Deep Learning para Business Analytics

Día 16 · Tema 6 — Adaptación de modelos

Cuándo fine-tunear y cuándo NO.

Eduardo C. Garrido-Merchán

Profesor Propio Adjunto, Métodos Cuantitativos, ICADE

ecgarrido@comillas.edu

Madrid, Día 16

ejemplos del día

Lo que vamos a cubrir hoy.

1. Cinco técnicas, comparadas

2. LoRA y SFT

3. RLHF y DPO

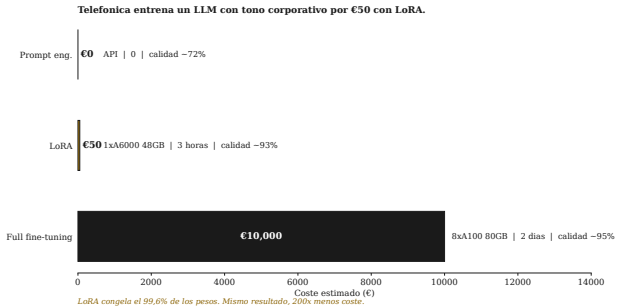
4. Árbol de decisión

Al final del día: sabes elegir entre RAG, LoRA, SFT, DPO y RLHF para un caso de negocio concreto.

Parte 1 / 4

Cinco técnicas, comparadas

Telefónica entrena un LLM con tono corporativo por €50 con LoRA.



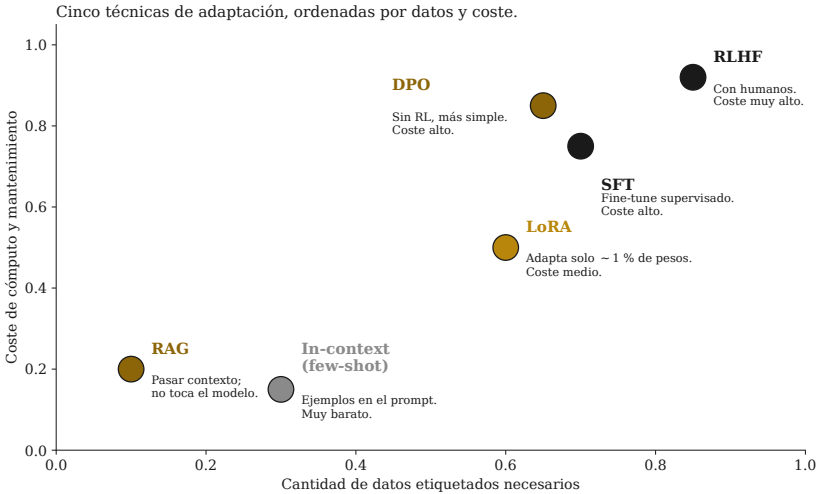
Contexto. 💰 Un banco o telco quiere que su LLM hable con tono corporativo. ¿Cuánto cuesta?

Comparación. Full fine-tuning: €10 000, 8×A100, 2 días. LoRA: €50, 1×A6000, 3 horas. Prompt engineering: €0 pero menor calidad.

Clave. ⚙️ LoRA congela el 99,6 % de los pesos y entrena adaptadores de rango bajo. Mismo resultado, 200× menos coste.

Idea clave. Para la mayoría de casos corporativos, LoRA domina el trade-off calidad-coste.

Cinco técnicas ⚖️ en un cuadrante de datos × coste.

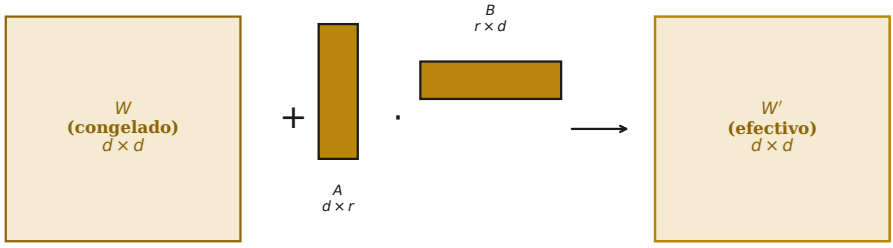


Parte 2 / 4

LoRA y SFT

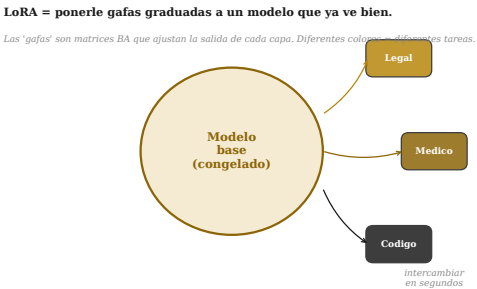
LoRA ⚙️: adapta dos matrices pequeñas, no toda la red.

