Rev. 1.0.0.0001(2026/3/30)

マルチモーダル AI
品質マネジメントガイドライン

第 1 版
(Revision 1.0.0)

2026 年 3 月 30 日

国立研究開発法人産業技術総合研究所

インテリジェントプラットフォーム研究部門
テクニカルレポート IPRI-TR-2026-01

サイバーフィジカルセキュリティ研究部門
テクニカルレポート CPSEC-TR-2026001

人工知能研究センター
テクニカルレポート

前書き

本ガイドラインは、国立研究開発法人産業技術総合研究所（産総研・AIST）が企業・大学等の有識者委員と共に構成した「機械学習品質マネジメント検討委員会」においてとりまとめたものである。本ガイドラインの検討は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の 2025 年度委託業務「AI の安全性確保に関する研究開発・検証等の推進事業／AI セーフティ強化に関する研究開発」(P25006)の一環として行われた。

本ガイドラインの内容に寄与した委員の意見は技術者としての個人の知見に基づくものであり、各々が所属する会社等の意見を代表するものではない。

本ガイドラインは、生成 AI を利用したシステム・サービスの開発を主導する企業等が、そのビジネス等への影響を踏まえて主体的にその採用の有無を選択し、共同開発者等と共に実践するものであって、法令・公的指針等との関係においては非拘束的なものである。本ガイドライン中で規範的 (normative) とされる規定は、あくまで本ガイドラインを任意に採用した場合に限り、規範的な意味を持つものである。

目次

Executive Summary

既存の予測型機械学習と生成 AI の品質マネジメントで確立されてきた、利用時品質を起点にシステムとしての品質を達成し、構成要素の外部品質目標を内部品質の評価と改善で実現するという骨格を踏まえつつ、マルチモーダルで顕在化する論点を整理して適用できる形にする。

マルチモーダル AI では、画像に写っているものとテキスト中の言葉の関係をどう扱うかが品質に直結しやすく、評価基準の合意の難しさが中心課題になりやすい。

同じ画像でも、どこを重要とみなすかや、どの言葉と対応させるかが用途や立場によって変わりやすく、正解の決め方が揺れると要求の定め方、評価の設計、運用時の合否判断が不安定になる。

そこで本書は、照応の対象範囲を共通領域と固有領域に分け、共通領域に対してクロスモーダル照応能力のレベルを用いて要求と評価の粒度をそろえるという前提を置く。

本書のアプローチでは、品質達成の対象を AI システムに置き、AI システム、機械学習要素、周辺ソフトウェアの役割分担と管理境界を明確にし、入力から最終提示までの手順と証跡を前提にする。

モデルのアーキテクチャを Modular、Native、Hybrid に整理し、表現空間の対応づけの設計と変更が照応性能に与える影響を踏まえて、受入れ評価と変更管理を含む評価計画を設計する。

データ設計では、照応レベルに応じて必要な学習用と評価用のデータを整備し、文や文節の単位で画像に根拠がある主張とない主張を区別できる形を用意して、評価可能性と再現性を確保する。

運用では、導入前の受入れ評価、段階導入、運用中の監視と再評価をライフサイクルとして扱い、入力分布の変化や失敗モードの傾向を継続的に把握し、必要な時点で改善と回帰評価を実施する。

周辺ソフトウェアは、プロンプト構築、RAG による出典参照、出力フィルタによる検査、外部ツールの呼び出しなどを通じてシステムとしての外部品質を補強する。

証跡として、入力識別子、プロンプトのバージョン、RAG の参照出典とスナップショット、モデルのバージョン、確信度、出力と出力フィルタの判定、最終判断の理由と運用者の介入記録を保存し、後から同一条件で結果をたどれる状態を維持する。

本書は、以上の前提と手順を、複数の適用事例を通じて具体化することを意図し、シンプルな構図の用途から、照応の深さや文脈依存性が増す用途へと段階的に示すことで、実務で迷いやすい判断点を共通の言葉で扱える状態を作る。

1 はじめに

本書は、マルチモーダル AI を現場で安心して使うために、何を前提にし、何を合意し、どこを検証するかを見失わないための道筋を提示する。

1.1 背景と問題意識

近年、テキストと静止画像を同時に扱うマルチモーダル AI が業務システムに組み込まれる場面が増え、画像とテキストの両方を入力として受け取り、両者を踏まえて判断や生成を行う用途が広がっている。その結果、画像に写っているものとテキスト中の言葉の関係をどう扱うかが、システムの品質に直結しやすくなっている。

