MULTIVERSE
COMPUTING

AI SC

10

JAVIER ALONSO MENCÍA

ML ENGINEER

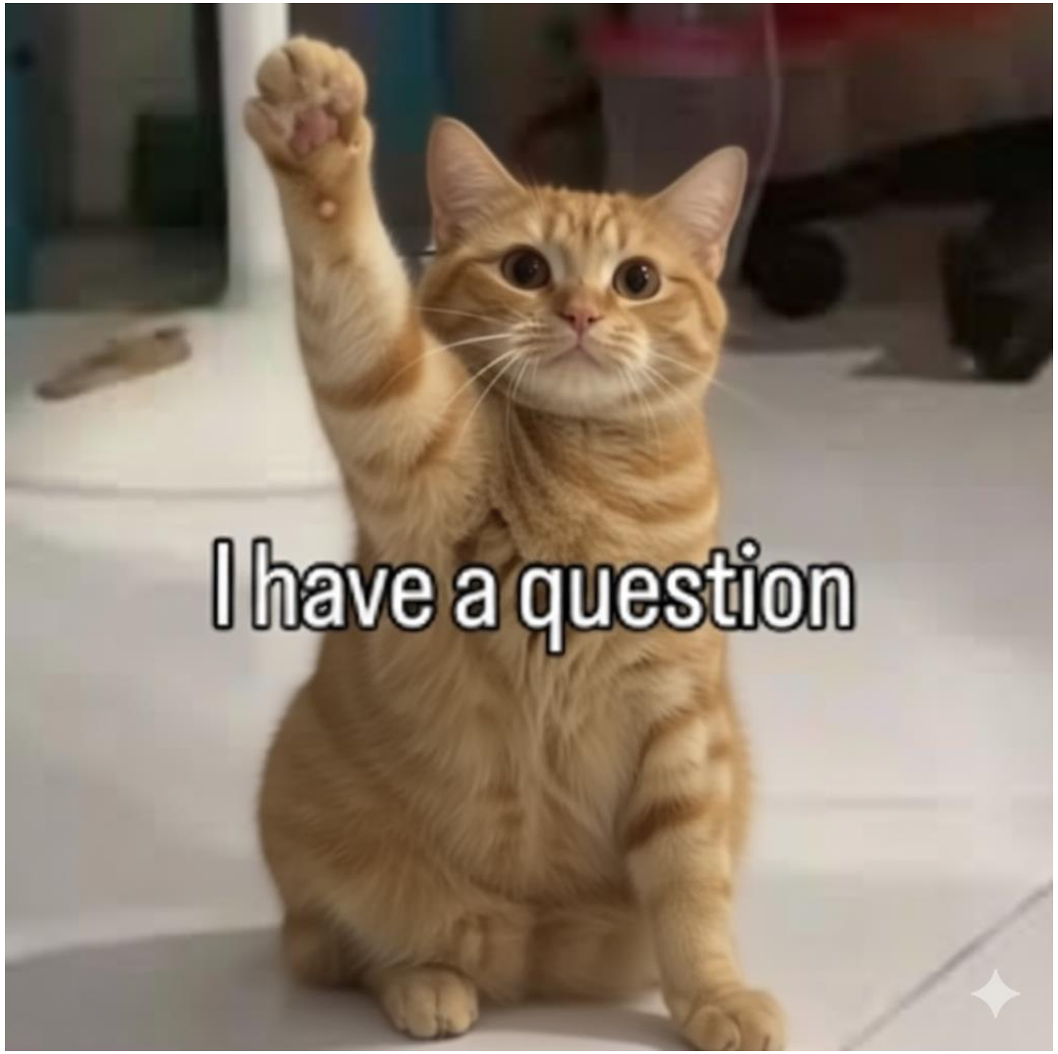
18

MULTIVERSE

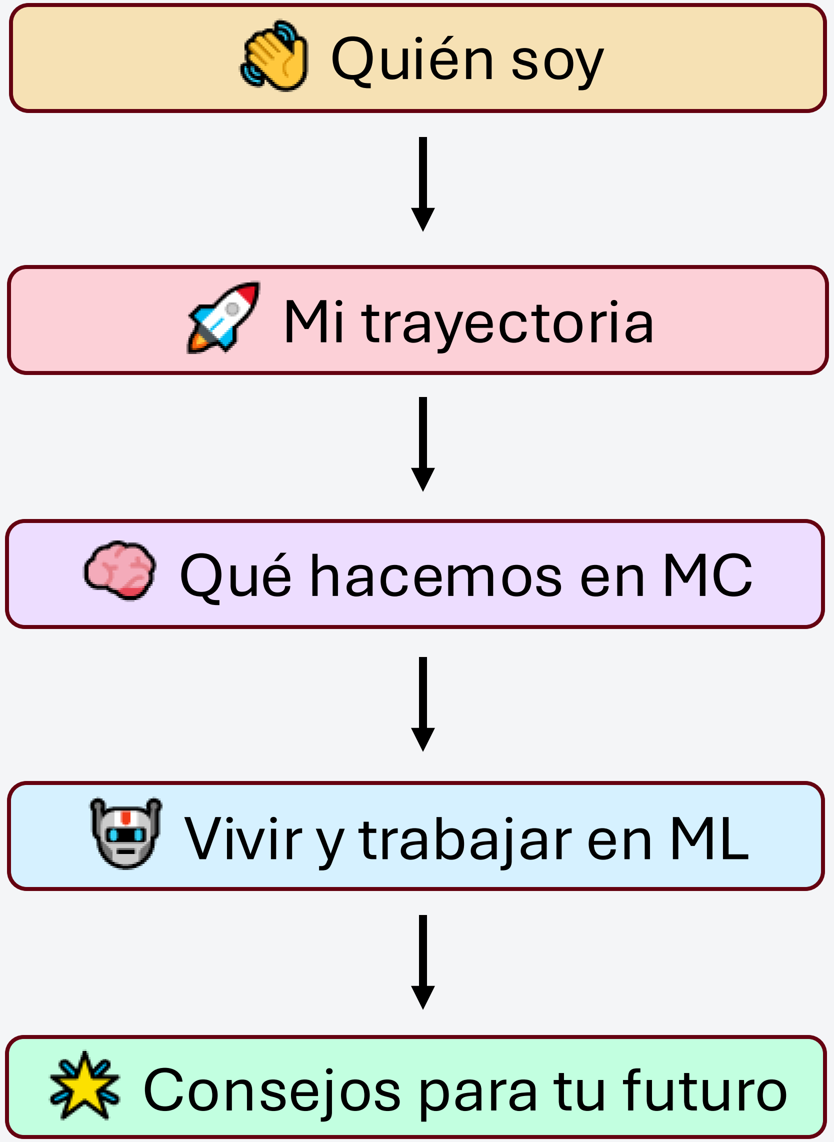
COMPUTING

Machine Learning en el mundo real:

