AI Congress Barcelona

20 - 21 - 22 OCT 2026

Auditori L'illa

Un esdeveniment de:

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per:

eurecat

Alineat amb l'Estratègia Catalunya IA 2030:

Generalitat de Catalunya

Segueix-nos a:

#AICongressBCN

aicongress.barcelona

Patrocinadors OR:

CaixaBank TECH

conversia compliance is our business

forvis mazars & infini DIGITAL EFFICIENCY

orange Empresas

raona coding corporate futures

softeng Maximize the cloud. Securely.

Patrocinadors PLATA:

apdcat Autoritat Catalana de Protecció de Dades

LABORAL kutxa

Col·laboradors:

Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CVC Centre de Visió per Computador

i2cat

UPC Universitat Politècnica de Catalunya Barcelona Tech

IDEAI Intelligent Data Science and Artificial Intelligence Research Center

AI Congress Barcelona

Un esdeveniment de: CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per: eurecat

Alineat amb l'Estratègia Catalunya IA 2030: Generalitat de Catalunya

20 - 21 - 22 OCT 2026
#AICongressBCN
aicongress.barcelona



Jezabel R. Garcia
Cap de recerca en IA
Multiverse Computing

KEYNOTE SPEAKER:

DELS GRANS MODELS A LA IA SOBIRANA: COMPRESSIÓ, EDGE AI I EL FUTUR DE LA INTEL·LIGÈNCIA ARTIFICIAL



Presenta:

Joan Mas
Director
CIDAI

Patrocinadors OR: CaixaBank TECH · conversia compliance is our business · forvis mazars & infini DIGITAL EFFICIENCY · orange Empresas · raona coding corporate futures · softeng Maximize the cloud. Securely.

Patrocinadors PLATA: apdcat Autoritat Catalana de Protecció de Dades · LABORAL kutxa

MULTIVERSE

COMPUTING

From big models to sovereign AI

Compression, Edge AI and the future of artificial intelligence.

Jezabel R. Garcia

AI Congress Barcelona, 20 October 2026

AT A GLANCE

The next revolution in AI is not a bigger model.

It is a model that runs where it adds value.

Limit. The race for size has hit memory, energy and cost: three makers control over 90% of DRAM and are shifting output to AI memory.

Shift. AI must run on devices, in enterprises, in critical infrastructure and in public administrations, often with no cloud at all.

Method. Compression plus healing keeps capability: on a standard Intel Xeon 6 CPU a compressed Llama 3.3 70B doubled throughput and kept over 97% of its accuracy.

Proof. Telefónica, MWC 2026 with Qualcomm, Intel and the Spanish AI gigafactory show the method from a phone to a datacentre.

Stake. Sovereign, efficient and sustainable AI are one thing: a model you can run, inspect and adapt yourself, in your own languages.

SECTION 01

The race for size, and why it stalls

Bigger models, bigger bills, and a memory supply chain that cannot keep up

The bill for ever-bigger models is arriving

$500B

Up to this much for the expansion of OpenAI's Stargate project with Oracle and SoftBank

>90%

Share of global DRAM made by three companies now shifting output to high-bandwidth memory for AI

HBM

The new bottleneck is high-bandwidth memory, not compute: supply is tight and prices are rising

ROMÁN ORÚS

Cloud costs soar and the numbers no longer add up.

Bigger does not scale: energy, cost and access all break first

Large models will not disappear, but they cannot be the only path.

- Energy: ever-bigger models hit an energy wall; the approach does not scale.
- Cost: memory scarcity and energy inefficiency raise prices for every user.
- Access: small companies, with the least access to advanced compute, are hit hardest.
- Nature: biological intelligence does far more with far less, so there is a long way to go.

ROMÁN ORÚS

We cannot keep doing AI the way we do it now and go to ever-bigger models; energetically it does not scale.

SECTION 02

AI moves to where it adds value

Devices, enterprises, critical infrastructure and public administrations

AI already runs where the connection drops and the hardware is small

Public figures: IEEE Spectrum, 3 September 2025; CompactifAI App, on the App Store and Google Play since 3 March 2026.

