CAN QUANTUM HARDWARE DIRECTLY IMPLEMENT NEURAL NETWORKS?

A first-principles feasibility, architecture, and thermodynamic assessment

Assessment date: 28 August 2026

Scope: large-scale neural inference, transformer architectures, quantum-native learning, energy, cooling, and water

Independent technical feasibility report

The established literature is distinguished from derivations and speculative proposals developed in this report. This document is not peer reviewed.

Abstract

Quantum mechanics and modern machine learning both use probability, but they use it in materially different ways. Neural-network weights are generally signed real parameters, not probabilities; quantum amplitudes are complex numbers whose squared moduli produce measurement probabilities. This distinction is not semantic. It determines which neural computations can be accelerated by interference and which apparent advantages disappear when classical data, classical weights, nonlinear activations, and classical outputs are included.

This report evaluates whether a neural network - especially a transformer used for large-language-model inference - can be implemented directly on quantum hardware so that waves accumulate weighted sums without explicitly executing every multiply-accumulate. The answer is mathematically yes but economically and architecturally conditional. A universal quantum computer can encode a vector in amplitudes and apply a block-encoded matrix or singular-value transformation. Under sparse, structured, or qRAM-style access assumptions, it can prepare a quantum state proportional to a layer output without forming every classical entry. However, an arbitrary trained model with P independently specified b-bit weights still contains Pb bits of programmable information. That information must reside in physical memory, couplers, a state-preparation circuit, a block-encoding oracle, or a compiled gate sequence. Exponential Hilbert-space dimension reduces address width, not the physical description complexity of arbitrary weights.

A second obstruction is output. Quantum computation is most compelling when the desired answer is a sample, expectation value, overlap, top item, or other compressed property. If every activation is required classically at fixed absolute precision, the readout cost is at least linear in output size; state tomography can be worse. Deep networks also require nonlinear maps. Closed-system quantum evolution is linear, so ReLU, GELU, softmax, normalization, and gating require reversible arithmetic, polynomial transformations, measurement-conditioned channels, dissipation, or postselection, each with precision and success-probability costs.

The resulting verdict is asymmetric. Running a current dense frontier transformer as a drop-in replacement on present quantum computers is not feasible. Even a future fault-tolerant machine does not automatically offer an end-to-end advantage for generic classically stored weights and full classical activations. Research is nevertheless justified in three narrower regimes: quantum-native generative models that measure a next-token sample directly; structured long-context or scientific models whose data and operators have efficient coherent representations; and reservoir, kernel, or associative-memory systems receiving data from quantum sensors. For the immediate objective of using waves to perform matrix products with low energy, classical coherent photonics and analog in-memory computing are better matched to the task than universal quantum computation.

Executive verdict

Current LLM architecture on current quantum hardware: no. Current dense LLM architecture on a hypothetical fault-tolerant computer with ideal qRAM: definable, but no credible generic energy or runtime advantage has yet been established. Structured, sample-only, quantum-data, and quantum-native models: scientifically worth researching. Near-term wave-based acceleration of ordinary neural networks: prioritize classical photonics and analog/in-memory hardware.

Contents

1. Decision summary and corrected premises

1.1 The high-level decision

The proposal contains a valid physical intuition: wave amplitudes can superpose, and detectors can return a result that depends on a sum over many paths without a digital processor separately storing every partial product. Interferometers, radio-frequency networks, analog crossbars, and quantum circuits all exploit this fact. The critical question is not whether a wave can represent a weighted sum; it can. The question is where the weights are physically embodied, how arbitrary input data are loaded, how nonlinear layers are realized, how errors and precision scale, what output is required, and how much wall-plug energy is consumed by the complete system.

For a generic dense neural layer y = phi(Wx + b), the linear part can be embedded in quantum dynamics. Yet three lower bounds survive any change of hardware: the model description must be present somewhere; a requested classical output must cross an interface; and entropy generated by measurement, reset, error correction, communication, and discarded information must ultimately be exported as heat. Quantum mechanics can change algorithmic scaling for carefully posed problems, but it does not erase these information and thermodynamic constraints.

1.2 Corrections to the motivating analogy

Motivating statement	Technical correction	Why it matters
Neural-network weights may literally be probabilities from 0 to 1.	Most learned weights are signed real or quantized values. Probabilities commonly appear after sigmoid, softmax, or an explicitly probabilistic parameterization.	Amplitude encoding is not a direct one-for-one replacement for ordinary weights.
A quantum wave stores probabilities.	A pure state stores complex probability amplitudes. Observable outcome probabilities are squared magnitudes, and relative phase drives interference.	Negative and complex interference can represent cancellations, but measurement does not reveal all amplitudes.
An LLM always selects the highest-probability token.	Generation may sample, use temperature, top-k or nucleus truncation, constrained decoding, speculative decoding, or deterministic selection.	A quantum sampler is naturally aligned with sampling one token, not necessarily with reconstructing the full logit vector.
Quantum hardware can sum without performing matrix multiplication.	The physical evolution is itself a linear transformation. It may implement the sum in one wave propagation, but the coefficients still reside in couplers, memory, phases, or a circuit.	The arithmetic can disappear as an explicit instruction stream while the hardware and I/O complexity remain.
AI power is becoming a terawatt-scale continuous load next year.	Authoritative projections are usually reported in TWh/year. The IEA base case of 945 TWh/year in 2030 equals about 108 GW average; LBNL projects 521-843 TWh/year for U.S. data centers in 2030, about 59-96 GW average. [14,15]	TWh/year and TW are different by a factor of 8,760 hours/year. The problem remains large, but units must be correct.
Cooling AI doubled U.S. water consumption.	The cited doubling is better understood as growth in data-center onsite water use, not total U.S. water withdrawals. National effects are modest in aggregate but can be severe in individual watersheds. [16-18]	Water must be evaluated by site, source, season, cooling design, and whether withdrawal or consumption is measured.

1.3 Scale of demand as of August 2026

The service scale is already large enough that modest per-token improvements matter. OpenAI reported that more than one billion people use ChatGPT. Google reported more than one billion monthly users of the Gemini app; it separately reported 2.5 billion monthly users of AI Overviews and more than one billion monthly users of AI Mode. These populations and product surfaces overlap and therefore cannot be added as distinct users. Google also reported more than 3.2 quadrillion tokens processed per month across its surfaces in May 2026. [19-21]

At this scale, the relevant systems metric is not a laboratory gate count but delivered useful work: joules per accepted output token at a specified model quality, latency percentile, availability, and utilization. Training, retrieval, networking, safety systems, image/audio generation, idle reserve, and data-center overhead must be accounted for separately rather than hidden in a single token number.

2. First-principles physics: what interference can and cannot buy

2.1 Classical linear layers and quantum state representations

A classical dense layer maps an input vector x in R^n to an output y in R^m:

```
y = phi(W x + b),        W in R^(m x n).
```

The matrix-vector product computes m inner products. Digital hardware typically performs multiply-accumulate operations, but optimized accelerators already fuse accumulation, tile data, reuse weights across a batch, and avoid materializing individual partial sums. The apparent simplicity of the algebra is not evidence that digital execution is

physically wasteful; modern performance is commonly limited by data movement, interconnect, and memory hierarchy rather than by the mathematical act of multiplication.

Amplitude encoding instead maps a normalized vector to a quantum state:

$$|x\rangle = (1 / ||x||_2) \sum_j x_j |j\rangle.$$

Only ceil(log2 n) address qubits are required for n amplitudes. This is the source of many exponential-looking dimensional advantages. It is also where misunderstandings begin. The state has n amplitudes, but an arbitrary list of n classical numbers is not prepared for free. Generic exact state preparation requires a circuit or data structure whose size is O(n); efficient preparation is available only when the vector has exploitable structure, is produced by a short quantum process, or is supplied through a physical qRAM-like oracle. [3,30]

2.2 A quantum matrix-vector product

A nonunitary matrix A can be embedded in a larger unitary U_A by block encoding:

$$(\langle 0^a| \otimes I)\, U_A\, (|0^a\rangle \otimes I) = A / \alpha, \qquad \alpha \geq ||A||.$$

Applying U_A to |0^a>|x> produces a branch proportional to A|x>/alpha. Measuring the ancillas and accepting the success branch prepares |Ax>, normalized. The success probability is

$$p_{success} = ||A x||_2^2 / (\alpha^2 ||x||_2^2).$$

Amplitude amplification can reduce the repetition cost from O(1/p_success) to O(1/sqrt(p_success)), so the state-preparation overhead contains the factor alpha ||x||/||Ax||. [2,31] Thus a matrix norm, condition number, or small success amplitude can replace the explicit dimension in the runtime. Deep composition can multiply these penalties unless each layer is carefully normalized and errors are controlled.