lo que no te cuentan en clase



I have a question



👋 Quién soy

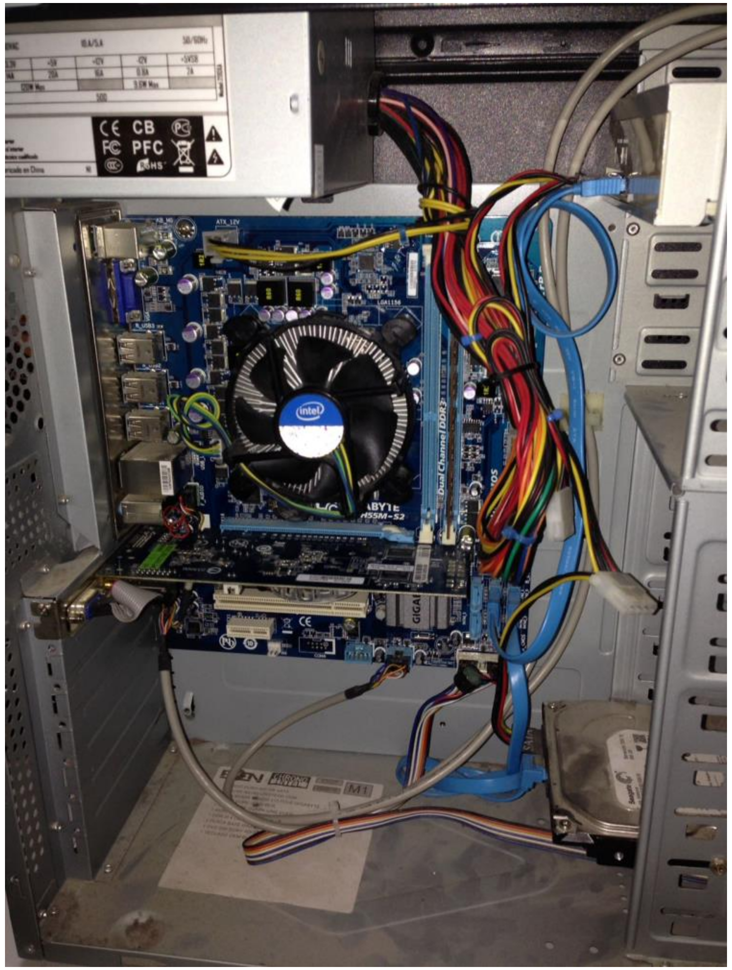
🚀 Mi trayectoria

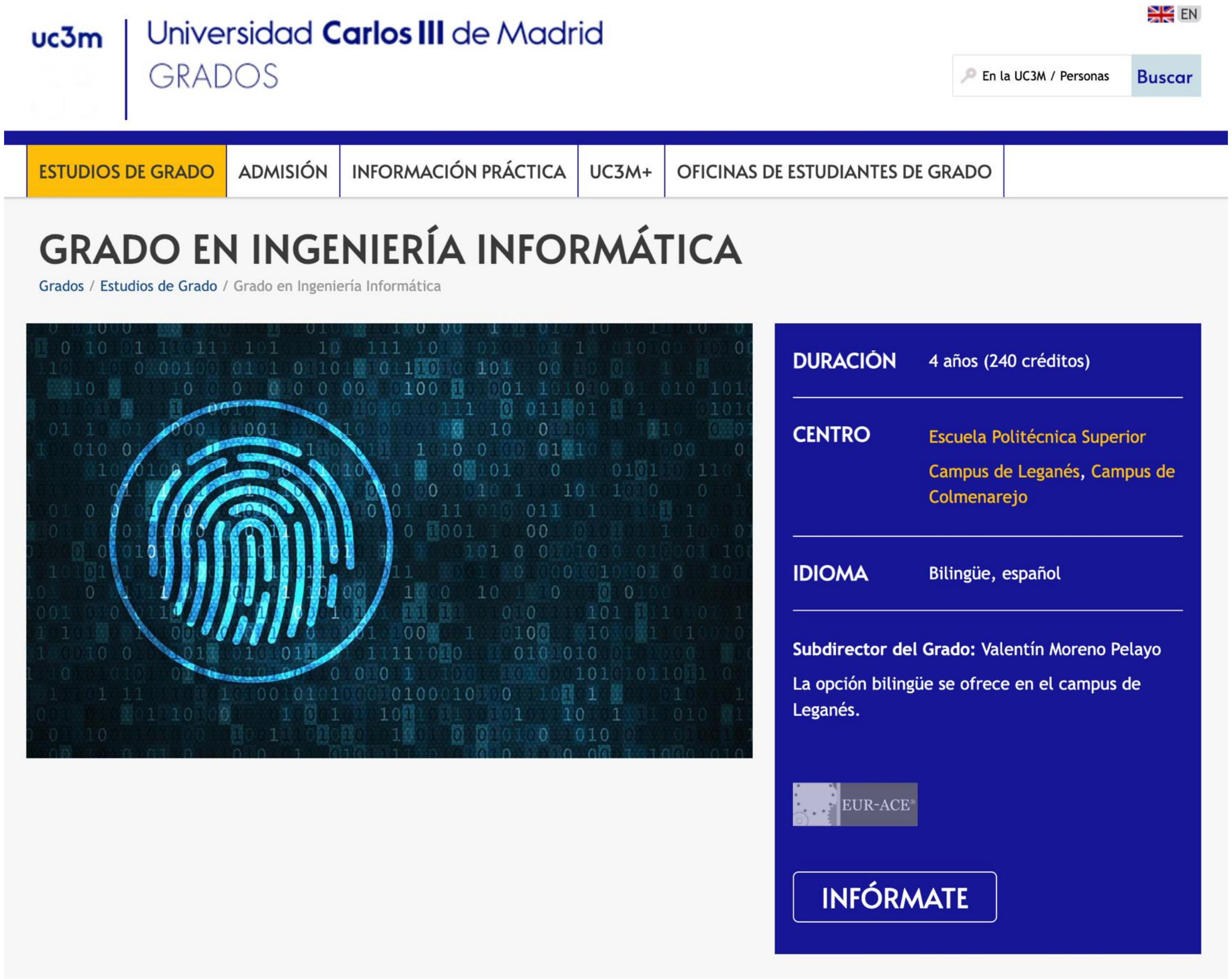
🧠 Qué hacemos en MC

🤖 Vivir y trabajar en ML

🌟 Consejos para tu futuro

Javier Alonso Mencía
Machine Learning Engineer
Multiverse Computing


CE CB
FC PFC
intel
Dual Channel DDR3
H55M-S2



uc3m | Universidad **Carlos III** de Madrid
GRADOS

GRADO EN INGENIERÍA INFORMÁTICA

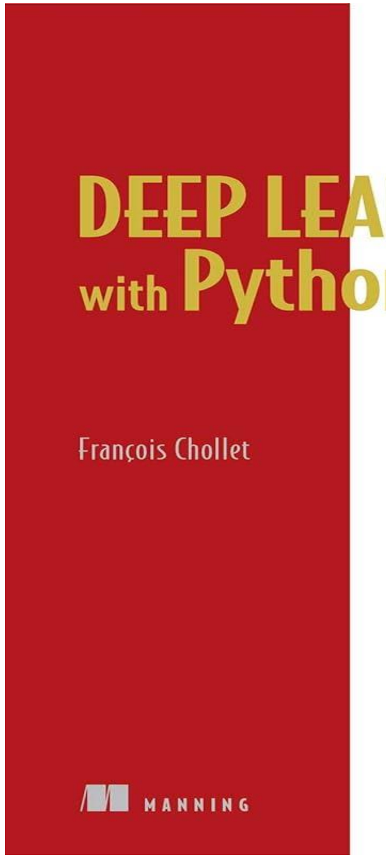
Grados / Estudios de Grado / Grado en Ingeniería Informática

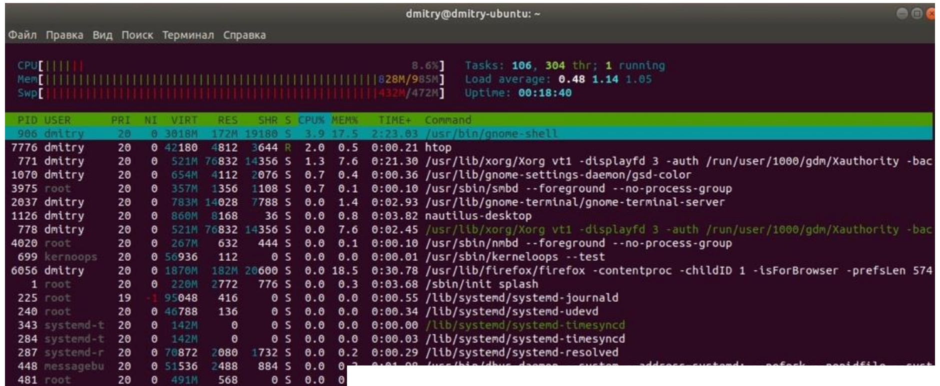
DURACIÓN	4 años (240 créditos)
CENTRO	Escuela Politécnica Superior Campus de Leganés, Campus de Colmenarejo
IDIOMA	Bilingüe, español

Subdirector del Grado: Valentín Moreno Pelayo

La opción bilingüe se ofrece en el campus de Leganés.

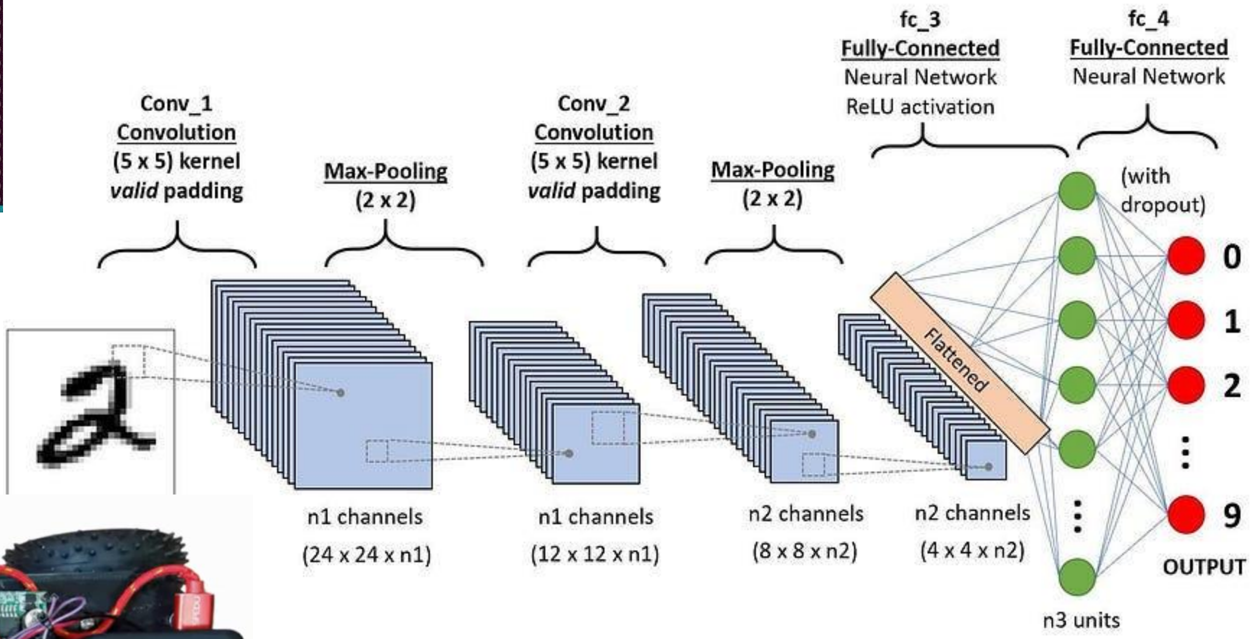
Ir a clase es solo el comienzo

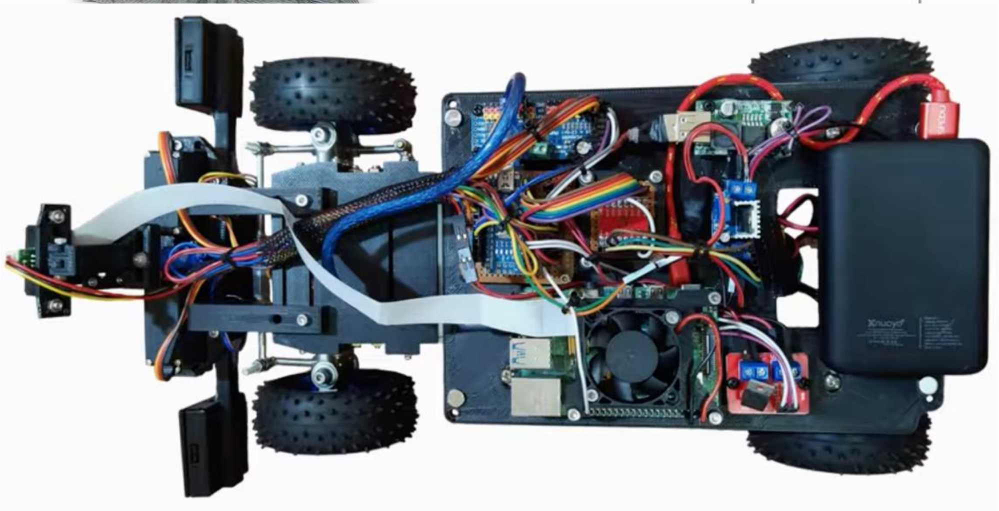




```
  CPU[|||||                                              8.6%]   Tasks: 106, 304 thr; 1 running
  Mem[||||||||||||||||||||||||||||||||||||||||||||||828M/985M]   Load average: 0.48 1.14 1.05
  Swp[||||||||||||||||||||||||||||||||||||||||||||||432M/472M]   Uptime: 00:18:40

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
  906 dmitry     20   0 3018M  172M 19180 S  3.9 17.5  2:23.03 /usr/bin/gnome-shell
 7776 dmitry     20   0 42180  4812  3644 R  2.0  0.5  0:00.21 htop
  771 dmitry     20   0  521M 76832 14356 S  1.3  7.6  0:21.30 /usr/lib/xorg/Xorg vt1 -displayfd 3 -auth /run/user/1000/gdm/Xauthority -bac
 1070 dmitry     20   0  654M  4112  2076 S  0.7  0.4  0:00.36 /usr/lib/gnome-settings-daemon/gsd-color
 3975 root       20   0  357M  1356  1108 S  0.7  0.1  0:00.10 /usr/sbin/smbd --foreground --no-process-group
 2037 dmitry     20   0  783M 14028  7788 S  0.0  1.4  0:02.93 /usr/lib/gnome-terminal/gnome-terminal-server
 1126 dmitry     20   0  860M  8168    36 S  0.0  0.8  0:03.82 nautilus-desktop
  778 dmitry     20   0  521M 76832 14356 S  0.0  7.6  0:02.45 /usr/lib/xorg/Xorg vt1 -displayfd 3 -auth /run/user/1000/gdm/Xauthority -bac
 4020 root       20   0  267M   632   444 S  0.0  0.1  0:00.10 /usr/sbin/nmbd --foreground --no-process-group
  699 kernoops   20   0 56936   112     0 S  0.0  0.0  0:00.01 /usr/sbin/kerneloops --test
 6056 dmitry     20   0 1870M  182M 20600 S  0.0 18.5  0:30.78 /usr/lib/firefox/firefox -contentproc -childID 1 -isForBrowser -prefsLen 574
    1 root       20   0  220M  2772   776 S  0.0  0.3  0:03.68 /sbin/init splash
  225 root       19  -1 95048   416     0 S  0.0  0.0  0:00.55 /lib/systemd/systemd-journald
  240 root       20   0 46788   136     0 S  0.0  0.0  0:00.34 /lib/systemd/systemd-udevd
  343 systemd-t  20   0  142M     0     0 S  0.0  0.0  0:00.00 /lib/systemd/systemd-timesyncd
  284 systemd-t  20   0  142M     0     0 S  0.0  0.0  0:00.03 /lib/systemd/systemd-timesyncd
  287 systemd-r  20   0 70872  2080  1732 S  0.0  0.2  0:00.29 /lib/systemd/systemd-resolved
  448 messagebu  20   0 51536  2488   884 S  0.0  0
  481 root       20   0  491M   568     0 S  0.0  0
  502 root       20   0  491M   568     0 S  0.0  0
  567 root       20   0  491M   568     0 S  0.0  0
  647 root       20   0  491M   568     0 S  0.0  0
  461 root       20   0  491M   568     0 S  0.0  0
  464 root       20   0 44752    64     0 S  0.0  0
  476 root       20   0  4552    36     0 S  0.0  0
  477 root       20   0 70668   400   176 S  0.0  0
  482 avahi      20   0 47256   980   764 S  0.0  0
  661 root       20   0  174M     0     0 S  0.0  0
  486 root       20   0  174M     0     0 S  0.0  0
  575 root       20   0  554M  1360     0 S  0.0  0
  580 root       20   0  554M  1360     0 S  0.0  0
  490 root       20   0  554M  1360     0 S  0.0  0
  549 root       20   0  417M    24     0 S  0.0  0
  558 root       20   0  417M    24     0 S  0.0  0
```