191 MB

SuperFly, compressed from SmolLM2, installed on an iPhone 14 Pro

115 tok/s

Generation speed of SuperFly on that phone, with no network at all

0 bytes

Sent to the cloud when the CompactifAI App runs a model offline on a phone or tablet

WHY IT MATTERS

Drones, satellites, cars and defence must decide without a connection; only a model that fits the device can.

Four places AI must run, and what each one needs

The four destinations named in the session abstract; the rest of the talk gives an example of each.

- Devices: cars, drones, phones and robots, with no guaranteed connection.
- Enterprises: on-premise, with private data that must not leave the building.
- Critical infrastructure: telecoms, energy and health, where latency and continuity matter.
- Public administrations: sovereign, accountable computing.

COMMON NEED

AI that is small, cheap and under your control.

SECTION 03

Compression, explained and measured

What the method does, and what it costs, in numbers

Compression finds what matters, removes the rest, then heals the model

Quantum-inspired tensor-network methods find the redundant correlations; a short targeted retraining recovers accuracy; quantisation stores each number in fewer bits.

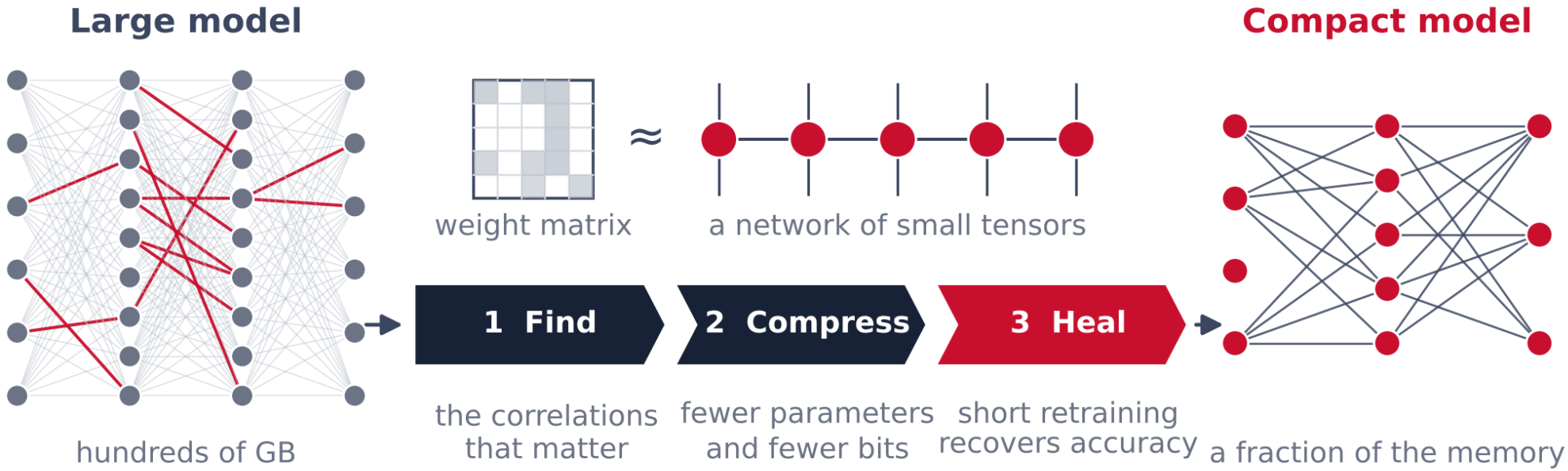


Figure 1. The compression pipeline. A smaller model runs on the device or on-premise, uses less energy and keeps the data in your hands.

Quasar 438B: a 148-expert specialist built from GLM-5.2

Multiverse technical note, 2026. Built for coding and agentic workloads; served through the CompactifAI API in FP8.

