LIONBRIDGE

AI データ分析による
モデルの評価

評価カテゴリー

大規模言語モデル (LLM) のパフォーマンスの評価には、言語生成の多面的な特性を捉えるための構造化されたアプローチが欠かせません。以下の評価カテゴリーは、品質、信頼性、ユーザーとの関連性といった観点からモデルの出力を評価するための包括的な枠組みを構築するものです。これらのカテゴリーを利用することで、単なる表面的な正しさにとどまらず、流暢さ、完全性、分野固有の専門用語、文化的適切性といったさまざまな要素を評価に反映させることができます。それによって、モデルの出力結果がユーザーの具体的な期待事項を満たし、ビジネス目標に合致し、倫理基準にも準拠しているかどうかを確実に評価できるようになります。

正確性

応答が事実として正確であり、誤りがないかどうかを測定します。

例: 法的な内容の応答において正しい法律を引用している。

完全性

質問のすべての部分が適切に回答されているかどうかを確認します。

例: 要約リクエストに対する応答において、すべての重要なポイントが網羅されている。

関連性

応答がテーマから外れていないか、質問に直接答えているかを評価します。

例: 製品説明に関連性のない情報が含まれていない。

一貫性/整合性

応答の内容に筋が通っており、矛盾がないかどうかを確認します。

例: 複数の段落で構成される応答の中で矛盾が発生していない。

流暢さ

文法的な正確さと文章の自然な流れを評価します。

例: 適切な文構造や句読点を使用している。

ハルシネーション (逆スコア)

モデルが事実を捏造しないかどうかを測定します。

(ハルシネーションが少ない = 反転してスコアが高い)

例: 製品の架空の機能をでっち上げていない。

用語管理

分野固有の用語や専門用語が正しく使用されているかどうかを評価します。

例: 医療関連の応答において正確な医療用語を使用している。

可読性 (読みやすさ)

テキストが対象オーディエンスにとってどの程度読みやすく理解しやすいかを測定します。

例: 一般の読者向けに明確で簡潔な表現を使っている。

文化的妥当性

応答が対象オーディエンスの文化的規範に配慮したものであり、その規範に照らして適切かどうかを評価します。

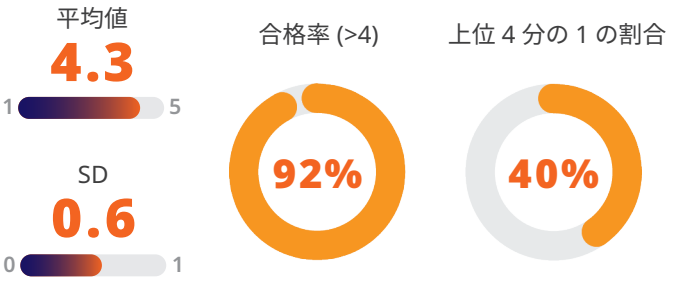
例: グローバルな環境で不適切になる可能性のある慣用句や言及を避けている。

これらのカテゴリーを体系的に適用することで (5 段階のリッカート尺度で評価)、お客様はモデルが優れている領域と、ファインチューニング、追加トレーニング、人間の監督などが必要そうな領域に関する実用的なインサイトを得ることができます。それにより、カスタマー サポート、コンテンツ生成、製品推薦、その他の分野などさまざまな用途において、AI を導入することで一貫した高品質な体験を創出できるようになります。

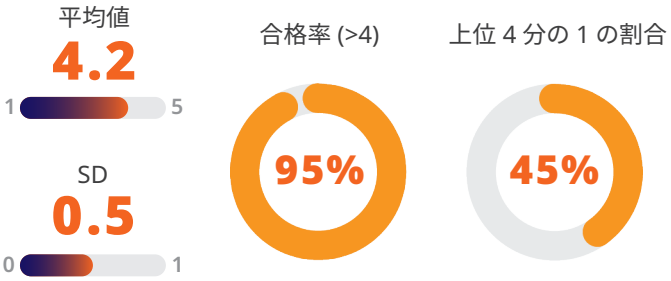
分析の例 | より深いインサイトを引き出す

ライオンブリッジのデータ サービスによる分析では、AI トレーニング プロセスの品質、一貫性、効率性に関する重要な情報を得ることができます。こうしたインサイトは、データの多様性、注釈付け担当者の合意レベル、モデルの強みと弱み、出力傾向のパターンなどを特定するのに役立つため、堅牢で偏りがなく高性能な AI システムを確保することができます。組織でこのような分析を活用すれば、データに基づく意思決定を行って、モデルのトレーニングを最適化し、注釈付けのワークフローを改善して、自信を持って AI の導入を進められるようになります。

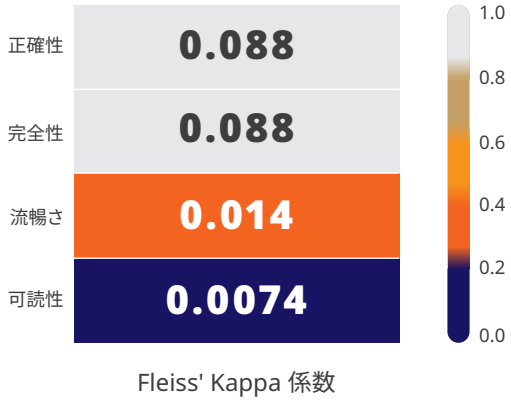
正確性



関連性



評価者間一致率 (Fleiss' Kappa)



Fleiss' Kappa 係数

評価カテゴリー間の相関関係

詳しくはこちら
LIONBRIDGE.COM