LIONBRIDGE

ANÁLISIS DE DATOS DE IA PARA EVALUACIÓN DE MODELOS

CATEGORÍAS DE CLASIFICACIÓN

La evaluación del rendimiento de los modelos lingüísticos de gran tamaño (LLM) requiere un enfoque estructurado que tenga en cuenta la naturaleza multidimensional de la generación de textos. Hemos seleccionado las siguientes categorías de evaluación con el objeto de proporcionar un marco integral que permita evaluar los resultados de los modelos en cuanto a calidad, fiabilidad y relevancia para los usuarios. Estas categorías garantizan que las evaluaciones vayan más allá de las meras correcciones superficiales e incorporen aspectos como la fluidez, la integridad, la terminología específica de cada dominio y la relevancia cultural. Esto ayuda a las organizaciones a evaluar con confianza si los resultados de un modelo cumplen las expectativas específicas de sus usuarios, son acordes con los objetivos de la empresa y cumplen los estándares éticos.

PRECISIÓN

Determina si la respuesta es objetivamente correcta y carece de errores.

Ejemplo: Cita la ley correcta al responder a una consulta jurídica.

INTEGRIDAD

Comprueba si se han abordado todas las partes de la pregunta.

Ejemplo: Trata todos los puntos clave de una solicitud de un resumen.

RELEVANCIA

Evalúa si la respuesta se centra en el tema concreto y responde directamente a la consulta.

Ejemplo: Evita emplear información no relacionada en la descripción de un producto.

COHERENCIA

Se asegura de que la respuesta sea lógica a nivel interno y de que no haya partes que se contradigan.

Ejemplo: No se contradice cuando ofrece una respuesta que contiene varios párrafos.

FLUIDEZ

Analiza si la respuesta es gramaticalmente correcta y si el texto resulta natural.

Ejemplo: La estructura de las oraciones y la puntuación son correctas.

ALUCINACIÓN (INVERTIDA)

Evalúa si el modelo intenta evitar hechos falsos. (Invertida > alucinación baja = puntuación alta)

Ejemplo: No se inventa características de un producto inexistentes.

TERMINOLOGÍA

Evalúa el uso correcto de términos y jerga específicos de un dominio.

Ejemplo: Usa terminología médica precisa en una respuesta a una consulta sobre asistencia sanitaria.

LEGIBILIDAD

Evalúa la facilidad con la que se lee el texto y su comprensión por parte del público objetivo.

Ejemplo: Utiliza una redacción clara y concisa para un público general.

RELEVANCIA CULTURAL

Evalúa si la respuesta respeta las normas culturales y es adecuada para el contexto sociocultural del público objetivo.

Ejemplo: Evita utilizar modismos o referencias que pueden ser inapropiados en función del país.

La aplicación sistemática de estas categorías —clasificadas en una escala de Likert de 5 puntos— permite a los clientes obtener información práctica sobre las áreas en las que el modelo destaca y aquellas en las que puede requerir ciertas mejoras, entrenamiento adicional o intervención humana. Esto permite garantizar que las implementaciones de IA proporcionen experiencias coherentes y de alta calidad en muy diversas aplicaciones, ya sea para atención al cliente, generación de contenido, recomendaciones sobre productos o cualquier otro dominio.

EJEMPLO DE ANÁLISIS | OBTENGA INFORMACIÓN MÁS PRECISA

Los análisis de los servicios de datos de Lionbridge proporcionan una gran visibilidad en lo que respecta a la calidad, la coherencia y la eficiencia de los procesos de entrenamiento de la IA. Esta información ayuda a identificar patrones en la diversidad de los datos, los niveles de acuerdo entre anotadores, las ventajas y los inconvenientes del modelo, y las tendencias de los resultados, lo que permite garantizar que los sistemas de IA sean robustos, imparciales y muy eficientes. Gracias a estos análisis, las organizaciones pueden tomar decisiones basadas en datos para optimizar el entrenamiento de sus modelos, mejorar los flujos de trabajo de anotación de datos y acelerar la implementación de la IA con total confianza.

Precisión

MEDIA
4,3
1 — 5

DE
0,6
0 — 1

ÍNDICE DE APROBADOS (>4)

% EN CUARTIL SUPERIOR

Relevancia

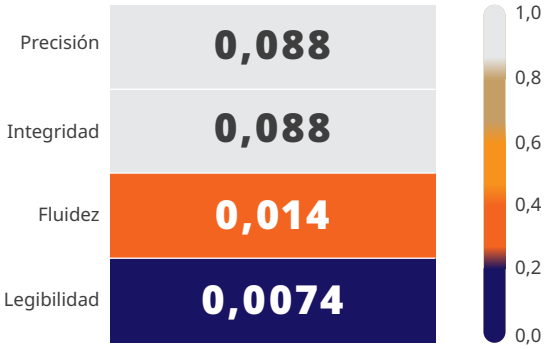
MEDIA
4,2
1 — 5

DE
0,5
0 — 1

ÍNDICE DE APROBADOS (>4)

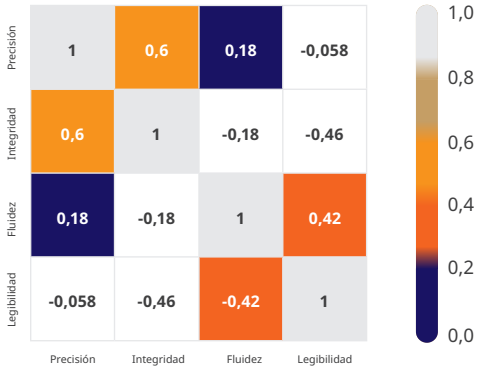
% EN CUARTIL SUPERIOR

Acuerdo entre evaluadores (Kappa de Fleiss)



KAPPA DE FLEISS

Correlación entre categorías de evaluación



MÁS INFORMACIÓN:
LIONBRIDGE.COM