人間による AI データ収集

選択すべき 5 つの理由

モデルをトレーニングするには、合成データ、人間が収集したデータ、あるいはその組み合わせなど、何らかのデータが欠かせません。ここでは、人間が収集したデータを (少なくとも全体の大部分で) 利用するほうが、モデルの性能と信頼性の向上につながりやすい理由をご紹介します。

1 品質

高品質なデータは通常、精度が高く最新であり、ノイズが少なく一貫性があり、構造も整っているほか、豊かな文脈情報も含んでいるため、モデルのパフォーマンス向上につながります。こうした特性を備えたデータを使用することで、モデルは言語や関連トピックをよりきめ細かく理解し、高度なパフォーマンスを発揮できるようになります。

例

ライオンブリッジはある eラーニング ソリューション プロバイダーを支援し、機械で文字起こしされた 300 本以上の動画の品質レビューを行いました。レビュー担当者は AI が生成した字幕を修正し、非常に精度の高い動画トランスクリプトを作り上げました。

2 多様性

多様性に富む世界規模の多言語人材ネットワークを擁するプロバイダーなら、実世界のシナリオをより幅広く網羅し、多様で偏りのない視点を反映したデータを提供することができます。

例

ある大手テクノロジー企業向けの案件で、ライオンブリッジは Aurora AI Studio™ と社内のリクルート チームを活用しながら、指定の感情で発話された 100 万文以上の録音データを収集しました。参加者は複数の言語や方言を話し、さまざまな属性を持つ人々で構成されていました。

3 スピード

適切なプロバイダーを選べば圧倒的なスピードを実現できます。ライオンブリッジの Aurora AI Studio は、世界中の 50 万名規模のコントリビューターと連携しています。また、当社の優れたリクルート チームにより、必要に応じてネットワーク外からもコントリビューターを確保することが可能です。タスクは数日、場合によっては数時間以内に対応が完了します。

例

ライオンブリッジはあるスマートフォン メーカーを支援するため、Aurora AI Studio を利用して、8 つの言語による「実生活の会話例」のデータセットを大規模に調達しました。それぞれの会話は最大 5 名の参加者によるもので、合計 20 万件以上の対話が収集されました。そのすべてのデータを、このプラットフォームの機能によって 4 週間以内に納品することができました。

4 コスト

合成データだけではモデル性能が低くなりがちで、追加の検証や、人間が収集したデータの追加的な確保が必要になることがあります。初めから人間が収集したデータを選ぶことで、コストを削減できます。

例

ライオンブリッジはあるオンライン動画サービス プロバイダーを支援し、多数の動画をさまざまな言語から英語に翻訳する作業を担当しました。翻訳者は、わいせつ・偏見・攻撃的・憎悪的といった不適切な内容をすべてチェックし、フラグ付けを行いました。こうした丁寧な確認とレビューによって LLM はすぐに不適切な内容を含めずに最適に動作するようになり、追加のデータ収集や検証の必要も生じませんでした。

5 開発プロセス

開発にはコストがかかります。開発者の時間や人件費を無駄にしないためにも、最初から人間が収集した信頼できるデータを選択しましょう。

例

ある AI 開発元をサポートするため、当社は複数の言語で必要になる大量の学習データを LLM に提供できる人間のレビュー担当者を手配しました。レビュー担当者はプロンプトに対する最適な応答を選択し、いくつかの要素に基づいてその応答を評価しました。

AI データ収集サービスにご関心があれば、ぜひ Lionbridge AI™ までお問い合わせください。詳しくはこちら: lionbridge.com/ja