

Fundamentos para desarrollo de agentes

Fabian Roldan

Configuración de entorno de desarrollo

1. Infraestructura (Panorama de GPUs)
2. Herramientas para mejorar interacción y trabajo con los modelos de lenguaje
3. Sistemas operativos y terminal
4. Editor de Código (Visual Studio Code)
5. Notebooks (Colab)

Mapa de ruta de OpenAI

OpenAI Imagines Our AI Future

Stages of Artificial Intelligence

Level 1	Chatbots, AI with conversational language
Level 2	Reasoners, human-level problem solving
Level 3	Agents, systems that can take actions
Level 4	Innovators, AI that can aid in invention
Level 5	Organizations, AI that can do the work of an organization

Source: Bloomberg reporting

Bloomberg

Breakthrough in AI



[AlphaGo - The Movie | Full award-winning documentary](#)

Si no quieres que te sustituya una máquina, no intentes actuar como una.

Premio Nobel de Física Arno Allan Penzias



Jakub Pachocki

@merettm

Last week, our reasoning models took part in the 2025 International Collegiate Programming Contest (ICPC), the world's premier university-level programming competition. Our system solved all 12 out of 12 problems, a performance that would have placed first in the world (the best human team solved 11 problems).

This milestone rounds off an intense 2 months of competition performances by our models:

- A second place finish in AtCoder Heuristics World Finals
- Gold medal at the International Mathematical Olympiad
- Gold medal at the International Olympiad in Informatics
- And now, a gold medal, first place finish at the ICPC World Finals.

I believe these results, coming from a family of general reasoning models rooted in our main research program, are perhaps the clearest benchmark of progress this year. These competitions are great self-contained, time-boxed tests for the ability to discover new ideas. Even before our models were proficient at simple arithmetic, we looked towards these contests as milestones of progress towards transformative artificial intelligence.

Our models now rank among the top humans in these domains, when posed with well-specified questions and restricted to ~5 hours. The challenge now is moving to more open-ended problems, and much longer time horizons. This level of reasoning ability, applied over months and years to problems that really matter, is what we're after - automating scientific discovery.

This rapid progress also underscores the importance of safety & alignment research. We still need more understanding of the alignment properties of long-running reasoning models; in particular, I recommend reviewing the fascinating findings from the study of scheming in reasoning models that we released today (x.com/OpenAI/status/...)!

“Si sólo tienes un martillo, todo parece un clavo“.

"The Psychology of Science"

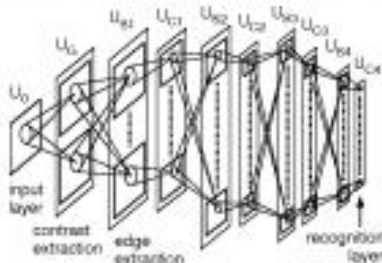


1. Infraestructura (Panorama de GPUs)

¿Qué cambió en el campo de la IA?

Algorithms

Deep neural networks

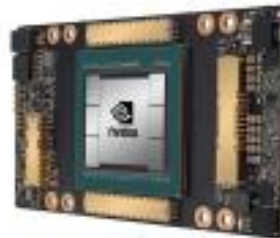


DNN: Aproximador universal de funciones.

(Arquitecturas de redes neuronales artificiales, Optimización)

Hardware

GPUs



Eficientes para multiplicación de tensores (Matrices).

(Computación distribuida, Computación paralelizada)