148

Experts per mixture-of-experts layer, down from 265, selected jointly because experts interact

NVFP4

Quantisation-aware variants in FP8 and NVFP4 for current and next-generation accelerators

438B

Parameters in the compressed specialist derived from the open-weights GLM-5.2

HEALING

Targeted retraining recovers and sharpens coding and agent behaviour after compression.

Measured, not promised: four numbers to take away

Compressed Llama 3.3 70B on an Intel Xeon 6737P CPU with vLLM; Multiverse and Intel, 23 July 2026.

2x

Output throughput, 2.00 to 3.86 tokens per second

-49%

Latency at one concurrent user, 5.1 s to 2.6 s

>97%

Accuracy retained across five benchmarks

1/2

Disk size, 130 GiB to 65 GiB

ASK EVERY VENDOR

What did compression cost on my task and my language? This is our answer.

On a standard CPU, compression doubled throughput and halved latency

1,024 input and 1,024 output tokens per request; concurrency from 1 to 256 users. Higher throughput and lower latency are better.

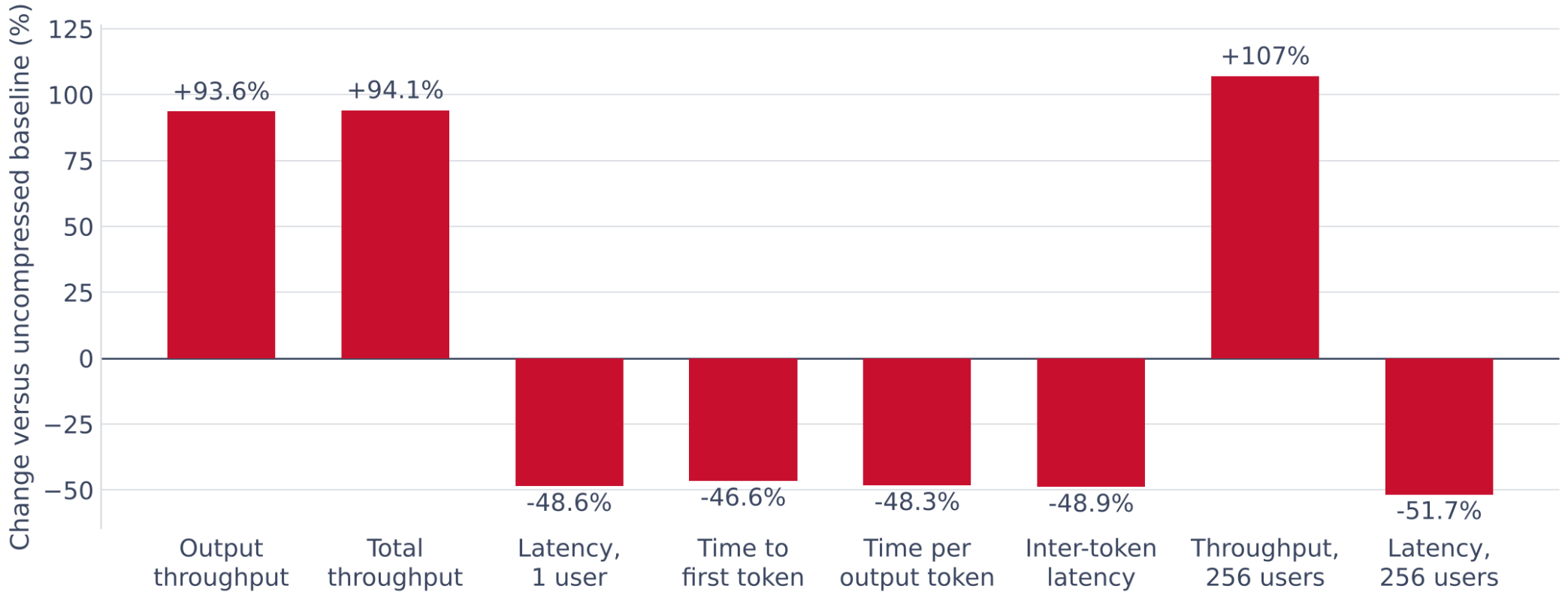


Figure 2. Change versus the uncompressed baseline. Throughput rises 94% at one user and 107% at 256 users; latency measures fall 47 to 52%.

