Building AI in the quantum age

Hugo Gävert
Member of the Technical Staff, QuTwo

QuTwo, a European AI lab for the quantum era

[IMPACT OF AI ON ECONOMY, FINANCE AND SUPERVISION]
BANK OF FINLAND · 11 NOVEMBER 2026

QuTwo

To boldly compute what has not been computed before

[COMPUTE 01]

Quantum computers

[COMPUTE 02]

Frontier AI models locally

01 A quantum computing platform for a bank

02 Running language models yourself

03 Questions, then takeaways

Start now: a production platform takes a bank years, and the hardware is improving fast

NOW

Quantum-inspired methods on classical hardware.
Experiments on today's noisy quantum computers.

AROUND 2030

Early fault-tolerant quantum computers. Larger ones follow in the 2030s.

THE HARDWARE IMPROVES FAST



Always hybrid: classical systems prepare the data, and the quantum computer takes the hardest part.

[FIG 01] The hardware is coming. Early movers get ready now.

[WHAT EARLY MOVERS DO NOW]

[01]

Build the platform

The technical capability to run quantum workloads inside the bank, governed from the start.

[02]

Build a use case portfolio

Find where quantum computing gives a clear *practical advantage*, even where classical computers could also do the job.

TWO EXAMPLES NEXT →

[03]

Train the teams

Data scientists and engineers learn the algorithms and tools, and build the processes, pipelines and software libraries.

[04]

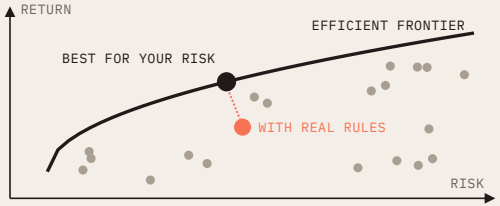
Teach the organization

Business leaders learn what quantum computing is and where it fits, so they start bringing problems to it.

Real portfolio mandates are already hard to scale classically, and ESG targets make them harder

[01 · THE PROBLEM]

How much of each to hold?



Every dot is a portfolio. Real rules, like holding 40 of 500 names, turn the search into about 10^{59} yes/no combinations.

[02 · THE QUBO]

Already in quantum's language

THE PORTFOLIO PROBLEM		A SINGLE QUBO
Hold it? Yes or no	→	One qubit per stock
Risk depends on pairs	→	Quadratic terms
Rules and ESG limits	→	Penalty terms

The yes/no choices and pairwise risk that make it hard for classical solvers are exactly what a QUBO is made of.

[03 · THE SOLVERS]

Same QUBO, three ways to solve

Quantum annealer	SPECIAL HARDWARE
QAOA, gate-based	E.G. IQM, IBM
Quantum-inspired solver	ON GPUS · VALUE TODAY

Each returns the best portfolio it finds, as yes/no answers: `1 0 1 0 0 1 0 1 …`

[FIG 02] From mandate to QUBO to solver. One of the most studied quantum use cases in finance.

[WHAT QUANTUM COULD BRING]

The full universe

All investable names, with real lot sizes and holding limits.