LoRA: aprende dos matrices pequeñas A, B con rango $r \ll d$ y suma al peso.



Si $r = 8$ y $d = 4096$: A, B son 32K + 32K = 64K parámetros frente a 16,7M del original.

LoRA = ponerle gafas graduadas a un modelo que ya ve bien.



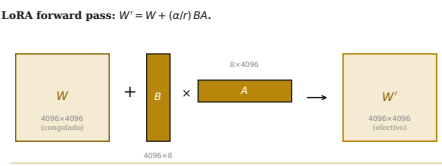
Analogía. 🧠 El modelo base es un cerebro que ya sabe mucho. Las "gafas" son matrices pequeñas *BA* que ajustan la salida de cada capa.

Multi-tarea. Diferentes colores de lente = diferentes tareas (legal, médico, código). El modelo base *no cambia*.

Intercambio. ⚙️ Puedes cambiar de "gafas" en segundos: cargas un adaptador distinto sin recargar el modelo de 7B parámetros.

Idea clave. Un solo modelo base sirve a N departamentos con N adaptadores LoRA.

LoRA forward pass: $W' = W + (\alpha/r)\,BA$.



Dimensiones. W es 4096×4096 (congelado). B es 4096×8, A es 8×4096. Solo 64K parámetros entrenables frente a 16,7M.

Ejemplo. $W=\left(\begin{smallmatrix}1&2\\3&4\end{smallmatrix}\right)$, $A=(0{,}1,\ 0{,}2)$, $B=\left(\begin{smallmatrix}0{,}5\\0{,}3\end{smallmatrix}\right)$, $\alpha=r=1$, $x=(1,1)^T$.

$$Wx = (3,\ 7)^T.$$

$$BAx = \left(\begin{smallmatrix}0{,}5\\0{,}3\end{smallmatrix}\right)(0{,}3) = (0{,}15,\ 0{,}09)^T.$$

$$W'x = (3{,}15,\ 7{,}09)^T.$$

Idea clave. El ajuste es una perturbación aditiva de rango bajo sobre la salida original.

Ejemplo a mano 💰: cuánto ahorra LoRA frente a SFT total.

Ejemplo a mano: coste comparado entre SFT total y LoRA $r = 8$ sobre Llama 7B.

Concepto	SFT total	LoRA r=8
Parámetros entrenables	8 000 M (Llama 7B)	32 M
GPU necesaria	8 × A100 80GB	1 × A100 40GB
Tiempo (10K ejemplos)	6 h	1 h
Coste cloud (estimado)	$240 por entrenamiento	$8 por entrenamiento
Mantenimiento mensual	Re-train completo	Cambiar A,B sin tocar W

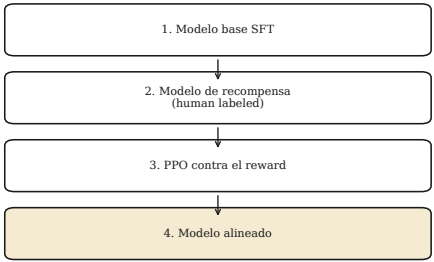
Parte 3 / 4

RLHF y DPO

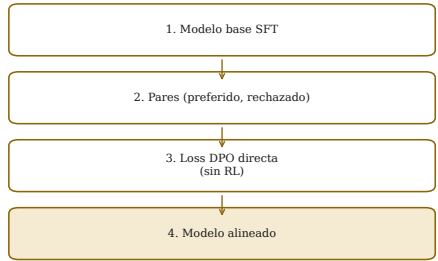
RLHF 🧠 clásico vs DPO 🎯 moderno.

RLHF entrena un modelo de recompensa primero; DPO se salta ese paso.

