Curso base

INTELIGENCIA ARTIFICIAL DESDE CERO

LECCIÓN 23
LLMs, ética y verificación avanzada

Cómo usar modelos de lenguaje con criterio para reconocer alucinaciones, sesgos y riesgos de privacidad,
verificar mejor las respuestas y aplicar mitigaciones prácticas en estudio y trabajo

Lección	23	**Tema**	LLMs, ética y verificación avanzada
Nombre del participante	_____	**Fecha**	_____

Actividad central

Actividad principal: analizar riesgos de uso de LLMs, revisar dos ejemplos guiados y construir un protocolo breve de verificación y mitigación.
Producto esperado: guía personal de uso responsable con protocolo breve de verificación y mitigación.

Propósito de la guía. Comprender cómo funcionan los modelos de lenguaje a nivel práctico, reconocer riesgos frecuentes como alucinaciones, sesgos y problemas de privacidad, y aplicar estrategias sencillas de verificación y mitigación antes de usar una respuesta en estudio, trabajo o vida diaria.

Idea central

Los modelos de lenguaje grandes, o LLMs, pueden producir respuestas muy útiles, rápidas y convincentes. Sin embargo, esa fluidez no garantiza que todo lo que afirman sea correcto, justo o seguro. Por eso su uso responsable exige criterio humano, revisión y verificación.

En esta guía vamos a estudiar cuatro focos clave: alucinaciones, sesgos, privacidad y mitigaciones. La meta no es desconfiar de toda respuesta, sino aprender a distinguir entre una ayuda valiosa y una salida que debe revisarse antes de usarse en decisiones, tareas, informes o documentos importantes.

¿Qué vamos a hacer en esta guía?

Primero construiremos una idea clara de qué es un LLM y por qué requiere cuidado. Después trabajaremos un primer ejemplo guiado para detectar una respuesta convincente pero dudosa. Luego desarrollaremos un segundo ejemplo guiado para mejorar prompts, proteger la privacidad y reducir sesgos. Finalmente cerraremos con una plantilla breve de verificación avanzada y una lista de recomendaciones prácticas.

Vocabulario mínimo

Concepto	Explicación sencilla
LLM	Modelo de lenguaje capaz de redactar, resumir, explicar y responder preguntas a partir de grandes cantidades de texto.
Alucinación	Respuesta incorrecta o inventada que puede sonar convincente aunque no esté verificada.
Sesgo	Tendencia a reproducir estereotipos, omisiones o tratamientos desiguales presentes en datos o instrucciones.
Privacidad	Cuidado con datos personales, sensibles o confidenciales que no deberían compartirse sin necesidad.
Verificación	Proceso de contrastar una respuesta con fuentes, contexto y sentido crítico antes de usarla.
Mitigación	Acciones concretas para reducir riesgo: pedir cautela, revisar, comparar y limitar datos sensibles.

¿Por qué un LLM exige cuidado?

Riesgo	Señal práctica	¿Por qué importa?
Alucinaciones	Da datos exactos, citas o fechas sin respaldo claro.	Podemos repetir errores con apariencia de autoridad.
Sesgos	Usa generalizaciones injustas o lenguaje estereotipado.	Puede afectar la calidad y la justicia de una decisión.
Privacidad	Pide o usa más datos personales de los necesarios.	Puede exponer información sensible del usuario o de terceros.
Exceso de confianza	Responde con tono muy seguro en temas complejos.	Hace más fácil aceptar una salida incorrecta sin revisar.

Idea clave para principiantes

Un LLM puede escribir con mucha seguridad y aun así equivocarse. Por eso no basta con que una respuesta suene bien: hay que comprobar si realmente se sostiene.

Ejemplo guiado 1: una respuesta convincente, pero dudosa

Imaginemos que un estudiante o trabajador pregunta a un asistente de IA por una fecha exacta, una norma, una cita o una cifra concreta. El asistente responde con gran seguridad y entrega detalles muy precisos. A primera vista la respuesta parece excelente. El problema es que todavía no sabemos si es verdadera.