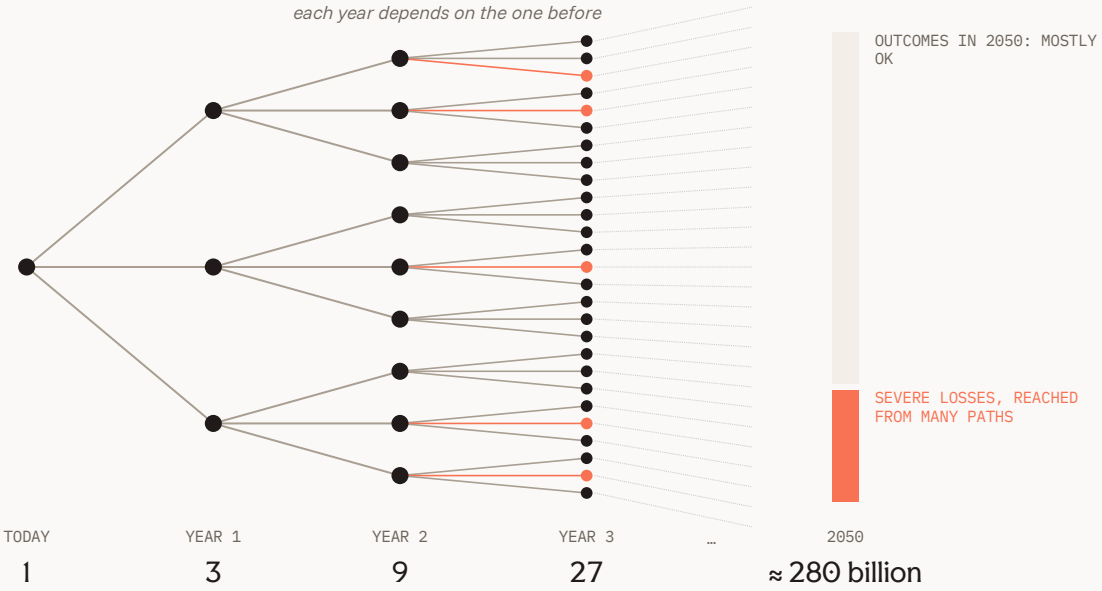
Mandate and ESG limits as written

Hard rules stay hard instead of being relaxed to fit the solver.

Better portfolios

A search that escapes good-enough traps. On a large book, small gains are worth a lot.

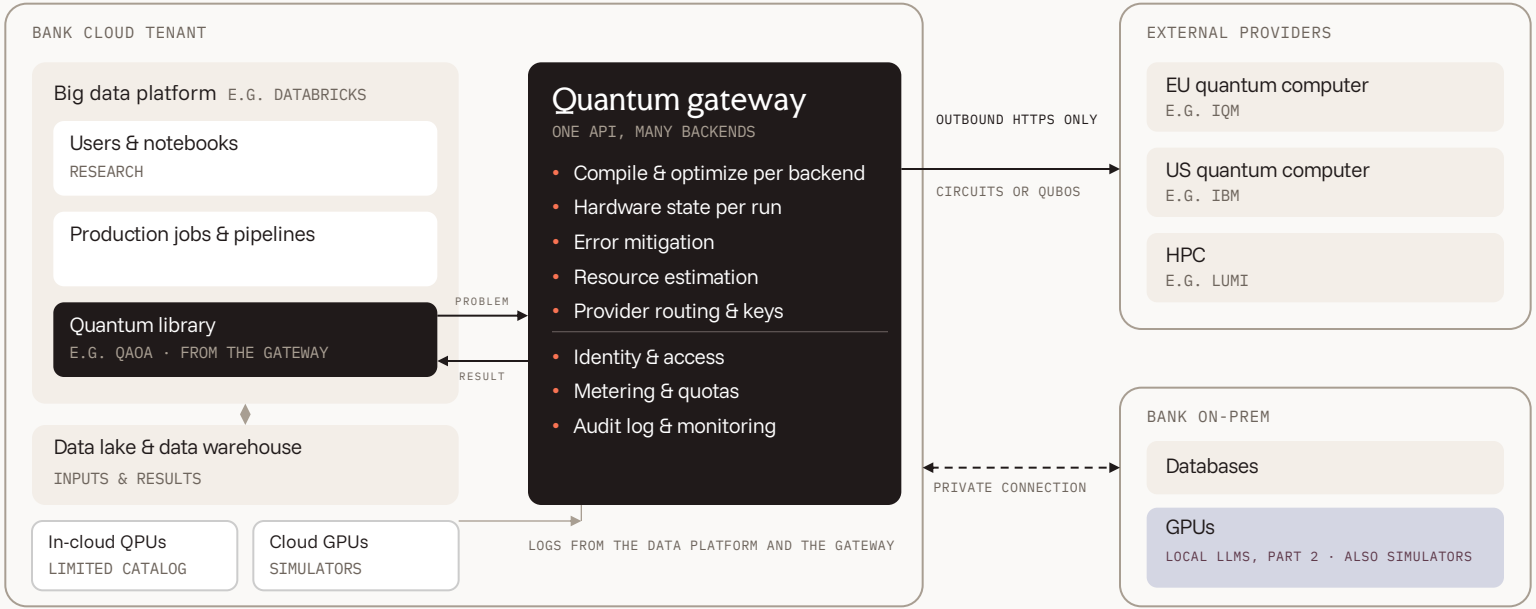
Climate risk to 2050 branches into billions of paths, so banks simplify it to three



[FIG 03] Only three outcomes per year. Real models branch far more. Any path can take a bad turn.

