Self-attention y bloque Transformer

El Transformer entero, paso a paso.

Conceptos clave

- **Self-attention**: Q, K, V vienen todos de la misma secuencia. Cada token atiende a todos.
- **Multi-head**: varias atenciones en paralelo con dim $d_k = d_{model}/h$. Coste total igual.
- Bloque encoder: → `LayerNorm` → Attn → residual → `LayerNorm` → MLP → residual →.
- Tres formas: encoder (BERT), decoder (GPT, Llama), enc-dec (T5, BART).

Ejemplo a mano

Self-attention en cinco pasos sobre 4 tokens, $d_{model} = 4$:

1. Setup: $X = \{x_1, x_2, x_3, x_4\} \in \mathbb{R}^{4\times 4}$.
2. Proyecciones: $Q = XW^Q$, $K = XW^K$, $V = XW^V$.
3. Scores: $S = QK^\top/\sqrt{d_k} \in \mathbb{R}^{4\times 4}$.
4. Softmax: $\alpha_{ij} = e^{S_{ij}}/\sum_k e^{S_{ik}}$.
5. Salida: $z_i = \sum_j \alpha_{ij} v_j \in \mathbb{R}^4$.

Multi-head: repetir con h conjuntos W^Q, W^K, W^V, concatenar y multiplicar por W^O.

Concepto → aplicación: tres formas del Transformer

Forma	Modelo emblemático	Tarea típica
Solo encoder	BERT, RoBERTa, DeBERTa.	Clasificación, NER.
Solo decoder	GPT-5, Llama 3, Mistral.	Generación autoregresiva.
Encoder-decoder	T5, BART, MarianMT.	Traducción, resumen.
Visión	ViT, Swin, ConvNeXt.	Clasificación imágenes.
Multimodal	CLIP, Flamingo, GPT-4V.	Imagen + texto.

Idea clave del Día 09. **El Transformer es la arquitectura única** que ha desplazado al SOTA en seis dominios (NLP, visión, audio, robótica, código, series). En 2026 el primer modelo que pruebes es siempre Transformer; las alternativas (Día 10) son optimizaciones para contextos largos.

Para llevarte: lectura Vaswani et al., *Attention is All You Need* (sec. 3 y 5); ejercicio: visualizar heatmaps por cabeza sobre una frase en español.