予測型の機械学習を対象とする品質マネジメント手法と、生成 AI を対象とする品質マネジメント手法は、すでにガイドラインとして提示されている。 これらのガイドラインは、品質達成の対象を AI システムに置き、利用時品質から構成要素の外部品質を定め、内部品質の評価と改善によって外部品質を達成し、利用時品質へ到達するという基本的な考え方を提供している。 しかし、これらの手法はマルチモーダル AI を主要な対象として想定していないため、同じ考え方をそのまま適用しようとすると、現場で迷いが生じやすい。

マルチモーダル AI では、評価基準の合意の難しさが目立つ。画像は情報量が多く、同じ画像でも、どこを重要とみなすかや、どの言葉と対応させるかが用途や立場によって異なりやすいため、正解の決め方が揺れやすい。 この揺れは、要求の定め方、評価の設計、運用時の合否判断のいずれにも影響し、品質マネジメントを進めにくくする。

そこで本書は、既存の二つのガイドラインの考え方を踏まえつつ、マルチモーダル AI で顕在化する課題を整理し、品質マネジメントに向けたアプローチを提示する。

1.2 本書が対象とするマルチモーダル AI の定義

本書では、複数の異なるデータ形式を入力として受け取り、かつ出力として生成し、それらを組み合わせて情報処理を行う AI システムをマルチモーダル AI と呼ぶ。ここでいうデータ形式とは、情報を表現し伝達する形式を指し、本書ではテキストと静止画像を主対象とする。

マルチモーダル AI としては、他にもさまざまなデータ形式 (以下モダリティと呼ぶ) を扱うものがある。また複数のモダリティを扱う AI の中には、出力としてラベルもしくは単一の予測値を出力とするもの (識別・予測 AI、 predictive AI)も考えられる。しかし、本書では、マルチモーダル AI のうち、出力としてテキストを生成する生成 AI (generative AI)を主な対象とする。

マルチモーダル AI の実現手法としては、モデル単体が複数モダリティを扱える場合に限らず、単一モダリティのモデルを複数組み合わせることで、システムとして複数モダリティを扱えるようにしたものも対象に含める。たとえば視覚と言語を扱うモデルと大規模言語モデルを組み合わせて、画像とテキストを入力として受け取り、テキストを出力するような構成も対象に含める。

1.3 想定読者と利用場面

本書の想定読者は、マルチモーダル AI を利用するシステムの企画、開発、導入、運用、評価に関与し、品質に関する合意や説明責任を担う立場の人である。主として、AI を利用するシステムの企画者とプロジェクト管理者、AI 開発者と提供者、品質保証担当者、調達担当者、監査に関与する担当者を想定する。研究者や技術者であっても、実運用を見据えて品質の考え方を整理したい場合に参照できる。

本書は、次のような場面で利用することを想定する。

- 導入可否の検討と適用範囲の決定
- 要求の整理と関係者間の合意形成
- システム構成と役割分担の整理
- 評価計画の策定と合否判断の根拠づけ
- 外部調達部品の選定と受入れ評価の準備
- 運用開始前の準備と運用中の変更管理
- 不具合発生時の状況整理と再発防止の検討
- 監査や説明のための記録と証跡の整理

これらの利用場面では、個別の実装技術の解説よりも、品質目標の立て方、評価の考え方、運用上の判断の筋道を揃えることが重要になる。本書は、そのための共通前提と整理の枠組みを提供し、関係者が同じ言葉で議論できる状態を作ることを狙いとする。

1.4 本書の構成と読み方

本書は、マルチモーダル AI システムの品質マネジメントを検討し実務に適用するために、前提の共有、課題の整理、手順の提示、適用事例の提示という順に章を配置する。前半で共通の前提と言葉をそろえ、後半で実務に落とし込む際の判断の筋道を示す。

本書の構成は次のとおりである。

第 1 章 本書の目的、対象範囲、想定読者、用語の定義
第 2 章 AI 品質マネジメントの一般論の要点
第 3 章 本書で扱うマルチモーダル AI の前提とシステム構成
第 4 章 マルチモーダル固有の課題整理 [1]
第 5 章 マルチモーダル AI に特化した品質マネジメント手順
付録 1 適用事例による具体化

読み方は、目的に応じて選ぶ。全体像を把握したい場合は上から順に読む。すでに AI 品質マネジメントの一般論に馴染みがある場合は、第 3 章から読み始めて対象領域の前提を確認し、課題と手順に進む。実務に適用する場合は、まず第 5 章で手順の全体像を確認し、そのうえで付録 1 を参照して自組織の用途に近い例を手掛かりに検討事項を具体化する。