Quantum singular-value transformation (QSVT) generalizes this construction. Given a block encoding, a degree-l polynomial can be applied to singular values using O(l) oracle calls. This is a powerful framework for matrix inversion, filtering, Hamiltonian simulation, and polynomial nonlinear transforms in a subspace. It does not construct an arbitrary dense block encoding; it assumes one. [2]

2.3 What “the wave computes the sum” really means

In a linear optical network or a quantum circuit, an output amplitude is a coherent sum over paths. That can eliminate the need to enumerate paths in software. Nevertheless, four conservation laws of computational complexity remain:

1. Coefficient embodiment. Each independent matrix coefficient must affect some physical degree of freedom: a memory cell, programmable phase, coupling, pulse parameter, or compiled gate. A completely unstructured m by n matrix therefore requires O(mn) configuration information even if propagation occurs in one pass.
2. Input embodiment. Arbitrary classical x must be converted into optical amplitudes, voltages, or a quantum state. The energy and latency of digital-to-analog conversion, modulation, state preparation, and normalization can dominate the core operation.
3. Precision and noise. Interference estimates an analog quantity. Shot noise, thermal noise, control error, loss, decoherence, and finite calibration determine the photons, repetitions, or error-correction cycles needed per useful bit.
4. Output embodiment. A detector returns a finite amount of classical information. Producing m output numbers with b reliable bits requires at least Omega(mb) output bits, regardless of how the internal sum was formed.

The correct comparison is therefore not “one wave event versus mn multiplications.” It is a complete physical implementation versus the best complete classical implementation, including memory, converters, routing, calibration, fault tolerance, and output.

2.4 New result I: a model-description lower bound

Model-description bound (derived here)

Consider a family of models with P independently selectable weights, each stored to b bits. There are 2^(P b) possible parameter strings. An exact deterministic programmable processor that can implement every member of this family must receive enough distinguishable program information to select among them. In a quantum program register, exact deterministic programming of distinct unitaries requires orthogonal program states; therefore the program Hilbert space must have dimension at least 2^(P b), or at least P b program qubits, unless the model description is instead compiled into fixed gates, couplers, or external memory. [32] Approximation and model structure can compress this bound, but an arbitrary dense model cannot be placed in O(log P) physical storage merely by amplitude encoding.

This statement does not imply that a quantum algorithm must use one logical qubit per weight during execution. A compiled circuit may stream or reuse instructions. It means that the model’s independent information must exist somewhere in the physical system. qRAM changes the access pattern: n address qubits can query N = 2^n cells in superposition, but the N memory cells and their coherent routing still exist. [3,4]

A closely related information-theoretic view uses the Holevo bound: q qubits cannot yield more than q classical bits of accessible information per copy. Although a quantum state has continuously valued amplitudes, those values cannot serve

as an exponentially readable classical database. Repeated use of a fixed, known program state and approximate operation creates subtleties, but it does not furnish free arbitrary model storage.

2.5 New result II: the output and normalization penalty

Suppose a quantum layer prepares the normalized state |y> = y/||y||. An l-infinity tomography theorem may recover each amplitude to error epsilon with a number of state preparations logarithmic in dimension at fixed epsilon. The meaningful precision for a classical neural activation is not generally fixed amplitude error. If typical unnormalized components y_i are O(1), then ||y|| = Theta(sqrt(d)), so normalized amplitudes are O(1/sqrt(d)). Recovering each original coordinate to absolute error delta requires epsilon = Theta(delta/sqrt(d)). Substitution into an O(log d / epsilon^2) tomography cost gives

```
number of state preparations = O(d log d / delta^2).
```

This derivation does not invalidate every l-infinity tomography use: the norm may be known, only large entries may matter, task loss may tolerate dimension-dependent error, or the output can stay quantum. It does show that fixed amplitude precision must not be confused with fixed activation precision. Full classical readout can restore a linear-in-d cost even when the state itself was prepared faster. General tomography lower bounds are also linear in Hilbert-space dimension for pure states at fixed trace-distance accuracy. [5]

2.6 Nonlinearity is not supplied by unitary evolution

A closed quantum system evolves linearly: |psi> maps to U|psi>, and density operators map linearly under a quantum channel. An element-wise ReLU or GELU followed by renormalization is nonlinear in amplitudes. It cannot be an exact deterministic unitary operation on an arbitrary unknown state. Quantum implementations therefore use one or more of the following:

- Reversible arithmetic in a computational-basis representation: explicitly calculate f(x_i) in ancillas, then uncompute garbage. This can be precise but uses many logical gates and qubits.
- Polynomial approximation and block-encoding techniques: approximate f over a bounded domain, with degree increasing as the function becomes sharper or required error decreases.
- Measurement, postselection, and feed-forward: conditional states can exhibit effective nonlinearity, but low-probability branches impose repetition or amplitude-amplification costs.
- Open-system or dissipative dynamics: nonlinear behavior can appear in reduced or normalized conditional descriptions, while the underlying density-matrix evolution remains linear.
- Hybrid classical nonlinearities: measure intermediate values, apply a classical activation, and reload them, forfeiting coherence and incurring I/O.

The need for nonlinearity is not merely an implementation inconvenience. It is central to the expressive and stable operation of deep networks. A quantum model should therefore be designed around native measurement and channel operations rather than attempting to reproduce every conventional activation verbatim.

2.7 No-cloning, branching, and cached state

Residual branches and attention reuse are easy classically because bit strings can be copied. An arbitrary unknown quantum state cannot be cloned. [33] Quantum algorithms can route, entangle, or recompute known prepared states, and error correction does not violate no-cloning because it encodes information redundantly without creating independent copies. Still, residual connections, fan-out, key-value reuse, and speculative branches require careful reversible constructions or repeated state preparation. The cost is often hidden when an algorithm assumes multiple coherent calls to a state-preparation oracle.

3. Mapping neural-network and transformer operations to quantum circuits

3.1 Dense and convolutional layers

For an amplitude-encoded vector, a dense linear layer is the most direct use of block encoding. Orthogonal or unitary layers are easier: they can be compiled as quantum gates, with a generic d by d unitary requiring O(d^2) continuous parameters and a comparable number of elementary rotations in standard decompositions. A general rectangular or nonunitary W requires dilation, linear combination of unitaries, singular-value transformation, or postselection.

Convolution can exploit sparsity and translation structure. A local kernel corresponds to a sparse Toeplitz operator; Fourier transforms diagonalize circular convolution. Both quantum and classical algorithms benefit. A claimed quantum advantage must therefore compare against FFTs, Winograd methods, low-rank kernels, sparsity, and specialized analog hardware. The most credible quantum use is not generic image convolution with classical pixels, but convolution-like operations on states already generated by a quantum sensor or simulation.

3.2 Transformer decomposition

```
Q = X W_Q,   K = X W_K,   V = X W_V
A = softmax((Q K^T) / sqrt(d) + mask),   H = A V
```

```
FFN(h) = W_2 GELU(W_1 h), followed by residual addition and normalization.
```

Every component has a formal quantum analogue, but each stresses a different resource. Projection matrices require weight oracles. The attention score matrix requires overlaps or products. Softmax is an element-wise exponential plus row normalization. The feed-forward network requires two dense matrices and a nonlinear activation. Layer normalization requires mean, variance, reciprocal square root, and rescaling. Residual connections require coherent addition or re-preparation.

3.3 Self-attention and long context

A quantum computer can estimate an overlap between normalized query and key states without explicitly multiplying all coordinates, provided both states are efficiently prepared. A swap test estimates squared overlap with O(1/epsilon^2) samples; coherent amplitude estimation can improve precision dependence toward O(1/epsilon) under deeper fault-tolerant circuits. This can be useful when one needs one global statistic or one selected key. It does not produce the entire N by N attention matrix cheaply.

