AI開発を支える
データセット作成・
データ整備の方法

高品質データは「作業」ではなく「設計」で作る：
現場に効く品質基準

2026年2月3日
株式会社キャリア・マム
ソリューション事業部　マネージャー　小尾　和美

Index

株式会社キャリア・マムについて

全国12万人のネットワークを活かし、
権利クリアで的確・安心な国産データを大量に提供する、
データプロダクションのプロフェッショナル集団です。

会社紹介

CareerMam AI Data Production

精度の高いAIは高品質データから

こんな課題を解決します

・教師データを社内で整備する体制が整っていない
・外注しても品質や納期が安定せず不安を感じる
・レアケース・専門・多言語など特殊データに対応できない
・チャットボットやRAG構築に必要なデータ整備ができない

全国12万人のネットワーク

権利クリアで安心なデータ提供。女性中心の豊富な人材プール

専門人材の最適編成

特殊要件にも柔軟対応できるチーム構築

トリプルチェック体制

多くの実績・知見を活かした高品質データ供給

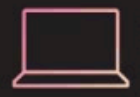
完全国内データ

すべて日本国内で収集・作成

国立研究所・国立大学・大企業での採用実績多数！
権利クリアな国産データを大量に提供可能！

全国12 万人のネットワークから、権利クリアで安心なデータ提供
豊富な専門人材から最適なチームを編成し、特殊要件にも柔軟対応
データはトリプルチェックの上納品、多くの実績・知見を活かした高品質データを供給

ソリューション



データ作成
質疑応答文作成等
タグ付け
判定

リアル
データ収集
音声・画像・動画

データ
クレンジング

固有表現抽出
会話コーパス

RAG 構築向け
データ整備

LLM 評価

インストラクション
チューニング
データ作成

データ
ディレクション
精度課題の把握
課題分析
対策立案

※すべて日本国内で収集・作成

1枚の画像を複数の人が説明したら

テキストアノテーションのチームを100名体制で組み、プロジェクトを進行するとします。画像の説明項目を決めて共有すれば問題ないでしょうか?

実は、そう単純ではありません。複数の人間が集まると、見方も判断基準も実に多様になるのです。

人による解釈のばらつき

同じ画像でも表現は様々

Aさんの記述:
「車が3台、向きはバラバラに駐車している。画面最前列に男性、女性、小学生女児が立っている」

Bさんの記述:
「雲ひとつない晴天。手前に親子。2階建ての住宅が連なる」

Cさんの記述:
「青空の下、3名とも視線が同方向へ向き、少し楽しそう。ベンチのある広めの歩道または広場」

ばらつきの原因と影響

個人の感じ方の違い
「楽しそう」の基準は人それぞれ

良し悪しの基準
何を重要と見なすかの判断

感情の基準
表情から読み取る感情の解釈

好みによる偏り
個人的な嗜好の影響

表現スキルの違い
語彙力や描写力の個人差

これは、AIのモデル構築において致命傷になります。
必要なのは、判断基準を一定にすること。つまり、
全員が同じ物差しで画像を見て、同じルールで言語化できる仕組みです。

科学的根拠: Image Specificity研究

研究データが示す事実

研究名: Image Specificity
著者: M. Jas & D. Parikh (2015, CVPR)

同じ画像を複数人に説明させると、ある画像では似たような説明が多くなる一方、別の画像では表現が大きく分かれることが実証されました。

この現象は「specificity(説明の一致度)」という指標で定量化され、**画像によっては説明の一致度に2〜3割程度の差が出る**ことが明らかになっています。

