leaders

www.aileaders-project.eu

EXERCÍCIO DE

Como desenvolver o seu próprio 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.

EXERCÍCIO DE CENÁRIO - Como desenvolver o seu próprio chatbot ético

01 Resumo

Tipo de REA

Exercício de cenário

Objetivo/Finalidade

Fornecer um protótipo prático e de baixo custo que mostre como construir e implementar um **chatbot ético** no Google Colab.

Resultados de aprendizagem esperados

O estudante será capaz de criar um chatbot funcional usando ferramentas fáceis de usar para iniciantes, incorporando princípios éticos de design.

Palavras-chave

- IA ética
- Chatbot
- Protótipo
- Google Colab
- DialoGPT
- Transformers
- Gradio UI
- Moderação de conteúdo
- Isenções de responsabilidade

Sugerido Metodológico Abordagem

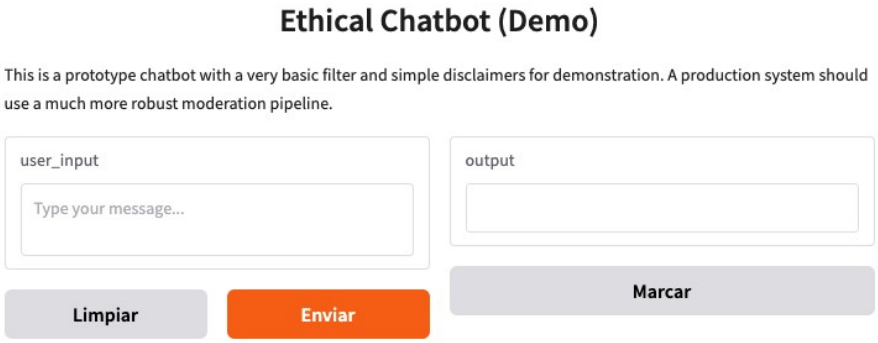
Aprendizagem baseada em problemas

02 Introdução

Os chatbots generativos estão a sair dos laboratórios de investigação e a entrar nos serviços de atendimento ao cliente, nas salas de aula e nos dispositivos pessoais, mas a sua rápida adoção ultrapassou o trabalho igualmente rápido em matéria de salvaguardas **éticas da IA**.

Os programadores que desejam lançar assistentes úteis enfrentam um problema fundamental: ***como adicionar camadas mínimas de moderação de conteúdo e transparência sem*** infraestruturas caras ou APIs proprietárias?

Nas próximas secções, forneceremos algumas diretrizes para conhecer os requisitos que um chatbot ético deve cumprir, bem como um exercício prático para desenvolver o seu próprio chatbot.



03 Apresentação das ferramentas

Ferramentas utilizadas neste exercício de cenário

01 Python

A linguagem de programação mais utilizada em Ciência de Dados

02 Um editor Python

Google Colab ou Jupyter Notebook

03 Bibliotecas e estruturas

- **Transformers (Hugging Face):** carregue e execute o modelo conversacional pré-treinado (DialoGPT-medium)
- **PyTorch (torch):** backend de deep learning para inferência de modelos
- **Gradio:** interface de utilizador web sem código para implementar e partilhar o chatbot

04 Técnicas aplicadas

- Configuração de dados e código num notebook Colab gratuito com GPU
- Carregamento de um modelo de diálogo pré-treinado e manutenção do estado da conversa
- Moderação de conteúdo: lista negra de palavras-chave (ou classificador de pipeline) para bloquear entradas tóxicas
- Detecção de tópicos sensíveis: verificação de palavras-chave → avisos automáticos (“consulte um profissional”)
- Geração de respostas do chatbot: amostragem do DialoGPT com contexto histórico
- Implantação rápida: interface Gradio de célula única que gera um link público para testes

04 Diretrizes sobre como desenvolver um chatbot ético

Requisitos para um chatbot ético

01 DEFINA O OBJETIVO E O ÂMBITO DO CHATBOT

Identifique o caso de uso:
- Determine o que o seu chatbot deve fazer (por exemplo, responder a perguntas de atendimento ao cliente,
- facilitar o apoio à saúde mental ou fornecer perguntas e respostas gerais).