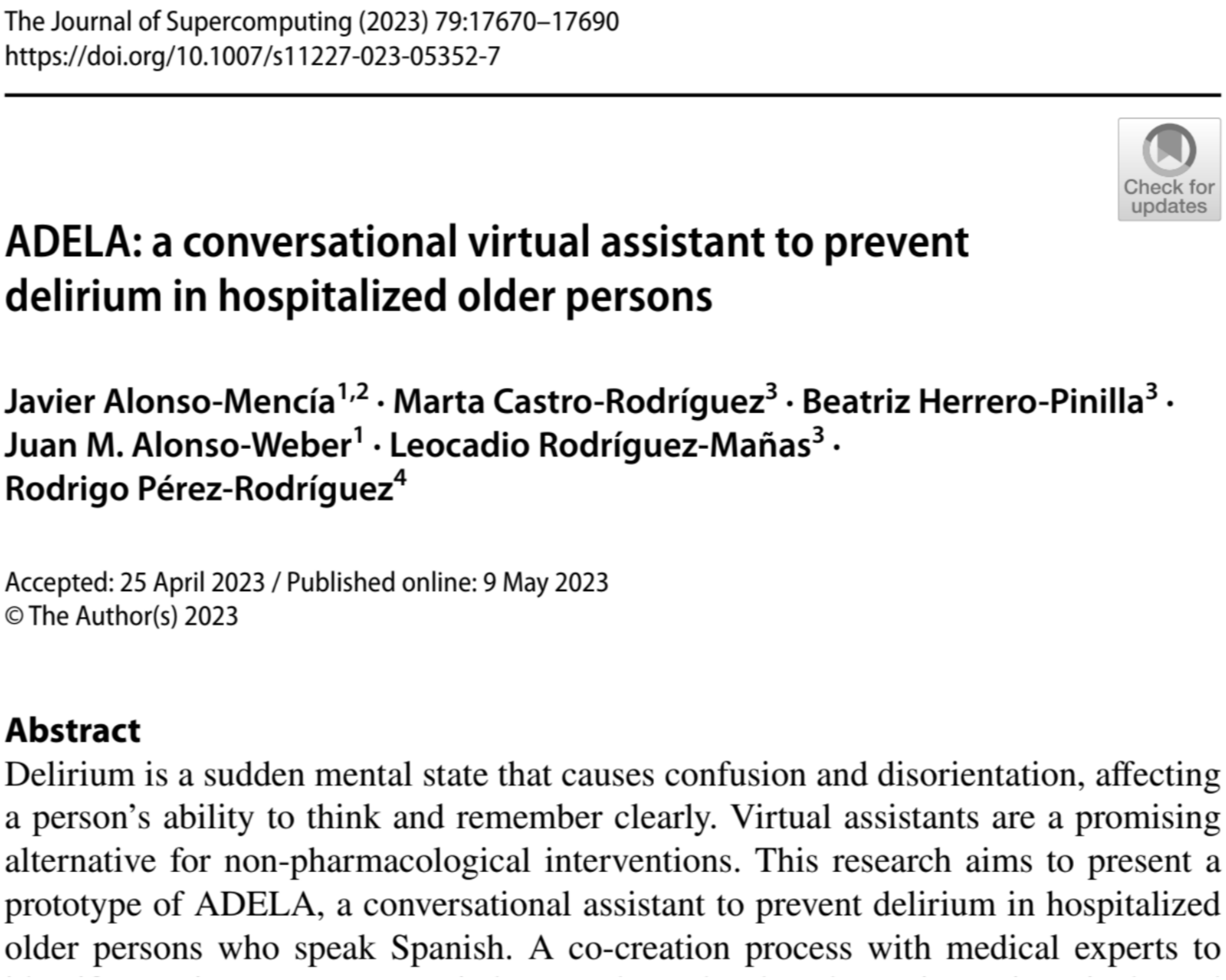




NVIDIA

Y ahora qué?

SO THIS IS AN ENTRY LEVEL POSITION
BUT IT REQUIRES 3-5 YEARS OF EXPERIENCE?
imgflip.com



The Journal of Supercomputing (2023) 79:17670–17690
https://doi.org/10.1007/s11227-023-05352-7

Check for updates

ADELA: a conversational virtual assistant to prevent delirium in hospitalized older persons

Javier Alonso-Mencía[1,2] · Marta Castro-Rodríguez[3] · Beatriz Herrero-Pinilla[3] · Juan M. Alonso-Weber[1] · Leocadio Rodríguez-Mañas[3] · Rodrigo Pérez-Rodríguez[4]

Accepted: 25 April 2023 / Published online: 9 May 2023

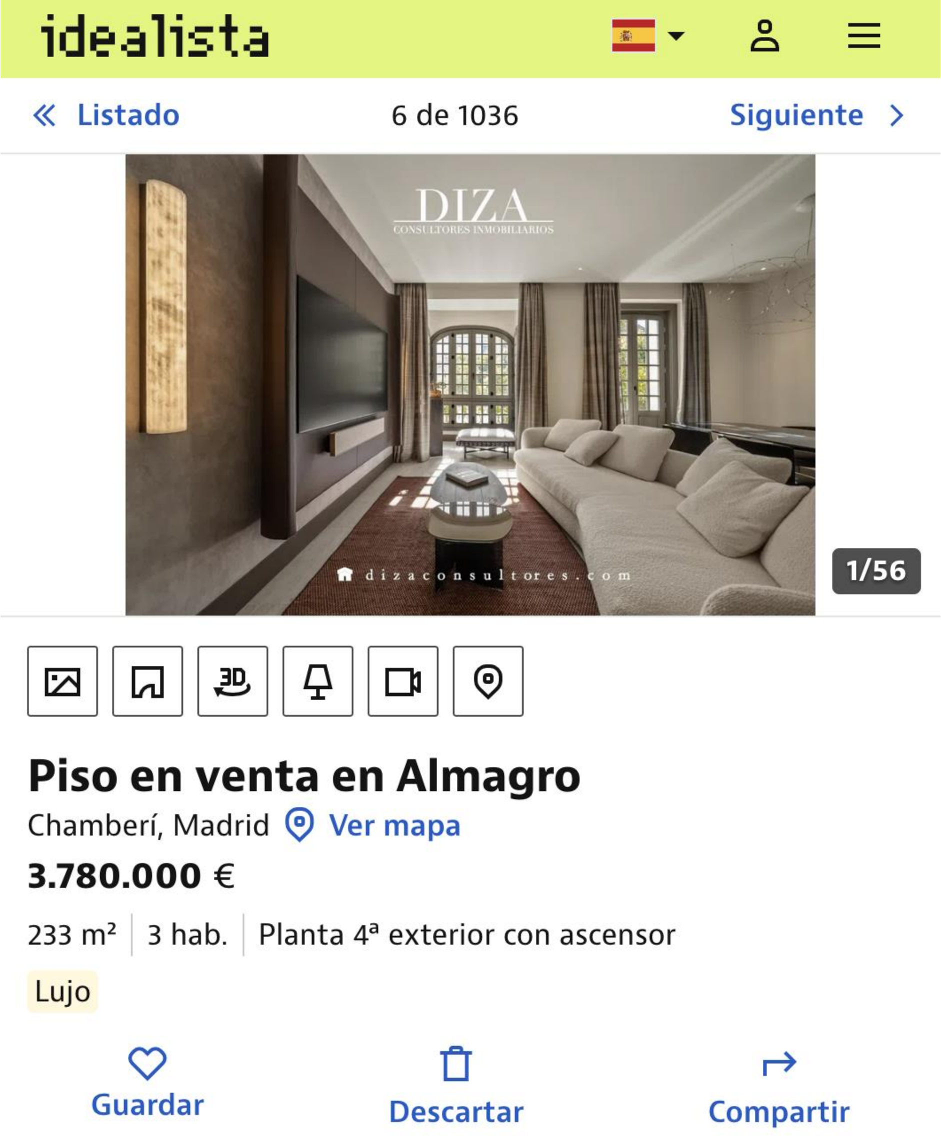
Abstract

Delirium is a sudden mental state that causes confusion and disorientation, affecting a person's ability to think and remember clearly. Virtual assistants are a promising alternative for non-pharmacological interventions. This research aims to present a prototype of ADELA, a conversational assistant to prevent delirium in hospitalized older persons who speak Spanish. A co-creation process with medical experts to

uc3m

SaludMadrid

FUNDACIÓN DE INVESTIGACIÓN BIOMÉDICA
Hospital Universitario de Getafe



Piso en venta en Almagro

Chamberí, Madrid Ver mapa

3.780.000 €

233 m² | 3 hab. | Planta 4ª exterior con ascensor

Lujo

Time Line

2020

uc3m

Bachelor

2021

uc3m

FUNDACIÓN DE INVESTIGACIÓN BIOMÉDICA
SaludMadrid Hospital Universitario de Getafe