1.5 用語集

用語集における定義の位置づけ

本用語集は、本書の議論を一貫させるために、本書の中で用いる語の意味を定めるものである。ここで示す定義は、広く一般的に認められた定義を主張するものではなく、本書の中での用法を確定することを目的とする。

マルチモーダル AI

本書では、複数のモダリティを入力として受け取り、かつまたは出力として生成し、それらを組み合わせて情報処理を行う人工知能をマルチモーダル AI と呼ぶ。本書の主対象は、テキストと静止画像を入力として受け取り、テキストを出力するマルチモーダル AI である。

AI システム

本書では、訓練済みモデルを含む構成要素と周辺ソフトウェアを組み合わせ、利用者に機能

[1] 本書では残念ながらマルチモーダル AI のセキュリティに関して触れることができていない。マルチモーダル AI 固有の脆弱性、攻撃、対策については多数の報告がある。これらの整理は本書の次版での課題である。

を提供するシステム全体を AI システムと呼ぶ。品質マネジメントの最終的な達成対象は AI システムであり、AI システムが利用者に提供すべき品質を利用時品質として扱う。

モダリティ
本書では、情報を表現し伝達する形式や様式をモダリティと呼ぶ。本書の主対象では、テキストと静止画像をモダリティとして扱う。

大規模言語モデル（LLM）
本書では、主として自然言語のテキストを入力として受け取り、テキストを出力として生成できる大規模な訓練済みモデルを大規模言語モデルと呼ぶ。LLM は単体で用いる場合に限らず、マルチモーダル AI システムの構成要素として組み込まれる場合も含めて扱う。

ビジョン言語モデル（VLM）
本書では、視覚情報としての静止画像と自然言語テキストを同時に扱い、両者を踏まえた理解や推論や生成を行う訓練済みモデルをビジョン言語モデルと呼ぶ。VLM は、モデル単体としてマルチモーダルである場合と、AI システムの構成要素として用いられる場合の両方を含めて扱う。

2 AI 品質マネジメント

本章は、AI 品質マネジメントに共通する基本的な考え方を、本書全体の前提として整理し、以降の章で参照するための土台を与える。本章で述べる考え方は産総研がこれまでに発行した AI 品質マネジメントガイドライン 2 件 [MLQM2023][MLQM2025] と共通している。

2.1 品質マネジメントの対象と階層

本書では、品質マネジメントの最終的な対象を AI システムに置く。AI システムが利用者に提供すべき品質を利用時品質と呼び、利用時品質を満たすことを品質マネジメントのゴールとして扱う。次に、AI システムを構成する要素のうち、訓練済みモデルを含む要素を機械学習要素と呼び、利用時品質を実現するために機械学習要素が備えるべき品質を機械学習要素の外部品質として定める

外部品質を達成するために、設計時や運用時に満たすべき事項を内部品質と呼ぶ。内部品質は、問題領域、データセット、モデル、プログラム、運用といった観点に現れ、内部品質を評価し改善することにより外部品質を間接的に実現し、外部品質の達成を通じて利用時品質へ到達するという構造で整理する。以降の節では、この階層構造を前提

に、外部品質の名称と内部品質の管理観点、外部調達部品を含む場合の扱い、ライフサイクルとしての運用の要点を簡潔に示す。

コラム　AI の品質をマネジメントするとは

品質マネジメントとは、出来上がったものを検査して合否を決めるだけの活動ではなく、何を良い品質とみなすかを先に定め、その達成に必要な手段を計画し、結果を確認し、必要に応じて改善するという一連の活動である。AI を使ったシステムでは、入出力の種類や利用状況が多様であり、動作のばらつきや想定外の振る舞いが利用者や事業に影響する可能性があるため、品質を偶然に任せず管理対象として扱う必要が生じる。

AI の現場では、モデルの精度を上げることだけが品質だと誤解されやすいが、実際には用途に応じて求められる品質は異なる。たとえば安全性の重みが大きい用途もあれば、誤りが少ない平均性能が重視される用途もあり、法や社会的要請や運用制約の影響も受ける。したがって、品質マネジメントでは、何を達成すべきかを用途と利用状況から明確にし、その達成を継続的に確認できる形にすることが重要になる。

