www.aileaders-project.eu

EJERCICIO DE

Cómo desarrollar tu propio chatbot ético

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

EJERCICIO DE ESCENARIO - Cómo desarrollar tu propio chatbot ético

01 Resumen

Tipo de REA

Ejercicio de simulación

Objetivo/Propósito

Proporcionar un prototipo práctico y de bajo coste que muestre cómo crear e implementar un **chatbot ético** en Google Colab.

Resultados de aprendizaje esperados

El alumno será capaz de crear un chatbot funcional utilizando herramientas aptas para principiantes, incorporando principios de diseño ético.

Palabras clave

- IA ética
- Chatbot
- Prototipo
- Google Colab
- DialoGPT
- Transformadores Hugging Face
- Gradio UI
- Moderación de contenidos
- Avisos legales

Sugerido Metodológico Enfoque

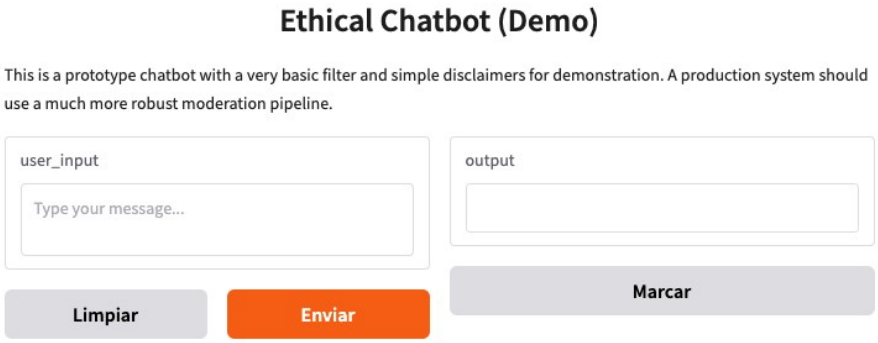
Aprendizaje basado en problemas

02 Introducción

Los chatbots generativos están pasando de los laboratorios de investigación a los servicios de atención al cliente, las aulas y los dispositivos personales, pero su rápida adopción ha superado el igualmente rápido trabajo en materia de salvaguardias **éticas de la IA**.

Los desarrolladores que desean ofrecer asistentes útiles se enfrentan a un problema clave: ¿***cómo añadir incluso unas mínimas capas de moderación de contenidos y transparencia sin*** una infraestructura costosa o API propietarias?

En las siguientes secciones proporcionaremos algunas pautas para conocer los requisitos que debe cumplir un chatbot ético, así como un ejercicio práctico para desarrollar su propio chatbot.



03 Presentación de herramientas

Herramientas utilizadas en este ejercicio de simulación

01 Python

El lenguaje de programación más utilizado en ciencia de datos

02 Un editor de Python

Google Colab o Jupyter Notebook

03 Bibliotecas y marcos

- **Transformers (Hugging Face):** carga y ejecuta el modelo conversacional preentrenado (DialoGPT-medium)
- **PyTorch (torch):** backend de aprendizaje profundo para la inferencia de modelos
- **Gradio:** interfaz de usuario web sin código para implementar y compartir el chatbot

04 Técnicas aplicadas

- Configuración de datos y código en un cuaderno Colab gratuito con GPU
- Carga de un modelo de diálogo preentrenado y mantenimiento del estado de la conversación
- Moderación de contenido: lista negra de palabras clave (o clasificador de canalización) para bloquear entradas tóxicas
- Detección de temas sensibles: escaneo de palabras clave → avisos legales automáticos («consulte a un profesional»)
- Generación de respuestas del chatbot: muestreo de DialoGPT con contexto histórico
- Despliegue rápido: interfaz Gradio de una sola celda que genera un enlace público para realizar pruebas

04 Directrices sobre cómo desarrollar un chatbot ético

Requisitos para un chatbot ético

01 DEFINE EL PROPÓSITO Y EL ALCANCE DEL CHATBOT

Identifique el caso de uso:
- Determine qué debe hacer su chatbot (por ejemplo, responder a preguntas de atención al cliente,
- facilitar el apoyo a la salud mental o proporcionar preguntas y respuestas generales).

Aclare las limitaciones:
- Sea explícito sobre lo que su chatbot no puede hacer (por ejemplo, proporcionar asesoramiento médico, gestionar transacciones financieras).

02 ESTABLECER DIRECTRICES ÉTICAS Y DE CONDUCTA

Desarrollar una «Carta ética» o política:
- Describa los principios que debe cumplir su chatbot (por ejemplo, evitar contenidos perjudiciales, respetar la privacidad de los usuarios).
- Incluya la equidad, la seguridad, la transparencia y el respeto por la diversidad.

Cree un «código de conducta» para su chatbot:
- ¿Cómo debe responder al contenido ofensivo o acosador?
- ¿Qué postura adopta ante la desinformación o la información errónea?
- ¿Cómo gestiona las solicitudes de datos personales o sensibles?