主要な相関データ

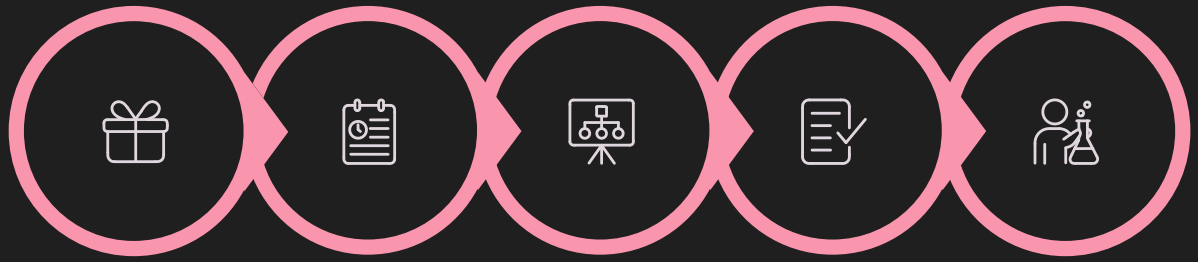
- ヒト評価と自動計算指標の相関: **Spearman ρ = 0.69** ($p < 0.01$)
- 記憶に残りやすさとの相関: **ρ = 0.33** ($p < 0.01$)
- 対象物の面積との相関: **ρ = 0.16** ($p < 0.01$)
- 対象の数とは相関なし: $\rho = -0.04$ ($p = 0.19$)

出典: arXiv:1502.04569

プロトタイプの構築

迷わない現場が品質を生む

プロトタイプの段階から要望を反映し、詳細項目まで一貫した判断基準を設定することで、100名のチームでも高品質なデータ作成が可能になります。



具体的な実施内容

- アノテーションサンプルを複数パターン提示
- 対応や表現をパターン表示で理解促進
- 解釈内容やプロセスを図示して可視化
- アノテーションチェック項目の細分リスト化
- 実際の試行を通じた検証と修正

このプロセスを通じて、個人の解釈のばらつきを最小化し、全員が同じ基準で作業できる環境を構築します。

納品物チェック項目例　～プロトタイプを案件に反映～

案件: 画像に映っている「事実」と「感情」をテキスト化する
(200文字以内)
制作文言チェック項目の例

チェック項目
指定の順番でテキスト化しているか
憶測で記載せず、事実のみをテキスト化しているか
右から左の順番で一連の説明ができるか
日本語としての客観性はあるか
感情表現に個人的な偏りがないか
色彩表現が統一基準に沿っているか

これらの細分化されたチェック項目（上記は項目の一部）により、アノテーターは自己チェックが可能になり、品質の均一化が実現します。

プロトタイプを進化させるプロジェクト進行

プロトタイプは一度作って終わりではありません。実際の作業とフィードバックを通じて継続的に進化させ、最終形仕様へと昇華させていきます。

アノテーション実施

構築した基準に基づいてデータ作成

フィードバック収集

納品データへの評価と課題抽出

チーム内共有

解釈の齟齬を解消する仕様の統一

プロトタイプ進化

ほころびを補う仕様への改善

プロトタイプから生まれる最終形仕様を適用したアノテーション・学習データは、想定レベルを上回る高品質化を実現します。このサイクルを回すことで、データの精度は指数関数的に向上していくのです。

だけど、それだけじゃうまくいかない–キーは「人」

どれほど優れたプロトタイプを構築しても、それをコントロールするのは「人」。個々のスキル、コントロール力、統率力–これらはプロトタイプ構築だけでは補えない、プロジェクト成功の重要な要素です。

テクノロジーの進化が加速する現代においても、最終的な品質を決定するのは人的体制なのです。では、理想的なコントロールがかなうプロジェクト体制とはどのようなものでしょうか?