The most plausible attention speedup is a search or sampling primitive: prepare a superposition of keys, encode a similarity-dependent amplitude, and amplify or sample high-similarity indices. Under coherent oracle access, Grover-type methods can reduce an O(N) scan to O(sqrt(N)) queries. The qRAM, normalization, score gap, and state-preparation costs remain. If all N scores or all N output vectors are requested, the advantage contracts or disappears.

Autoregressive inference is more favorable than full-sequence training in one respect: at each step, only one new token must be produced. Classical systems maintain a key-value cache and compute one new query against N cached keys. A quantum associative attention module might therefore target one sampled retrieval or one weighted aggregate per token rather than materializing an attention matrix. This is a narrower and more defensible research goal than an end-to-end quantum replacement of every transformer layer.

3.4 Softmax, GELU, normalization, and precision

Softmax can be interpreted as Gibbs normalization. Quantum Gibbs-state preparation and polynomial amplitude transformations can approximate exponentials, but complexity depends on the dynamic range of scores, approximation degree, success probability, and partition function. A sharp distribution can be easier to sample than to reconstruct, yet small changes in logits near a decision boundary can require high precision.

GELU and normalization can be approximated over bounded intervals. The quantum-transformer proposal of Guo et al. constructs element-wise functions of block-encoded matrices and a complete single-layer output state. Its informal complexity is approximately O-tilde(sqrt(N) d log^2(1/epsilon)) oracle uses under assumptions including efficient block encodings and favorable matrix norms. For multiple layers and all tokens, it reports approximately O-tilde(k N^(3/2) d). [7] These are query complexities, not gate counts or wall-clock/energy estimates.

3.5 Classical output, vocabulary projection, and token sampling

A conventional LLM forms logits z = W_vocab h over a vocabulary of size V, then samples or selects a token. Reconstructing all V logits on a quantum device incurs output cost. A quantum-native decoder can instead implement a Born distribution over a log2(V)-qubit output register and measure one token. This is one of the strongest architecture-level matches between quantum measurement and generative inference: the service ultimately needs a token sample, not necessarily a classical list of all probabilities.

The challenge moves to training and controllability. The measured distribution must match the target conditional distribution over many contexts; safety constraints, log-probability evaluation, beam search, exact constrained decoding, and calibration may require more than one sample. Nonetheless, avoiding full-logit tomography is a legitimate objective for a quantum-native model.

3.6 Training is harder than fixed inference

Training a P-parameter model conventionally produces P gradients or parameter updates. If those updates are exported to a classical optimizer, an Omega(P) interface cost returns. Parameter-shift gradients typically require at least two circuit evaluations per parameter per batch, multiplied by shots; stochastic estimators reduce evaluations but increase variance. Quantum gradient algorithms can improve oracle-query complexity in specialized coherent models, but they assume reversible loss oracles and still face precision, loading, and model-update costs.

Variational circuits can suffer barren plateaus in which gradient variance decreases exponentially with qubit count, circuit expressivity, depth, global cost functions, or noise. [6] Structured local costs and problem-informed circuits can mitigate this, and dissipative or convolutional architectures may avoid specific plateau theorems. It remains unsafe to assume that a circuit expressive enough to match a large language model will also be trainable at scale.

For this reason, the most credible early program is fixed or lightly tuned quantum inference using a classically trained, aggressively structured model; quantum-native training should initially use small output spaces, local losses, layerwise objectives, and in-situ parameters that avoid full gradient readout.

3.7 Operation-by-operation assessment

Transformer primitive	Quantum construction	Conditional advantage	Dominant obstruction

Transformer primitive	Quantum construction	Conditional advantage	Dominant obstruction
Embedding/state load	Amplitude, angle, basis, or qRAM encoding	Logarithmic address register	Generic state preparation or qRAM hardware is O(data size).
Dense projection	Unitary compilation, block encoding, LCU, QSVT	Can prepare |Wx> without listing entries	Arbitrary W must be stored/compiled; normalization and success amplitude.
Dot product	Swap/Hadamard tests, amplitude estimation	Dimension-independent query count after state preparation	State preparation and precision; one scalar per measurement task.
Attention search	Amplitude amplification over keys	Potential O(sqrt(N)) queries for one item/sample	qRAM, score gap, coherent oracle, and repeated autoregressive calls.
Softmax	Gibbs preparation or polynomial amplitude transform	Direct sampling may avoid full vector	Partition function, dynamic range, polynomial degree, postselection.
GELU/ReLU	Reversible arithmetic or polynomial transform	Possible while remaining coherent	Quantum dynamics is linear; accuracy/gate cost.
Layer norm	Coherent mean/variance arithmetic	No classical intermediate needed	Precision, reciprocal/square-root circuits, ancillas.
Residual/fan-out	LCU, controlled routing, recomputation	Coherent addition possible	No-cloning and state-preparation reuse.
Vocabulary output	Born measurement or amplitude search	One sampled token can be native	Training/calibration; constrained decoding; log-prob access.
Backpropagation	Parameter shift, adjoint/coherent gradients	Special oracle-model speedups	Shots, barren plateaus, P-parameter update interface.

4. Assessment of the research literature and hardware state

4.1 Quantum linear algebra: powerful but output-restricted

The HHL linear-system algorithm is the canonical example. For a sparse, well-conditioned N by N matrix with efficient oracle access, it prepares a state proportional to A^(-1)b in time polynomial in log N and the condition number. The original paper explicitly targets expectation values associated with the solution rather than outputting the full solution vector. [1] This output restriction is not a footnote; it is the pattern that makes exponential dimensional compression possible.

QSVT and block-encoding methods significantly generalize the toolbox and improve polynomial dependencies. [2] They provide the right language for a fault-tolerant quantum neural layer. Their performance is meaningful only after specifying: how the input state is prepared; how the matrix oracle is physically realized; the block-encoding scale alpha; the condition numbers and spectral ranges; the degree required for nonlinear approximations; the success probability; and the requested output.

4.2 qRAM and the input model

The bucket-brigade qRAM architecture uses O(log N) active switches per query to address N cells in superposition, but the memory still contains O(N) cells and a coherent routing tree. [3] Robustness analyses show that large qRAMs can face stringent error requirements for algorithms such as quantum search. [29] No large, fault-tolerant, high-bandwidth qRAM suitable for billion-parameter models exists today.

A fair classical comparison must grant analogous sampling/query access when a quantum algorithm assumes it. Tang's dequantization of a quantum recommendation algorithm showed that strong l2 sampling access can give a classical randomized algorithm only polynomially slower than the quantum method, eliminating a claimed exponential advantage. [4] This does not prove that all quantum machine learning dequantizes. It does require that every proposal state its data-access model and compare to classical algorithms with the same structure.

4.3 Quantum transformer proposals

Guo et al. present the most direct fault-tolerant construction of transformer components located in this review. The algorithm prepares an amplitude encoding of one token's transformer output. Under normalized weight matrices, input norm scaling, and efficient block encodings, the stated single-layer query cost is O-tilde(sqrt(N)d log^2(1/epsilon)); a multilayer architecture that obtains all token vectors and reloads them is O-tilde(k N^(3/2)d). The authors explicitly note qRAM/input assumptions, dense-weight costs without qRAM, tomography, and that training is future work. [7]

The proposal is important because it converts a vague "quantum transformer" idea into explicit subroutines and norms. It does not yet establish an end-to-end practical speed or energy advantage. In the no-qRAM dense-weight construction discussed in its supplement, loading O(d^2) elements takes O(d^2) time and ancillas, and the block-encoding factor can scale as O(d). For multiple layers, classical vectors are reconstructed and reloaded, which reintroduces output/input costs. The report's normalization-precision derivation in Section 2.5 further narrows the regime in which l-infinity tomography remains cheap at meaningful activation accuracy.

Quixer is a different, quantum-native transformer-like model based on linear combinations of unitaries and singular-value transformation. It was tested by classical simulation on a small practical language-modeling task and reported

performance competitive with an equivalent classical baseline, with quantum resource estimates. [8] It is evidence that transformer-inspired inductive biases can be translated into quantum circuits; it is not evidence of LLM-scale quantum advantage.