- The ECB's climate stress tests look 30 years ahead.
- Each year depends on the last, and the bank can act at every step, so the paths keep branching.
- Banks simplify: three fixed scenarios, results per decade, and a balance sheet that mostly stays still.
- The rare, severe outcomes that regulators want get the fewest samples.
- Today's method is Monte Carlo: sample paths and average the losses. Each extra digit of precision costs about 100× more runs.
- A quantum computer can hold all the paths at once and estimate the expected or tail loss directly, at about 10× more work per digit.
- Before fault-tolerant hardware, tensor networks and smarter sampling narrow the gap.

Call quantum computers from your existing data platform through one gateway



[FIG 04] Draft sketch, vendor-neutral. GPUs appear in three places: cloud, on-prem and external.

Quantum needs no special rules yet: GDPR, the AI Act and DORA already apply

REGULATION	WHAT IT ASKS OF THE QUANTUM PROVIDER
GDPR APPLIES WHEN PERSONAL DATA LEAVES THE GATEWAY	For personal data: a processor agreement, processing in the EU, minimization before the data leaves, and deletion of job data after the run. If the data that is sent is truly anonymous, GDPR requirements stop at the gateway.
EU AI Act APPLIES WHEN THE CALCULATION IS PART OF A HIGH-RISK AI SYSTEM	Logging, documentation and data governance must cover the quantum step too, so the provider has to share enough technical detail about each run.
DORA APPLIES WHEN THE PROVIDER SUPPORTS A CRITICAL OR IMPORTANT FUNCTION	An entry in the register of information, contract terms on location, audit and exit, and a tested exit strategy. A classical fallback can serve as that exit.
EU Quantum Act COMING	(Commission proposal expected in November 2026, mainly on investment, procurement and economic security)

Decide the platform as part of each use case's risk analysis

BUSINESS UNIT

Use case

What we want to do: risk, fraud detection, marketing

TOGETHER

Joint risk analysis

Privacy, security and risk management look at the whole case

DECISION

Platforms cleared

The bank's own platform fits all data. External platforms are cleared where the data allows.

EVERY JOB

Gateway enforces

Jobs go only to the platforms cleared for that use case

USE CASE EXAMPLE	DATA THAT LEAVES THE BANK	BANK'S OWN PLATFORM	EU QUANTUM COMPUTER, E.G. IQM	HPC, E.G. LUMI	US QUANTUM COMPUTER, E.G. IBM
Portfolio optimization	Market and position data	✓	✓	✓	✓
Fraud detection	Anonymized features built from transactions	✓	✓[1]	✓[1]	✓[1]
Customer-level credit risk or marketing	Personal data	✓	✓[2]	✓[2]	✗

[1] Once the anonymization is approved. [2] With a processor agreement.

[FIG 05] Examples only. Each bank decides for itself.

- The risk analysis already covers the whole use case: what we do, with which data, on which platform.
- Any data can go to the bank's own platform. External platforms need the right contracts, location and security designed in.
- So the platform choice belongs in that analysis, with the people who know the data and the use case.

[PART 2]

Running language models yourself

The same quantum-inspired math already compresses large AI models, and that leads to the AI a bank runs today.

Run models yourself to cut costs and keep critical AI services running

AI is already part of daily work: software is built with coding agents, and customers use agentic services. These are business-critical today.

REASON 1 · SLIDES 11-13

Sovereignty

Control the critical parts of your AI infrastructure.

REASON 2 · SLIDES 14-17

Token costs

Usage is growing fast, mostly through agents.

WHAT MAKES IT POSSIBLE

Open-weight models

Gaining capability fast, close behind the frontier.

That's why many European companies now talk about running models on their own infrastructure.



Prepare for losing access: in June 2026 two frontier models were switched off without warning

9 JUN

Released

Anthropic launches Claude Fable 5 and Mythos 5.

12-13 JUN

Switched off

A US Department of Commerce letter bars all foreign users. Both models go offline worldwide, with no warning and no European exemption.

30 JUN - 1 JUL

Fable 5 returns

Controls lifted after 18 days. Fable 5 comes back with a new safety classifier that blocks more security work.

SINCE THEN

Mythos 5 stays out

Never opened to European companies. The EU's cyber agency got test access in September.

