¿Cómo se asegura OpenAI de que ChatGPT no genere contenido inapropiado o peligroso?

OpenAI tiene varias medidas para asegurarse de que ChatGPT no genere contenido inapropiado o peligroso:

1. **Ajuste fino**: Después del preentrenamiento, ChatGPT se ajusta en un conjunto de datos más pequeño que es generado con la ayuda de revisores humanos. Estos revisores siguen pautas proporcionadas por OpenAI para calificar las posibles salidas del modelo para una variedad de entradas. A través de este proceso de ajuste fino, el modelo aprende a generar respuestas que estén alineadas con las pautas y expectativas de OpenAI.
2. **Pautas de revisión**: Los revisores humanos siguen pautas estrictas que les indican que no deben completar solicitudes para generar contenido ilegal, peligroso, dañino, ofensivo, difamatorio, violento, sexualmente explícito o de cualquier otra manera inapropiado.
3. **Filtros de moderación**: OpenAI también utiliza filtros de moderación para bloquear ciertos tipos de contenido inapropiado o peligroso. Estos filtros no son perfectos y pueden tener falsos positivos y falsos negativos, pero ayudan a reducir la probabilidad de que ChatGPT genere contenido inapropiado.
4. **Retroalimentación y mejoras continuas**: OpenAI se toma muy en serio la retroalimentación de los usuarios y trabaja continuamente para mejorar y perfeccionar sus modelos y sistemas basándose en esta retroalimentación.