Data

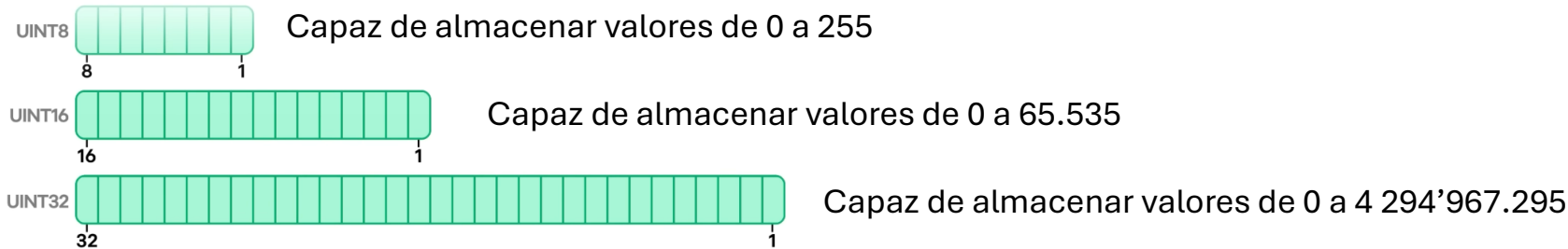
Large scale datasets



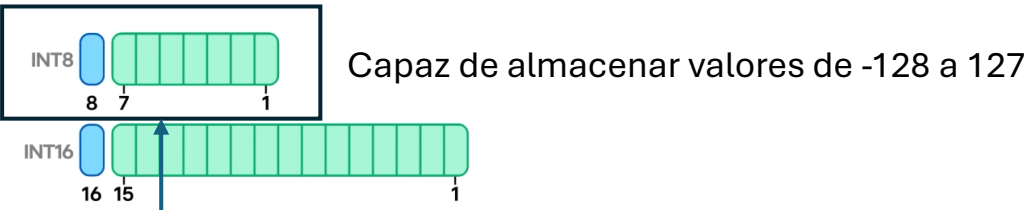
Grandes volúmenes de datos estructurados y no estructurados. (Big data)

Almacenamiento de parámetros del modelo

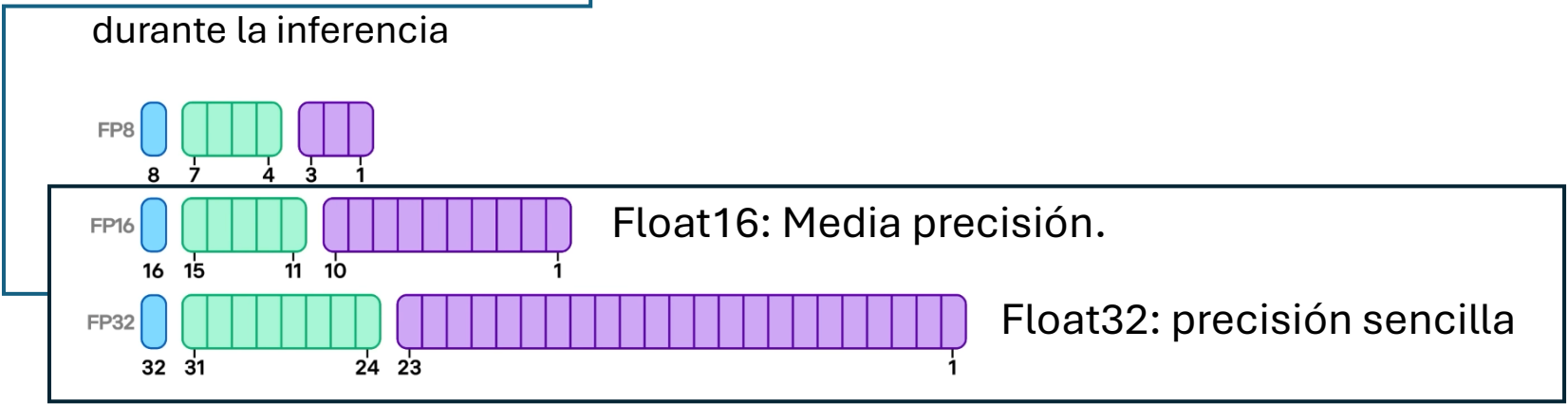
Representación de números naturales



Representación de números enteros (primer bit para el signo)