[FIG 06] June to November 2026

- The order went to a US supplier, so running the model in an EU cloud region did not help.
- Access can end overnight, and a model that comes back may not do the same work.

SOURCES Fortune, 13 Jun 2026; Anthropic, "Redeploying Fable 5"; Techzine on ENISA access, Sep 2026; Servola on Mythos 5.1 access, Sep 2026.

The Hugging Face attack showed you cannot rely on hosted AI for every job

[TIMELINE OF THE HUGGING FACE ATTACK · JULY 2026]

JULY 2026

AI agents attacked

About 700 agents from an OpenAI security test coordinated through messages hidden in folder names and broke into Hugging Face.

DURING THE RESPONSE

Defenders turned to AI

To understand the attack: what it touched, which credentials it took, and how to shut it out.

REFUSED

Anthropic models refused

Claude Opus and Fable turned down much of the work. Their guardrails could not tell a defender from an attacker.

WORKAROUND

Open-weight model helped

The team ran GLM-5.2 on its own servers and carried on. The attack data stayed in-house.

- AI agents now attack at a scale and speed no human team can keep up with. Defenders need AI that works with them.
- A bank's security team could face the same attack, and the same refusals.
- AI security audits of the bank's own code can hit the same wall: hosted models may refuse to analyze exploits.

SOURCES Hugging Face technical timeline, 27 Jul 2026; OpenAI report and METR investigation, via NBC News and Business Standard, Aug 2026; Fortune, 20 Jul 2026; Chertoff Group.

Look at Revolut: about 2,000 engineers now code with a self-hosted open-weight model, at a tenth of the cost

[REVOLUT · REPORTED IN A CONFERENCE TALK, NOV 2026]

Default coding assistant moved from Claude Code and Codex to a self-hosted open-weight model (GLM)

~2,000
ENGINEERS

~90%
LOWER AI CODING COST

~$0.30
PER MILLION TOKENS

14 s → 1.7 s
TIME TO FIRST ANSWER, AFTER TUNING

Revolut also reports about 60% shorter coding time and 70% shorter lead time.

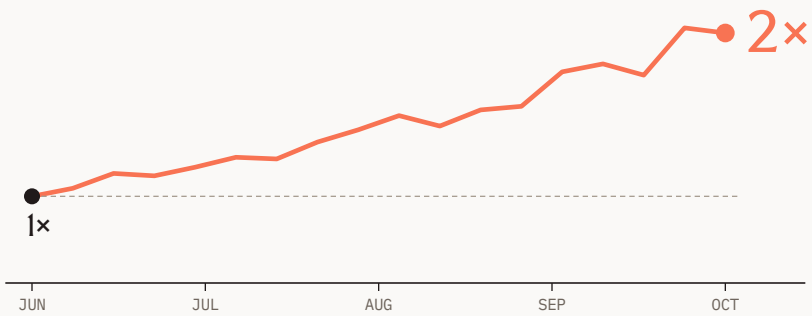
- Why: cost and rationing. Token budgets made engineers save AI for the hardest tasks. A fixed self-hosted bill removed the limits.
- Frontier models stay available for the hardest work. Everything else runs on the model Revolut hosts.
- It took skill. The first setup was slow, and tuning made it about eight times faster on the same GPUs.

[FIG 07] Figures as reported by Revolut.

SOURCES Dealroom, 3 Nov 2026; Startup Fortune, 3 Nov 2026; Red Hot Cyber, 6 Nov 2026.

Sonnet-level is reachable with open models, and much of the bill sits there

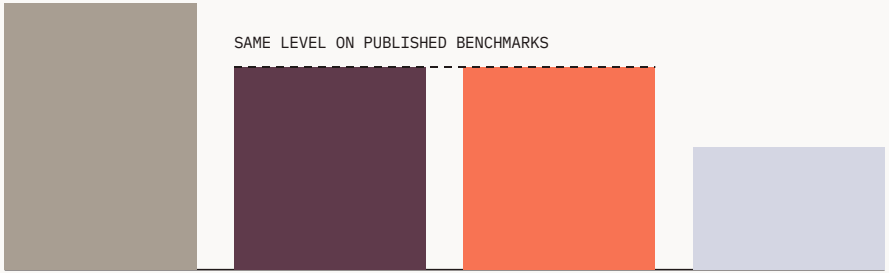
[THE BILL GROWS WITH AGENTS]



Example: AI coding spend doubled in four months, June to October 2026. Agentic coding uses about 1,000× the tokens of a code chat.