Esclareça as limitações:
- Seja explícito sobre o que o seu chatbot não pode fazer (por exemplo, fornecer aconselhamento médico, lidar com transações financeiras).

02 ESTABELECER DIRETRIZES ÉTICAS E COMPORTAMENTAIS

Desenvolva uma «Código Ético» ou política:
- Descreva os princípios que o seu chatbot deve seguir (por exemplo, evitar conteúdos prejudiciais, reconhecer a privacidade do utilizador).
- Inclua justiça, segurança, transparência e respeito pela diversidade.

Crie um «Código de Conduta» para o seu chatbot:
- Como ele deve responder a conteúdos odiosos ou de assédio?
- Qual é a sua postura em relação a informações erradas ou desinformação?
- Como ele lida com solicitações de dados pessoais ou confidenciais?

Configure mecanismos de fallback:
em cenários em que o chatbot receba instruções ilegais ou prejudiciais, programe-o para responder com uma recusa ou uma alternativa segura (como indicar profissionais qualificados ao utilizador).

03 TREINO DO MODELO

Obtenha dados de formação éticos:
- Certifique-se de que os dados não infringem a privacidade ou os direitos autorais.
- Procure obter dados representativos que incluam diversos idiomas, regiões e culturas para minimizar o viés.
- Filtrar e pré-processar:
- Remova ou identifique conteúdos potencialmente prejudiciais ou tendenciosos.
- Considere usar ferramentas prontas para uso ou filtros internos para identificar discurso de ódio, conteúdo explícito ou dados pessoais.

Seleção do modelo:
- Equilibre a complexidade do modelo com a interpretabilidade e o uso de recursos.
- A aprendizagem por reforço a partir do feedback humano (RLHF) ou abordagens semelhantes podem ajudar a refinar o comportamento do modelo após o treino.

04 PROPORCIONE TRANSPARÊNCIA AOS UTILIZADORES

Divulgue a natureza do seu chatbot:

- Informe claramente aos utilizadores que eles estão a interagir com um sistema baseado em IA.
- Forneça avisos sobre quaisquer limitações e possíveis imprecisões.

Explique a utilização dos dados:

- Informe aos utilizadores como os seus dados estão a ser armazenados, processados e utilizados para melhorar o sistema.
- Cumpra os regulamentos de proteção de dados (por exemplo, GDPR, CCPA) e torne a política de privacidade facilmente acessível.

Ofereça uma «saída» fácil:

- Informe aos utilizadores que eles podem interromper ou cancelar a recolha de dados a qualquer momento.
- Forneça um mecanismo para excluir ou tornar anónimos os dados deles quando solicitado

05 LIDAR COM CASOS EXTREMOS E TÓPICOS SENSÍVEIS

Recusa ou conclusão segura:

- Para solicitações ilegais, prejudiciais ou altamente sensíveis (por exemplo, aconselhamento médico, jurídico ou financeiro), o chatbot deve:
- Fornecer um aviso legal e informações parciais com referências, ou
- Encaminhar o utilizador para ajuda profissional.

Contexto de saúde mental ou automutilação:

- Prepare respostas seguras: expresse empatia, incentive o utilizador a procurar ajuda profissional e partilhe detalhes de contacto para serviços de apoio (quando possível).

Cenários de emergência:

- Deixe claro que o chatbot não está equipado para lidar com emergências.
- Forneça instruções ou detalhes de contacto para serviços de emergência se sinais reconhecíveis de perigo aparecerem.

06 CUMPRIR OS REQUISITOS LEGAIS E ÉTICOS

Conformidade regulamentar:

- Mantenha-se atualizado sobre a legislação relativa à IA, privacidade de dados e serviços digitais nas regiões onde o seu chatbot irá operar.
- Implemente mecanismos robustos de proteção de dados e consentimento.

Considerações sobre responsabilidade:

- Defina claramente quem é responsável quando o bot fornece informações incorretas ou prejudiciais.
- Publique avisos detalhando o escopo do chatbot (por exemplo, "não é um profissional médico licenciado").

Mantenha uma visão ética clara

- Manter um chatbot ético é um processo contínuo. Reveja periodicamente os seus princípios e certifique-se de que estão alinhados com as regulamentações emergentes, os valores sociais e o feedback da comunidade. A consistência e a transparência sobre as atualizações ajudam a construir confiança. Acompanhe as métricas de desempenho e incorpore ciclos de feedback dos utilizadores para alcançar uma melhoria contínua.