Over 97% of the original accuracy survives compression

Same configuration as Figure 2; higher is better. The crimson series is the compressed model.

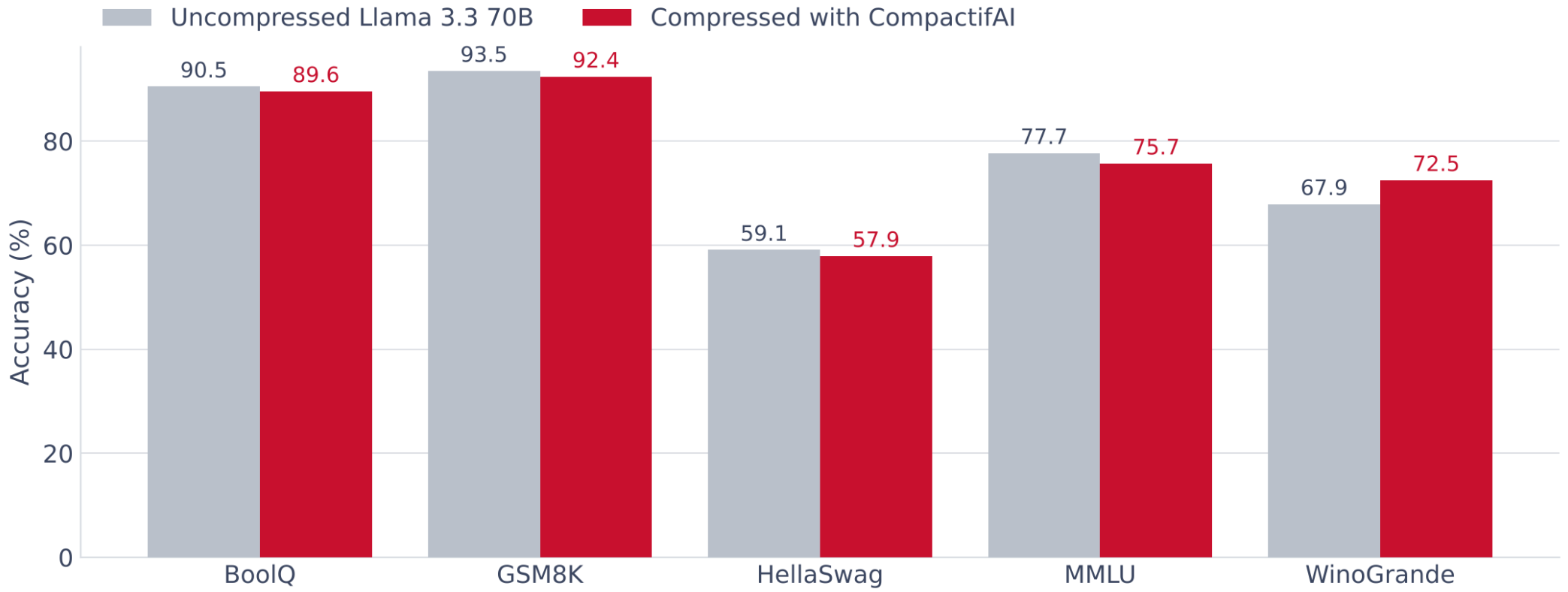


Figure 3. Accuracy by benchmark. Four benchmarks lose under 2.5 points; WinoGrande gains 6.9 points, a side effect of the healing pass.

SECTION 04

Industrial applications over time

From a telco's customer service to a national datacentre

Telefónica: 80% smaller models, up to 75% less energy, on its own network

Llama 3.1 8B and 3.3 70B, compressed and fine-tuned for the assistant supporting Movistar customer-service agents; 12 November 2025.

- 80% smaller models, with the quality of responses maintained.
- Up to 75% less energy than the uncompressed models, because they run on premise in Telefónica's own network.
- The central offices run on 100% renewable electricity, so CO2 emissions fall as well.
- Open-source base models, in line with Telefónica's openness and neutrality goals.



