LIONBRIDGE

AI DATA ANALYTICS FOR MODEL EVALUATION

RATING CATEGORIES

Evaluating the performance of large language models (LLMs) requires a structured approach that captures the multi-dimensional nature of language generation. The following evaluation categories have been selected to provide a comprehensive framework for assessing model outputs across quality, reliability, and user relevance. These categories ensure that evaluations go beyond surface-level correctness, incorporating aspects such as fluency, completeness, domain-specific terminology, and cultural relevance. This allows organizations to confidently assess whether a model's outputs meet the specific expectations of their users, align with business objectives, and adhere to ethical standards.

ACCURACY

Measures if the response is factually correct and free of errors.

Example: Citing the correct law in a legal answer.

COMPLETENESS

Checks whether all parts of the question are fully addressed.

Example: Covering all key points in a summary request.

RELEVANCE

Assesses if the response stays on topic and directly answers the query.

Example: Avoiding unrelated information in a product description.

CONSISTENCY

Ensures the response is internally logical and free of contradictions.

Example: Not contradicting itself within a multi-paragraph answer.

FLUENCY

Evaluates the grammatical correctness and natural flow of the language.

Example: Using proper sentence structure and punctuation.

HALLUCINATION (INVERTED)

Measures whether the model avoids making up facts.

(lower hallucination = higher score when inverted)

Example: Not inventing nonexistent product features

TERMINOLOGY

Assesses the correct use of domain-specific terms and jargon.

Example: Using precise medical terminology in a healthcare response.

READABILITY

Measures how easy the text is to read and understand for the target audience.

Example: Clear, concise wording for a general audience.

CULTURAL RELEVANCE

Evaluates if the response is sensitive to and appropriate for the cultural norms of the target audience.

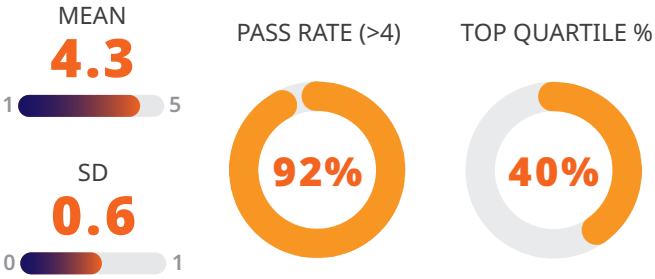
Example: Avoiding idioms or references that may be inappropriate globally.

By systematically applying these categories — rated on a 5-point Likert scale — clients gain actionable insights into areas where the model excels and where it may require fine-tuning, additional training, or human oversight. This ensures that AI deployments deliver consistent, high-quality experiences across various applications, whether in customer support, content generation, product recommendations, or other domains.

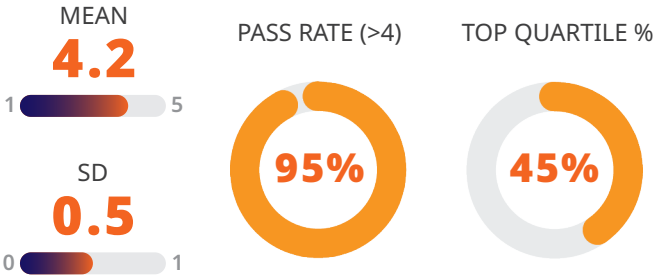
EXAMPLE ANALYTICS | UNLOCK DEEPER INSIGHTS

Analytics from Lionbridge's data services provide critical visibility into the quality, consistency, and efficiency of AI training processes. These insights help identify patterns in data diversity, annotator agreement levels, model strengths and weaknesses, and output trends — ensuring AI systems are robust, unbiased, and high-performing. By leveraging these analytics, organizations can make data-driven decisions to optimize model training, improve annotation workflows, and accelerate AI deployment with confidence.

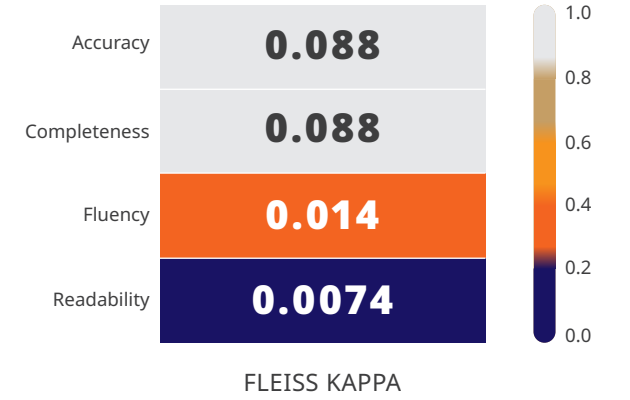
Accuracy



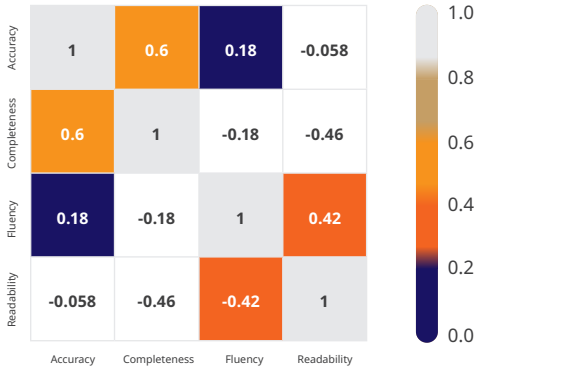
Relevance



Inter-Rater Agreement (Fleiss' Kappa)



Correlation Between Evaluation Categories



LEARN MORE AT
LIONBRIDGE.COM