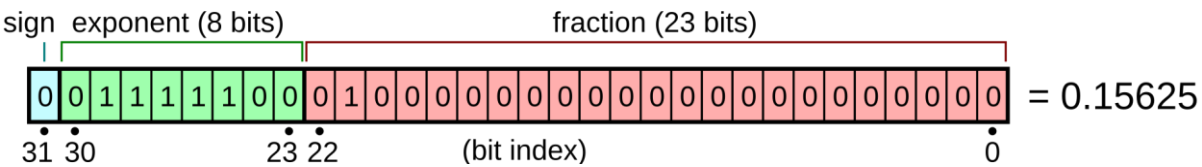


Cuantización de modelos durante la inferencia




Almacenamiento de parámetros del modelo

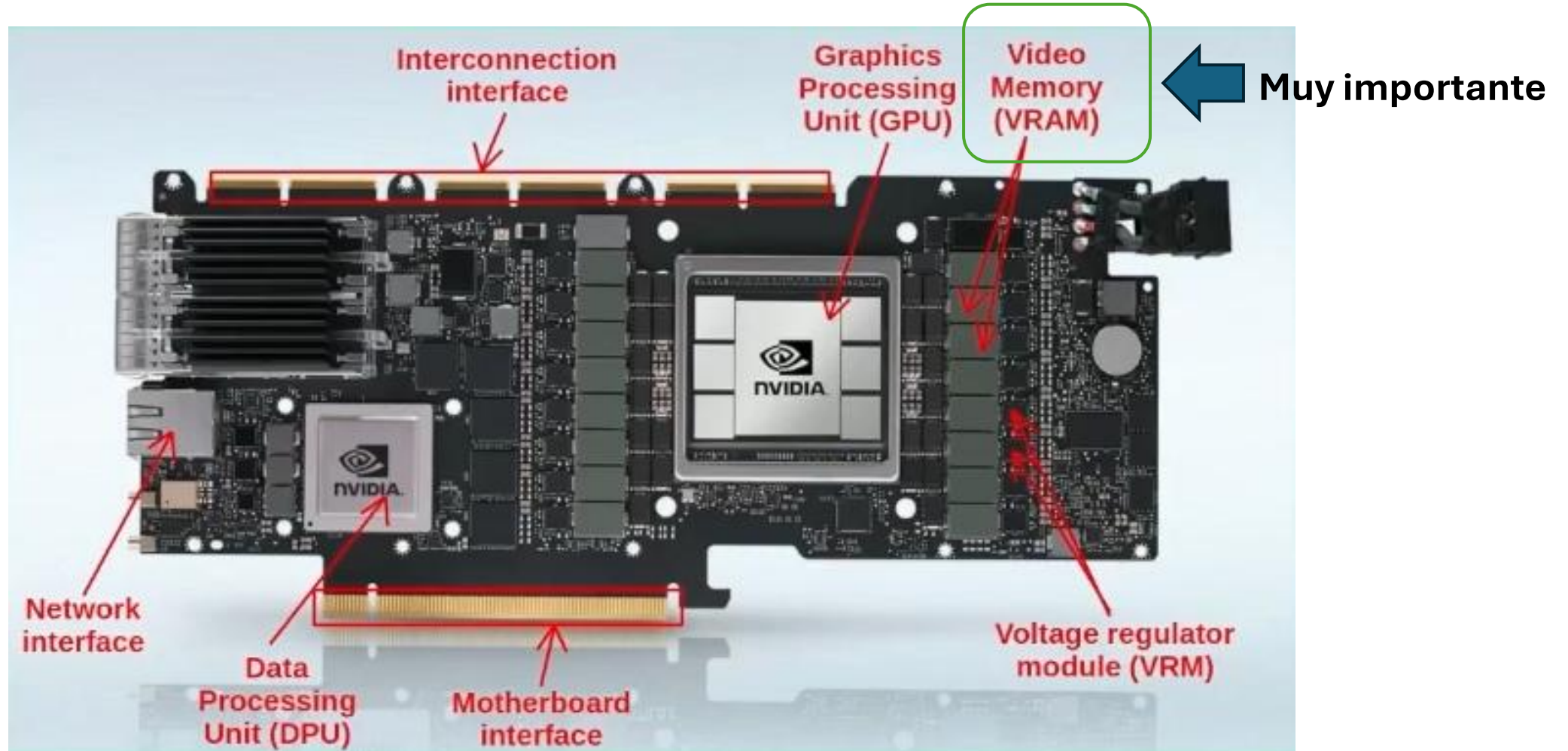
Float32: precisión sencilla



	Representación (número de bits)				
Tipo	Signo	Exponente	Significante	Total	Tamaño
Medio (half)	1	5	10	16	2 bytes (16 bits)
Simple (single)	1	8	23	32	4 bytes (32 bits)
Doble (double)	1	11	52	64	8 bytes (64 bits)

 Parámetros  Peso de todos los parámetros del modelo

Arquitectura GPU



Recursos de GPU para modelos

Un modelo de 7B parámetros puede requerir **~112GB VRAM** para un **entrenamiento completo**, lo que normalmente significa clusters multi-GPU.