4.4 Variational quantum neural networks, kernels, and reservoirs

Variational quantum neural networks use parameterized circuits as trainable feature maps or models. They are experimentally accessible at small scale and can represent functions that are difficult to simulate in selected settings. Their central unresolved questions are trainability, noise, data loading, generalization, and whether the same structure that avoids barren plateaus also makes the circuit classically simulable. [6,27]

Quantum kernels can estimate overlaps in a high-dimensional feature space; a useful advantage requires a feature map that is both hard to simulate classically and well aligned with the data. Concentration can make kernels nearly constant in large Hilbert spaces, and classical kernel approximations are strong. These methods are better viewed as specialized small-data learners than as direct replacements for dense deep networks.

Quantum reservoir computing uses fixed many-body dynamics as a high-dimensional nonlinear temporal feature generator, with only a classical readout trained. An experimental correlated-spin reservoir reported high-accuracy temporal prediction with a nine-spin platform. [26] Reservoirs avoid deep parameterized circuits and are a credible niche for quantum sensors, dynamical systems, and edge temporal processing, though they do not provide the programmable knowledge capacity of a large language model.

4.5 Hardware status as of 28 August 2026

Quantum hardware has crossed important error-correction milestones but remains many engineering orders from an AI-scale fault-tolerant system. Google's Willow experiment demonstrated below-threshold surface-code memories, including a distance-7 code using 101 physical qubits with a logical error rate of 0.143 percent per correction cycle. [9] This is a logical-memory result, not a large register of universal logical qubits capable of billions of deep gates.

Quantinuum reported, in a vendor account of a 2026 result, up to 94 error-detected or 48 error-corrected logical qubits from 98 physical trapped-ion qubits, and used 64 error-detected logical qubits for a magnetism simulation. [11] Encoding definitions and accepted/discarded runs matter when comparing such counts. IBM's public roadmap targets a fault-tolerant system in 2029 and is building modular cryogenic infrastructure; roadmap targets are not current capabilities. [10]

No current platform provides the combination required by a full transformer: large fault-tolerant logical memory, high-rate non-Clifford gates, coherent high-bandwidth access to tens of gigabytes of model data, fast state preparation, low-latency classical decoding, and energy-efficient continuous service. Hardware progress makes algorithmic resource accounting timely, not deployment imminent.

5. Resource estimates for a current dense model

5.1 A 70-billion-parameter reference model

Use a dense 70-billion-parameter decoder model as an illustrative scale, not as a claim about any proprietary system. With 4-bit stored weights, the raw parameter payload is

```
70 x 10^9 weights x 4 bits = 2.8 x 10^11 bits = 35 GB.
```

The address of one of 70 billion weights needs only about 36 qubits; addressing individual bits needs about 39. This logarithmic address width is often presented as exponential compression. The memory itself still needs 280 billion reliable bit cells or an equivalent compiled description. A fault-tolerant quantum oracle must also route queries coherently and return values with sufficient precision while not decohering the address superposition.

For a dense batch-one token step, the weight matrices contribute on the order of one multiply-accumulate per parameter, roughly 70 billion MACs, excluding attention, key-value cache movement, embeddings, normalization, and routing. Mixture-of-experts models use only an active subset, and batching amortizes weight movement. The estimate is deliberately transparent so that different model architectures can be substituted.

5.2 Why logical-qubit counts alone are misleading

Amplitude encoding of an activation with d = 16,384 components uses 14 address qubits, plus ancillas. A context of N = 1,048,576 tokens needs 20 address qubits. These numbers sound small. They do not count the qRAM cells, block-encoding workspaces, arithmetic registers for fixed-point values, surface-code patches, routing, magic-state factories, error decoders, or repeated state preparations. A practical quantum resource estimate must report all of the following:

- logical data and ancilla qubits;
- T-count, T-depth, Clifford count, and measurement depth;
- block-encoding and state-preparation circuit costs, not merely oracle calls;
- physical qubits per logical qubit at a target algorithmic failure probability;
- magic-state throughput and factory footprint;
- qRAM/memory cells, interconnect, and coherent bandwidth;
- classical syndrome decoding throughput and control electronics;
- wall-clock latency, repetition/postselection rate, and wall-plug power.

No published quantum-transformer resource estimate currently specifies this complete stack for a frontier-scale model. Query complexity is scientifically useful, but it is insufficient for an energy or deployment decision.

5.3 New result III: full-layer classical output imposes a bandwidth floor

Classical interface bound (derived here)

If an inference layer must return M activation values to a classical system at b reliable bits each, the interface must transmit at least M b bits. At interface bandwidth B bits/s, latency is at least M b / B; at interface energy e_IO joules/bit, output energy is at least M b e_IO. A quantum algorithm can beat this only by retaining the state coherently for later quantum processing, by returning a sparse/compressed answer, or by sampling a small output. Therefore an end-to-end quantum network should not repeatedly cross the quantum-classical boundary between conventional layers.

For a sequence activation matrix with N tokens and d features, full output is Nd values. Even if a quantum circuit creates the normalized state in polylogarithmic time under an oracle model, producing all Nd fixed-precision values cannot be polylogarithmic. The requirement is avoided only when subsequent layers remain quantum or the task needs a much smaller statistic.

5.4 Error accumulation and fault tolerance

A useful LLM token requires a long sequence of operations with a very low aggregate failure probability. If a logical circuit uses G gates and the acceptable probability of an uncorrected fault per token is eta, a crude union-bound target is p_L << eta/G. For G in the billions or more, p_L may need to be around 10^(-12) or lower, depending on fault correlations and verification. Current logical-memory error rates are far above this, though code distance can in principle suppress errors below threshold.

Surface-code suppression costs physical qubits and cycles. Non-Clifford arithmetic and QSVT polynomials consume magic states. Deep amplitude estimation trades fewer samples for longer coherent circuits. A quantum algorithm that reduces asymptotic queries but requires enormous T-depth may lose on latency and energy. The appropriate resource metric is therefore spacetime volume - physical qubits times code cycles - plus full-system overhead.

5.5 Conditional complexity versus realized complexity

Layer of accounting	Typical attractive statement	What must be added
Mathematical dimension	n qubits represent 2^n amplitudes.	State-preparation description, normalization, and requested observable.
Oracle query	One block-encoding call applies A in polylog dimension.	Circuit for oracle, memory cells, routing, alpha, precision, and success probability.
Logical circuit	O(G) fault-tolerant gates.	Code distance, magic states, decoder, routing, and failure budget.
Physical hardware	Microwave/laser gate energy is small.	Refrigeration, control, readout, calibration, idle, networking, and repetitions.
Application	Quantum output state obtained.	Classical token, accuracy, latency, service availability, and retraining.

6. Energy, waste heat, refrigeration, and water

6.1 Thermodynamics: reversible logic is not a free service

Landauer's principle gives the minimum heat for irreversible erasure of one bit at temperature T:

$$E_L = k_B T \ln 2 = 2.87 \times 10^{-21} \text{ J at } 300 \text{ K}.$$

This is a lower bound on logically irreversible erasure, not on every multiply, memory access, photon, or quantum gate. A unitary quantum gate is logically reversible and can in principle dissipate far less than a conventional irreversible operation. An operational inference service nevertheless measures, resets, corrects errors, discards ancillas, decodes syndromes, updates classical control, and emits outputs. These processes export entropy. Today's energy use is many orders above the Landauer floor because speed, noise margin, communication, and device physics dominate.

As an illustration, one million physical qubits measured once per microsecond would generate order 10^12 raw syndrome bits per second. Erasing that many bits at the 300 K Landauer limit would dissipate only about 2.9 nanowatts. Actual readout chains, analog-to-digital conversion, data transport, real-time decoding, pulse generation, and refrigeration would consume vastly more. This gap shows both the long-run thermodynamic opportunity and the irrelevance of quoting Landauer alone for near-term engineering.

6.2 Classical 70B inference: transparent sensitivity calculation

The largest batch-one cost is often moving weights, not executing arithmetic. The following table is a sensitivity calculation, not a benchmark of a specific accelerator. It assumes all 280 billion stored weight bits are fetched once per token. Real systems change this through batching, caching, compression, sparsity, and multi-chip communication.