ENTERPRISE

The same model, on the operator's own hardware, at a quarter of the energy.

Health and finance: live demos at MWC 2026 ran faster on less memory and power

Compressed open-source LLM on a Qualcomm Cloud AI100 Ultra accelerator, Mobile World Congress, March 2026; figures are 'up to' and accuracy was unchanged.

- Emergency medical reporting: 93% faster responses, 44% higher throughput, 45% less memory, 21% less power.
- On-premise RAG chatbot over confidential financial documents: 35% faster, 54% higher throughput, 45% less memory, 14% less power.
- The confidential data never leaves the building.

MULTIVERSE COMPUTING × Qualcomm

PRIVACY

A smaller model on your own hardware is a privacy decision as much as a cost decision.

Every public deployment cut size, memory or energy by 20 to 80 per cent

Reported reductions versus the uncompressed model; 'up to' figures as published.

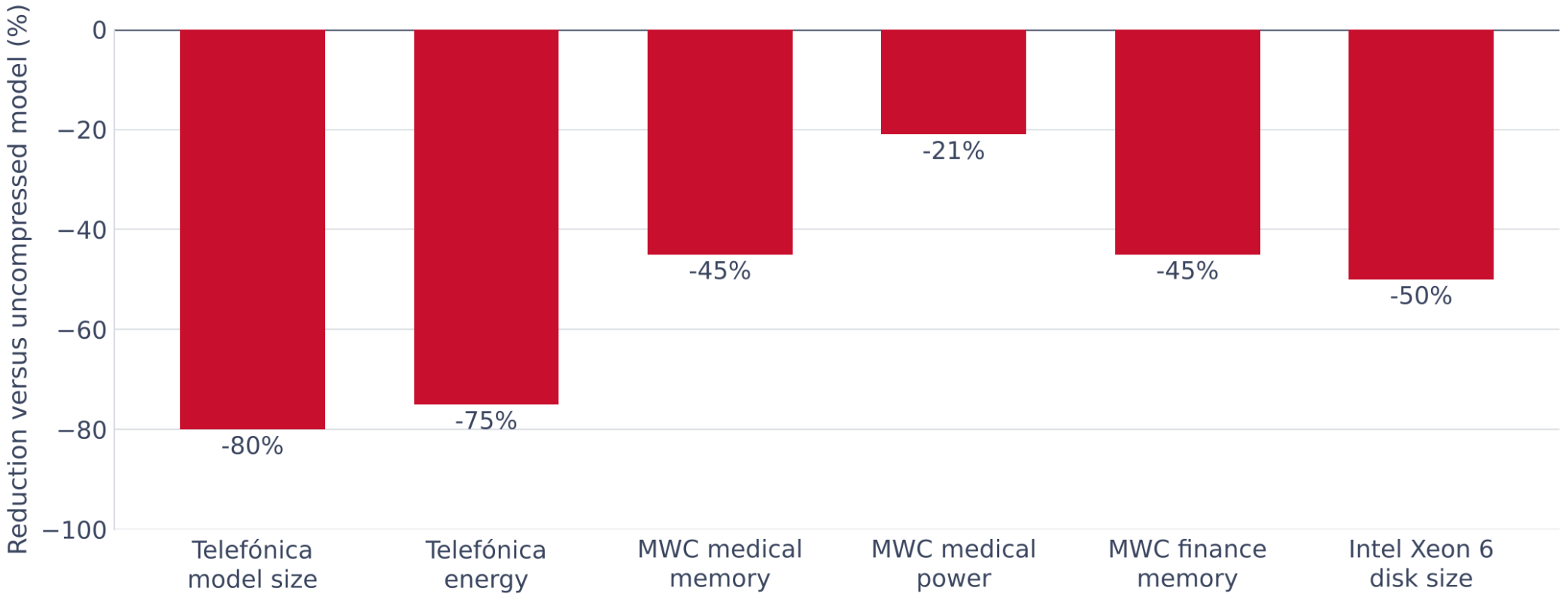


Figure 4. Telefónica (November 2025), MWC with Qualcomm (March 2026) and Intel Xeon 6 (July 2026). Sources: Multiverse press releases.