05 Atividades práticas

Descrição da solução (siga o caderno fornecido)

Este exercício irá guiá-lo sobre como desenvolver um protótipo de chatbot construído inteiramente com ferramentas gratuitas: um caderno Google Colab que carrega um modelo Hugging Face Transformers (por exemplo, DialoGPT), envolve-o com regras leves de moderação de conteúdo e avisos legais para tópicos sensíveis e apresenta o resultado através de uma interface Gradio sem código. Em menos de uma hora, você verá como passar do conceito a um link partilhável, provando que a implementação rápida e o design responsável não precisam ser mutuamente exclusivos. Fornecemos um **caderno Google Colab** pronto a usar que o orienta passo a passo na implementação de um chatbot ético. O caderno está estruturado em secções claramente identificadas para que você possa compreender e modificar cada camada da solução.

Chamada à ação

Abra o notebook Ethical Chatbot Demo, execute cada secção por ordem e, em poucos minutos, terá um chatbot ativo completo com salvaguardas éticas básicas, pronto para demonstrar aos colegas de equipa ou iterar para uso em produção.

06 Conclusão

Principais insights

01 Definir claramente o que o chatbot deve e não deve fazer é a melhor proteção contra o uso indevido no futuro.

02 Os preconceitos nos dados de treino propagam-se diretamente para o comportamento do modelo; conjuntos de dados diversificados e que respeitem a privacidade são imprescindíveis.

03 Um filtro de moderação em tempo real detecta diferentes modos de falha.

04 A transparência é imprescindível. Os utilizadores que sabem que estão a falar com uma IA compreendem os limites, controlam os seus dados e são mais tolerantes com erros ocasionais.

05 A governança deve ser contínua, para garantir não apenas um bom desempenho, mas também o cumprimento dos requisitos legais e éticos.

Lições importantes

01 É mais fácil suavizar salvaguardas excessivamente rígidas do que reforçar a moderação depois que resultados prejudiciais chegarem aos utilizadores.

02 Acompanhe as taxas de resultados prejudiciais, a equidade demográfica e a satisfação do utilizador e, em seguida, vincule as atualizações do modelo a metas concretas.

03 Modelos menores e interpretáveis, com barreiras de proteção claras, geralmente superam modelos maiores de caixa preta em termos de segurança e conformidade.

04 Ferramentas e políticas falham sem uma mentalidade de equipa que priorize o bem-estar do utilizador em detrimento da velocidade ou da novidade

05 Incorpore a ética em cada sprint.

.

07 Referências

- Comissão Europeia. (2024). Lei da Inteligência Artificial da UE: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
- Documentação do Hugging Face Transformers: https://huggingface.co/docs/transformers/en/index
- Documentação do Gradio: https://www.gradio.app/docs

08 Material complementar

Ficheiro Jupyter Notebook (IPYNB)

Está incluído um Python Notebook detalhado com código comentado que orienta cada etapa do processo

O notebook contém

Configuração do ambiente	Instale transformers, torch e gradio; verifique a GPU.
Carregamento do modelo	Descarregue o modelo conversacional de código aberto DialoGPT-medium (ou qualquer modelo que você especificar) do Hugging Face e inicialize um buffer de conversação.
Filtro de moderação básico	Implemente uma lista negra simples de palavras-chave *e* mostre como trocar por um pipeline de toxicidade do Hugging Face para uma filtragem mais forte.
Proteções para tópicos sensíveis	Verifique as entradas do utilizador em busca de palavras-chave médicas, jurídicas ou financeiras e insere um aviso automático *do tipo «Consulte um profissional»*.
Lógica de geração de respostas	Codifique o prompt do utilizador, anexe o histórico de chat e faça uma amostragem do modelo para produzir uma resposta contextualizada.
Implementação da interface do utilizador Gradio	Envolve a função de resposta numa interface web sem código; ao executar a célula, é apresentado um link público que pode ser partilhado instantaneamente.
Dicas de personalização	Os comentários em linha mostram onde inserir um modelo diferente, expandir a lista de filtros ou adicionar análises.

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.