さらに、多くの場面では AI システムの一部に外部調達した部品が含まれ、その開発工程に直接関与できないことがある。大規模言語モデルなど、開発に多大なリソースを必要とする汎用基盤モデルは多くの企業では外部調達の対象となる。この場合でも品質マネジメントの対象をシステムに置けば、周辺ソフトウェアの設計や受入れ評価、運用時の監視と更新管理によって、システムとしての品質目標に到達しやすくなる。外部調達部品を使う場合には、開示情報を確保し、評価と運用に必要な情報が継続的に入手できる体制を整えることが有効になる。

本書はこの考え方に基づき、AI システムが利用者に提供すべき品質を起点にして、構成要素の品質目標と管理対象を整理し、評価と改善を繰り返す枠組みを示す。以降の節では、詳細な手順やチェック項目は既存の各ガイドラインに委ねつつ、最小限の要点を示す。

コラム　仕様中心から品質目標中心へ

従来ソフトウェアの品質保証では、設計者が仕様を定め、その仕様に沿って動作することを確認すれば品質を管理できるという前提が成り立っていた。そのため、品質保証の中心は仕様の明確化と、仕様に基づく検証の積み重ねに置かれてきた。

機械学習を用いる AI は、人が仕様として動作の詳細を十分に書き下せない処理を実現するために用いられることが多い。この場合、仕様と動作の一致を検証するだけでは、必要な品質を達成できているかを説明する根拠が不足しやすい。すなわち、従来の仕様中心の枠組みだけで品質を管理しようとすると、検証できない部分が残りやすい。

そこで、仕様そのものではなく、用途と利用状況に基づいて必要な品質を先に定め、その目標を達成することを品質マネジメントの中心に置く発想が必要になる。本書では、AI システムが利用者に提供すべき品質を利用時品質として定め、そこから構成要素が備えるべき品質へと落としていく考え方を採る。

この発想は、外部調達した部品を含むシステムにも適用できる。部品の開発工程に直接関与できない場合でも、システムとしての品質目標を定め、構成要素の外部品質と内部品質を管理し、必要な証跡を残すことで、品質の達成と説明責任の両方を成立させやすくなる。本書の以降の節では、この考え方を前提に、詳細は既存の各ガイドラインを参照しつつ、品質目標中心のマネジメントを実務に落とし込むための要点を整理する。

2.2 外部品質と内部品質の関係

外部品質は、利用時品質を実現するために機械学習要素が備えるべき品質として定め、目標レベルを設定したうえで達成を目指す。外部品質は、内部品質を評価し改善することで間接的に実現されるという前提に立ち、内部品質の状態を継続的に把握して、必要な対策を講じることで外部品質の目標に到達する。

外部品質特性は、次の名称で整理する。

- リスク回避性
- AI パフォーマンス
- 公平性
- プライバシー
- AI セキュリティ

内部品質は、外部品質に直接影響する管理観点として定義し、問題領域、データ、モデル、プログラム、運用にまたがる観点を対象にする。代表的には、問題領域分析とデータ設計の十分性、データセットの被覆性と均一性と妥当性、モデルの正確性と安定性、モデル外のプログラムの信頼性、運用開始時に確保した品質の維持性を管理対象として

扱う。内部品質の評価と改善は、外部品質の目標レベルを達成するための手段であり、どの内部品質を優先して改善すべきかは、用途とリスクの性質に応じて決定する。

2.3 外部調達部品を含む場合の一般的な方針

AI システムの構成要素の一部を外部から調達して利用する場合、当該部品の開発工程や学習過程に直接関与できないことがある。典型的なケースは、世界最先端の汎用基盤モデルを調達して使う場合で、モデル利用者はモデル提供者が提示する UI や API を介してモデルを利用することしかできない。

このとき品質マネジメントの対象は部品単体ではなく AI システムに置き、利用時品質の達成に必要な外部品質目標をシステム設計で定め、システムとして合否を判定できる状態を整えることが重要である。

この前提のもとでは、周辺ソフトウェアが品質達成の主要な手段となる。具体的には、プロンプト構築、RAG による出典参照、出力フィルタによる検査、外部ツールの呼び出し、最終提示の制御をシステム側の責務として設計し、部品のふるまいを前提にせずに合否を満たす経路を確保する。外部調達部品を選定するときは、推奨用途、推奨入力の形式、学習データの性質、評価結果などの開示が十分なものを優先し、受入れ評価と運用監視に必要な情報が継続的に入手できることを確認する。