Datacentre scale: efficiency is the cheapest capacity you can buy

Intel Xeon 6 (23 July 2026) and Qualcomm Dragonfly AI200 and AI250 (5 August 2026).

- Compressed 70B models now serve on standard CPUs with mainstream open-source tooling: vLLM, PyTorch, Hugging Face.
- Models specialised for Qualcomm Dragonfly accelerators free headroom on existing deployments.
- Operators serve more requests, or run more models, without adding accelerators or expanding the datacentre.

FOR THE CFO

More capacity on the same hardware, with lower power per request.

SECTION 05

Research and sovereign AI

Where the advances come from, and why they matter for Europe

The advances come from research, because each model generation breaks yesterday's methods

Multiverse research lines: compression and efficiency; evaluation and failure analysis; alignment.

- Mixture of experts, reasoning models and agents each needed a new compression recipe.
- Healing began as a research idea; it now recovers, and sometimes improves, accuracy after compression.
- We publish and collaborate with universities: European talent and ideas are part of sovereignty.

Today's AI runs on algorithms from the 1940s to the 1960s. Whoever invents the next ones sets the rules.

THESIS

Sovereign AI has to speak our languages, so compression is evaluated per language

Spanish, Catalan, Basque, Galician and every EU language.

- Most models and benchmarks are English-first.
- Compression can affect languages unequally, and aggregate scores hide it.
- We evaluate per language and per item, not only on the average.

QUESTION

Efficient for whom? That is a research question, and a sovereignty question.

Sovereignty matters because AI is geopolitically strategic

Román Orús, El Español, 28 May 2026 (translated from Spanish).

Europe has traditionally been slow and has missed trains such as the cloud and social networks. Here, better late than never: there is no alternative.

AI is a geopolitically strategic technology, and the regions that fall behind will end up irrelevant. We cannot depend completely on American or Chinese technology.

IMPLICATION

Europe must build its own capabilities, from algorithms and models to infrastructure and hardware.

The Spanish AI gigafactory puts that capability in Tarragona and Madrid

Public-private consortium backed by the Government of Spain, preparing its bid for one of Europe's AI gigafactories; Multiverse, 1 July 2026.

- Up to EUR 5 billion of investment; sites at Móra la Nova (Tarragona) and Madrid.
- Partners: SETT, Telefónica, ACS, Santander, the Government of Catalonia, and Multiverse as technology partner.
- Multiverse contributes the datacentre software stack: GPU orchestration, AI services and sovereignty controls.



ENRIQUE LIZASO, CEO

Not only powerful, but also efficient and sovereign.

Sovereign, efficient, sustainable: compression delivers all three

Where does your AI run? What does each query cost in memory and energy? What did compression cost on your task and your language?

Jezabel R. Garcia, Research Manager, Multiverse Computing

JG

AI Congress Barcelona

20 - 21 - 22 OCT 2026

Auditori L'illa

Un esdeveniment de:

CIDAI Centre of Innovation for Data tech and Artificial Intelligence

Organitzat per:

eurecat

Alineat amb l'Estratègia Catalunya IA 2030:

Generalitat de Catalunya

Segueix-nos a:

#AICongressBCN

aicongress.barcelona

Patrocinadors OR:

CaixaBank TECH

conversia compliance is our business

forvis mazars & infini DIGITAL EFFICIENCY

orange Empresas

raona coding corporate futures

softeng Maximize the cloud. Securely.

Patrocinadors PLATA:

apdcat Autoritat Catalana de Protecció de Dades

LABORAL kutxa

Col·laboradors:

Barcelona

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

CVC Centre de Visió per Computador

i2cat

UPC Universitat Politècnica de Catalunya Barcelona Tech

IDEAI Intelligent Data Science and Artificial Intelligence Research Center