Este primer ejemplo no busca demonizar la IA. Busca enseñar una pausa crítica: antes de copiar o usar una respuesta, debemos preguntarnos qué parte parece un hecho confirmado, qué parte parece inferencia y qué parte necesita verificación externa.

Paso	Qué debe hacer el estudiante
Paso 1	Leer la respuesta completa y marcar los datos más delicados: cifras, fechas, nombres, normas o citas.
Paso 2	Separar lo que parece un hecho de lo que parece una interpretación o una suposición.
Paso 3	Pedir al modelo que indique su nivel de certeza y que advierta posibles dudas o límites.
Paso 4	Contrastar la información con una fuente confiable: documento oficial, material del curso o sitio reconocido.
Paso 5	Redactar una versión prudente de la respuesta, aclarando lo que sí está verificado y lo que aún falta revisar.

Señales de alerta en una respuesta dudosa

Señal	Cómo leerla
Demasiada precisión sin apoyo	Puede ser una invención presentada con tono seguro.
Fuente vaga o inexistente	No basta con mencionar una fuente; hay que poder verificarla.
Lenguaje muy absoluto	Frases como *siempre*, *nunca* o *totalmente* pueden ocultar matices importantes.
Falta de contexto	Una respuesta puede ser parcialmente cierta, pero incompleta para el caso real.

Prompts de IA generativa para acompañar el ejemplo 1

Prompt	¿Para qué sirve?
Revisa esta respuesta y separa hechos, inferencias y dudas usando lenguaje sencillo.	Ayuda a ordenar lo que parece seguro y lo que no.
Indica qué afirmaciones de este texto requieren verificación externa antes de usarse.	Permite detectar puntos sensibles.

Redáctame una versión prudente de esta respuesta sin inventar datos ni citas.	Sirve para escribir con mayor responsabilidad.
Dime cómo verificar esta afirmación con fuentes confiables y qué tipo de fuente debo buscar.	Fomenta un uso más crítico de la IA.

Mini protocolo de verificación para el ejemplo 1

1. ¿Qué afirmación concreta quiero comprobar?
2. ¿La respuesta trae una fuente real y verificable?
3. ¿Coincide con el contexto de mi tarea o de mi trabajo?
4. ¿Qué parte debo dejar en duda hasta revisar mejor?
5. ¿Cómo redacto una versión final más prudente?

Ejemplo guiado 2: mejorar un prompt para proteger privacidad y reducir sesgos

Ahora pasemos a una situación cotidiana. Una persona desea usar un chatbot para resumir información de clientes, estudiantes o usuarios. Sin darse cuenta, pega nombres completos, números de documento, teléfonos y otros datos que no son necesarios para la tarea. Además, formula una instrucción sesgada, por ejemplo pidiendo conclusiones apresuradas sobre grupos de personas o decisiones delicadas.

El objetivo de este segundo ejemplo es aprender a reescribir el prompt. No se trata solo de pedir una mejor respuesta; se trata de pedirla de una manera más segura, más justa y más responsable.

Elemento del prompt	Versión riesgosa	Versión mejorada
Datos personales	Incluye nombres, teléfonos y números de identificación.	Usa datos anonimizados o variables generales.
Lenguaje	Pide conclusiones apresuradas o estereotipadas.	Pide análisis prudente, neutral y basado en la información disponible.
Objetivo	No deja claro para qué se usará la respuesta.	Aclara el propósito y limita el alcance del análisis.
Verificación	Acepta la primera respuesta sin control.	Exige advertencias, límites y pasos de revisión.

Pasos del ejemplo 2

1. Identificar qué datos del prompt son innecesarios o demasiado sensibles para compartir.
2. Reemplazar nombres y detalles personales por etiquetas generales o datos anonimizados.
3. Reescribir la instrucción para pedir un análisis neutral, respetuoso y prudente.
4. Pedir al modelo que señale límites, posibles sesgos y necesidades de verificación.

