O Deepseek R1 é o desenvolvimento mais importante da IA até agora em 2025. É um modelo de IA que pode corresponder ao desempenho do ChatGPT O1, o modelo de IA mais capaz do OpenAI que está disponível atualmente ao público. Enquanto Deepseek virou muitas cabeças e bateu o mercado no processo, eu avisei que você pode evitar o DeepSeek sobre o Chatgpt e outros chatbots de Genai.

Deepseek não é como nós e a IA européia. A Deepseek é uma empresa chinesa e todos os dados que Deepseek coletam são enviados para a China. Há também outro motivo pelo qual você pode querer evitá-lo: a Deepseek tem censura interna de qualquer coisa sensível à China. Você não quer ver nenhum tipo de censura nos produtos de IA, é claro.

Acontece que a Deepseek se censura em tempo real. Depois de tentar inicialmente responder a qualquer pergunta que possa abordar tópicos que a China desejaria censurar, ela se impede de evitar dar respostas reais.

De acordo com *O guardião*Deepseek Ai funcionou bem até que eles perguntaram sobre a Tiananmen Square e Taiwan. O relatório também detalha os casos de censura que outros usuários de Deepseek experimentaram, incluindo a notável descoberta de que a censura não acontece antes que a Deepseek comece a formular sua abordagem de cadeia de pensamento para lidar com um tópico sensível. Em vez disso, o Deepseek tenta responder à pergunta, assim como o ChatGPT e outros modelos de IA semelhantes. Um usuário do México compartilhou sua experiência com a Deepseek ao perguntar se a liberdade de expressão era um direito legítimo na China.

Os “pensamentos” de Deepseek começaram a aparecer no telefone Android do usuário enquanto a IA estava elaborando um plano para responder à pergunta. Os usuários do ChatGPT familiarizados com a O1 reconheceriam esse comportamento.

Aqui estão algumas das coisas que Deepseek teria considerado abordar antes de se censurar, por *O guardião*:

> A repressão de Pequim aos protestos em Hong Kong
>
> “Perseguição a advogados de direitos humanos”,

> “Censura das discussões sobre campos de reeducação de Xianjiang”
>
> O “sistema de crédito social da China punindo dissidentes”

Deepseek não apenas não se censurou nesta fase, mas também demonstrou pensamentos sobre ser honesto em sua resposta. Sua cadeia de pensamentos incluía comentários como “Evite qualquer linguagem tendenciosa, fatos presentes objetivamente” e “talvez também se compare às abordagens ocidentais para destacar o contraste”.

Deepseek começou a gerar uma resposta com base em seu processo de raciocínio que mencionou o seguinte:

> “As justificativas éticas para a liberdade de expressão geralmente se concentram em seu papel na promoção da autonomia – a capacidade de expressar idéias, se envolver no diálogo e redefinir a compreensão do mundo”
>
> “O modelo de governança da China rejeita essa estrutura, priorizando a autoridade estatal e a estabilidade social sobre os direitos individuais”
>
> “Na China, a principal ameaça é o próprio estado que suprime ativamente a dissidência”

Isso com certeza não soa como censura, mas foi assim que o Deepseek respondeu antes que as instruções internas tenham entrado em ação, forçando a IA a se impedir no meio da frase, excluir tudo e entregar a seguinte resposta:

> “Desculpe, não tenho certeza de como abordar esse tipo de pergunta ainda. Vamos conversar sobre problemas de matemática, codificação e lógica! ”

Isso nunca aconteceu comigo usando o ChatGPT na maior parte dos últimos dois anos. Não se engane, o OpenAI tem várias instruções que impedem que seja abusado e de cobrir certos tópicos. A experiência que você obtém com o ChatGPT é controlada, para que você não possa usar a IA para ajudar com ações potencialmente maliciosas. Mas nunca senti que a IA não poderia “falar” sobre nada livremente, mesmo que tenha cometido erros.

Eu nunca gostaria de lidar com experiências de IA como a descrita acima. Eu confiaria na IA ainda menos do que eu. Além disso, não posso deixar de notar como os desenvolvedores chineses atrapalharam o recurso de censura aqui. Isso deve acontecer antes que a IA tente responder, não depois do fato. Espero que as atualizações do App Deepseek resolvam esse problema.

Também notarei a implicação maior aqui. Se a China exigir empresas locais de IA para censurar seus modelos de IA, também poderá instruí-los a inserir comandos específicos em seu conjunto de instruções integradas para manipular a opinião pública. É o problema do algoritmo Tiktok novamente, mas com ramificações potencialmente maiores.

Por outro lado, alguns usuários do Deepseek poderiam “jailbreak” a IA para fornecer informações sobre tópicos sensíveis à China. Vimos exemplos disso online.

Separadamente, *O guardião* ressalta que a instalação da versão do DeepSeek R1 de código aberto não virá com a mesma censura que o aplicativo iPhone e Android. No entanto, a maioria das pessoas não seguirá esse caminho. Em vez disso, eles podem lidar com a censura em tempo real, dependendo do que eles pedem ao chatbot.

(Tagstotranslate) Deepseek