, scalable metric that maps directly onto the cost structure AI buyers already manage. Architecture decides who can deliver on it. The shallow, high-fidelity, fully connected circuits that produce the favorable energy scaling are ion trap's home turf.

IonQ does not win because of one paper. It is positioned to win the narrative because the architecture it bet on is the architecture the energy question rewards, and because the founding evidence was generated on its own machine. In the Hard ROI Era, the company that owns the energy framing owns the conversation. Watch the break-even number. The day it crosses into commercial qubit counts is the day this stops being a tailwind and starts being a market.

# What to Watch: Why This is a Tailwind for IonQ

IonQ is one of two commercial leaders in trapped ion, alongside Quantinuum, and the only pure-play public name in the modality. Four things compound the energy narrative into a directional tailwind.

The result ran on IonQ hardware. The first credible energy-to-solution validation in quantum AI was produced on a Forte Enterprise system. When the metric the market eventually cares about is demonstrated on your platform, that is category positioning that cannot be bought.

1. **Forte Enterprise is built for the datacenter.** With 36 qubits, Forte Enterprise is data center-ready, modular, and built for scalable hybrid workflows and enterprise deployment. The energy thesis only matters if the hardware can sit where AI workloads already live. This one is engineered to.

2. **The roadmap leans into the quality-first game.** IonQ's modular strategy links high-fidelity ion traps via photonic interconnects rather than chasing a monolithic qubit-count scale. Recent acquisitions, Oxford Ionics for high-density 2D ion traps and Lightsynq for photonic interconnects, target the exact bottleneck, scaling without sacrificing the fidelity and connectivity that make the energy case work.

3. **It maps to a metric the buyer already owns.** Enterprise and hyperscale AI buyers are already managing power as a first-order constraint. An energy-to-solution advantage is not a quantum-literacy pitch. It is a line item that buyers can model today, which shortens the distance between quantum capability and quantum procurement.

4. **IonQ is already productizing the framing.** The 24 percent and 34-qubit figures now anchor IonQ's own customer-facing materials, positioning the company not just as the platform the founding result ran on but as the entity carrying the category flag. When a vendor moves a research result onto the sales deck this fast, it is a leading indicator of intent to own the narrative.

Hold the discipline the moment demands. One result on a narrow task does not move a company's fundamentals, and the break-even point sits above the qubit counts in commercial systems today. The tailwind is in the framing, not yet the revenue. But framing is where category leadership gets won. If energy-to-solution becomes the metric the quantum-for-AI conversation organizes around, the company whose architecture is built for it, and on whose hardware the founding result was measured, starts the next phase already holding the narrative.

## Other Insights from Futurum

Five Layers of the AI Cake

From Proof of Concept to Inference ROI: Overcoming the Five Failure Modes of Production AI with Nebius Token Factory

Arm at the Center of the AI & Data Center Revolution

# About Us

## About the Authors

Daniel Newman is the CEO of The Futurum Group. Living his life at the intersection of people and technology, Daniel works with the world's largest technology brands exploring Digital Transformation and how it is influencing the enterprise. From the leading edge of AI to global technology policy, Daniel makes the connections between business, people, and tech that enable companies to benefit most from their technology investments. Daniel is a top-5 globally ranked industry analyst, and his ideas are regularly cited or shared in television appearances on CNBC, Bloomberg, the Wall Street Journal, and hundreds of other sites around the world. A 7x Best-Selling Author, including his most recent book, "Human/Machine." Daniel is also a Forbes and MarketWatch (Dow Jones) contributor. An MBA and Former Graduate Adjunct Faculty, Daniel is an Austin, Texas, transplant after 40 years in Chicago. His speaking takes him around the world each year as he shares his vision of the role technology will play in our future.

## About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

## Copyright & Use License

### Copyright Notice



### License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization including partners, resellers, external consultants, etc., in any media format

Posting on a website which is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

### Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.