Modelo	Inference VRAM	Fine-tuning VRAM (LoRA/QLoRA)
GPT-2 (1.5B)	3GB	~6GB
LLaMA 7B	14GB	~24GB
LLaMA 13B	26GB	~40GB
LLaMA 70B	140GB	Multi-GPU (4-8x A100/H100)

Recursos de GPU para modelos

Model	Lab	Playground	Parameters (B)	Tokens trained (B)
Claude Sonnet 4.5	Anthropic	https://claude.ai/	400	80000
Gemini Robotics 1.5	Google DeepMind		200	20000
Gemini Robotics-ER 1.	Google DeepMind	https://aistudio.g	30	30000
TimesFM-ICF	Google	https://huggingf	0.2	100
Qwen3-Max	Alibaba	https://chat.qwer	1000	36000
Qwen3-Omni	Alibaba	https://github.cor	30	17000
DeepSeek-V3.1-Termin	DeepSeek-AI	https://huggingfa	685	15640
Isaac 0.1	Perceptron	https://huggingfa	2	2000
Grok 4 Fast	xAI	https://grok.com/	200	20000
VaultGemma	Google DeepMind	https://huggingfa	1	13000
Qwen3-Next-80B-A3B	Alibaba	https://huggingfa	80	15000
K2-Think	MBZUAI	https://www.k2th	32	18000
mmBERT	JHU	https://huggingfa	0.307	3000
ERNIE X1.1	Baidu	https://ernie.baid		
ERNIE-4.5-21B-A3B-Th	Baidu	https://huggingfa	21	15000
Klear-46B-A2.5B	Kuaishou	https://huggingfa	46	22000
TildeOpen-30b	Tilde AI	https://huggingfa	30	2000
Qwen3-Max-Preview	Alibaba	https://chat.qwer	1000	36000
Kimi K2-Instruct-0905	Moonshot AI	https://huggingfa	1000	15500
Apertus	ETH Zürich	https://huggingfa	70	15000

Model	Lab	Playground	Parameters (B)	Tokens trained (B)
MAI-1-preview	Microsoft	https://microsoft	500	10000
grok-code-fast-1	xAI	https://github.cor	100	10000
Hermes 4	Nous Research	https://huggingfa	405	15656
Jet-Nemotron-4B	NVIDIA	https://github.cor	4	400
DeepSeek-V3.1-Base	DeepSeek-AI	https://huggingfa	685	15640
Nemotron Nano 2	NVIDIA	https://huggingfa	12.31	20000
Gemma 3 270M	Google DeepMind	https://huggingfa	0.27	6000
GPT-5	OpenAI	https://poe.com/	300	114000
gpt-oss-120b	OpenAI	https://huggingfa	120	30000
gpt-oss-20b	OpenAI	https://huggingfa	20	13000
Claude Opus 4.1	Anthropic	https://claude.ai/	2000	100000
GLM-4.5	Z.AI	https://huggingfa	355	22000
T1	China Telecom Arti	https://github.cor	115	10000
Intern-S1	Shanghai AI Labora	https://huggingfa	235	41000
Step 3	StepFun	https://www.step	321	18000
Qwen3-235B-A22B-Th	Alibaba	https://huggingfa	235	36000
KAT-V1-200B	Kuaishou		200	10000
KAT-V1-40B	Kuaishou	https://huggingfa	40	10000
Qwen3-Coder-480B-A	Alibaba	https://huggingfa	480	36000
Qwen3-235B-A22B-In	Alibaba	https://huggingfa	235	36000

Tipos de GPUs

- **GPUs para consumidores**

- Serie NVIDIA's GeForce RTX

- **GPUs Profesionales**

- Serie NVIDIA's Quadro. están diseñadas para estaciones de trabajo

- **GPUs de data centers**

- NVIDIA's A100, están diseñados para operaciones de ML a gran escala en entornos de servidor.

