Adaptación corporativa: SFT, LoRA, RLHF, DPO

Cuándo ajustar el modelo y cuándo NO.

Conceptos clave

- Escalón de coste y datos: in-context < RAG < **LoRA** < SFT total < **DPO** < RLHF.
- LoRA aprende dos matrices pequeñas A, B con rango $r \ll d$. Resto del modelo congelado.
- DPO ha desplazado a RLHF: no requiere modelo de recompensa, solo pares (preferido, rechazado).
- Regla práctica: **empieza por RAG**; sube a LoRA solo si RAG no captura tono o formato.

Ejemplo a mano

LoRA con $r = 8$ sobre Llama 7B (matrices de atención $d = 4096$):

$$\text{params LoRA} = 2 \cdot (d \cdot r) = 2 \cdot 4096 \cdot 8 = \mathbf{65\,536} \quad \text{frente a } d \cdot d = 16{,}7\,\text{M del SFT.}$$

Reducción $\sim 256\times$ en parámetros entrenables. Tiempo: 1 h en 1 GPU A100 (vs. 6 h en 8 GPUs A100 para SFT total). Coste cloud: 8 \$ vs. 240 \$ por entrenamiento. Mantenimiento: cambias A, B sin tocar W.

Concepto → aplicación: cuándo cada técnica

Técnica	Caso típico	Datos	Coste
RAG	Conocimiento corporativo cambiante.	Documentos.	Bajo
LoRA	Tono y formato de marca, dominio cerrado.	1K–10K ejemplos.	Medio
SFT total	Cambio drástico de tarea o lenguaje.	50K+ ejemplos.	Alto
DPO	Alinear comportamiento con preferencias.	Pares preferencia.	Alto
RLHF	Producto crítico con humanos.	Recompensa labeled.	Muy alto

Idea clave del Día 16. Para el 90 % de casos empresariales basta con **RAG + LoRA**. RLHF es solo para producto en frontera (Claude, GPT). El árbol de decisión empieza siempre por RAG.

Para llevarte: lectura Hu et al., *LoRA*; práctica: fine-tune con LoRA un Llama 3.2 1B sobre 500 pares (pregunta, respuesta) corporativos.