運用では、外部調達部品のバージョンと周辺ソフトウェアのバージョンを一体として管理し、変更時には回帰試験とロールバックの手順を準備する。証跡として、入力識別子、プロンプトのバージョン、RAG の参照出典とスナップショット、モデルのバージョン、確信度、出力と出力フィルタの判定結果を保存し、後から同一条件で結果をたどれるようにする。これにより、外部調達部品に由来する不確実性が残る状況でも、AI システムとしての品質の説明責任を果たしやすくなる。

2.4 ライフサイクルとしての品質マネジメント

品質マネジメントは導入前の一回限りの評価ではなく、設計、開発、導入、運用の各段階で評価と改善を繰り返し、運用期間を通じて品質を維持する活動として扱う。設計段階では利用時品質と外部品質の目標を確定し、内部品質の管理観点を定める。導入段階では受入れ評価により合否を判定し、運用段階では入力分布や運用条件の変化を監視し、必要な時点で再評価と改善を実施する。運用開始時に確保した品質を継続的に維持

するという観点は内部品質として位置づけ、品質の劣化を早期に検知して対処できる体制を前提とする。

以上、本章では AI 品質マネジメントの一般論の骨格を示した。以後の章はこの骨格を前提としてマルチモーダル AI 特有の事項を説明する。第 3 章では本書が対象とするマルチモーダル AI の前提と範囲を定め、AI システムと機械学習要素の境界を含む最小構成とモデルアーキテクチャの類型を位置づける。第 4 章ではマルチモーダルに固有の課題を整理し、第 5 章で課題に対応する品質マネジメントの手順を示す。付録 1 章では適用事例を通じて、前章までの考え方を具体のユースケースに当てはめる際の見取り図を与える。

3　マルチモーダル AI

本章は、本書で扱うマルチモーダル AI の前提と範囲を定め、以降の章で議論する品質上の課題とその扱いに共通する土台を与える。ここで示す枠組みは、後続の章における用語と議論の整合を保つための参照点として用いる。

3.1　概観

本章は、以後の議論で用いるマルチモーダル AI の前提と範囲を定める。対象はテキストと静止画像を基本モダリティとするシステムとし、生成を主用途に据える。予測型の利用形態は必要に応じて触れるが、詳細は本書の品質マネジメント各章に委ねる。エージェント的な自律反復や外界の状態変更を伴う機構は範囲外とする。

本章の構成は次のとおりである。まず 3.2 節で、AI システムと機械学習要素の境界を明確にした最小構成を示す。以降の章で参照可能な共通前提とする。続く 3.3 節では、モデル側のアーキテクチャを Modular、Native、Hybrid の三つに整理し、どこで表現空間の対応づけを行い、どの範囲のパラメータを学習で更新する設計なのかという観点で位置付ける。

3.2　システムアーキテクチャ

本節は、テキストと静止画像を扱うマルチモーダル AI の最小構成を定め、その上で第 4 章後半のモデルアーキテクチャの位置付けと、後章で扱う工程・評価の前提をそろえる。モデルと組合せて AI システムを構成する周辺ソフトウェアはプロンプト構築や

参照機能（RAG）、出力の検査と提示、証跡の記録を担い、機械学習要素はモデル本体として内部で表現空間の対応づけを含む推論を担う。二者の境界を明確にし、入力からモデル、モデルから出力、出力から出力フィルタ、出力フィルタから最終提示へと進む直列の流れを基本とする。

入力は、画像とテキスト、および行わせたい処理を記述したプロンプトとする。プロンプトはテンプレート適用や引数埋め込みなどの前処理によって構成され、以後の推論に一貫した形式で渡される。ここでの前処理は周辺ソフトウェアの責務であり、モデル内部の挙動に依存しない形で再現できる必要がある。

参照機能として、RAG を用いて出典に基づく根拠情報を取得し、プロンプトに組み込む。参照対象はテキストだけでなく画像文書も含められるが、いずれも参照は一方向であり、モデルの入出力の直列な流れを乱さないようにする。以降の議論を複雑にしすぎないために、外部ツールは情報検索や計算を行う小さな呼び出しに限り、このシステムの外側の状態は変えない範囲にあえて限定する。

機械学習要素は、マルチモーダルのモデル本体として、画像と言語の内部表現を橋渡しする表現空間の対応づけを内部に持つ。構成や学習の仕方は後節で述べるが、アーキテクチャにかかわらず、この内部機構が推論全体の挙動に影響する点は共通である。AI システムとは明確に境界を引き、モデル入出力のインタフェースを安定に保つ。