Assumption	Low	Central	High	Consequence
System-level weight movement energy	2 pJ/bit	5-10 pJ/bit	20 pJ/bit	0.56, 1.4-2.8, or 5.6 J/token for weights
Low-precision MAC energy	0.05 pJ/MAC	0.1-0.5 pJ/MAC	1 pJ/MAC	0.0035, 0.007-0.035, or 0.07 J/token for 70B MACs
Facility/network/KV/idle overhead	Small at high utilization	Comparable fraction	Dominant at low utilization	Widens system result beyond core estimate
Illustrative delivered range	about 0.6 J/token	about 2-5 J/token	about 8+ J/token	Not universal; batching can be lower and long-context/reasoning can be higher

The calculation identifies a design target: reducing weight traffic and improving utilization can matter more than reducing the arithmetic energy by another factor of ten. This favors in-memory compute, larger batches, model compression, structured matrices, local edge models, and optical/analog cores whose coefficients remain resident.

6.3 Global inference scenarios

For U users, q queries per user per day, t generated-and-processed tokens per query, and E_tok joules per token, annual energy is

```
E_year [TWh] = U q t E_tok 365 / (3.6 x 10^15).
```

Scenario	Annual inference energy	Average power
1B users; 5 queries/day; 500 tokens/query; 5 J/token	1.27 TWh/year	0.145 GW
3B users; 5 queries/day; 500 tokens/query; 5 J/token	3.80 TWh/year	0.434 GW
3B users; 5 queries/day; 500 tokens/query; 50 J/token	38.0 TWh/year	4.34 GW
3B users; 10 queries/day; 1,000 tokens/query; 10 J/token	30.4 TWh/year	3.47 GW

These scenarios are not forecasts. They omit training, search/retrieval, image/video/audio generation, safety and ranking models, storage, networking, redundancy, idle headroom, and edge robots. They show that per-token energy and workload definition dominate the total. Billions of users do not automatically imply a terawatt of continuous inference load; a tenfold energy or token-count change changes the answer tenfold.

For context, the IEA estimated global data-center electricity use at about 415 TWh in 2024 and projects about 945 TWh in 2030 in its base case. LBNL's June 2026 reference case projects U.S. data centers at 649 TWh in 2030, with a 521-843 TWh uncertainty range. [14,15] These totals cover far more than language-model inference.

6.4 Heat rejection

Almost all electrical energy entering a data center ultimately becomes heat in the building or its electrical supply chain. To first order, a facility drawing P_electric rejects approximately P_electric of heat, with location and timing shifted by chilled-water loops, evaporative cooling, heat recovery, or remote electricity generation. A 1 GW continuous computational load is therefore also approximately a 1 GW heat-rejection problem.

Power usage effectiveness (PUE) multiplies IT energy by facility overhead. Water usage effectiveness (WUE), often expressed in liters per kWh, converts facility energy to onsite water consumption for a specified cooling system. Neither is a universal constant. LBNL found workload-level water use varying by more than 10,000-fold due to server efficiency, grid water intensity, utilization, cooling, climate, inactive capacity, and refresh cycle. [17]

6.5 Quantum refrigeration and control energy

For a refrigerator moving heat Q_c from cold temperature T_c to ambient T_h, the ideal Carnot work satisfies

```
W_min / Q_c = T_h / T_c - 1.
```

At 300 K ambient and 10 mK, the ratio is approximately 29,999. Removing 10 microwatts at the mixing chamber therefore requires at least 0.30 W even in an ideal reversible refrigerator; 100 microwatts requires at least 3.0 W. Real dilution refrigerators require substantially more wall power because of finite efficiency, staged cooling, pumps, compressors, wiring heat leaks, radiation, and warm electronics. At 4 K, the Carnot ratio is still 74.

Cryoelectronics studies report W-class cooling at 4 K but only tens to hundreds of microwatts at 10-20 mK, and demonstrated cryo-CMOS control blocks can dissipate milliwatts per actively controlled physical qubit at 4 K. [13] Consequently, moving control closer to qubits reduces cable count but creates a severe cooling-power constraint. Superconducting logic and aggressive multiplexing may help; they do not eliminate readout, decoding, or warm-stage energy.

A 2026 full-system analysis emphasizes that NISQ energy can be dominated by repeated sampling from error mitigation, whereas fault-tolerant energy shifts toward physical-qubit overhead, continuous syndrome processing, and magic-state resources. [12] A quantum speedup must be large enough to overcome this fixed and scaling overhead.

6.6 Break-even energy-throughput requirement

Let a complete quantum system draw P_Q watts while delivering R accepted tokens per second. It beats a classical system consuming E_C joules per token only if P_Q/R < E_C, or R > P_Q/E_C. The following values ignore differences in model quality and latency and therefore represent only an energy threshold.

Quantum system power	Classical 1 J/token	5 J/token	10 J/token	50 J/token
1 MW	1,000,000 tok/s	200,000 tok/s	100,000 tok/s	20,000 tok/s
10 MW	10,000,000 tok/s	2,000,000 tok/s	1,000,000 tok/s	200,000 tok/s

A fault-tolerant machine with low utilization would be especially disadvantaged because refrigeration and control consume power while useful circuits wait. High-throughput batching may be difficult for quantum algorithms that require long coherent, sequential autoregressive state evolution. Energy advantage is therefore not implied by a low-energy qubit gate.

6.7 Water: correcting the national-scale claim

The 2024 LBNL data-center report estimated approximately 66 billion liters of direct onsite U.S. data-center water consumption in 2023 and scenarios around 145-275 billion liters in 2028. [16] This is serious growth, especially where facilities cluster in water-stressed basins. It does not mean total U.S. water consumption doubled. USGS estimated total U.S. withdrawals at 322 billion gallons per day in 2015 - about 4.45 x 10^14 liters per year. [18] Direct data-center water is well below one tenth of one percent of that withdrawal total, although withdrawals and consumptive use are not identical metrics.

National percentages can conceal local harm. Evaporative cooling consumes water rather than returning it promptly; drought timing matters; municipal potable supplies differ from reclaimed water; and electricity generation creates an indirect water footprint. A defensible AI water calculation must report the data-center location, WUE, grid mix, temporal matching, source quality, and whether the measure is withdrawal or consumption. Dry cooling reduces water but can increase electricity and capital requirements, creating a water-energy trade-off.

6.8 The closest practical embodiment of the original wave idea: classical photonics

Linear optics naturally computes matrix-vector products. Beam splitters and phase shifters implement unitary transforms; diffraction and wavelength/spatial multiplexing implement dense parallel sums; photodetectors convert optical intensity to electrical signals. No nonclassical state is required for this. A 2026 photonic matrix processor reported approximately 20 aJ/MAC of optical-core energy at 96.4 percent benchmark accuracy. Another analog optical computer projected about 500 TOPS/W at 8-bit precision, and a rack-integrated photonic tensor processor demonstrated PyTorch-connected inference. [22-24]

Optical-core numbers are not wall-plug numbers. Lasers, modulators, DACs, ADCs, transimpedance amplifiers, control, calibration, memory, and nonlinearity can dominate. A quantum-limited stochastic optical-neural-network experiment explicitly notes that end-to-end system energy is the preferred metric and that electronics can dominate once optical energy approaches a photon per MAC. [25] Even with that caveat, classical photonics aligns better than gate-based quantum computing with the goal of performing ordinary neural sums as wave propagation.

A shot-noise estimate gives scale. At 1550 nm, one photon carries about 1.28 x 10^(-19) J. Resolving an analog scalar to roughly b bits under a simple Poisson model requires signal-to-noise of order 2^b and thus order 2^(2b) detected photons. For b = 8, that is about 65,536 photons or 8.4 fJ per detected scalar before loss and electronics. Neural inference can tolerate noise and lower precision, so this is not a universal MAC floor; it illustrates why precision determines optical energy.

