LIONBRIDGE

KI-DATENANALYSE FÜR DIE MODELL-EVALUIERUNG

BEWERTUNGSKATEGORIEN

Zum Evaluieren der Leistung großer Sprachmodelle (Large Language Models, LLM) ist ein strukturierter Ansatz erforderlich, der die multidimensionale Natur der Spracherzeugung berücksichtigt. Die folgenden Evaluierungskategorien wurden ausgewählt, um einen umfassenden Rahmen für die Bewertung der Ausgaben von Modellen in Bezug auf Qualität, Zuverlässigkeit und Benutzerrelevanz zu schaffen. Diese Kategorien stellen sicher, dass Evaluierungen nicht nur die oberflächliche Richtigkeit erfassen, sondern Aspekte wie Sprachgewandtheit, Vollständigkeit, fachspezifische Terminologie und kulturelle Relevanz einbeziehen. Erst dann können Unternehmen und Organisationen fundiert beurteilen, ob die Ausgaben eines Modells den spezifischen Erwartungen ihrer Benutzer entsprechen, an den betrieblichen Zielen ausgerichtet sind und ethischen Standards genügen.

GENAUIGKEIT

Mit dieser Kategorie wird gemessen, ob eine Antwort sachlich richtig und fehlerfrei ist.

Beispiel: Nennung des richtigen Gesetzes in einer juristischen Antwort.

VOLLSTÄNDIGKEIT

Mit dieser Kategorie wird geprüft, ob alle Teile der Frage vollständig beantwortet wurden.

Beispiel: Berücksichtigung aller wichtigen Aspekte beim Erstellen einer Zusammenfassung.

RELEVANZ

Mit dieser Kategorie wird bewertet, ob die Antwort eine direkte Antwort auf die Frage darstellt.

Beispiel: Vermeidung irrelevanter Informationen in einer Produktbeschreibung.

KONSISTENZ

Eine Antwort ist konsistent, wenn sie in sich logisch und widerspruchsfrei ist.

Beispiel: Vermeidung von Widersprüchen in einer aus mehreren Absätzen bestehenden Antwort.

SPRACHGEWANDTHEIT

Mit dieser Kategorie werden die grammatische Richtigkeit und der natürliche Sprachfluss bewertet.

Beispiel: Verwendung richtiger Satzstrukturen und Zeichensetzung.

HALLUZINATION (INVERTIERT)

Mit dieser Kategorie wird gemessen, ob das Modell das Erfinden von Fakten vermeidet (weniger Halluzinationen = höherer Wert, weil Kehrwert).

Beispiel: Kein Erfinden nicht existenter Produktmerkmale.

TERMINOLOGIE

Diese Kategorie bewertet die Verwendung fachspezifischer Begriffe und fachspezifischen Jargons.

Beispiel: Verwendung präziser medizinischer Terminologie in einer Antwort zur Gesundheitsfürsorge.

LESBARKEIT

Mit dieser Kategorie wird der Lesefluss des Textes und seine Verständlichkeit für die Zielgruppe gemessen.

Beispiel: Klare, prägnante Formulierungen für das allgemeine Publikum.

KULTURELLE RELEVANZ

Mit dieser Kategorie wird bewertet, ob die Antwort einfühlsam und für die kulturellen Normen der Zielgruppe angemessen ist.

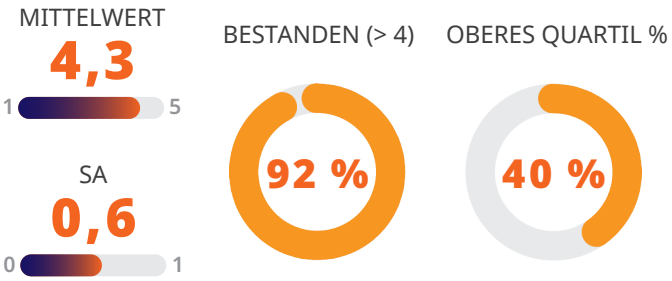
Beispiel: Vermeidung von Redewendungen oder Anspielungen, die global gesehen unangemessen sind.