出力は、まず周辺ソフトウェアの出力フィルタで検査する。ここでは規則や別モデルを用いて不適切な断定や整合性の欠落を抑え、その後に最終提示へ渡す。出力フィルタはモデル本体と独立に改修できるように設計し、運用時の調整や段階導入を支えやすくしておく。

情報の流れは、入力からモデル、モデルから出力、出力から出力フィルタ、出力フィルタから最終提示へと直列に進む。RAG や外部ツールの参照は片方向とし、モデル入力側に統合する。循環や自律的な反復は扱わず、同期の一回呼び出しで完結させる。これにより、アーキテクチャを複雑化させずに役割分担を保てる。

再現性と監査のため、周辺ソフトウェアで証跡を記録する。具体的には、入力識別子、プロンプトのバージョン、RAG の参照出典とスナップショット、モデルのバージョン、確信度、出力と出力フィルタの判定などを、後から同じ入力と設定で結果をたどれる粒度で保全する。これらは後続の品質マネジメントで用いるが、位置付けとしてはシステム構成要素に属する。

3.3 モデルアーキテクチャ

本節では、テキストと静止画像を扱うマルチモーダル AI モデルのアーキテクチャ類型として Modular, Native, Hybrid を概説する。これらは本書における定義であり、広く合意されたものではない[2]。いずれの類型でも、画像と言語の内部表現を橋渡しする表現空間の対応づけが不可欠であり、ここでの設計（どこで・どのように対応づけるか）がシステム全体のふるまいに大きく影響する。

3.3.1 Modular (モジュール型)

各モダリティを専用エンコーダで埋め込みに変換し、写像層（プロジェクタ）を介して言語モデルに入力する構成とする。開発時は、事前学習済みのエンコーダや言語モデルを原則として凍結し、写像層や小規模アダプタのみを学習対象にする設計を基本とする。必要に応じてエンコーダを部分的に更新する変種はあるが、基調は凍結再利用である。表現空間の対応づけの出来が品質の要となり、段階ごとの出力を追跡しやすいという運用上の利点をもつ。

3.3.2 Native (ネイティブ型)

複数モダリティを単一のトークン列として同一モデルで処理する統合構成とする。学習時はモデル全体を更新対象とし、モダリティ間の相関をエンドツーエンドで獲得する。事前学習済みのベースを起点にする場合でも、運用要件を満たす段階では全体または大部分を再学習する前提が強い。深い相互作用を捉えやすい一方で、学習や運用に必要な資源は大きく、内部動作の解釈は相対的に難しくなる。

3.3.3 Hybrid (ハイブリッド型)

構造は Modular に近く写像層を用いるが、学習は Native に寄せ、モダリティを同時に学習して一体最適化の効果を取り込む。凍結と更新を部位ごとに組み合わせ、写像層

[2] [Shucor 2025] ではマルチモーダル AI のアーキテクチャをモダリティ融合のタイミングと事前学習の有無の 2 軸で分類している。本書の Modular は遅い融合、事前学習あり、Native は早い融合、事前学習なしに概ね対応するが、事前学習なしをすべて native と呼ぶ場合や、一部事前学習ありでも native と呼ぶ文献も見られる。

や一部エンコーダは更新し、言語モデルは凍結とするなど、実装方針に応じた配分を設ける。改修余地と一体学習の効用の折り合いを付けやすい反面、どの部分を凍結しどの部分を更新するかの方針を明確にしないと振るまいが不安定になりやすい。

4 マルチモーダル AI の品質マネジメントの難しさ

4.1 モダリティ間の照応

テキスト内の代名詞解決のように、記号列の内部で参照関係を結び直すだけでは済まない場面がある。画像とテキストが併置されるタスクでは、二つのモダリティのあいだで対象や出来事を対応させる照応が必要になり、説明文生成、質問応答、検索、適法性判定などの基盤となる。ここで必要とされる照応は、対象の有無の一致にとどまらず、対象ごとの属性の帰属や対象間の関係の一致、場合によっては状況に関する抽象的判断までを含む。この照応を行う能力をここでは「クロスモーダル照応能力」と呼ぶ。さらに 4.2 節において、クロスモーダル照応能力に 4 つのレベルを導入する。

クロスモーダル照応は、あらゆる情報に対して成立するわけではない。画像とテキストの双方で同一の対象や事象を表現できる「共通領域」と、特定のモダリティでのみ自然に表現できる「固有領域」を区別する必要がある。共通領域の例としては、写っている物体やその位置関係、個数、色などが挙げられる。固有領域の例としては、画像からは直接確認できない素材や由来の情報、歴史的背景や制度上の概念などがある。抽象概念は多くの場合テキスト側での記述が中心となり、細かな質感や微妙な形状差は画像側に依存することが多い。