Masters
Researcher

2022

EA

Data Science
intern

2023

idealista

Data Scientist

2024

idealista

Data Scientist

2025

MULTIVERSE
COMPUTING

Machine Learning
Engineer

Key Highlights

TEAM

+180 people

+35 nationalities

20% PhD[1]

30% women

FINANCIAL

$65M+

Min 2025 AnRR

Targeting $**150M+ by EOY**

17x

YoY AnRR

Growth

IP

+200

Patents[2]

+40

Research Publications

FUNDING

$250M

Series B closed June 2025,

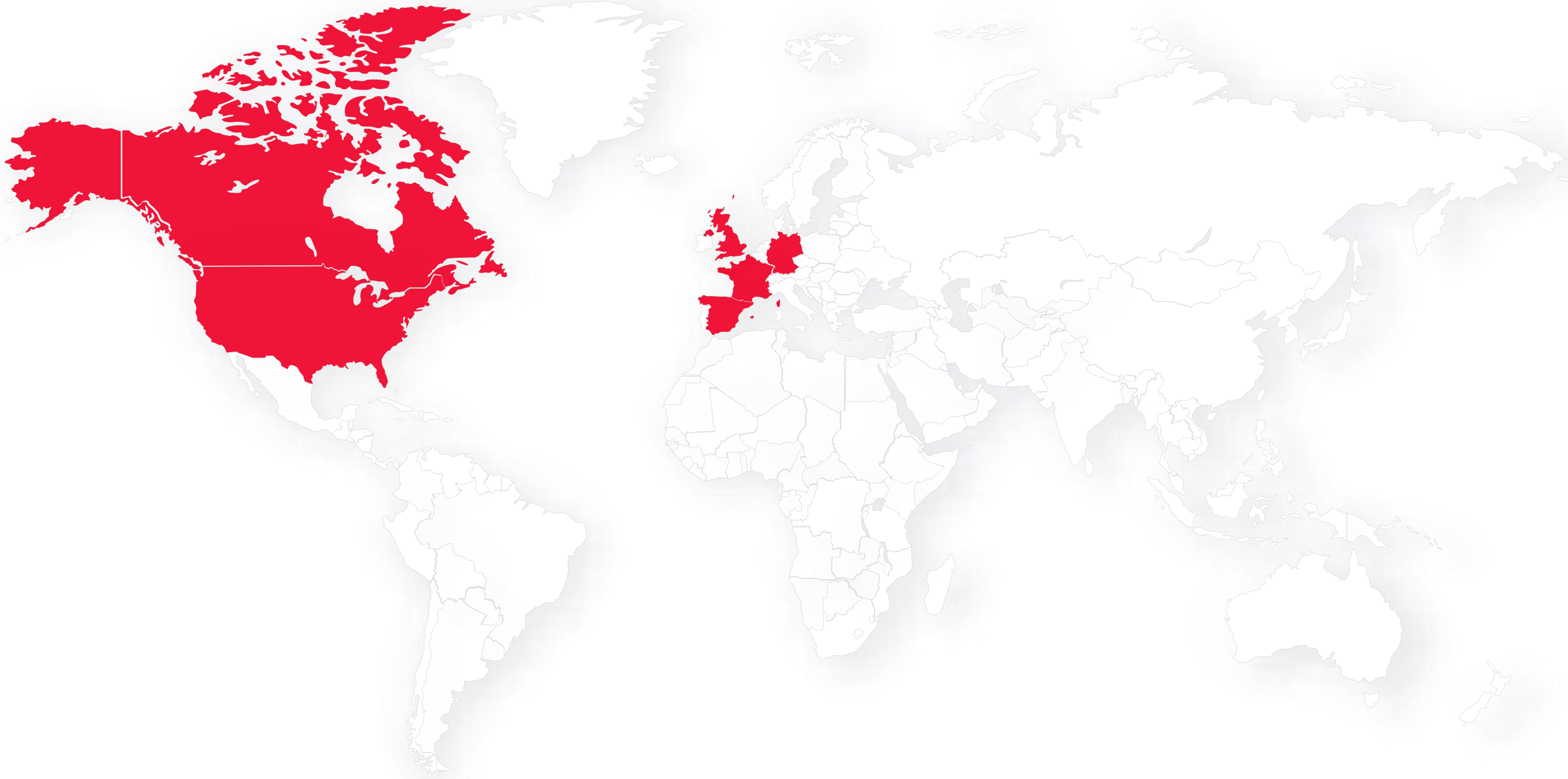
the largest quantum-AI round in Europe

Source: Company information as of September 2025

[1] As of March 2024 | [2] As of June 2025

Our locations

- Spain - HQ
- ● Canada
- ● USA
- ● UK
- ● Germany
- ● France



Application Sectors

Finance

CA CaixaBank BANK OF CANADA BANQUE DU CANADA BASF We create chemistry ally

BBVA MOODY'S Point72 Itaú Gipuzkoako Foru Aldundia Diputación Foral de Gipuzkoa Santander

Manufacturing

Navantia masermic senspain SAFETY SOLUTIONS ikerlan MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE COPRECI

DANOBAT FAGOR aingura IIoT BOSCH

Energy

CAF REPSOL acciona Iberdrola CELSA GROUP

Healthcare & Life Sciences

CIC biomaGUNE MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE Clínic Barcelona VIVE biotech

Chemistry

Universidad del País Vasco Euskal Herriko Unibertsitatea tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE REPSOL

Cybersecurity

elecnor SALTO inspiredaccess CounterCraft Allianz bdr BUNDESDRUCKEREI

Hydrogen

tecnalia MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE REPSOL IDEA INGENIERIA

Defense

LEONARDO indra AIRBUS

Pure Engineering

ZF

Aerospace

avs hispasat esa

Others

GARRIGUES

Todo esto está muy bien… pero, ¿qué hacéis en Multiverse?

Our Technology:

CompactifAI

Multiverse Computing

AI Model Compression

Harness the Power of Tensor Networks

CompactifAI

AI Model Compressor

Reduce the number of parameters of AI models without compromising accuracy, making them more accessible, affordable and sustainable.

Cost-saving

Slash Computational and Infrastructure Costs

Private

Run Anywhere: on Premise, Cloud, or Any Device

Small

Less GPU Memory and Storage

Fast

Faster training and Inference

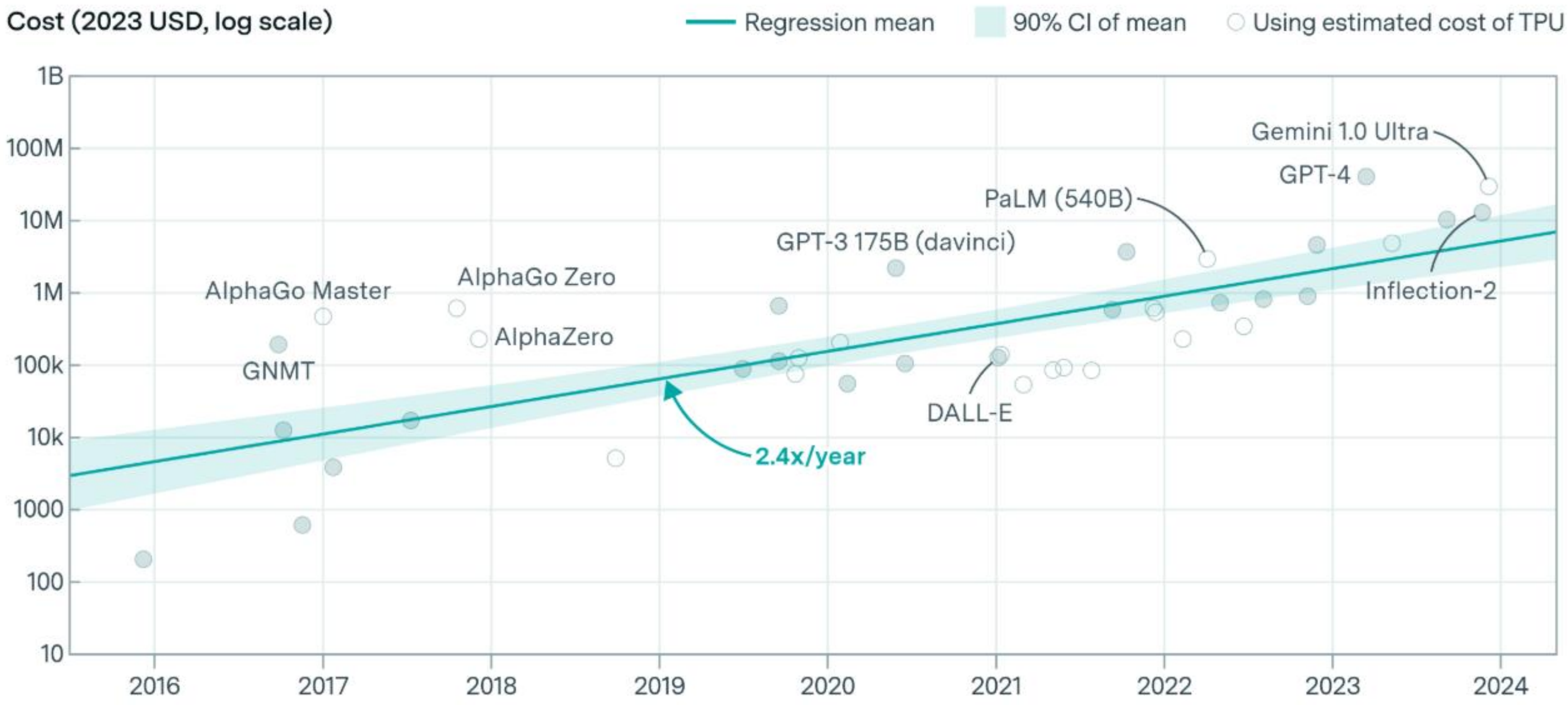
Efficient

Less Energy Consumption and Less CO_2 emission

Skyrocketing AI Compute Costs

The cost of training notable AI models has grown by a factor of 2.4x per year for the past eight years, **suggesting that the largest models will cost over a billion dollars by 2027.**

Amortized hardware and energy cost to train notable AI models over time



Ben Cottier, Robi Rahman, Loredana Fattori, Nestor Maslej and David Owen (2024), "How Much Does It Cost to Train Frontier AI Models?". Published online at epoch.ai. Retrieved from: 'https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models' [online resource]. Updated on January 13th , 2025

MULTIVERSE
COMPUTING

Multiverse Computing is a **global leader in compressed AI** powered by quantum-inspired tensor networks, universalizing affordable and efficient AI across cloud, edge or on-prem.