7. Feasibility by hardware platform and use case

Platform / approach	Natural operation	Fit to current dense neural nets	Energy and scaling issue	Feasibility judgment (2026)
Superconducting gate-based FTQC	Fast unitary gates, QSVT, amplitude estimation	Formal implementation possible after block encoding	mK refrigeration, QEC, wiring, decoding, magic states, no large qRAM	Current LLM: infeasible. Long-term structured subroutines: research-worthy.
Trapped-ion gate-based	High-fidelity all-to-all gates, long coherence	Good small variational/reservoir circuits; slow deep arithmetic	Laser/control overhead, gate latency, modular interconnect	Small quantum-native models; no credible dense LLM path now.
Neutral atoms	Large arrays, analog Hamiltonians, reconfigurable interactions	Reservoirs, kernels, optimization; gate model emerging	Optical control, loss, measurement, QEC maturity	Promising for analog/structured learning, not drop-in inference.
Photonic quantum computing	Interference, bosonic sampling, CV/GKP possibilities	Linear transform natural; deterministic non-Gaussian operations difficult	Loss, source/detector efficiency, fault-tolerance overhead	Quantum-native sampling research; ordinary MVM better done classically.
Quantum annealing / Ising	Energy minimization and sampling	Maps to Boltzmann/associative models, not feed-forward transformer	Embedding overhead, limited connectivity, thermalization/mixing	Niche optimization or sampling; no generic LLM inference.
Quantum reservoir / open system	Many-body temporal dynamics and observables	Avoids copying conventional layers; small trainable readout	Measurement shots, memory time, reproducibility, task specificity	Credible niche for quantum sensors and temporal edge data.
Classical coherent photonics	Wave-based MVM in one propagation	Direct fit to dense/structured linear layers	Converters, memory, calibration, nonlinearity, laser power	Highest near-term relevance to low-energy neural inference.
Analog in-memory electronic compute	Ohmic/current summation in crossbars	Direct fit to MVM and resident weights	Device variation, ADC/DAC, endurance, precision	High practical relevance; compare alongside photonics.

7.1 Feasibility of specific goals

Goal	Physics feasibility	Engineering feasibility	Research priority
Run an unmodified frontier dense transformer on a NISQ device	In principle tiny fragments only	No: qubits, depth, loading, noise, and output are prohibitive	Low
Run an unmodified transformer on future FTQC with qRAM	Yes, formally	Unknown-to-low: huge coherent memory/oracle and QEC stack	Medium as resource-analysis research; low as deployment bet
Quantum single-layer/one-token inference with structured operators	Yes	Conditional on efficient state/oracle preparation and small output	High
Quantum long-context associative retrieval	Potential quadratic query reduction	Requires coherent memory and favorable score gaps	High, with explicit qRAM accounting
Quantum-native next-token Born sampler	Natural measurement output	Training and fault tolerance unproven	High-risk, high-value
Quantum reservoir for temporal or quantum sensor data	Natural	Small demonstrations feasible	High for niche applications
End-to-end quantum training of a frontier LLM	No physical prohibition	No credible resource/trainability path	Low until smaller models show scaling
Wave-based low-energy ordinary matmul	Natural in classical optics/electronics	Active engineering field with prototypes	Very high

8. Speculative quantum-native learning architectures

Quantum mechanics cannot literally escape matrices: finite-dimensional states, observables, channels, and Hamiltonians are linear operators. “Not built on matrices” can therefore only mean that the model is specified by local interactions, circuits, channels, or physical geometry rather than by explicit dense arrays and repeated digital matmul. This change can still be profound. Local Hamiltonians may generate a function through dynamics; measurement may emit a sample directly; dissipation may find a fixed point; and training may tune a small set of physical couplings.

The proposals below are research architectures, not claims of demonstrated quantum advantage. Each is formulated to match a native quantum operation and to avoid full classical intermediate activations.

8.1 Proposal A: Measured Quantum Autoregressive Channel (MQAC)

Encode context c in a quantum memory or a state prepared by a structured classical-to-quantum map. Apply a parameterized fault-tolerant channel E_theta and measure a vocabulary register:

```
p_theta(v | c) = Tr[ Pi_v E_theta(rho_c) ].
```

One measurement returns the next token v. The model never forms a full classical logit vector. It can update a persistent quantum context state conditioned on v and repeat. This architecture uses the Born rule as the decoder rather than treating quantum amplitudes as hidden classical activations.

Potential advantage: output cost is O(log V) measured bits for one sampled token; interference can combine many latent paths; a persistent state may implement recurrent memory. Main barriers: loading long classical context; training probabilities for rare observed tokens; estimating log-likelihoods; maintaining coherent state across long generations; no-cloning for parallel decoding; safety constraints; and fault tolerance. A practical test would use a vocabulary of 16-256 symbols and compare exact negative log-likelihood, calibration, samples per token, and joules per accepted token against tensor-network and recurrent classical baselines.

8.2 Proposal B: Hamiltonian Feature-Transport Network (HFTN)

Replace dense layers with learned local Hamiltonians and polynomial spectral filters. For layer l, evolve under a sparse/local H_l and apply a QSVT polynomial f_l:

```
|psi_(l+1)> = Normalize[ f_l(H_l) |psi_l> ].
```

Weights are local couplings and a short list of polynomial phases rather than a dense matrix. Long-range mixing emerges from time evolution. Measurement-conditioned channels between blocks introduce effective nonlinearity. The model is “matrix-free” operationally while remaining linear-algebraic mathematically.

Potential advantage: parameter count and hardware description scale with locality rather than Hilbert-space dimension; Hamiltonian simulation is a native strength of quantum computers. The best domains are scientific sequences, molecular states, lattices, and data already represented by local quantum degrees of freedom. Main barriers: expressivity for language, trainability, spectral normalization, long simulation time, and the possibility that low-entanglement versions are efficiently classically simulable.

8.3 Proposal C: Coherent Associative Attention (CAA)

Store or generate key states |k_i> and value states |v_i> under coherent indexed access. Given query |q>, use an overlap-dependent oracle and amplitude amplification or Gibbs sampling to prepare

```
|Psi> proportional to sum_i g(<q|k_i>) |i>|v_i>.
```

Measure an index for retrieval, or coherently aggregate a value observable. This does not reproduce the full attention matrix. It uses quantum search for the operation an autoregressive model actually needs: one relevant retrieval or one compact aggregate.

Potential advantage: O(sqrt(N)) query scaling for a marked or well-separated match, and no classical list of scores. Main barriers: qRAM, repeated state preparation, overlap precision, score gaps, updating the memory, and comparison with approximate-nearest-neighbor indexes and product quantization. The go/no-go benchmark should include physical memory construction and best classical vector-search baselines.

8.4 Proposal D: dissipative quantum recurrent field

Use an open-system evolution with learned local generators:

```
d rho/dt = -i [H_theta, rho] + sum_mu (L_mu rho L_mu^dagger - 1/2 {L_mu^dagger L_mu,
                                        rho}).
```

Inputs modulate boundary drives; outputs are local measurements or a steady-state distribution. Dissipation supplies attractors, forgetting, and effective nonlinearity at the level of normalized observables. Training can tune only boundary couplings or readout, leaving the reservoir fixed.

Potential advantage: natural temporal processing with no deep compiled circuit and possible robustness to noise. Main barriers: mixing time, reproducibility, controllability, readout shots, and whether a classically simulable open-system model

matches performance. This is a plausible architecture for edge sensors and robotics where data arrive as continuous physical signals, not as web-scale textual corpora.

8.5 Proposal E: quantum Gibbs language model

Define a conditional local Hamiltonian H_theta(c, y) whose low-energy distribution assigns probability to output symbol or sequence y:

```
p_theta(y | c) proportional to Tr[ Pi_y exp(-beta H_theta(c)) ].
```

A quantum annealer, analog simulator, or Gibbs-preparation circuit samples outputs. This resembles a quantum Boltzmann machine or energy-based model rather than a transformer. Quantum tunneling or coherent transitions might improve mixing in selected rugged landscapes.

Main barriers are the hard general problem of Gibbs-state preparation, partition-function estimation, model training, embedding overhead, and no broad evidence that quantum mixing accelerates the distributions relevant to language. The architecture is more credible for combinatorial planning, structured control, and constrained generation than for open-domain text.