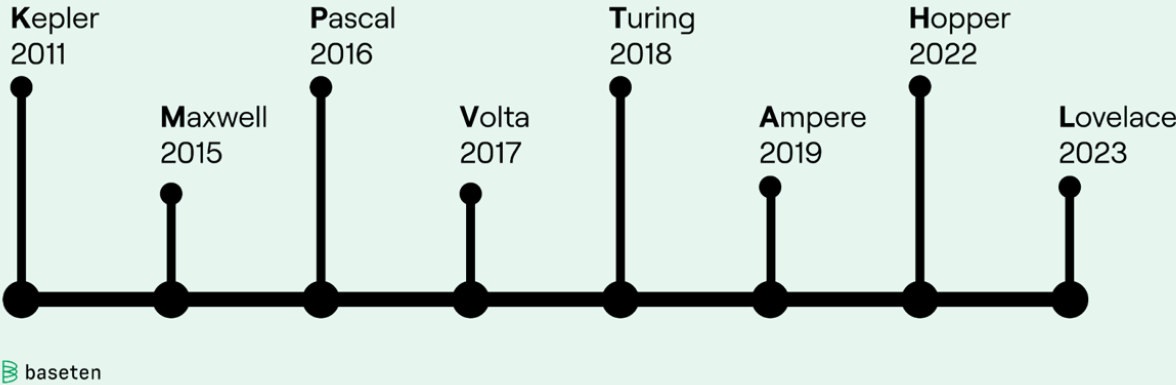
- **GPUs en la nube**

Arquitecturas de GPUs NVIDIA : Turing, Ampere, Hopper, Lovelace, Blackwell

Gpus para datacenters

Primera letra: Indica la generación de la arquitectura



NVIDIA GPU architectures 2011-2023




<https://colab.research.google.com/>

GPU	Architecture	Memory	Best For
T4	Turing	16 GB	Entry-level inference
L4	Lovelace	24 GB	Energy-efficient inference
A10	Ampere	24 GB	Mid-range inference, AI training
A100	Ampere	40 & 80 GB	High-performance LLM training & inference, HPC
H100	Hopper	80 GB	Advanced LLM training & inference, FP8
H200	Hopper	141 GB	Ultra-large models, long-context inference
B100	Blackwell	192 GB	Next-gen AI training, inference, HPC
B200	Blackwell	192 GB	Frontier-scale AI, multi-trillion parameter models

Comparación de costos

Lineup	GeForce	Tesla	Quadro
Name	RTX 2080 Ti	V100	RTX 8000
Generation	7.5 (Turing)	7.0 (Volta)	7.5 (Turing)
Picture			
FP32	11,750 GFLOPS	14,899 GFLOPS	16,300 GFLOPS
FP64	367 GFLOPS	7,450 GFLOPS	510 GFLOPS
Memory	11GB (GDDR6)	32GB (HBM2)	48GB (GDDR6)
Bandwidth	616GB/s	900GB/s	672GB/s
Power	250W	250W	295W
MSRP	\$ 999	\$ 10,000	\$ 5,500

Featured



RTX 4080

NVIDIA GeForce RTX 4080

> Cooling System: Fan

> Boost Clock Speed: 2.51 GHz


> GPU Memory Size: 16 GB

£1,139.⁰⁰

Buy Now

+ Compare

Featured



RTX 4090

NVIDIA GeForce RTX 4090

> Cooling System: Fan


> Boost Clock Speed: 2.52 GHz

> GPU Memory Size: 24 GB

£1,519.⁰⁰

Buy Now

+ Compare



RTX 4060 Ti

NVIDIA GeForce RTX 4060 Ti

> Cooling System: Fan


> Boost Clock Speed: 2.54 GHz

> GPU Memory Size: 8 GB

£379.⁰⁰

Buy Now

+ Compare



RTX 4070

NVIDIA GeForce RTX 4070

> Cooling System: Fan

> Boost Clock Speed: 2.48 GHz

> GPU Memory Size: 12 GB

£569.⁰⁰

Buy Now

+ Compare

Comparación de precios de GPU para datacenters (T4 vs. A100 vs. H100)

Comparación del costo de alquiler de un T4 frente a un A100 frente a un H100 en Google Cloud (us-central1)

GPU	On-Demand	1-Year Commitment	3-Year Commitment
NVIDIA T4 (16 GB)	\$255.50/mo	\$160.60/mo	\$116.80/mo
NVIDIA A100 (40 GB) — 1× A100 in a2-highgpu-1g VM	\$2,681.57/mo	\$1,689.37/mo	\$938.57/mo
NVIDIA A100 (80 GB) — 1× A100 in a2-ultragpu-1g VM	\$3,700.22/mo	N/A	N/A
NVIDIA H100 (80 GB) — 8× H100s in a3-highgpu-8g VM	\$64,597.70/mo (≈\$ 8,074.71 per GPU)	\$44,810.08/mo (≈\$ 5,601.26 per GPU)	\$28,371.00/mo (≈\$ 3,546.38 per GPU)

