



LIONBRIDGE

ANALISI DEI DATI AI PER LA VALUTAZIONE DEI MODELLI



CATEGORIE DI VALUTAZIONE

Per valutare le prestazioni dei Large Language Model (LLM) è necessario un approccio strutturato che prenda in considerazione la natura multidimensionale della generazione di contenuti linguistici. Le categorie di valutazione seguenti sono state selezionate per offrire un quadro completo per la valutazione dell'output dei modelli in base a qualità, affidabilità e rilevanza per gli utenti. Queste categorie permettono di effettuare valutazioni che vanno oltre la mera correttezza formale, includendo aspetti come la fluidità, la completezza, l'uso di terminologia specifica del settore e la rilevanza culturale. In questo modo, le organizzazioni possono determinare con sicurezza se gli output di un modello soddisfano le aspettative specifiche degli utenti, sono allineati agli obiettivi aziendali e rispettano gli standard etici.



ACCURATEZZA

Determina se la risposta è corretta dal punto di vista fattuale e priva di errori.

Esempio: citazione di una legge appropriata in una risposta legale.



COMPLETEZZA

Controlla se viene data risposta a tutte le questioni poste in una domanda.

Esempio: inclusione di tutti i punti chiave in una richiesta di riepilogo.



RILEVANZA

Verifica che la risposta rimanga in tema e fornisca le informazioni richieste.

Esempio: assenza di informazioni non correlate in una descrizione di un prodotto.



COERENZA

Assicura che la risposta abbia una logica interna e sia priva di contraddizioni.

Esempio: assenza di contraddizioni in una risposta articolata in più paragrafi.



FLUIDITÀ

Valuta la correttezza grammaticale e la naturalezza del linguaggio.

Esempio: uso corretto della struttura delle frasi e della punteggiatura.



ALLUCINAZIONE (INVERSA)

Valuta se il modello è in grado di evitare di inventare le informazioni.

(Meno allucinazioni = punteggio più alto nella scala inversa)

Esempio: assenza di informazioni inventate su caratteristiche inesistenti di un prodotto.



TERMINOLOGIA

Analizza l'uso corretto del gergo e dei termini specifici del dominio.

Esempio: uso di terminologia medica precisa in un contesto sanitario.



LEGGIBILITÀ

Misura la facilità con cui il pubblico di riferimento può leggere e comprendere il testo.

Esempio: formulazione dei concetti in linguaggio chiaro e conciso per un pubblico generale.



RILEVANZA CULTURALE

Valuta se la risposta è sensibile e appropriata rispetto alle norme culturali del pubblico di riferimento.

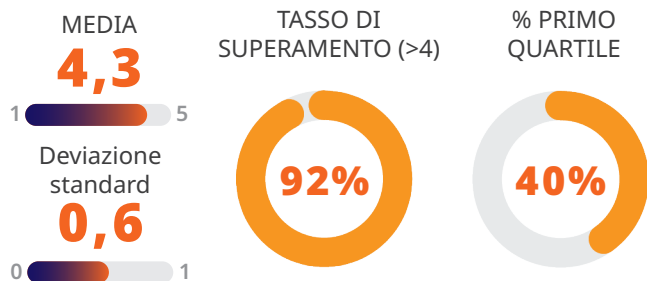
Esempio: assenza di modi di dire o riferimenti inappropriati a livello globale.

Applicando sistematicamente queste categorie, valutate su una scala Likert a 5 punti, i clienti ottengono informazioni di utilità pratica in merito alle aree in cui il modello eccelle e a quelle che invece necessitano di perfezionamento, ulteriore addestramento o supervisione umana. In questo modo è possibile assicurarsi che l'AI offra esperienze coerenti e di alta qualità in vari ambiti applicativi, come l'assistenza clienti, la generazione di contenuti, le raccomandazioni di prodotti e altri ambiti.

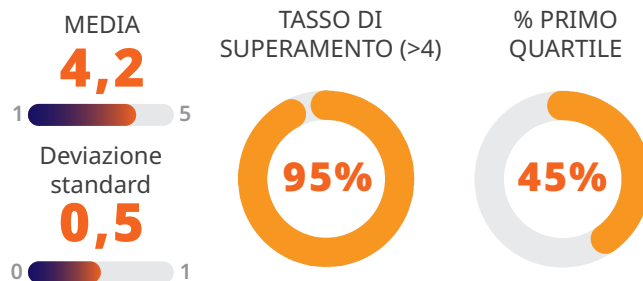
ESEMPIO DI ANALISI | COME OTTENERE INFORMAZIONI PIÙ APPROFONDITE

Le analisi fornite dai servizi di dati di Lionbridge offrono una visibilità cruciale su qualità, coerenza ed efficienza dei processi di addestramento dell'AI. Queste informazioni approfondite aiutano a identificare schemi ricorrenti nella diversità dei dati, nei livelli di concordanza tra annotatori, nei punti di forza e debolezza dei modelli e nei trend degli output, per assicurare che i sistemi di intelligenza artificiale siano affidabili, imparziali e ad alte prestazioni. Grazie a queste analisi, le organizzazioni possono prendere decisioni basate sui dati per ottimizzare l'addestramento dei modelli, migliorare i flussi di lavoro di annotazione e accelerare l'implementazione dell'AI in tutta sicurezza.

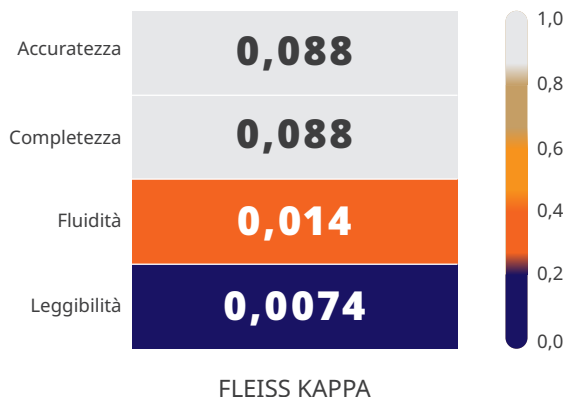
Accuratezza



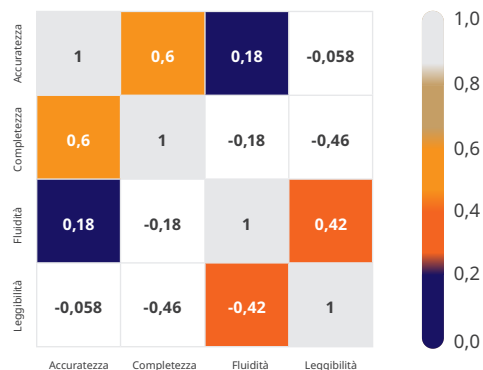
Rilevanza



Accordo tra valutatori (Fleiss' Kappa)



Correlazione tra categorie di valutazione



PER SAPERNE DI PIÙ:
LIONBRIDGE.COM