8.6 Proposal F: continuous-variable photonic Born network

Encode information in optical modes and apply interferometers, squeezing, displacement, non-Gaussian elements, and adaptive photodetection. Gaussian linear optics provides scalable transformations; photon counting supplies a discrete output distribution. The device can be room-temperature in parts and naturally parallel in frequency or space.

A purely coherent-state implementation is largely classical analog photonics, which may be exactly what energy-efficient inference needs. Genuine quantum advantage requires nonclassical states or non-Gaussian measurements robust to loss. Fault-tolerant photonic quantum computing may eventually support such models, but current loss, source, detector, and feed-forward requirements are severe. This path should be evaluated side by side with classical optical neural networks, not assumed quantum by virtue of using photons.

8.7 Proposal G: quantum-data foundation models

The strongest general case for quantum machine learning may be data that are quantum from birth: many-body states, quantum sensor streams, spectroscopy, scattering amplitudes, and quantum-control histories. Classical loading disappears, and the output may be an observable, phase, sample, or control action rather than a full state description. A quantum foundation model could learn reusable latent dynamics or measurement policies across families of quantum systems.

This does not immediately reduce the energy of public chatbots, but it targets a domain where quantum information is not forced through a classical bottleneck. Scientific and sensing applications may establish the first robust quantum-learning advantages and generate techniques later useful for hybrid AI.

8.8 Architecture selection principles

Principle	Design implication
Make the output a sample or small observable.	Do not perform full tomography merely to imitate a classical activation vector.
Keep data quantum or efficiently generable.	Avoid generic classical amplitude loading; use quantum sensors, procedural states, sparse/local operators, or small learned encoders.
Put parameters in local structure.	Use Hamiltonians, repeated cells, tensor networks, Fourier/Givens factors, or shared circuits rather than arbitrary dense weights.
Use measurement/dissipation intentionally.	Treat them as nonlinear computational primitives, not only as errors.
Co-design training with hardware.	Local losses, layerwise learning, fixed reservoirs, and in-situ parameters reduce gradient/readout burden.
Compare against dequantized and analog baselines.	Any quantum structure may also enable a faster classical randomized, tensor-network, photonic, or in-memory method.

9. Research program, benchmarks, and stop/go criteria

9.1 An “advantage contract” for every project

Every quantum-AI proposal should begin with a written contract specifying the problem, access model, output, accuracy, and full-system accounting. At minimum:

1. Task and baseline: exact dataset, model quality metric, context length, vocabulary, and best classical GPU/TPU, tensor-network, photonic, analog, and randomized baselines.
2. Input contract: where data and weights physically reside; time and energy to prepare states or construct qRAM/oracles; whether the cost is amortized across tokens.

3. Output contract: sample, top-k, expectation, or full vector; absolute/relative precision; number of repetitions; classical bandwidth and post-processing.
4. Quantum resource contract: logical qubits, T-count/T-depth, measurements, oracle circuit, postselection, code distance, physical qubits, decoder and magic-state throughput.
5. Service contract: accepted tokens per second, first-token latency, tail latency, availability, calibration downtime, and batch size.
6. Energy/water contract: metered wall-plug energy, cooling, control electronics, network, embodied hardware boundaries, PUE/WUE, and local grid/water context.

9.2 Recommended experiments

Experiment	Purpose	Success threshold
Compile a 10^4-10^6 parameter normalized transformer block to a logical circuit.	Expose oracle, arithmetic, QSVT polynomial, T-count, and output costs.	Complete reproducible surface-code estimate; no black-box qRAM calls.
Tiny MQAC next-token model on 16-64 logical qubits.	Test direct Born sampling, training stability, and persistent context.	Match tensor-network baseline quality; report shots and energy per token.
CAA long-context retrieval benchmark.	Test sqrt(N)-style retrieval under realistic state preparation.	Beat optimized ANN/vector database at matched recall after memory construction is included.
Quantum reservoir on sensor/robot time series.	Exploit physical dynamics and small trained readout.	Outperform classical reservoirs at equal sensing bandwidth and total energy.
Wall-plug quantum energy instrumentation.	Measure cryostat, controls, decoder, and repetitions, not pulse energy only.	Auditable joules per accepted result with uncertainty bars.
Photonic/analog transformer block.	Pursue the near-term wave-computing opportunity.	At least 10x energy efficiency at matched accuracy and useful precision, including converters.

9.3 Stop/go criteria

A project should not advance on asymptotic query complexity alone. Suggested go criteria for a practical inference program are: matched task quality; at least a 10x credible path in wall-plug energy or a decisive capability advantage; at least a 2x latency or throughput benefit after I/O; physically specified state preparation; and robustness under a realistic error-correction model. A result that only shifts O(P) work into an unspecified oracle, requires full tomography between layers, or reports only optical/qubit-core energy should be classified as exploratory physics rather than an AI infrastructure solution.

Negative results are valuable. Demonstrating that a model-description, qRAM, precision, or cooling term dominates can redirect investment toward structured models, classical photonics, or quantum-data applications. The research field benefits from pre-registered resource accounting and publication of unfavorable comparisons.

9.4 Suggested portfolio

For an energy-focused program whose objective is useful AI service rather than quantum technology for its own sake, a rational portfolio would weight near-term classical photonics, analog in-memory compute, low-bit digital architecture, and model sparsity most heavily; maintain a smaller but substantial program in fault-tolerant structured quantum subroutines and long-context retrieval; and fund a high-risk exploratory program in quantum-native Born, reservoir, and dissipative models. The relative allocation should change only when metered end-to-end results cross the stop/go thresholds above.

10. Final conclusions

There is a means to do matrix multiplication on a quantum computer. Indeed, every unitary quantum evolution is a matrix-vector multiplication, and block encodings extend this to nonunitary matrices. Interference can accumulate contributions without a digital sequence of scalar MAC instructions. That mathematical fact does not by itself make a quantum neural network efficient.

For an arbitrary current neural network, the dominant obstacle is not that matrix multiplication is intellectually simple. It is that the model has enormous independent description complexity, the data and weights are classical, the nonlinear layers are not unitary, and useful outputs are classical. The exponential state space of n qubits gives an implicit vector of length 2^n, but arbitrary preparation, arbitrary programmable weights, and full readout each restore costs proportional to the information involved. Fault tolerance adds physical-qubit, syndrome, magic-state, and refrigeration overhead.

Accordingly, a current frontier transformer cannot be run usefully on present quantum hardware. A future fault-tolerant machine could implement a normalized transformer under qRAM/block-encoding assumptions, but there is no established generic path to lower wall-plug energy or latency than highly optimized digital, photonic, or in-memory accelerators. The plausible quantum advantage region is narrower: structured operators; data already quantum or procedurally generated; very long search spaces with small outputs; and tasks whose desired result is a sample or observable rather than a full activation vector.

The most important architectural opportunity is to stop imitating a classical network. A quantum-native autoregressive model can measure a next token directly; a Hamiltonian network can learn local generators; a reservoir can exploit natural many-body dynamics; and coherent associative attention can target one retrieval rather than an attention matrix. These ideas have real scientific merit, but they must be tested against dequantized classical methods and complete energy accounting.

For the immediate energy and cooling problem, the practical answer to “encode weights as waves and observe the sum” is more likely classical photons or analog currents than fault-tolerant qubits. Quantum research remains justified because a successful sample-native or quantum-data architecture could open a qualitatively different computing regime. It should be pursued as a targeted, falsifiable program, not as an assumption that quantum parallelism automatically replaces dense matmul.

Bottom line

Worth researching: yes - for structured, sample-only, quantum-data, reservoir, and associative tasks, with rigorous full-stack benchmarks. Feasible as a drop-in engine for current LLM architecture: no on current hardware, and not yet shown to be advantageous even on hypothetical fault-tolerant hardware. Best near-term route to wave-based low-energy inference: classical photonic and analog in-memory accelerators, combined with model compression and reduced data movement.

Appendix A. Compact derivations

A.1 Program-information counting

Let each of P parameters take one of 2^b values. The model family has M = 2^(Pb) members. A deterministic exact programmable gate array with program states |p_i> implementing distinguishable unitaries U_i requires orthogonal program states for distinct exact operations under the Nielsen-Chuang no-programming result. Hence dim(H_program) >= M and q_program >= log2 M = Pb. A fixed-function compiled device escapes a separate program register only because its circuit or physical couplings encode the same information. Approximate models replace Pb with the rate-distortion description length of the model family.