TOP100 Fastest-growing startups in Europe
sifted
2025

'Future Unicorn' Award
DIGITALEUROPE
2024

We developed CompactifAI, an AI model compression technology to enhance AI system performance by reducing LLM size

CompactifAI
Multiverse Computing

98% compression rates of LLMs[1]
4x-12x speed
50-80% cost savings

For some clients, this translates to **half-a-billion $ savings p.a.**

We offer the market's **leading AI models as**

LLaMA by Meta · OpenAI · deepseek
MISTRAL AI_ · ultralytics YOLO

and more…

+ we launched our CompactifAI API in 2025, offering original and compressed models at an **unbeatable price**

Available on
aws marketplace

We count with leading in-house compressed **NanoModels to run anywhere**

SuperFly	**94M parameters**
ChickBrain	**3.2B parameters**

Deliver top-tier performance on ultra-light hardware, enabling **true on-device and edge AI**

> **"Buzzy AI startup Multiverse creates two of the smallest high-performing models ever"**
> TechCrunch

[1] As of June 2025

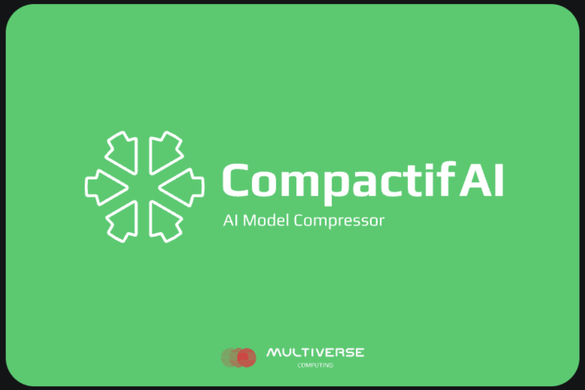
DeepSeek R1 Slim is now live, replacing **DeepSeek R1 0528**.
Whisper v3 is live as our first speech-to-text endpoint.
Chat Completions now supports multi-modal input via **Mistral 3.1 Small**.

Welcome to the CompactifAI API

Build intelligent applications with our state-of-the-art language models

Quickstart →

API Reference



Get started with CompactifAI

Introduction

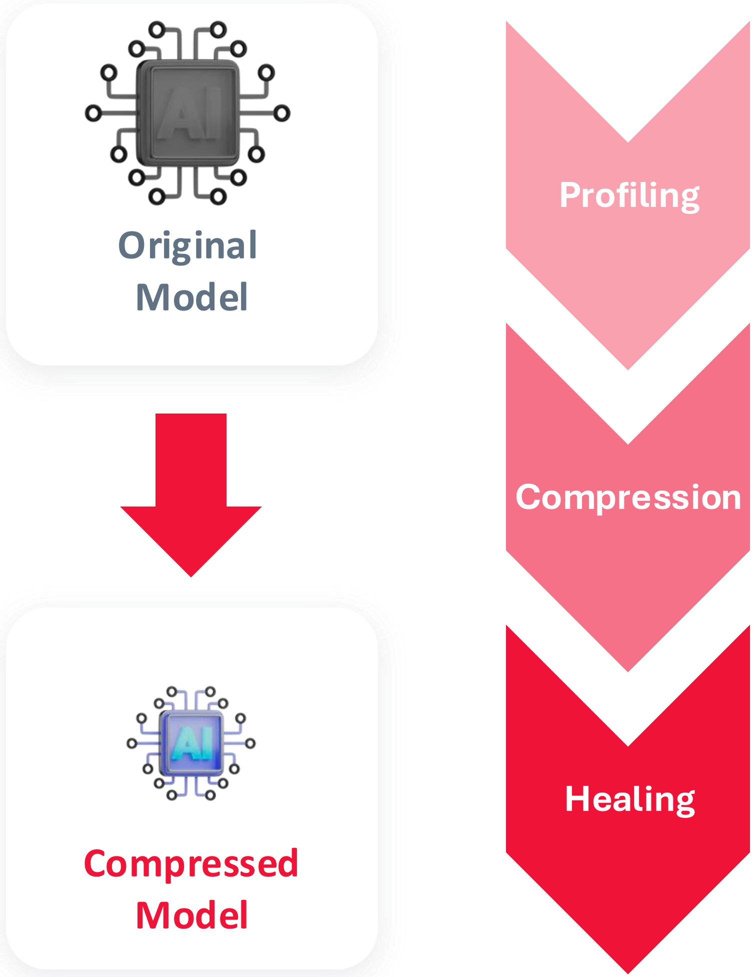
Learn about the CompactifAI API and its capabilities

Pricing

See our pricing for the CompactifAI API

pero.. ¿cómo hacéis la compresión?

Compression Steps



Original Model

Compressed Model

Profiling

Investigating the Model
+
Analyze Sensitivity of Layers
to Compression

Compression

Tensor Network Compression

Healing

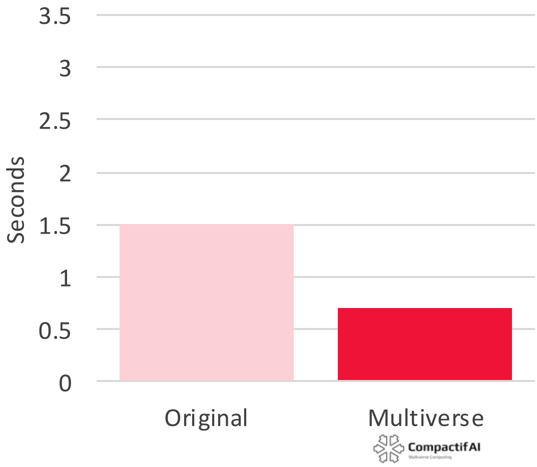
Small Training to Recover Accuracy

Speed

From the first token to the full response — in record time. CompactifAI significantly reduces latency and boosts throughput, delivering faster performance for real-time applications, even on resource-constrained hardware.

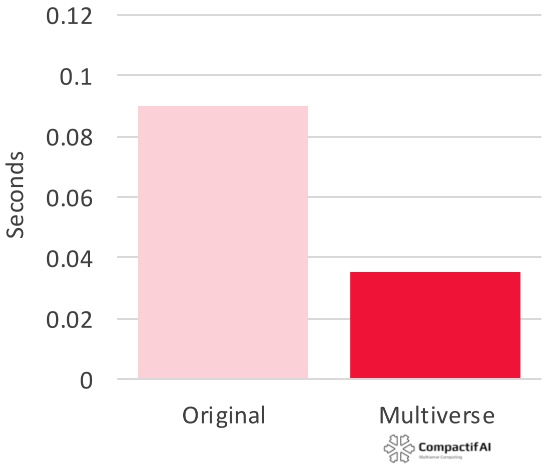
Llama 3.3 70B

Time to First Token[1]



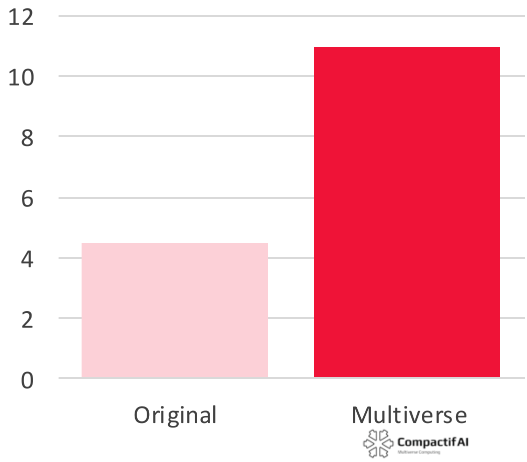
Microsoft Phi 4

Time to First Token[1]



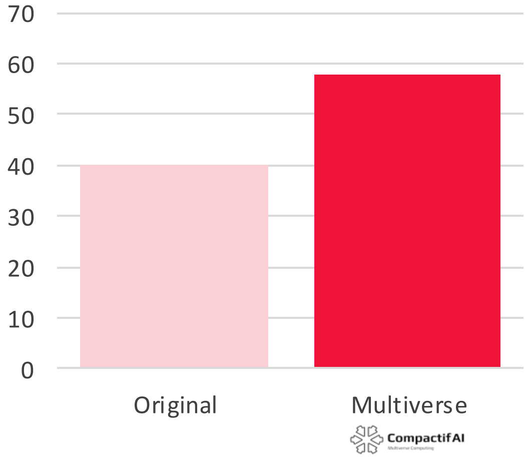
Llama 3.3 70B

Tokens per Second[1]



Microsoft Phi 4

Tokens per Second[1]



↓53%
Reduction in TTFT
vs. original model

↓61%
Reduction in TTFT
vs. original model

Up to
2.4X
Increase in Speed
vs. original models

Llama 3.3 70B compressed model by MVC offers a **nearly 145% faster throughput** than the original Meta LLM.

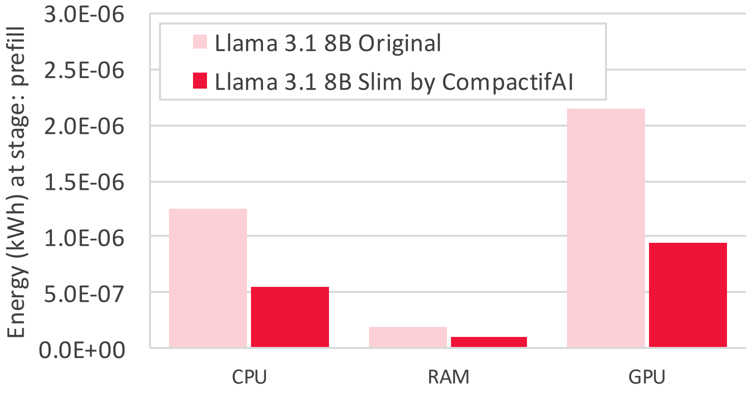
[1] Real-time performance evaluation for MVC Compressed vs Original models. Source: External benchmarks by a leading enterprise hardware provider, 2025.

Cost Efficiency & Sustainability

Achieve more with less. Cut your AI expenses and lower your carbon footprint using ultra-efficient LLMs. Our compressed models operate on smaller hardware, consume less power, and provide enterprise-level performance.