Die systematische Anwendung dieser Kategorien und deren Bewertung anhand einer Likert-Skala mit fünf Antworten liefert Kunden nützliche Einblicke, indem Bereiche identifiziert werden, in denen das Modell besonders gut funktioniert oder in denen möglicherweise eine bessere Konfiguration, zusätzliches Training oder menschliche Aufsicht erforderlich ist. Auf diese Weise wird sichergestellt, dass KI-Implementierungen in unterschiedlichen Einsatzbereichen – Kundensupport, Contenterstellung, Produktempfehlungen und andere – konsistente, hochwertige Erfahrungen schaffen.

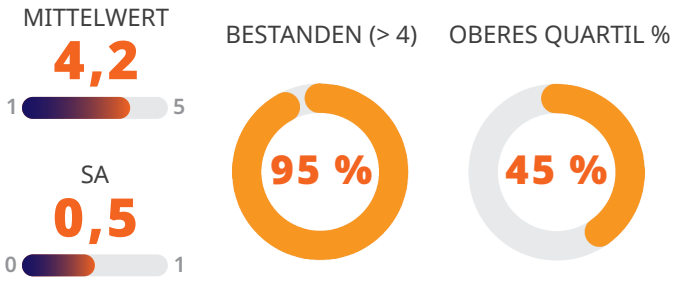
BEISPIELANALYSE | EINBLICKE GEWINNEN

Analysen von Lionbridge-Datenservices liefern wichtige Erkenntnisse zu Qualität, Konsistenz und Effizienz von KI-Trainingsprozessen. Dank dieser Erkenntnisse können Muster in Bezug auf Datenvielfalt, konsistente Datenannotation, Stärken und Schwächen des Modells sowie Ausgabetrends identifiziert werden, um die zuverlässige und leistungsstarke Funktion eines KI-Systems ohne Bias (Voreingenommenheiten) sicherzustellen. Auf Basis der Analysen können Unternehmen datengestützte Entscheidungen treffen, um das Modelltraining zu optimieren, Annotationsworkflows zu verbessern und die KI-Bereitstellung zu beschleunigen.

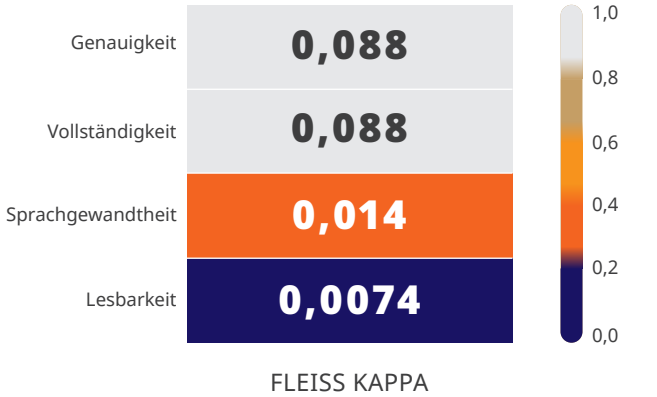
Genauigkeit



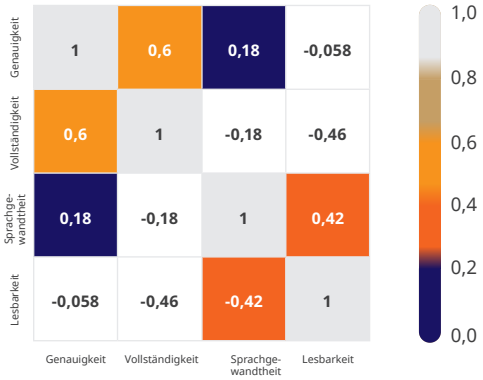
Relevanz



Übereinstimmung der Bewerter (Fleiss Kappa)



Korrelation der Evaluierungskategorien



ERFAHREN SIE MEHR AUF
LIONBRIDGE.COM