A.2 Output precision from normalized amplitudes

Let y in R^d have components of order one and norm ||y|| = c sqrt(d). The state amplitude a_i = y_i/(c sqrt(d)). If tomography returns |a_i - a_i_hat| <= epsilon and the norm is known, then |y_i - y_i_hat| <= c sqrt(d) epsilon. Setting this below delta gives epsilon <= delta/(c sqrt(d)). A tomography query count proportional to log(d)/epsilon^2 becomes proportional to c^2 d log(d)/delta^2. Sparsity or a different error metric can change this result; fixed normalized-amplitude error is not fixed original-coordinate error.

A.3 Block-encoding success factor

For U_A whose top-left block is A/alpha, applying U_A to |0>|x> produces the desired ancilla-zero branch with norm ||Ax||/(alpha||x||). Direct postselection needs expected repetitions alpha^2||x||^2/||Ax||^2. Amplitude amplification reduces this to order alpha||x||/||Ax|| calls. If alpha is chosen near a Frobenius norm that grows with dimension, or if Ax is small, the apparent logarithmic address advantage can be overwhelmed.

A.4 Energy scenario identities

One TWh equals 3.6 x 10^15 J. Average GW equals annual TWh divided by 8.76. A device drawing P watts at throughput R tokens/s uses P/R joules/token. A cooling system with WUE w L/kWh and facility energy E kWh consumes wE liters onsite, subject to the definition of WUE and local operating conditions.

Appendix B. Assumptions and uncertainty

Quantity	Value used	Status / sensitivity
Reference model parameters	70 billion	Illustrative dense model; not a statement about a named proprietary model.
Weight precision	4 bits	Raw stored payload only; scales linearly with precision.
Weight movement energy	2-20 pJ/bit	Scenario range, not a measured universal constant; includes system-level variation.
MAC energy	0.05-1 pJ/MAC	Scenario range for low-precision arithmetic; excludes memory and facility overhead.
Photonic precision estimate	N_photons approximately 2^(2b)	Simple shot-noise scaling for an analog scalar; architecture and task tolerance can differ.
Quantum refrigerator temperatures	10 mK cold, 300 K ambient	Carnot lower bound only; real wall power is higher.
User scenarios	1B or 3B nominal users	Not a forecast; overlapping products and variable use make distinct active-user totals uncertain.
Water comparison	USGS 2015 withdrawals	National withdrawal benchmark; not directly equal to consumptive use or local scarcity.

References

[1] Harrow, A. W., Hassidim, A., and Lloyd, S. “Quantum algorithm for solving linear systems of equations.” Physical Review Letters 103, 150502 (2009). doi:10.1103/PhysRevLett.103.150502.

[2] Gilyen, A., Su, Y., Low, G. H., and Wiebe, N. “Quantum singular value transformation and beyond: exponential improvements for quantum matrix arithmetics.” Proceedings of STOC 2019; arXiv:1806.01838.

[3] Giovannetti, V., Lloyd, S., and Maccone, L. “Quantum random access memory.” Physical Review Letters 100, 160501 (2008); and “Architectures for a quantum random access memory,” Physical Review A 78, 052310 (2008).

[4] Tang, E. “A quantum-inspired classical algorithm for recommendation systems.” STOC 2019, 217-228. doi:10.1145/3313276.3316310.

[5] Scharnhorst, T., Spilecki, J., and Wright, J. “Optimal lower bounds for quantum state tomography.” arXiv:2510.07699 (2025).

[6] McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R., and Neven, H. “Barren plateaus in quantum neural network training landscapes.” Nature Communications 9, 4812 (2018).

[7] Guo, N. et al. “Quantum Transformer: Accelerating model inference via quantum linear algebra.” arXiv:2402.16714v3 (2025).

[8] Khatri, N., Matos, G., Coopmans, L., and Clark, S. “Quixer: A Quantum Transformer Model.” arXiv:2406.04305 (2024).

[9] Google Quantum AI and Collaborators. “Quantum error correction below the surface code threshold.” Nature 638, 920-926 (2025). doi:10.1038/s41586-024-08449-y.

[10] IBM. “Quantum 2026 - IBM Technology Atlas” and IBM newsroom announcement on modular cryogenic systems, 19 August 2026. Roadmap statements are prospective.

[11] Quantinuum. “Skinny Logic: Quantum Codes Go on a Diet.” 4 March 2026. Vendor description of associated logical-qubit work.

[12] Niu, S. et al. “Estimating the Energy Consumption of Quantum Computing from a Full System Aspect.” arXiv:2605.09580v2 (2026).

[13] Kawabata, S. “Integration and Resource Estimation of Cryoelectronics for Superconducting Fault-Tolerant Quantum Computers.” arXiv:2601.03922v2 (2026).

[14] Smith, S. J. et al. United States Data Center Energy Usage Report: 2025 Update. Lawrence Berkeley National Laboratory, June 2026. doi:10.71468/P1RP4F.

[15] International Energy Agency. Energy and AI. IEA, Paris (2025), including “Energy demand from AI.”

[16] Shehabi, A. et al. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory (2024). doi:10.71468/P1WC7Q.

[17] Lei, N., Lu, J., Shehabi, A., and Masanet, E. R. “The water use of data center workloads: A review and assessment of key determinants.” Resources, Conservation and Recycling 219, 108310 (2025). doi:10.1016/j.resconrec.2025.108310.

[18] Dieter, C. A. et al. Estimated Use of Water in the United States in 2015. U.S. Geological Survey Circular 1441 (2018).

[19] OpenAI. “From asking to doing: How the world is putting ChatGPT to work.” 6 August 2026.

[20] Google. “More than 1 billion people are using the Gemini app every month.” 11 August 2026.

[21] Google. “I/O 2026: Welcome to the agentic Gemini era.” 19 May 2026.

[22] Luan, X. et al. “Single-shot matrix-matrix photonic processor based on spatial-spectral hypermultiplexed parallel diffraction.” Nature Communications (2026). doi:10.1038/s41467-026-68452-x.

[23] Kalinin, K. P. et al. “Analog optical computer for AI inference and combinatorial optimization.” Nature (2025). doi:10.1038/s41586-025-09430-z.

[24] Meyer, B. H. et al. “Deep neural network inference on an integrated, reconfigurable photonic tensor processor.” Nature Communications 17, 3396 (2026). doi:10.1038/s41467-026-71599-2.

[25] Ma, S.-Y., Wang, T., Laydevant, J. et al. “Quantum-limited stochastic optical neural networks operating at a few quanta per activation.” Nature Communications 16, 359 (2025). doi:10.1038/s41467-024-55220-y.

[26] Hou, Y. et al. “High-Accuracy Temporal Prediction via Experimental Quantum Reservoir Computing in Correlated Spins.” Physical Review Letters 136, 120602 (2026); arXiv:2508.12383.

[27] Beer, K. et al. “Training deep quantum neural networks.” Nature Communications 11, 808 (2020). doi:10.1038/s41467-020-14454-2.

[28] Shao, C. “Quantum Algorithms to Matrix Multiplication.” arXiv:1803.01601 (2018).

[29] Arunachalam, S., Gheorghiu, V., Jochym-O’Connor, T., Mosca, M., and Srinivasan, P. “On the robustness of bucket brigade quantum RAM.” New Journal of Physics 17, 123010 (2015); arXiv:1502.03450.

[30] Suresh, K. and Suresh, S. “PyEncode: An Open-Source Library for Structured Quantum State Preparation.” arXiv:2603.28259 (2026).

[31] Brassard, G., Hoyer, P., Mosca, M., and Tapp, A. “Quantum Amplitude Amplification and Estimation.” Contemporary Mathematics 305, 53-74 (2002); arXiv:quant-ph/0005055.

[32] Nielsen, M. A. and Chuang, I. L. “Programmable Quantum Gate Arrays.” Physical Review Letters 79, 321-324 (1997). doi:10.1103/PhysRevLett.79.321.

[33] Wootters, W. K. and Zurek, W. H. “A single quantum cannot be cloned.” Nature 299, 802-803 (1982).