Tab 1

Llama 3.1 and Scaling Issues

Picture this: It's launch day for Meta's new AI, Llama 3.1. You have spent months collecting data and optimizing code to outperform OpenAI with a sophisticated model. You load up your computer, start running the training… and you see the timeline: 4,486 years. Never fear, you are part of a multibillion-dollar company, so you can get 15,999 more GPUs to lower the timeline to 3 months via parallel computing. However, the energy required to train the AI is comparable to powering a small city. This is not an exaggeration that would make sense in a sci-fi novel. This is real life, and we are hitting the limits of our GPUs' physical capabilities. There will be a point in the near future where we have so much data that we run out of GPUs to train AI models to write the next line of code or generate images more accurately.

Why do we need so much time to train these AI models? Training an AI model means we need to convert datasets into matrices. Then, once we have another one of these matrices, we multiply them together, as shown below.

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} = \sum_{k=1}^{n} x_k \begin{bmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{n,k} \end{bmatrix}$$

(Source: LinkedIn)

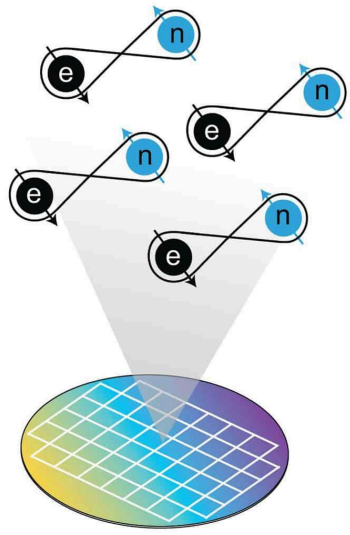
As you can see, they can get really long. For example, we might need to multiply a 1000 × 1000 matrix by another matrix of a similar size. This means about a billion calculations (FLOPs) for a single layer. Now imagine what happens when we have thousands of layers. This is when the fans sound like jet engines.

Here are a few statistics of power costs and their effect on the environment. In 2024, global data center consumption was about 415 Terawatt-Hours, which is 1.5% of global electricity use. Nearly 40% of the electricity is for cooling. For each kilowatt-hour needed for cooling, we need about 2 liters of water. Doing the math, that is about 332 billion liters of water for cooling.

We can make AI smarter. But based on the above information, throwing more GPUs at it won't work forever. We need to rethink how we process information. Enter quantum computing.

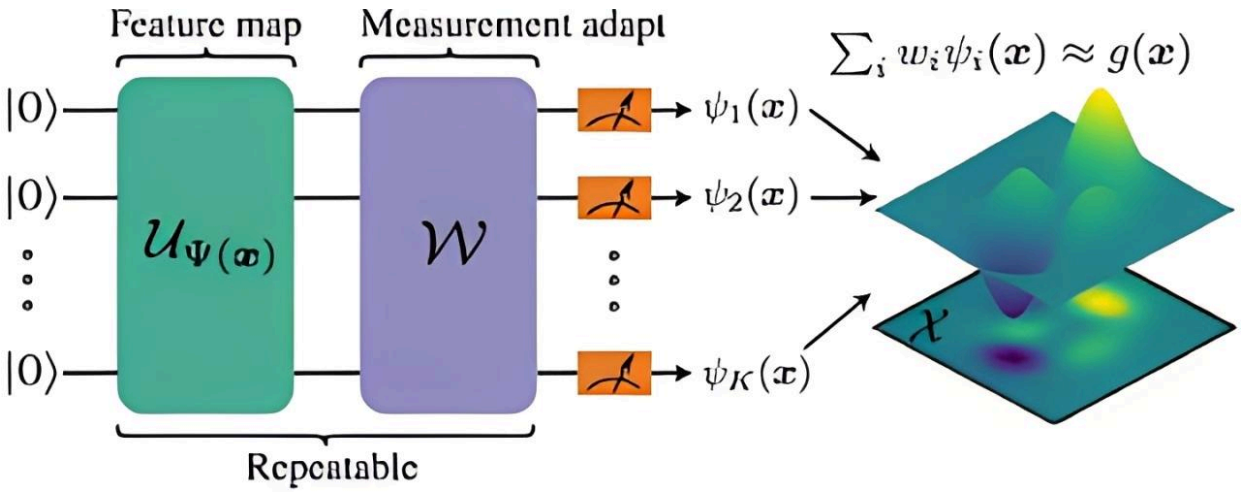
Quantum Computing

How is quantum computing different from a regular computer? Classical computers use bits, while quantum computers use something called a qubit. A qubit can be a 0 or a 1 at the same time, like deciding a coin flip while it is still in the air. This is called superposition. If qubits become linked, they can influence each other instantly, like twins. This means quantum computers can explore many solutions simultaneously.



(Source: IEA)

Remember those giant matrix multiplications? Instead of using fixed matrices, quantum computing can encode it into quantum states using something called quantum feature maps (see image below).