実践事例

ディレクションの重要性–設問設計の事例

クライアント要望

「質問を100件作ってください」

キャリア・マムの対応

単に100件の質問を作るのではなく、**設問設計をディレクション**します。

1. 質問の種類を定義
2. 各種類ごとの構成を設計
3. 具体的な作成ルールを明文化
4. サンプルを複数パターン提示

ディレクションの有無による違い

ディレクション なし

アノテーターは、個人解釈・思い込みで作成。
それぞれの視点で解釈し、バラバラな質問が生成される。
品質にムラが出る。

ディレクション あり

明確な基準と構成に基づいて作成。
全員が同じ理解で作業し、均一で高品質な質問が生成される。

これがあるかないかでクオリティが全く変わります。

理想のプロジェクト体制図–キャリア・マムの場合

各層で品質を担保。貴社との窓口は担当社員に一本化することで、コミュニケーションの齟齬を防ぎ、スムーズなプロジェクト進行を実現します。

過去の業務実施事例

実績例

1

PPTテキスト化

- 期間：3ヶ月
- 人数：**350名**
- 対応件数：**41,000件**
- 対応内容：図・表などを含むスライドについて、自然文で 検索ができるように説明文章を作成する業務。 スライド上の情報に対して、主語や接続詞など省略されている場合、語句を補い、内容を全て 網羅し、複雑な図表もテキスト化。

2

質検応答作成業務

- 期間：3ヶ月
- 人数：**150名**
- 対応件数：**62,000件**
- 対応内容：200文字程度の短文から30~50文字程度の任意の箇所を指定し、「質問」と「解答」、「書き 換え質問」「質問の言い換え」、「削り質問」 「質問の語尾変更」作成する作業。

3

スポーツ映像の字幕へのタグ付け

- 期間：3ヶ月
- 人数：**50名**
- 対応件数：**38時間**
- 対応内容：複数競技の試合の実況映像(録画)を文字に起こして、発言のジャンルや話者特定等、 指定のタグをつける作業。

4

医療画像アノテーション

- 期間：1.5ヶ月
- 人数：**50名**
- 対応件数：**11,000件**
- 対応内容：レントゲン画像上の指定部位をセグメンテーション する作業。

5

生成AI評価補助業務

- 期間：3ヶ月
- 人数：**85名**
- 対応件数：**6,000件**
- 対応内容：生成AIが作成した質問回答文の評価を6項目 の評価項目に照らし合わせ、5段階で妥当性を 細かく精査。3以下の評価については修正案を 作成する業務。

6

チャットボット用データ整備

- 期間：2ヶ月
- 人数：**10名**
- 対応件数：**600ページ**
- 対応内容：既存データの診断・修復と文書ルールの整備、 AI-OCRによるテキスト化とマークダウン構造化 を実施。

AI評価案件の取り組み例：AI相談ボット（仮）

開発と評価プロセス

想定される利用状況に基づいた評価項目を策定し、継続的な品質向上を目指しました。

【評価プロセス全体図】

このプロセスを通じて、AIの応答品質と実用性を多角的に検証します。

1. ユーザー募集・選定
   専門性の異なるユーザー5名を募集・
   選定しアサイン
2. 会話実施
   各ユーザーが1週間で10ターン以上の会話を実施
3. 評価項目策定
   利用想定に基づく評価項目・ポイントを提示
4. AI回答評価
   他ユーザー3名をアサインし、AI回答を評価項目に沿って評価

【ユーザー選定と会話実施】

属性・専門性が異なる5名を選定し、多様な視点を確保
1週間で10ターン以上の対話を行い、実際の利用感を収集
会話内容はログとして記録し、後続評価に活用

【評価項目の策定と評価体制】

利用状況を想定した評価ポイント例：
回答の正確性
専門用語の理解度
対話の自然さ・一貫性
ユーザー満足度
3名の評価者が独立して評価を実施し、多面的な評価を実現します。

まとめ：多角的評価でAI品質を高める
専門性の異なるユーザー参加による実践的評価
明確な評価項目に基づく定量・定性評価の両立

まとめ

↑
開発を支える学習データの作成の鍵は
ディレクション力

判断基準を作る（プロトタイプ）

開発–データ作成の一体化が必要

まずは小ロットトライアルで、
基準と品質を一緒に固めませんか？

ご清聴ありがとうございました

株式会社キャリア・マム
〒206-0033
東京都多摩市落合1-46-1ココリア多摩センター5F

eigyou@mail.c-mam.co.jp
https://corp.c-mam.co.jp/

高品質なAI学習データで、未来を創る

プロトタイプ構築から体制整備、継続的な品質改善まで、AI開発を支える学習データ作成をトータルサポート。12万人のネットワークと豊富な実績で、貴社のAI開発を成功に導きます。