5. Guardar una versión final del prompt que pueda reutilizarse con seguridad en contextos parecidos.

Prompts de IA generativa para acompañar el ejemplo 2

Prompt	Utilidad pedagógica
Reescribe este prompt para proteger la privacidad y eliminar datos personales innecesarios.	Ayuda a aprender anonimización básica.
Corrige este prompt para que use lenguaje neutral y reduzca posibles sesgos.	Refuerza el cuidado ético en la redacción.
Dime qué partes de este pedido pueden ser injustas, invasivas o poco prudentes.	Permite revisar el riesgo antes de usar el modelo.
Construye una versión final de este prompt con objetivo claro, límites y pasos de verificación.	Sirve para crear un modelo de trabajo reutilizable.

Plantilla breve de verificación avanzada

Plantilla adaptable de verificación avanzada

Rol: Actúa como asistente prudente y verificable.
Tarea: Responde de forma clara y útil.
Instrucción 1: Si no estás seguro, dilo explícitamente.
Instrucción 2: Separa hechos, inferencias y dudas.
Instrucción 3: No inventes fuentes, cifras ni citas.
Instrucción 4: Señala posibles sesgos o límites.
Instrucción 5: Evita pedir datos sensibles innecesarios.
Instrucción 6: Sugiere cómo verificar la respuesta antes de usarla.
Formato de salida: respuesta breve, nivel de certeza, riesgos detectados y pasos de verificación.

¿Qué aprendemos al comparar los dos ejemplos?

Aspecto	Ejemplo 1	Ejemplo 2
Problema principal	Posible alucinación o dato dudoso	Privacidad y sesgo en la formulación del prompt
Acción clave	Verificar y redactar con prudencia	Reescribir y limitar información sensible

Resultado esperado	Respuesta más confiable	Uso más seguro y responsable del modelo
Aprendizaje central	No aceptar automáticamente una salida convincente	No pedir de cualquier manera lo que queremos resolver

Buenas prácticas para esta lección

Recomendación	¿Por qué importa?
No confiar solo en el tono seguro de la respuesta	La seguridad aparente no garantiza verdad ni calidad.
Pedir niveles de certeza y límites	Ayuda a detectar cuándo hace falta revisar más.
Proteger datos sensibles	Reduce riesgos de privacidad y uso inadecuado de información.
Revisar sesgos y lenguaje estereotipado	Mejora la justicia y la calidad del análisis.
Usar la IA como apoyo, no como sustituto del criterio humano	Favorece decisiones más prudentes y responsables.

Producto esperado

El producto esperado en esta guía es una guía personal de uso responsable, presentada en una hoja o en un notebook corto, donde el estudiante explique con palabras sencillas qué es un LLM, identifique riesgos principales y construya un pequeño protocolo de verificación y mitigación para los contextos en que piensa usar la IA.

Lista de verificación

Revisión	Criterio
☐	¿Se explicó con palabras sencillas qué es un LLM?
☐	¿Se identificaron alucinaciones, sesgos y riesgos de privacidad?
☐	¿Se propusieron al menos tres acciones concretas de mitigación?
☐	¿Se trabajó un ejemplo donde la respuesta parecía buena, pero necesitaba verificación?
☐	¿Se reescribió un prompt para proteger mejor la privacidad y reducir sesgos?
☐	¿Se redactó una guía personal clara y aplicable de uso responsable?

Cierre

Trabajar con LLMs de forma responsable significa combinar utilidad con prudencia. Un buen usuario no es quien acepta todo lo que la IA produce, sino quien sabe preguntar mejor, detectar riesgos, proteger información sensible y verificar antes de usar una salida en contextos importantes. Con esa actitud crítica y práctica, la inteligencia artificial se convierte en una ayuda mucho más valiosa para estudiar, trabajar y tomar mejores decisiones.