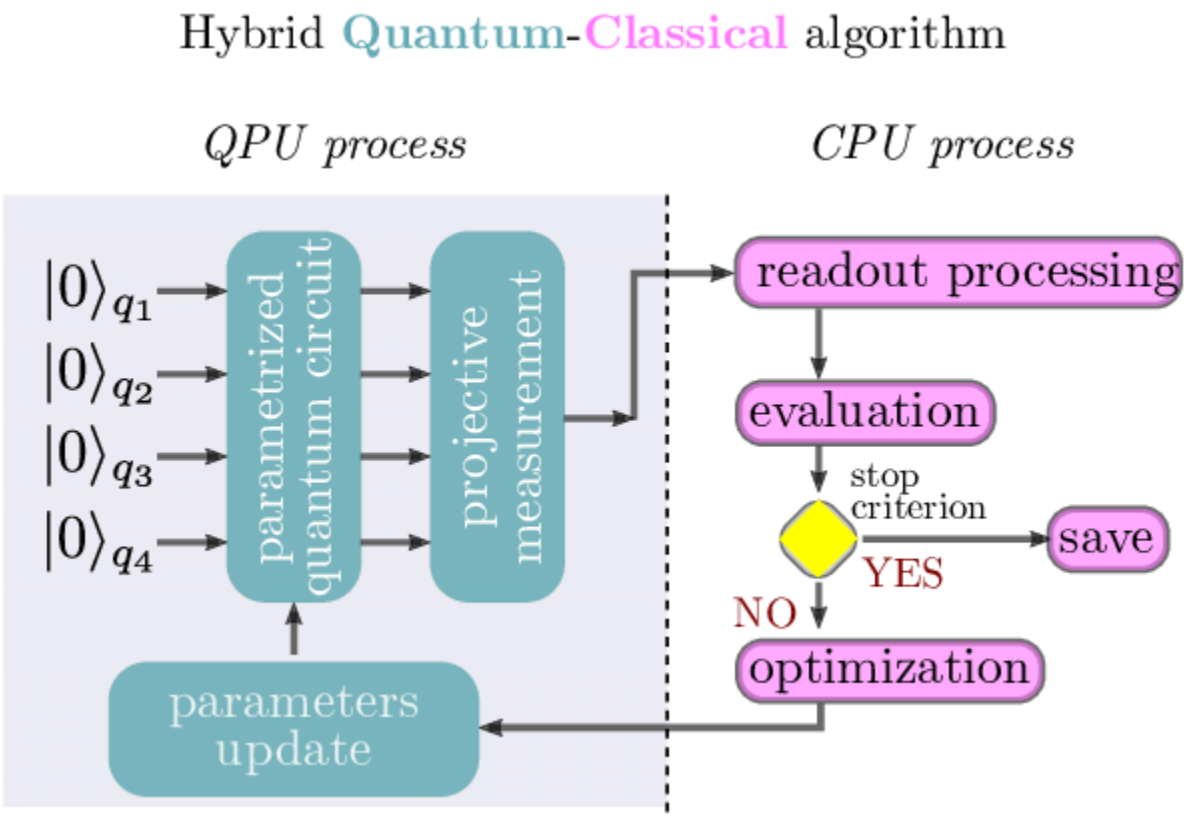
(Source: PhysicsWorld)

The data is embedded into a Hilbert space, which has more dimensions than any GPU can handle. The results are patterns that look random in the classical world are crystal clear. Imagine stacking 2D images vs. projecting the images into a 3D hologram: A much more efficient way of examining the data. Before we explore what happens if we succeed, we need to figure out what is in the way today. This all sounds amazing and all, but why do we not have this already? In 1 word: noise.

Today's Reality

Quantum computing is a powerful tool, but today's versions are far from perfect. The biggest challenge so far is noise, which are tiny disturbances to qubits because of heat, vibration, or electromagnetic interference. These issues may seem small, but they can change a qubits state and cause the system to be inaccurate. If this happens, the qubit can lose its quantum state, or it "decoheres." This is a limiting factor on the amount of time a quantum computer can run before information goes away. Error rates are another major issue. In classical computers, a transistor flipping by mistake rarely happens. However, quantum computing is based on probability, which includes a chance of error. This probability is a feature instead of a bug, and the longer the algorithm runs, the more likely it is to fail. Another issue is stability. For qubits to be stable, they need to be at near absolute zero. The temperatures make scaling hard. For example, building 1,000 reliable qubits means a lot more than simply adding more, it requires keeping every single qubit perfectly synchronized and protected.

Since we do not have perfect quantum computers, researchers combined imperfect quantum computers with perfect classical computers to make hybrid models. The quantum components handle the hard math, while the classical computer adjusts parameters and fixes errors.