この区別によって、実務で行うべきことが変わる。共通領域は画像からテキスト、テキストから画像への相互参照が可能であり、照応の評価対象となる。一方、固有領域はモダリティ固有の能力として扱い、照応の評価とは分けて設計と評価を行う。両者を混在させると、過剰な要求や不当な不合格を招き、品質マネジメントの判断が揺らぐ。したがって、要求を定める段階で共通領域と固有領域の境界を明示し、照応の対象範囲と、次節で定義する照応レベルを参照した判定の基準を文章で固定しておくことが望ましい。設計や評価の段階では、この区別をそのまま引き継ぎ、共通領域に対しては照応の合否を、固有領域に対してはモダリティ固有の指標を用いて判定する。運用においても、両領域の扱いを混同しないことが、合否判断の一貫性を保つ前提となる。

4.2 照応レベル

本節では、画像とテキストのあいだで対象や出来事を対応させる力を「クロスモーダル照応能力」と呼び、その達成段階を「照応レベル」（以下「レベル」）として定義する。レベルは、要求の明確化、データ設計、評価設計、運用上の合否判断における共通言語として用いる。評価の対象は共通領域に属する事柄とし、固有領域に属する事柄は照応の評価対象から切り離す。

- Level 1 は対象の有無とカテゴリの認識を扱うレベルであり、画像に写る事物が存在するかどうかを正しく判定し、要求で定めたカテゴリ体系に従って適切なクラスラベルを付与する(例:犬が写っているか､車か自転車かを誤らない)。
- Level 2 は属性の帰属を扱うレベルであり、どの属性がどの対象に属するかを正しく結び付ける（例：赤いのは車、丸いのは時計)。
- Level 3 は事象間の関係を問うレベルであり、対象どうし、あるいは対象とフレームとの関係を正しく記述できることを求める(例:男性の右側に机がある、画像の下半分に花が見える、箱の中に本が入っている)。
- Level 4 は抽象的認識を問うレベルであり、状況や意図、概念的なラベル付けなど、具体的所見の組み合わせを超えた解釈の妥当性を扱う（例：群衆の様子から混乱が生じていると認識するか)。

以上の定義は､タスクの目的に応じて必要最小限のレベルを選び取るための座標である。どのレベルを満たせば業務が成立するかをあらかじめ記述し、共通領域に対しては Level 1 から Level 3 を中心に、やむを得ず抽象的判断が不可避な場合に限って Level 4 を適用する。これにより、過不足のない要求と評価可能性を備えた品質マネジメントの基盤を整える。

4.3 照応失敗の類型

本節は、クロスモーダル照応能力の照応レベルに沿って、画像とテキストの対応がどのように崩れるかを整理する。マルチモーダル AI に特有の失敗だけを対象とし、単一モダリティだけで完結する誤り（画像分類の単純な取り違えなど）とは区別する。

Level 1 では、対象の有無やカテゴリの認識がモダリティ間で一致しない失敗が生じる。テキストが指示する対象を画像内で見つけられない、またはテキストが主張するカテゴリと画像に写っている事物のカテゴリが一致しないといった不整合が代表である。

Level 2 では、属性の帰属が誤る。画像内に複数の対象が写っているときに、テキストで述べた属性を誤った対象に結び付ける、あるいは画像で裏付けられない属性を断定的に述べるといった誤りが典型である。たとえば、赤い車と白い家がある画像に対して「赤い家」と記述するような取り違えが該当する。

Level 3 では、対象どうし、または対象とフレームとの関係が誤る。相対位置や包含、順序や役割といった関係の記述が、画像で観察できる関係と一致しない失敗が生じる。たとえば、「男性の右側にある机」「画像の下半分に見える花」といった表現が、実際の位置関係やフレーム基準と合致しない場合がこれに当たる。

Level 4 では、抽象的判断の読み誤りが生じる。具体的所見の組み合わせを超えて状況や意図を解釈する際に、文脈を取り違えることで、同じ画像にもかかわらず解釈が不適切に転ぶ。たとえば、祭りの写真を暴動の様子と理解するような抽象的な誤認識がこれに該当する。

以上の失敗は、照応レベルが上がるほど文脈や対象間の構造に依存しやすく、共通領域と固有領域の切り分けが曖昧なままでは評価も設計も揺らぐ。次節以降では、これらの失敗を前提に、要求、データ、評価、運用の各工程で照応の合否を安定して判定できるように、レベルに沿った前提と記述を整える。

