Sesión 2 — Prompt engineering y ética

- Estructura: rol, objetivo, contexto, formato
- Técnicas: few-shot, instrucciones, formato de respuesta
- Riesgos: sesgos, alucinaciones y límites de los LLMs
- Uso responsable: privacidad, derechos de autor, transparencia

Objetivos de la sesión

- Fundamentos del prompt engineering: estructura, claridad, contexto
- Técnicas básicas: few-shot prompting, instrucciones, formato de respuesta
- Sesgos, alucinaciones y límites de los LLMs
- Uso responsable de IA generativa: privacidad, derechos de autor, transparencia

Fundamentos del prompt engineering

- Claridad: explica la tarea en lenguaje simple y directo
- Contexto: incluye datos, restricciones, audiencia y tono
- Desambiguación: define qué NO hacer y qué ignorar
- Verificabilidad: pide evidencia, fuentes o formato validable

Estructura recomendada (RTF)

- Rol: quién es el asistente (p. ej., "experto en…")
- Tarea: qué debe lograr (objetivo medible)
- Formato: cómo debe responder (estructura, longitud, validación)

Ejemplo breve:

Rol: editor técnico. Tarea: resume el texto en 5 viñetas para ejecutivos. Formato: JSON con claves `bullets` (lista) y `riesgos` (lista).

Técnicas básicas

- Zero-shot: consigna directa, sin ejemplos
- Few-shot: de 1 a 5 ejemplos representativos y variados
- Cadena de instrucciones: pasos secuenciales con criterios de éxito
- Restricciones explícitas: longitud, estilo, idioma, validación de esquema
- Auto-chequeo: “si no hay suficiente evidencia, di ‘no sé’ y pide más datos”

Ética y uso responsable

- Derechos de autor y atribución
- Datos sensibles y PII
- Transparencia y verificabilidad
- Sesgos y representaciones: revisar lenguaje y cobertura
- Alucinaciones: preferir respuestas verificables y con fuentes

Plantilla de prompt

- Contexto
- Instrucciones claras
- Ejemplos (opcional)
- Formato de salida validable

Sugerencia: indicar audiencia, tono, límites de longitud y campos obligatorios.

Evaluación e iteración

- Define criterios: precisión, cobertura, estructura, trazabilidad
- Itera en ciclos cortos: cambiar 1–2 variables por vez
- A/B testing con variantes de rol/ejemplos/formatos
- Mide con rúbrica simple o script (`ejercicios/ex_prompt_eval.py`)

Taller breve

- Itera un prompt en 3 ciclos (zero-shot → few-shot → formato validable)
- Mide calidad con `ex_prompt_eval.py` (F1, estructura, adecuación, requisitos)
- Usa las plantillas en `prompts_demo.md` para la demo en vivo