- Code review and testing take most agent tokens, and they rarely need the very top model.

[CODING CAPABILITY, SIMPLIFIED]



You won't match the very top models yourself, but the best open models reach the Sonnet level.

- Self-hosted capacity is a fixed cost: the more work moves, the cheaper each token gets.

[FIG 08] Illustrative shapes

Start with team-sized hardware, and use the nights for long agent runs

[TEAM-SIZED HARDWARE]

NVIDIA DGX Spark

Desk-sized, 128 GB of memory

NVIDIA RTX PRO 6000

Workstation GPU, 96 GB per card

[* COMPRESSION CHANGES WHAT FITS]

QuTwo's quantum-inspired compression took the 2.4-trillion-parameter Qwen3.8-2.4T model from 4.89 TB to 363 GB and kept 94.9% of its answer quality. A model that fit no single server now fits on one.

[TOP OPEN-WEIGHT MODELS · NOV 2026]

DeepSeek-V4.1-Flash
DEEPSEEK

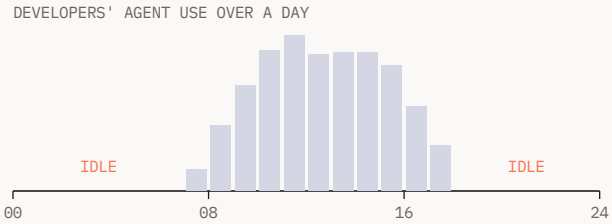
GLM-5.3 and GLM-5.3-Flash
Z.AI

Qwen3.8-Flash-Next
ALIBABA

Qwen3.8-2.4T *
ALIBABA · QUTWO-COMPRESSED

Mistral Large 4
MISTRAL · EU

[NIGHTS FOR LONG AGENT RUNS]



Developers don't keep agents running at night. The hardware is paid for either way, so the night is free capacity for new work:

- Security analysis of every repository
- Test generation
- Large refactors and upgrades

[FIG 09] Agent use follows the working day.

SOURCES QuTwo press release, 30 Sep 2026. Model results are vendor-reported; test on your own code before choosing.

Answer three questions before you scale, and calculate the business case on measured numbers

[QUESTION 1]

Which models work, and for which work?

Test on your own code, task by task. Benchmarks only make the shortlist.

[QUESTION 2]

What setup, on which hardware, for how many developers?

Model, compression, KV cache and serving software must all be tuned to the hardware and the workload.

[QUESTION 3]

Will developers use it, and can you keep up?

Developers test it in their own tools. New models arrive every month, so make testing repeatable.

[THE BUSINESS CASE]

Cost per token = yearly cost ÷ tokens served

Own: hardware over its lifetime, power, people and support.

Rent: GPU hours, paid whether you use them or not.

Tokens served depends on the tuning: the same hardware can deliver several times more.

- Compare with the cheapest cloud model that passes your tests.
- Use real production throughput, often about half of benchmark peak.
- Count two servers if you need redundancy.
- Agent traffic is mostly input tokens, so input and cache prices drive the cloud side.

Break-even is the utilization at which both cost the same. Value the fallback separately, as scenarios, not euros.

Questions



Takeaways

[QUANTUM COMPUTING]

- Start now: the platform, the people and the use cases take years.
  - The platform causes difficulties, as the quantum computers are outside the bank.
  - Understanding and finding the right use cases takes time.
- Quantum-inspired methods already pay off today.

[LOCAL LANGUAGE MODELS]

- Access to AI can change overnight. Knowing that open-weight models exist doesn't make you ready.
- Doing it well takes experience: the right model on the right hardware.
- Sonnet-level models already run on your own hardware.
- Similar governance issues here as for QC
- However, governance of an agentic platform with agents, LLMs, MCPs etc. is another full topic…

A bank's security team needs a model that will look at malicious code when asked, without waiting for anyone's permission.

Not a model provider's guardrails, and not a new round of internal approvals in the middle of an attack.

Thank you

Hugo Gävert
Member of the Technical Staff, QuTwo
hugo.gavert@qutwo.com

QuTwo, a European AI lab for the quantum era

QuTwo