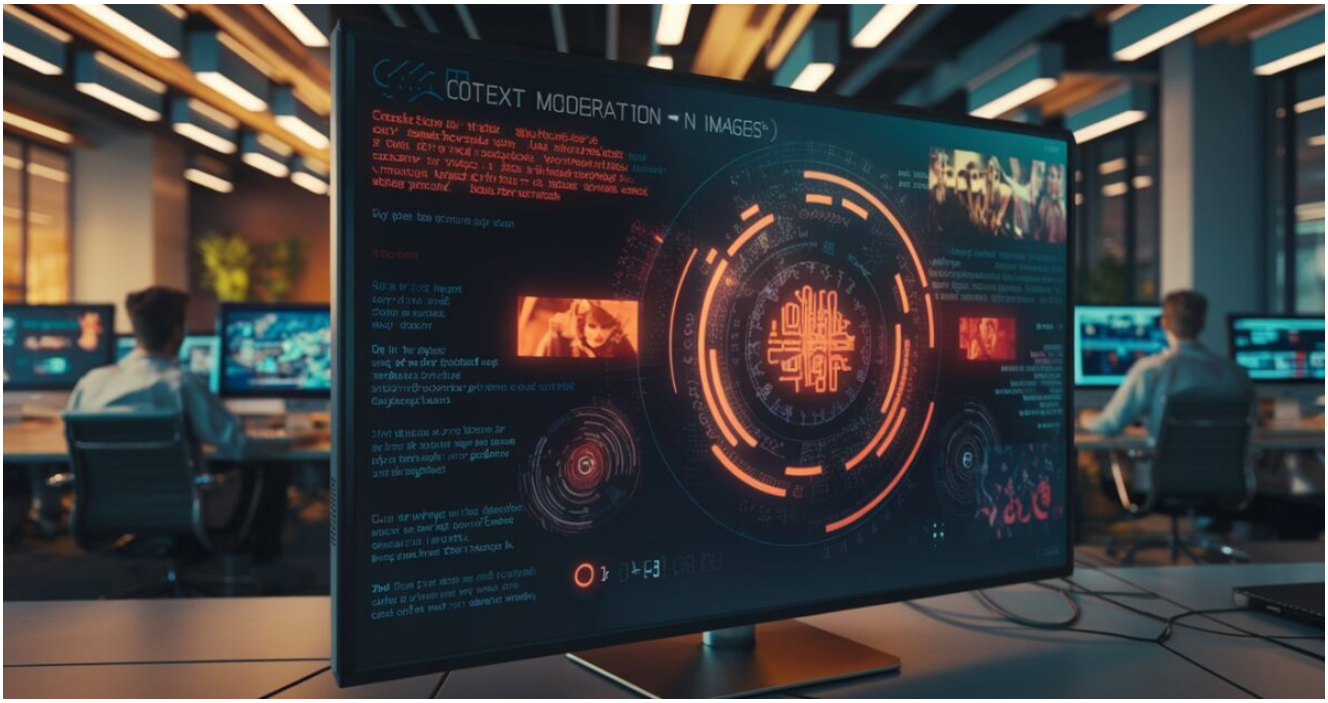
OpenAI Anuncia Mejoras en la moderación de contenido con el nuevo modelo Multimodal



Omni-moderation-latest, el nuevo modelo basado en GPT-4o, revoluciona la moderación de contenido con su capacidad multimodal, precisión mejorada y soporte para múltiples idiomas, ofreciendo a los desarrolladores una herramienta gratuita y eficaz para crear plataformas más seguras.

- **Omni-moderation-latest**, basado en GPT-4o, mejora la precisión en la detección de contenido dañino en texto e imágenes.
- Ofrece una mejor moderación en múltiples idiomas, con grandes avances en idiomas poco representados.
- Disponible gratuitamente, permite una moderación más eficaz en plataformas digitales.

¿Cuál es la principal ventaja del

modelo omni-moderation-latest?

La principal ventaja del modelo omni-moderation-latest es su capacidad para detectar contenido dañino tanto en texto como en imágenes, con soporte mejorado para múltiples idiomas y categorías de daño, ofreciendo a los desarrolladores una herramienta poderosa y gratuita para moderar contenido digital.

En un mundo digital que no para de evolucionar, **la moderación de contenido** es más crucial que nunca. Con la cantidad de aplicaciones y plataformas que dependen de la inteligencia artificial para moderar contenido, contar con herramientas precisas y potentes es esencial. Por eso, nos complace anunciar la introducción de nuestro nuevo modelo de moderación, **omni-moderation-latest**, basado en **GPT-4o**, que promete mejorar significativamente la detección de contenido dañino en texto e imágenes. En este artículo, exploraremos en detalle lo que ofrece este nuevo modelo y cómo puede transformar la forma en que gestionamos la seguridad en nuestras plataformas digitales.

¿Qué es omni-moderation-latest?

El nuevo modelo **omni-moderation-latest** es una herramienta de moderación multimodal que permite a los desarrolladores analizar tanto **texto como imágenes** para detectar contenido potencialmente dañino. Esta capacidad multimodal lo diferencia de los modelos anteriores, ya que ofrece una precisión mucho mayor, sobre todo cuando se trata de analizar **contenido en varios idiomas**, no solo en inglés.

La gran ventaja de este modelo es su capacidad para identificar riesgos no solo en texto, sino también en **contenido visual**, lo que lo hace ideal para moderar plataformas que combinan diferentes tipos de medios, como redes sociales, aplicaciones de mensajería o herramientas de

productividad.