(Source: ResearchGate)

Variational Quantum Algorithms (VQAs) such as Variational Quantum Eigensolvers (VQEs) and Quantum Approximate Optimization Algorithms (QAOAs) can tolerate noise, dealing with 1 of the many issues of quantum computing. VQEs are mostly used to find the lowest energy state of molecules for chemistry and material science, and QAOA solves large-scale optimization problems. Another vocab word: quantum annealing, which is a technique that uses quantum mechanics to find the best solution for complex problems (basically QAOA).

Here is a real-world example. D-Wave's Quantum Annealing System has up to 40% better performance in optimization than classical computers. The system uses 1000s of connected qubits to explore many possible combinations, and then uses a classical system to find the best results. It is not going to be the best method forever, but it shows that quantum computing can have an application in real-world tasks.

A solution that IonQ uses is trapped-ion processors to increase coherence time, so more calculations can be made for longer periods of time. Another solution comes from IBM and Rigetti, which uses superconducting qubits (qubits from superconducting circuits to do math) to make reliable gate-based systems. Gate-based systems are systems that use quantum gates/operations in the Hilbert space to change qubits and do math.

I found a researcher named Samuel Bosch, who did part of a PhD at MIT in quantum machine learning. He used MIT supercomputers to simulate quantum computing and showed that it can outperform classical computers for small datasets. In one project, he used quantum kernel methods, which are types of algorithms that map data into the Hilbert space by encoding classical data into quantum states. The result was that the model could separate data points that classical machines couldn't. This experiment showed the potential of quantum computing, which can also be shown in this quote by Jensen Huang, the CEO of NVIDIA.

> ***"NVIDIA will be a relic of the past."***
> ***- Jensen Huang***

This quote implies that even with NVIDIA producing the most powerful GPUs in the world, it will soon be replaced. The only logical possibility of that is quantum computers. He also talked about how the quantum rush is similar to the GPU rush, where companies had to compete for limited funding of something that was obviously going to be the next big thing. However, what separated NVIDIA was that it was able to leverage the need for GPUs in the gaming industry. Quantum computing has to do the same thing, which is leveraging a popular industry. Now, as promised, let's see what happens if quantum computing actually delivers in the future.

What Happens if we Succeed?

When we reach fully functional quantum AI with stable hardware and faultless qubits, we will have more possibilities than ever before in human history. Instead of using millions of GPU hours, terawatts of power, and large cooling farms, using up the limited fresh water on Earth, we can have quantum AI do the same amount of training using a small amount of energy. What can happen in months can happen in hours. Since quantum systems can represent exponentially large states, we can use less data to increase complexity. That means that the future of AI models can go from growing "bigger" with more data to smarter with relationships with less data.

With quantum AI, we can access entirely new fields of science that normal AI cannot hope to achieve today. We can simulate molecular interactions down to the electron, design new materials for batteries or superconductors, or even model biological interactions down to the molecule. We can train AI on data from the past, but we can also use simulated realities to discover physics laws, new drugs, and brand-new energy sources that don't even exist yet.

This is another technological arms race. The winner will redefine the possibilities in every field that requires predictions and math, such as research, defense, finance, and technology.

More AI models can be trained with minimal costs, increasing innovation in more fields than the ones listed above.

This power can create a societal imbalance. If only a few nations/corporations control access to high-level quantum technology, they control the next phases of innovations and intelligence. Models can be cheap to train, but only for people who have quantum access. For everyone else, it will be very expensive and impossible to replicate the sophisticated levels of the quantum class. This is why governments, startups, and research labs are all competing to change who controls the innovation.

AI isn't limited by intelligence. It's limited by hardware. Right now, we can say "add more GPUs," but that uses too much energy, water, and time. We need a different thought process.

Quantum AI is the different thought process. It mirrors the uncertainty of nature through probability. We can rebuild intelligence and innovation from the ground up in a new realm of math. Intelligence could be more efficient, more sustainable, and more human.

Sources:

https://arxiv.org/abs/2505.23860?

https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

https://arxiv.org/abs/2411.06863

https://arxiv.org/abs/2505.14295?

https://arxiv.org/abs/2504.07396?

https://arxiv.org/abs/2510.01797?

https://www.mdpi.com/2673-2688/6/8/175?

https://www.youtube.com/watch?v=sQSQBYHR0ms

https://www.youtube.com/watch?v=P7_SfxRrXTE&t=554s

https://www.youtube.com/watch?v=8I7IKrLP6Hk

https://www.youtube.com/watch?v=NqHKr9CGWJ0

https://www.youtube.com/watch?v=aFDbO0CZFto&t=2s

https://www.linkedin.com/pulse/magic-matrices-machine-learning-recommender-systems-mohamed-el-refaey/

https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

https://physicsworld.com/a/physicists-entangle-qubits-in-a-semiconductor-at-room-temperature/

https://www.researchgate.net/figure/Hybrid-quantum-classical-algorithm-for-data-driven-quantum-circuit-learning-DDQCL_fig1_330576706