Llama 3.1 8B

Energy Consumption[1]

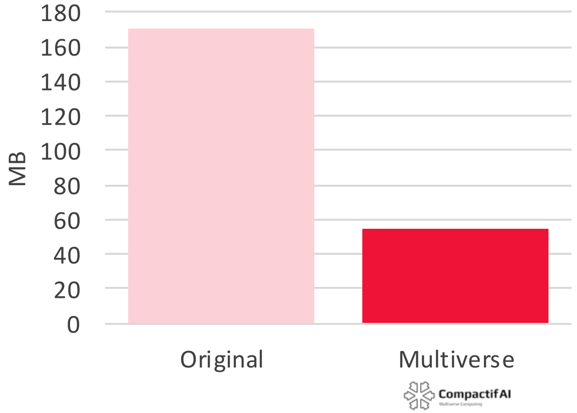


↓50%
GPU energy consump.
vs. original model

CompactifAI's slim model **halves the energy demand of the original Llama 3.1 8B**, making it ideal for sustainable, cost-effective AI deployments.

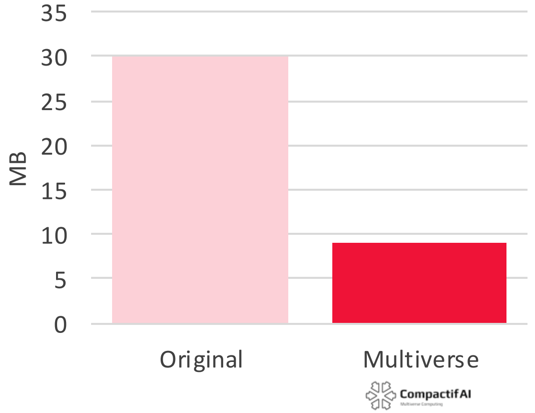
Llama 3.3 70B

RAM Usage[2]



Microsoft Phi 4

RAM Usage[2]



Up to
↓70%
Average RAM usage
vs. original models

MVC models deliver significantly better speed and comparable accuracy — achieving faster inference and stronger results without compromising performance.

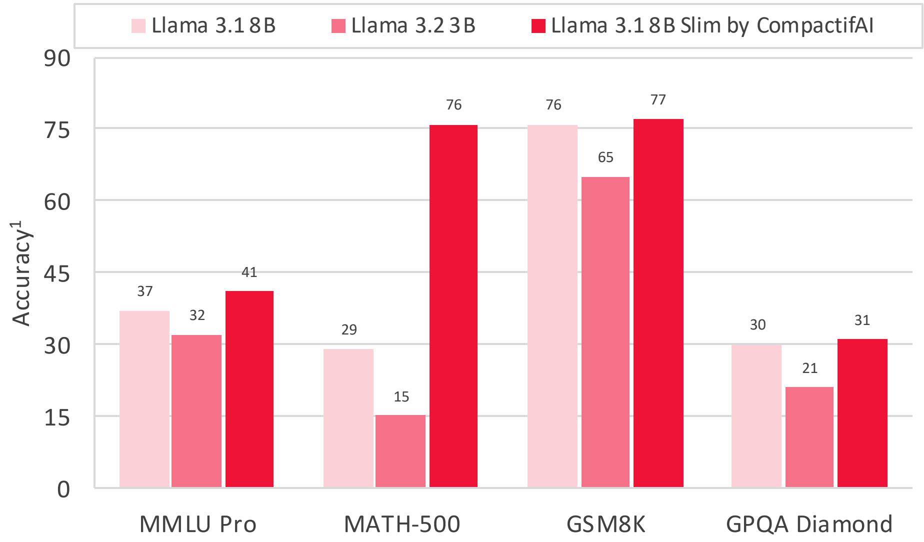
[1] Energy consumption in kWh (Prefill) on NVIDIA L4. Source: Technical report from Sngular

[2] DRAM usage across various sequences – MVC Slim model vs Original versions. Source: External benchmarks by a leading enterprise hardware provider, 2025.

Accuracy

Exceptional performance guaranteed. Our Slim models consistently match or outperform the accuracy of original models on key benchmarks—demonstrating that smaller truly means smarter.

Llama 3.1 8B



Up to

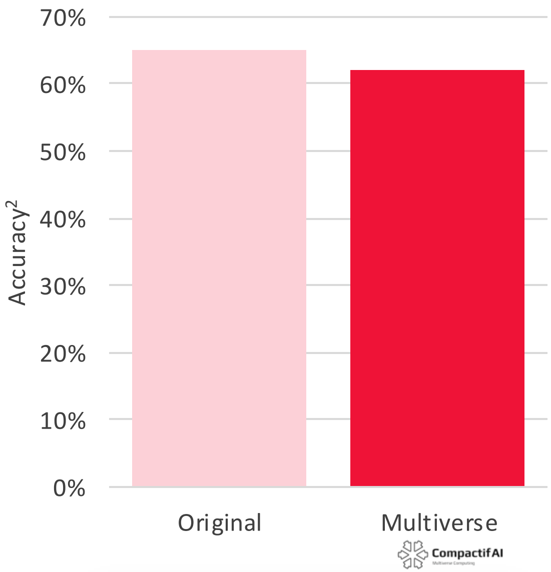
↑107%

Higher Accuracy

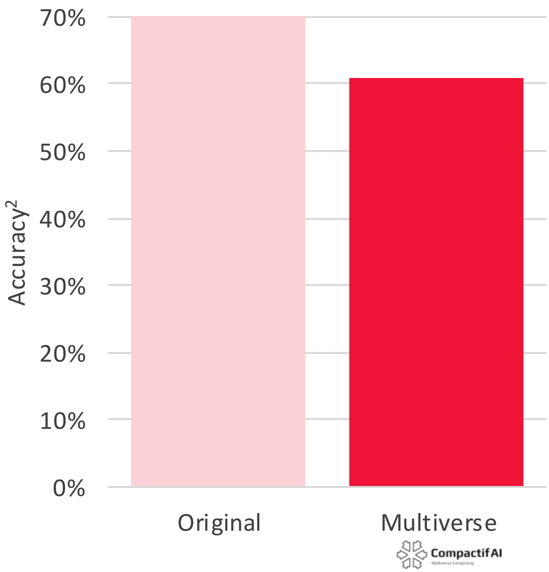
vs. Meta's model

CompactifAI Slim model outperform Llama 3.2 3B by **more than double in accuracy** on key reasoning tasks like MATH-500.

Llama 3.3 70B



Microsoft Phi 4



On average

6%

Minimal Accuracy Drop

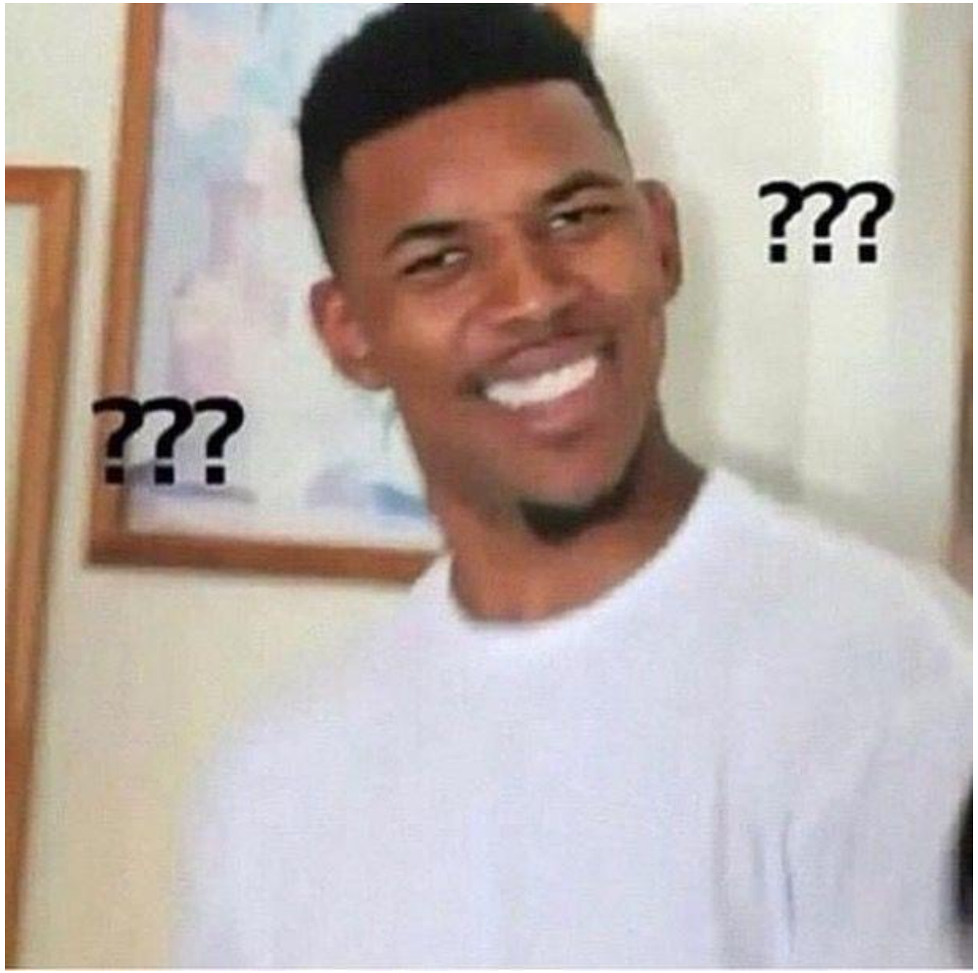
vs. Original models

On average, MVC models maintain comparable accuracy to original versions, with only a minimal drop of up to 6 percentage points — while **delivering superior performance in speed.**