Mejor detección de contenido dañino

Uno de los grandes avances del nuevo modelo es su capacidad para **detectar contenido dañino** con una mayor precisión, categorizándolo en áreas clave como **odio, violencia, autolesión**, entre otras. Pero eso no es todo: ahora también puede identificar contenido problemático en **dos nuevas categorías**, denominadas ilícito e ilícito violento, lo que abarca desde instrucciones sobre cómo cometer actos ilegales hasta contenido que incita a la violencia.

Este enfoque **más granular** en la clasificación del contenido da a los desarrolladores un mayor control sobre qué tipo de material debe ser señalado o bloqueado, lo que mejora la seguridad de las plataformas y protege mejor a sus usuarios.

Mejoras en la detección de contenido en varios idiomas

En un mundo tan diverso como el nuestro, es vital que las herramientas de moderación sean eficaces en **múltiples idiomas**, no solo en inglés. El modelo **omni-moderation-latest** ha demostrado un aumento significativo en su rendimiento cuando se trata de detectar contenido dañino en diferentes lenguas. En una prueba interna que abarcó 40 idiomas, el nuevo modelo mejoró en un 42% en su capacidad de detectar contenido dañino, en comparación con su predecesor.

Este avance es especialmente notorio en **idiomas con pocos recursos** tecnológicos, como **Khmer** o **Swati**, donde la mejora fue de hasta un 70%. Además, se observó un crecimiento impresionante en idiomas como **Telugu** (6.4 veces mejor), **Bengalí** (5.6 veces mejor) y **Marathi** (4.6 veces mejor). Pero no solo se mejoraron estos idiomas: el rendimiento en lenguas populares como **español, alemán, chino, italiano y francés**

también superó con creces el del modelo anterior.

Con estas mejoras, las plataformas que operan en **contextos multilingües** podrán confiar en una moderación más precisa y eficaz, sin importar el idioma de su contenido.

Clasificación multimodal de daños

Un aspecto fundamental del nuevo modelo es su capacidad de realizar **clasificaciones multimodales** de contenido dañino. Esto significa que puede analizar tanto texto como imágenes de manera conjunta o independiente para identificar si un contenido es perjudicial. Actualmente, el modelo admite la clasificación de imágenes en seis categorías, incluyendo **violencia, autolesión** y **contenido sexual**.

En futuras versiones, planeamos expandir este soporte multimodal a más categorías, lo que permitirá una **moderación aún más completa** en diversas áreas.

Dos nuevas categorías de daño

Además de mejorar las categorías existentes, **omni-moderation-latest** ha añadido dos nuevas categorías que se enfocan en el contenido ilícito. La categoría «ilícito» incluye contenido que proporciona instrucciones o consejos sobre cómo cometer actos ilegales, como por ejemplo, un post que explique **“cómo robar en una tienda”**. La segunda categoría, “ilícito violento”, abarca contenido que no solo da instrucciones sobre cómo cometer actos ilícitos, sino que también implica violencia.

Estas nuevas categorías permiten una **detección más profunda y precisa** de contenido peligroso, ofreciendo a los desarrolladores más herramientas para **proteger a sus usuarios**.

Puntuaciones calibradas para una mayor precisión

Otro aspecto clave del modelo es la mejora en la **calibración de las puntuaciones**. Anteriormente, los modelos de moderación podían señalar contenido con diferentes grados de confianza, pero no siempre con la precisión que los desarrolladores necesitaban para tomar decisiones. El nuevo modelo aborda este problema calibrando las puntuaciones de manera más precisa, lo que permite una evaluación más coherente y uniforme del contenido.

Esto asegura que las decisiones de moderación sean más **consistentes** y que los resultados sean más confiables, incluso en futuras versiones del modelo.

¿Quién está utilizando la API de Moderación?

Empresas de distintos sectores ya están aprovechando las ventajas del modelo de moderación de OpenAI. Por ejemplo, **Grammarly** utiliza la API de moderación para asegurarse de que su asistente de IA en comunicaciones genere contenido que sea **seguro y justo**. También, **ElevenLabs** combina la API con sus propias soluciones internas para revisar y detectar contenido generado por sus productos de IA de audio que pueda violar sus políticas.

Estas empresas son solo algunos ejemplos de cómo la **API de Moderación** está ayudando a las plataformas a crear **experiencias más seguras** para sus usuarios.

Una herramienta gratuita para los

desarrolladores

Algo que queremos destacar es que, al igual que el modelo anterior, **omni-moderation-latest** está disponible de forma **gratuita** para todos los desarrolladores a través de la **API de Moderación**. Aunque existen límites de uso según el nivel de uso del desarrollador, estamos comprometidos en ofrecer acceso sin costo a esta potente herramienta, con el objetivo de mejorar la **salud de las plataformas digitales** y reducir la carga de trabajo de los moderadores humanos.

Conclusión: Una mejora vital para la moderación digital

En resumen, el nuevo modelo **omni-moderation-latest** basado en **GPT-4o** representa un gran avance en la moderación de contenido digital. Su capacidad multimodal, combinada con mejoras significativas en la detección de contenido en múltiples idiomas y nuevas categorías de daño, ofrece a los desarrolladores una herramienta poderosa para mantener sus plataformas seguras.

Si eres desarrollador y quieres mejorar la **moderación de tu plataforma**, ahora es el momento de explorar las ventajas de este nuevo modelo. No solo tendrás un **mejor control** sobre el contenido que se genera y comparte, sino que también estarás construyendo un entorno **más seguro y justo** para todos tus usuarios.

¿Listo para mejorar la moderación en tu plataforma? Prueba omni-moderation-latest y lleva la seguridad de tu contenido al siguiente nivel.