RLHF (clásico, 2022)

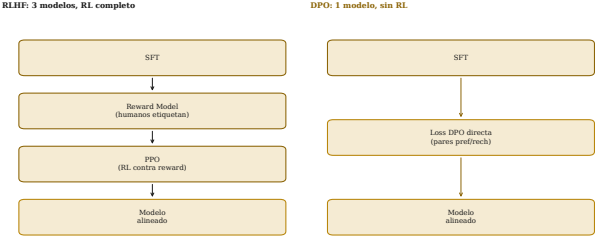


DPO (moderno, 2023)



DPO: alineación sin reward model ni RL.

DPO absorbe el reward model dentro de la política: sin RL, misma calidad.



Pipeline. 🧠 RLHF necesita 3 modelos (SFT → RM → PPO). DPO necesita 1 (SFT → DPO).

Loss DPO.

$$\mathcal{L} = -\mathbb{E}[\log \sigma(\beta\,\Delta)]$$

donde $\Delta = \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{SFT}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{SFT}}(y_l|x)}$.

Gradiente. 🎯 Empuja π_θ hacia y_w (preferido) y lejos de y_l (rechazado). DPO absorbe el reward model dentro de la política.

Idea clave. Mismo resultado que RLHF, sin inestabilidades de PPO, un solo modelo.

Concepto → aplicación: cuándo cada técnica.

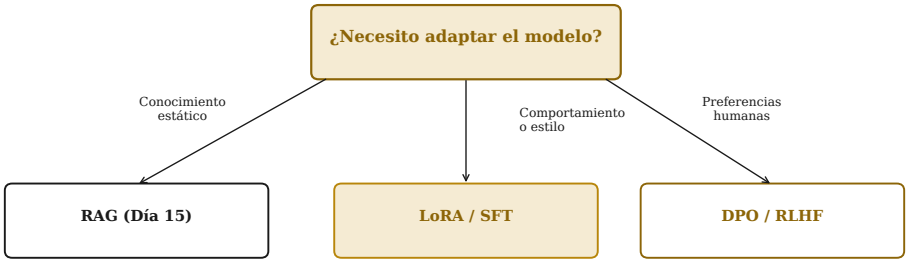
Técnica	Caso típico	Datos necesarios	Coste
RAG 🤖	Conocimiento corporativo cambiante.	Documentos.	Bajo.
LoRA ⚙️	Tono y formato de marca, dominio cerrado.	1K–10K ejemplos.	Medio.
SFT total	Cambio drástico de tarea o lenguaje.	50K+ ejemplos.	Alto.
DPO 🎯	Alinear comportamiento con preferencias.	Pares (preferido,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

Idea clave. Para el 90 % de casos empresariales basta con RAG + LoRA. RLHF es solo para frontera.

Parte 4 / 4

Árbol de decisión

Decidir ⚖️: ¿qué adaptar y cómo?



Regla práctica: empieza por RAG; sube a LoRA solo si RAG no captura el estilo o el formato.

Lo que el alumno se lleva a casa de este día.

Concepto 1. Hay un escalón claro de coste 💰 y datos: in-context < RAG < LoRA < SFT total < DPO < RLHF.

Concepto 2. LoRA ⚙️ ahorra dos órdenes de magnitud frente a SFT total con calidad similar.

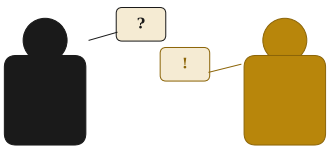
Concepto 3. DPO ha desplazado a RLHF como técnica moderna de alineamiento 🎯 por preferencias.

Concepto 4. El árbol de decisión ⚖️ empieza por RAG y solo sube cuando RAG no captura tono o formato.

Para la práctica. Fine-tunear con LoRA un Llama 3.2 1B sobre 500 pares (instrucción, respuesta) corporativos sintéticos.

Próximo día. Día 17 (Tema 6.3): multi-agente en producción.

Pausa para debatir: ¿Tiene sentido hablar de .el modelo de la empresa X¿



La pregunta. Si todos los modelos parten de Llama o Mistral y aplican LoRA, ¿hay "modelo propio.º es mercadotecnia?

Posición A. Es mercadotecnia: el peso de la diferencia está en los datos.

Posición B. Es real: LoRA cambia comportamiento y formato lo suficiente.

Reglas. 5 minutos, dos turnos de palabra por bando, móviles boca abajo. Yo modero, no participo.

Hasta el Día 17.

Multi-agente real, con CRM, ERP y data warehouse.

Eduardo C. Garrido-Merchán

ecgarrido@comillas.edu