Establece mecanismos de respaldo:
en situaciones en las que se le pidan al chatbot instrucciones ilegales o perjudiciales, diseña para que responda con una negativa o una alternativa segura (como remitir al usuario a profesionales cualificados).

03 FORMACIÓN DEL MODELO

Obtenga datos de entrenamiento éticos:
- Asegúrese de que los datos no infrinjan la privacidad ni los derechos de autor.
- Busque datos representativos que incluyan diversos idiomas, regiones y culturas para minimizar los sesgos.
- Filtrar y preprocesar:
- Elimine o etiquete el contenido potencialmente dañino o sesgado.
- Considere la posibilidad de utilizar herramientas comerciales o filtros internos para identificar el discurso de odio, el contenido explícito o los datos personales.

Selección del modelo:
- Equilibre la complejidad del modelo con la interpretabilidad y el uso de recursos.
- El aprendizaje por refuerzo a partir de la retroalimentación humana (RLHF) o enfoques similares pueden ayudar a refinar el comportamiento del modelo después del entrenamiento.

04 OFRECER TRANSPARENCIA A LOS USUARIOS

Revele la naturaleza de su chatbot:

- Informe claramente a los usuarios de que están interactuando con un sistema impulsado por IA.
- Proporcione avisos legales sobre cualquier limitación y posibles inexactitudes.

Explique el uso de los datos:

- Informe a los usuarios sobre cómo se almacenan, procesan y utilizan sus datos para mejorar el sistema.
- Cumpla con las normativas de protección de datos (por ejemplo, el RGPD o la CCPA) y facilite el acceso a la política de privacidad.

Ofrezca una «salida» fácil:

- Informe a los usuarios de que pueden detener o excluirse de la recopilación de datos en cualquier momento.
- Proporcione un mecanismo para eliminar o anonimizar sus datos cuando lo soliciten.

05 GESTIONAR CASOS EXTREMOS Y TEMAS DELICADOS

Rechazo o finalización segura:

- En el caso de solicitudes ilegales, perjudiciales o muy delicadas (por ejemplo, asesoramiento médico, jurídico o financiero), el chatbot debe:
- Proporcionar un aviso legal y información parcial con referencias, o
- Dirigir al usuario a ayuda profesional.

Contexto de salud mental o autolesiones:

- Preparar respuestas seguras: expresar empatía, animar al usuario a buscar ayuda profesional y compartir los datos de contacto de los servicios de apoyo (cuando sea posible).

Situaciones de emergencia:

- Deje claro que el chatbot no está equipado para manejar emergencias.
- Proporcione instrucciones o datos de contacto de los servicios de emergencia si aparecen señales reconocibles de peligro.

06 CUMPLIR CON LOS REQUISITOS LEGALES Y ÉTICOS

Cumplimiento normativo:

- Manténgase al día de la legislación relativa a la IA, la privacidad de los datos y los servi digitales en las regiones en las que operará su chatbot.
- Implemente mecanismos sólidos de protección de datos y consentimiento.

Consideraciones sobre responsabilidad:

- Defina claramente quién es responsable cuando el bot proporciona información incorrecta o perjudicial.
- Publique avisos legales que detallen el alcance del chatbot (por ejemplo, «no es un profesional médico titulado»).

Mantenga una visión ética clara

- Mantener un chatbot ético es un proceso continuo. Revise periódicamente sus principios y asegúrese de que se ajustan a las nuevas normativas, los valores sociales y los comentarios de la comunidad. La coherencia y la transparencia en las actualizaciones ayudan a generar confianza. Realice un seguimiento de las métricas de rendimiento e incorpore los comentarios de los usuarios para lograr una mejora continua.

05 Actividades prácticas

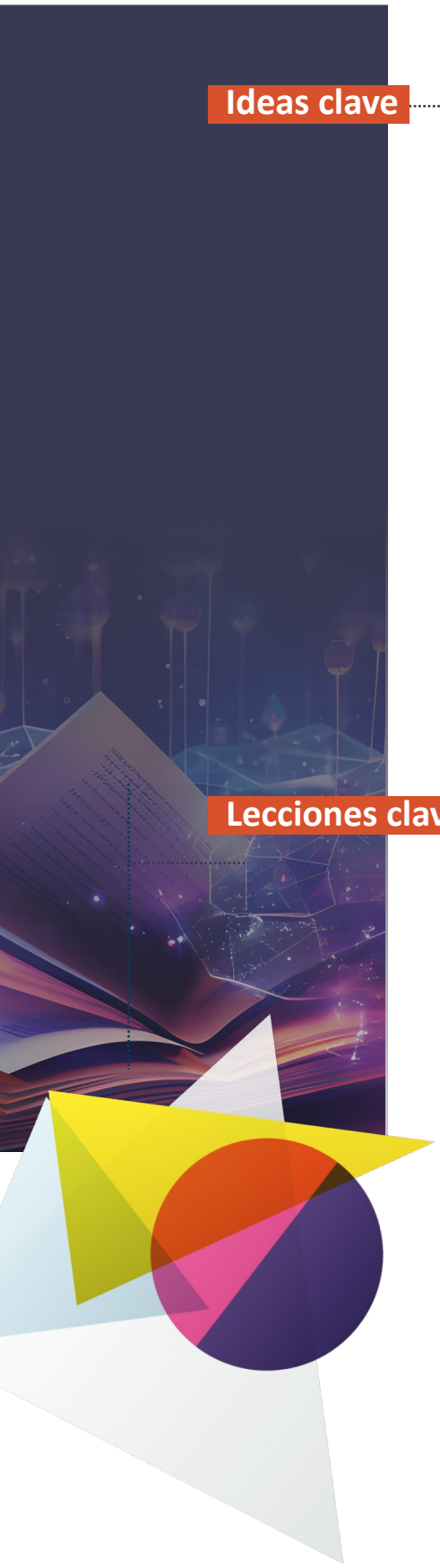
Descripción de la solución (siga el cuaderno proporcionado)

