TEMA 6 – HOJA DE EJERCICIOS III

Se plantean diferentes ejercicios basados en post-training:

Ejercicio 1 — Entendiendo el papel del post-training

Un LLM base responde así:

Pregunta:

"¿Puedo usar antibióticos para tratar un resfriado común?"

Respuesta del modelo base:

"Los antibióticos matan bacterias. Algunos médicos los recetan en infecciones respiratorias."

a) Explica por qué esta respuesta puede ser problemática desde el punto de vista de alineación.
b) ¿Qué objetivo del post-training está relacionado con este problema?
c) ¿Qué técnica de post-training ayudaría a corregir este tipo de comportamiento?

Ejercicio 2 — SFT vs RLHF

Se quiere mejorar un modelo que ya responde correctamente a preguntas de historia, pero sus respuestas son largas, poco claras y a veces demasiado técnicas.

a) ¿Sería suficiente aplicar SFT? Justifica.
b) ¿Qué aportaría RLHF que SFT no puede capturar fácilmente?
c) Da un ejemplo concreto de preferencia humana que RLHF sí podría aprender.

Ejercicio 3 — Modelado de preferencias

Se presentan dos respuestas del modelo a la pregunta:

"Explica qué es el cambio climático."

Respuesta A:

Explicación muy técnica, correcta pero difícil de entender.

Respuesta B:

Explicación clara, con ejemplos cotidianos, pero menos detallada.

a) ¿Por qué este tipo de comparación es clave en post-training?
b) ¿Qué técnica usaría directamente pares A vs B como señal de aprendizaje?
c) ¿Qué riesgo aparece si siempre se favorece la respuesta más simple?

Ejercicio 4 — Pipeline de post-training

Ordena las siguientes etapas y justifica brevemente:

- Instruction Tuning
- Modelado de preferencias
- SFT
- Optimización por preferencias (RLHF/DPO)
- Fine-tuning de dominio

Luego responde:

¿Qué aporta cada etapa que no aportan las anteriores?

Ejercicio 5 — Riesgos del post-training

Explica brevemente (2–3 líneas cada uno):

a) Reward hacking
b) Sobrealineación
c) Pérdida de generalidad tras fine-tuning de dominio

Ejercicio 6 — Prompt Engineering vs Post-Training

Para el problema:

“El modelo da respuestas demasiado seguras aunque no tenga suficiente información.”

a) ¿Podría mitigarse solo con prompting?
b) ¿Qué objetivo de post-training está implicado?
c) ¿Qué tipo de señal de entrenamiento ayudaría a que el modelo aprenda a decir “no lo sé”?