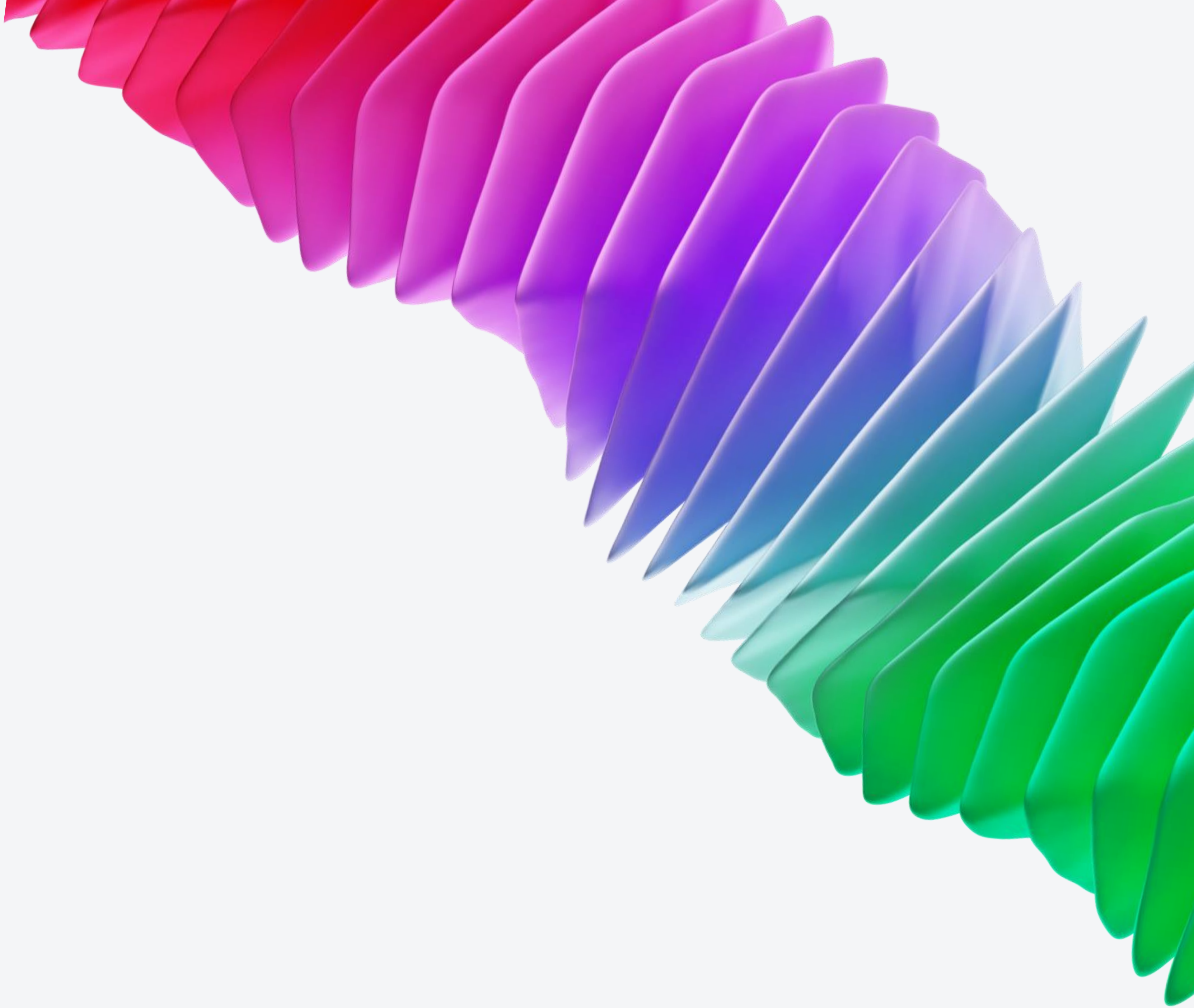
[1] Source: MVC internal benchmark, 2025

[2] Source: External benchmarks by a leading enterprise hardware provider, 2025.

¿Solo comprimis modelos? ¿Quién los usa?



Success Stories



Compressed LLM Application in Telecommunications

Use case: Internal Customer Service Chatbot for 8000 Sales Representatives

Llama 3.1 8B

Industry: Telecommunications

Client: IBEX 35 telecommunications provider with global presence

Goal: Compress LLM to use it within an internal RAG chatbot for agents in client's stores.

Cost reduction	Speed
↓50%	↑46%
incl storage requirements cut by 78%, energy cost cut by 63%, inference cost down 29%	Reduced time to first token, plus extra 26% reduced in-token latency

Compression achieved	Accuracy loss
80%	0%
60% fewer parameters plus quantization	Vs original model when using optimized prompts

Reduced costs

associated with the use of expensive API services (GPT, Gemini, etc.)

Speedup

Reduced latency – faster responses

Efficiency

Low power consumption

Security

Increased data privacy

Compressed Computer Vision Application in Defense & Military (1/2)

Use case: Satellite Image Processing of 670K km2 for Object Detection

YOLO v8-x

Industry: Defense

Client: Transnational defense and security organization

Goal:

Accelerate inference and reduce infrastructure costs.

Keep model accuracy as high as possible.

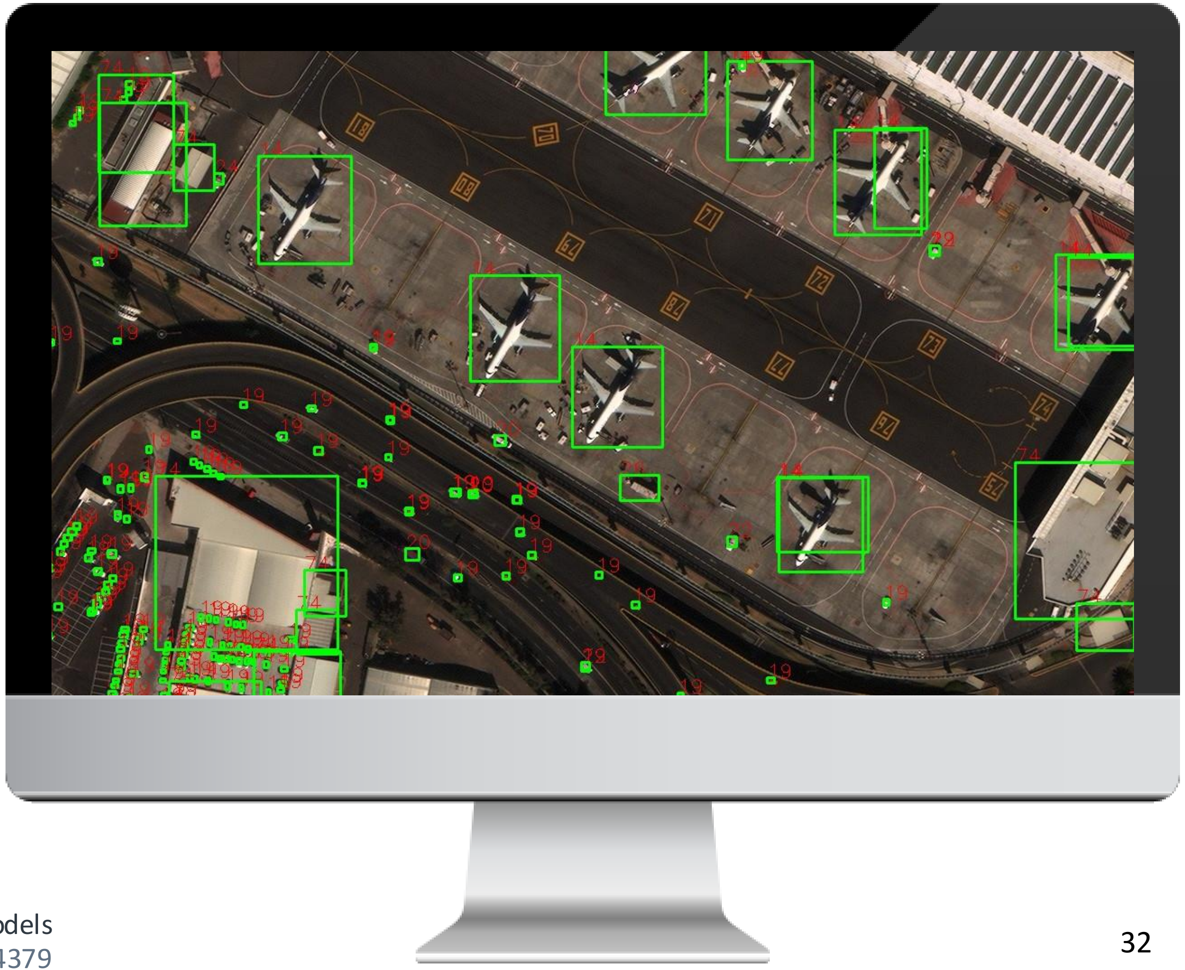
Compression of the computer vision model **YOLOv8-x.**

Use it to process high-resolution satellite imagery.

- 10 cm/pixel resolution.
- 670,000 km2 surface to analyze.
- Refresh rate: 4 times per hour.

Images in 3 different spectral bands: RGB, Infrared, SAR

Example image



See next slide for more info.

Tensor network compressibility of convolutional models
https://arxiv.org/pdf/2403.14379

Compressed Computer Vision Application in Automotive

Use case: On-Edge In-Car Virtual Assistant Using Compressed TextToSpeech AI Model

Style TTS

Industry: Automotive

Client: Leading European automotive manufacturer

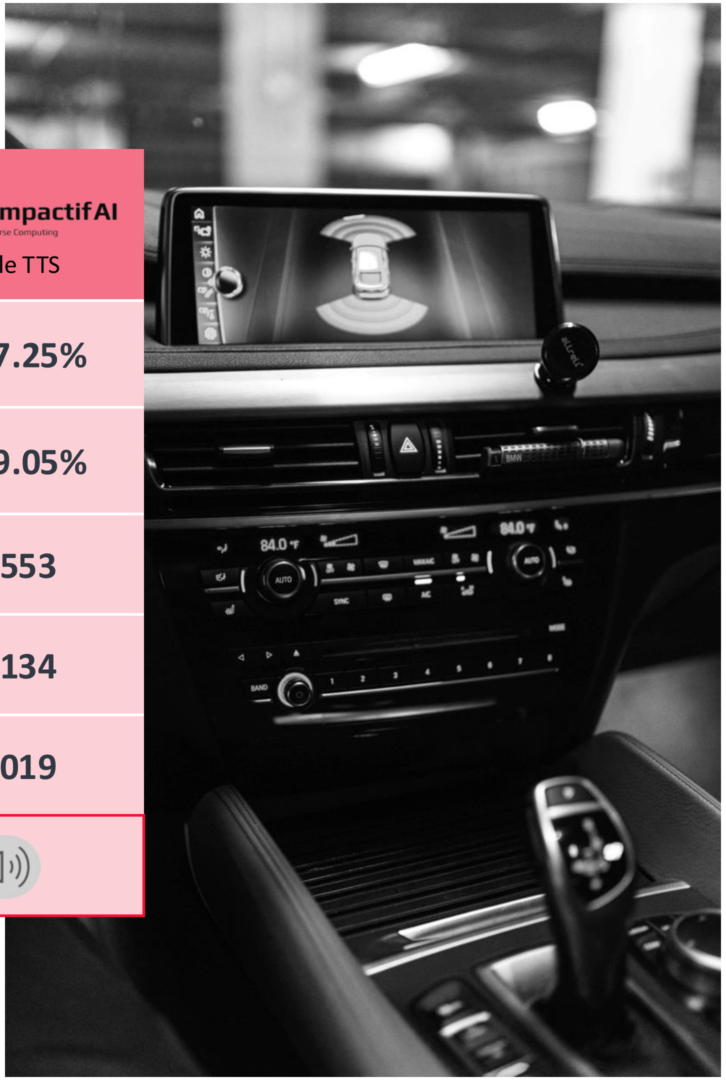
Goal:

- Compress the model as much as possible with no noticeable loss in the quality of the assistant's voice.
- Preliminary results looking promising -37% parameter reduction with no drop in audio quality.