Este ejercicio le guiará en el desarrollo de un prototipo de chatbot creado íntegramente con herramientas gratuitas: un cuaderno de Google Colab que carga un modelo Hugging Face Transformers (por ejemplo, DialoGPT), lo envuelve con reglas ligeras de moderación de contenido y avisos legales para temas delicados, y muestra el resultado a través de una interfaz de usuario Gradio sin código. En menos de una hora, verá cómo pasar del concepto a un enlace compartible, lo que demuestra que la implementación rápida y el diseño responsable no tienen por qué ser mutuamente excluyentes. Proporcionamos un **cuaderno de Google Colab** listo para usar que le guía línea por línea a través de la implementación de un chatbot ético. El cuaderno está estructurado en secciones claramente etiquetadas para que pueda comprender y modificar cada capa de la solución.

Llamada a la acción

Abre el cuaderno de demostración del chatbot ético, ejecuta cada sección en orden y, en cuestión de minutos, tendrás un chatbot en vivo con medidas de seguridad éticas básicas, listo para mostrar a tus compañeros de equipo o para iterar para su uso en producción.

06 Conclusión



Ideas clave

01 Definir claramente lo que el chatbot debe y no debe hacer es la mejor protección contra un uso indebido en el futuro.

02 Los sesgos en los datos de entrenamiento se propagan directamente al comportamiento del modelo; es imprescindible contar con conjuntos de datos diversos y que respeten la privacidad.

03 Un filtro de moderación en tiempo real detecta diferentes modos de fallo.

04 La transparencia es imprescindible. Los usuarios que saben que están hablando con una IA comprenden los límites, controlan sus datos y son más indulgentes con los errores ocasionales.

05 La gobernanza debe ser continua, para garantizar no solo un buen rendimiento, sino también el cumplimiento de los requisitos legales y éticos.

Lecciones clave

01 Es más fácil suavizar las medidas de seguridad excesivamente estrictas que aplicar medidas de moderación después de que los resultados perjudiciales hayan llegado a los usuarios.

02 Realice un seguimiento de las tasas de resultados perjudiciales, la equidad demográfica y la satisfacción de los usuarios, y luego v incule las actualizaciones del modelo a objetivos concretos.

03 Los modelos más pequeños e interpretables con barreras de protección claras suelen superar a los modelos más grandes de caja negra en materia de seguridad y cumplimiento.

04 Las herramientas y las políticas fracasan sin una mentalidad de equipo que priorice el bienestar de los usuarios por encima de la velocidad o la novedad

05 Incorpore la ética en cada sprint.

07 Referencias

- Comisión Europea. (2024). Ley de Inteligencia Artificial de la UE: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
- Documentación de Hugging Face Transformers: https://huggingface.co/docs/transformers/en/index
- Documentación de Gradio: https://www.gradio.app/docs

08 Material complementario

Archivo Jupyter Notebook (IPYNB)

Se incluye un cuaderno Python detallado con código comentado que guía a través de cada paso del proceso.

El cuaderno contiene

Configuración del entorno	Instala transformers, torch y gradio; comprueba la GPU.
Carga del modelo	Extrae el modelo conversacional de código abierto DialoGPT-medium (o cualquier modelo que especifiques) de Hugging Face e inicializa un búfer de conversación.
Filtro de moderación básico	Implementa una sencilla lista negra de palabras clave *y* muestra cómo intercambiar una canalización de toxicidad de Hugging Face para obtener un filtrado más potente.
Barreras de protección para temas sensibles	Analiza las entradas del usuario en busca de palabras clave médicas, legales o financieras e inserta automáticamente un aviso *de «Consulte a un profesional»*.
Lógica de generación de respuestas	Codifica la solicitud del usuario, añade el historial de chat y toma muestras del modelo para producir una respuesta que tenga en cuenta el contexto.
Implementación de la interfaz de usuario de Gradio	Envuelve la función de respuesta en una interfaz web sin código; al ejecutar la celda, se obtiene un enlace público que se puede compartir al instante.
Consejos de personalización	Los comentarios en línea muestran dónde conectar un modelo diferente, ampliar la lista de filtros o añadir análisis.

Sigue nuestro viaje

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.