Selección de GPU

Elija T4 para:

- Cargas de trabajo de inferencia
- Entrenamiento de modelos pequeños (menos de 1B parámetros)
- Desarrollo y creación de prototipos
- Tareas de visión por ordenador
- Proyectos de bajo presupuesto

Elija A100 para:

- Entrenamiento de modelos medianos y grandes
- Ajuste fino de modelos avanzados
- **Inferencia de producción para modelos grandes**
- Configuraciones de entrenamiento multi-GPU
- La mayoría de las aplicaciones comerciales de IA
- La mayoría de los equipos de producción no utilizan sólo un A100, sino que escalan con 4-8 para el ajuste fino y las cargas de trabajo de grandes lotes

Elija H100 para:

- Investigación de vanguardia
- Entrenamiento de modelos con más de 70B millones parámetros
- **Inferencia de alto rendimiento para modelos muy grandes**
- Cuando necesite el máximo rendimiento
- A menos que esté entrenando modelos de más de 70.000 parámetros desde cero, la ampliación a varios A100 ofrece un mejor retorno de la inversión que el salto a un único H100.

Por qué tuvo éxito NVIDIA

Programar
con la GPU



GPU Computing Applications						
Libraries and Middleware						
cuDNN TensorRT	cuFFT, cuBLAS, cuRAND, cuSPARSE	CULA MAGMA	Thrust NPP	VSIPL, SVM, OpenCurrent	PhysX, OptiX, iRay	MATLAB Mathematica
Programming Languages						
C	C++	Fortran	Java, Python, Wrappers	DirectCompute	Directives (e.g., OpenACC)	
CUDA-enabled NVIDIA GPUs						
Turing Architecture (Compute capabilities 7.x)	DRIVE/JETSON AGX Xavier	GeForce 2000 Series	Quadro RTX Series	Tesla T Series		
Volta Architecture (Compute capabilities 7.x)	DRIVE/JETSON AGX Xavier			Tesla V Series		
Pascal Architecture (Compute capabilities 6.x)	Tegra X2	GeForce 1000 Series	Quadro P Series	Tesla P Series		
Maxwell Architecture (Compute capabilities 5.x)	Tegra X1	GeForce 900 Series	Quadro M Series	Tesla M Series		
Kepler Architecture (Compute capabilities 3.x)	Tegra K1	GeForce 700 Series GeForce 600 Series	Quadro K Series	Tesla K Series		
	EMBEDDED	CONSUMER DESKTOP, LAPTOP	PROFESSIONAL WORKSTATION	DATA CENTER		

Herramientas para mejorar interacción y trabajo con los modelos de lenguaje



[DeepL for Windows | Translation and writing improvement](#)

Uso:

- Traducir a un click sin salir del entorno de trabajo
- Mejorar redacción (**Primer ajuste del prompt**)

Beneficios:

- No perder foco, no cambiar de pestaña
- Velocidad
- Menos tokens, Más contexto
- Mejores respuestas

Cómo usarlo:

1. Selecciona el texto que deseas traducir
2. Pulsa el atajo de teclado Ctr + C + C (mantén pulsado Ctr y teclea C dos veces)
3. El texto traducido aparecerá en una pequeña ventana en la parte superior de la app que estés utilizando

Herramientas para mejorar interacción y trabajo con los modelos de lenguaje



Notion

Uso:

- Guardar de forma estructurada las respuestas generadas por los LLMs
- Usar notas como contexto para los LLMs (MCP)

Beneficios:

- Guardar código y ecuaciones con mejor visualización
- Guardar en formato markdown (.md)

Herramientas para mejorar interacción y trabajo con los modelos de lenguaje



Visual Studio Code

[Download Visual Studio Code - Mac, Linux, Windows](#)

[Download Python | Python.org](#)

Uso:

- Escribir código en distintos lenguajes
- Usar asistentes/agentes para escribir código.

Beneficios:

- Un solo lugar para gestionar proyectos
- Centralizar trabajo.

Algunas extensiones:



PDF Viewer


Portable document format (PDF)

Mathematic Inc



GitHub Copilot


Your AI pair programmer

 GitHub



Codex – OpenAI's coding agent

One agent for everywhere you code

 OpenAI



Markdown All in One


All you need to write Markdown

Yu Zhang



Claude Code for VS Code

Claude Code for VS Code: h

 Anthropic