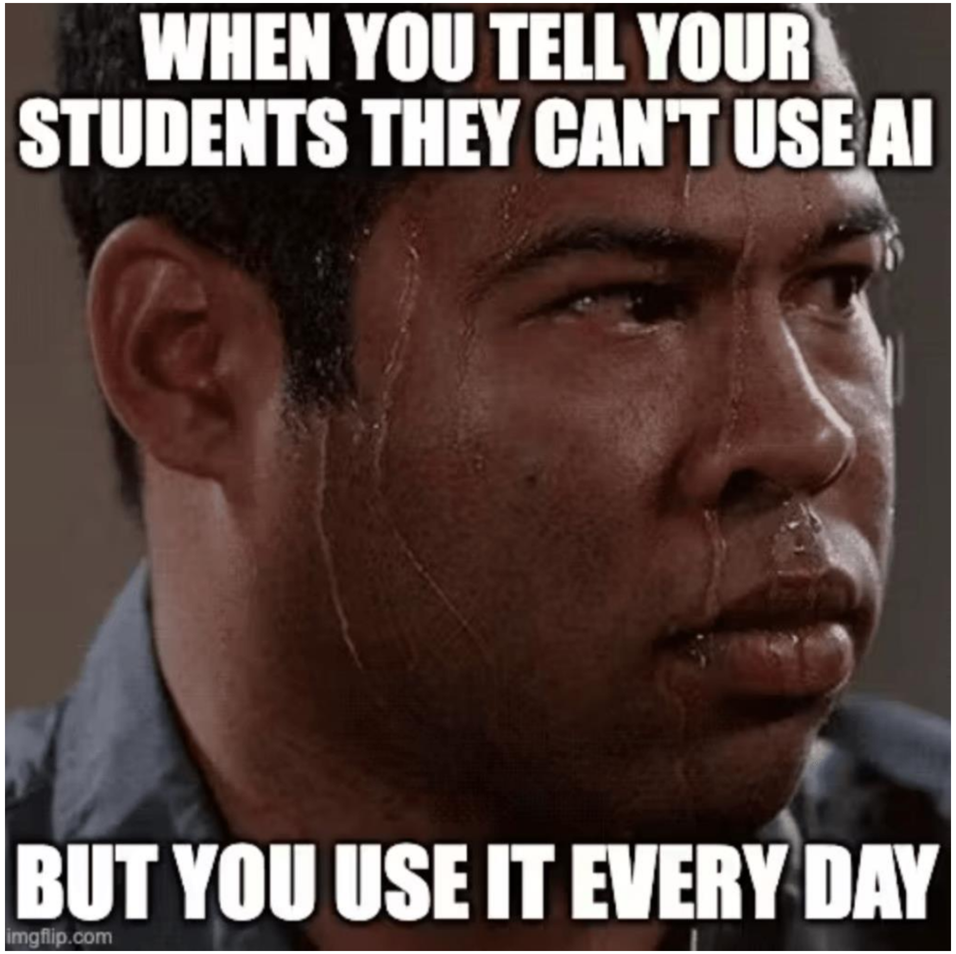
Next Steps

Continue compressing further, and combine CompactifAI with other techniques such as quantization.

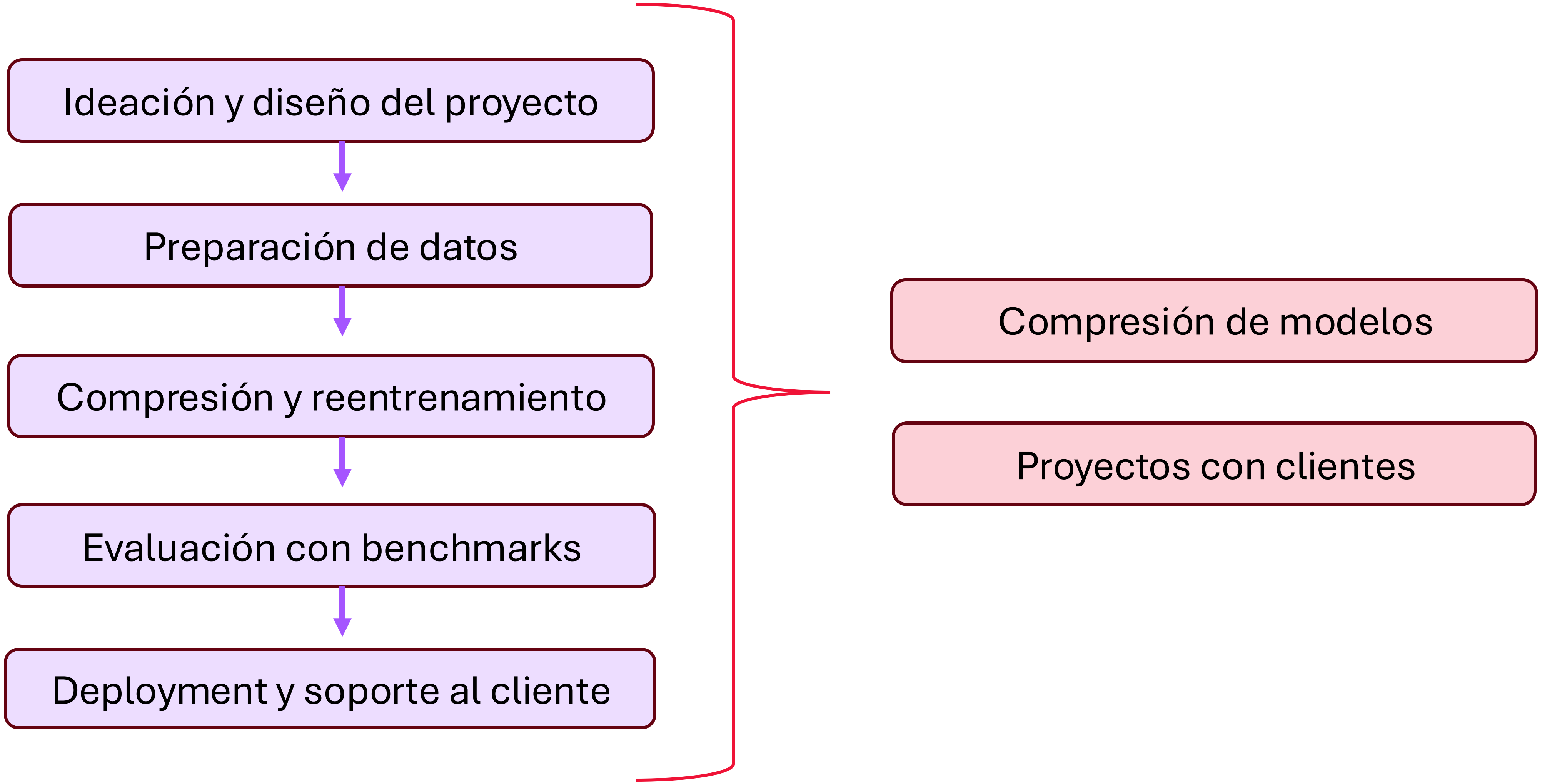
	Base Model Style TTS	**CompactifAI** Style TTS
Parameter Reduction	-	**↓37.25%**
Memory Reduction	-	**↓29.05%**
Accuracy: NISQA MOS	4.507	**4.553**
Accuracy: NISQA Noise	4.275	**4.134**
Accuracy: DNSMOS	4.017	**4.019**
Audio Sample	🔊	🔊



pero... ¿en qué trabajas tú?



Ser ML Engineer en Multiverse Computing - cómo trabajamos


Ideación y diseño del proyecto
Preparación de datos
Compresión y reentrenamiento
Evaluación con benchmarks
Deployment y soporte al cliente
Compresión de modelos
Proyectos con clientes

Equipo, herramientas y retos en Multiverse Computing

👥 Equipo y colaboración

→ Estructura multidisciplinar: MLOps, DevOps, Research, Servicios.
→ Comunicación constante entre equipos.

🧰 Herramientas y lenguajes

→ Python, PyTorch, frameworks de NVIDIA.
→ LangChain, MindEx y librerías de agentes.
→ Infraestructura y MLOps internos.

⚡ Retos y habilidades

→ **Técnicos:** escalabilidad, eficiencia, coordinación.
→ **Personales:** adaptación, comunicación, gestión del tiempo.
→ **Claves:** curiosidad, pensamiento crítico, aprendizaje continuo.

¿Es un buen momento para estudiar ML/IA?

Retos del sector tecnológico

⚙ Retos actuales

→ **Boom de la IA:** cientos de startups, inversión y crecimiento acelerado.
→ **Incertidumbre:** el auge sigue, pero nadie sabe cuánto durará (**burbuja ¿?)**
→ **Avance tecnológico rápido**, aunque el paradigma se mantiene estable.
→ **Desafíos éticos y de talento:** falta de profesionales y necesidad de una IA responsable.

💬 Lo que me ha sorprendido/cambiado mi forma de pensar

→ Las **habilidades sociales** son tan importantes como las técnicas.
→ **Adaptarse a cada persona y contexto** es clave.
→ La colaboración y la empatía marcan la diferencia en los proyectos.

Cómo mantenerse actualizado

🔁 Aprender siempre
La tecnología cambia cada pocos años: seguir aprendiendo es parte del trabajo.

⚙️ Actualizarse constantemente
Surgen nuevos modelos, herramientas y enfoques continuamente.

🧪 Probar cosas nuevas
Participar en proyectos, cursos o retos técnicos mantiene tu mente activa.

🤝 Aprender de otros
Eventos, comunidades y networking te mantienen al día y conectado.

¿Y si volviese a empezar?

Qué haría diferente y habilidades clave

🕰️ Qué haría diferente

→ Repetiría una carrera similar (**IA / Datos / Informática / etc**)

→ Haría un **máster** y luego consideraría un **doctorado fuera de España** por mejores condiciones.

🌱 Habilidades clave (soft skills)

→ **Comunicación y habilidades sociales:** ir a charlas, hablar con ponentes y participar en eventos.

→ **Networking:** crear contactos y aprovechar oportunidades.

→ **Proyectos personales:** diferénciate y demuestra iniciativa.

→ **No todo es técnica:** las soft skills abren más puertas que un currículum impecable.

Máster, prácticas y primer trabajo

🎓 **Elige con propósito**
El sector es competitivo, elige un máster o proyecto que te **haga crecer**.

💼 **Empieza con prácticas**
Las prácticas son la mejor puerta de entrada al sector.

🧭 **Gana experiencia real**
Una vez dentro, es más fácil conseguir un puesto fijo o cambiar de empresa.

🚀 **Haz lo que te motive**
No elijas por nota: busca lo que **te apasione y te rete**.

🧪 **Aprende haciendo**
Crea mini proyectos, investiga y prueba nuevas herramientas.

One last advice

In a world driven by AI, your greatest advantage is still your ability to connect, communicate, and stay human.

En un mundo impulsado por la IA, tu mayor ventaja sigue siendo tu capacidad para conectar, comunicarte y mantener tu humanidad.

"La curiosidad es la clave del éxito"

MULTIVERSE
COMPUTING

https://multiversecomputing.com/join-us

@javieralonso9

javilonso9@gmail.com

Javier Alonso Mencía
Machine Learning Engineer