5 マルチモーダル AI の品質マネジメント

本章は、マルチモーダル AI に特有の品質マネジメントの進め方を示し、AI システムの利用時品質を起点に照応の要求水準と外部品質・内部品質の関係を明確にする。続く各節では、役割分担と管理境界から評価と運用までを段階的にたどり、表現空間の対応づけに配慮しつつ実装へ落とし込む道筋を提示する。

5.1 概観

本章は、第 3 章で定めた枠組み（AI システムの利用時品質を起点に、要素の外部品質目標を定め、内部品質の評価・改善で到達する）を前提に、対象をマルチモーダル AI に限定して適用の道筋を示す。最終の合否は AI システムとしての照応に対して判定し、表現空間の対応づけは照応の達成可能性に関わる内部機構として扱う。

以降では、役割分担と管理境界、問題領域分析、アーキテクチャ依存性、データ設計・作成、モデル品質マネジメント、周辺ソフトウェア、運用・保守・継続的改善、評価フ

レーム、ガバナンスと調達、文書化と導入準備の順に、マルチモーダル AI の品質マネジメントを段階的に整理する。各節は、利用時品質を最終判定の拠り所としつつ、外部品質と内部品質を結ぶ証跡と再評価の要点に絞って記述する。

5.2 役割分担と管理境界

本節は、AI システム、機械学習要素、周辺ソフトウェアの責務と権限を区分し、利用時品質を最終判定の対象とする前提のもとで、外部品質の達成をどの要素で担うかを定める。AI システムは品質達成の主対象であり、合否は照応（Level 1, 2, 3、必要に応じて Level 4）に対して判定する。機械学習要素は外部品質の一部（照応性能）を直接担い、表現空間の対応付けの内部機構は、その達成可能性を左右するが、合否の根拠としては照応評価を優先する。周辺ソフトウェアは、プロンプト構築、RAG、出力フィルタ、ポリシー適合確認、外部ツール連携、確信度に応じた保留や人へのエスカレーションを通じて、システムとしての外部品質を補強し、モデルへの介入が難しい場合でも合否を満たす経路を提供する。

管理境界は、入力から最終判断までの工程で明確化する。入力の受領、前処理、プロンプトと入力構造の決定、RAG の出典選択、推論の実行、出力フィルタによる検査、最終判断と記録の一連を手順化し、どの工程を機械学習要素が自律的に担い、どの工程を周辺ソフトウェアが統制するかを定義する。共通領域（画像とテキストの双方で表現できる対象）に属する照応は機械学習要素の主責とし、固有領域（片方のモダリティに依存する判断）や断定抑制、出典提示、適法性・ポリシー適合などは周辺ソフトウェア側の責務として設計する。これにより、共通領域の対応は照応評価、固有領域はモダリティ固有の評価で扱うという試験設計の分離が可能になる。

調達と改修の前提も、役割分担に組み込む。外部調達が基本となるモデルについては、受入れ評価と運用監視をシステム側で実施し、必要に応じて表現空間の対応付けに関わる要素に限定的に関与する（写像層の再学習や画像エンコーダの追加学習など）。関与できない部分は周辺ソフトウェアの改定（プロンプト様式、RAG の索引、出力フィルタの規則）で補う。変更管理では、モデルのバージョンと周辺ソフトウェアのバージョンを一体で扱い、変更単位ごとに補助評価（表現対応の適合度など）と照応評価を併用した回帰試験を計画する。

証跡と説明責任は、境界の合意と一体で運用する。最小集合として、入力識別子、前処理の手順、プロンプトとその版、RAG の出典、モデルのバージョン、推論設定、出力と出力フィルタの判定、最終判断の理由、運用者の介入記録を保存し、工程ごとの責任者と承認権限を文書化する。これにより、外部品質の合否、内部品質の評価結果、運用判断の三者を再現可能に連結し、監査や回帰試験に耐える体制を維持できる。

5.3 問題領域分析

問題領域分析は、AI システムの利用時品質を起点に、マルチモーダル AI が担うべき機能の射程、出力の様式、照応の要求水準を定める作業である。まず、利用者、利用場面、合否基準、説明責任の要件を明らかにし、画像とテキストのどの対応（有無、属性の帰属、事象間関係、抽象的判断）を扱うかを決める。照応の要求水準は Level 1, 2, 3, 4 で表し、どの水準が達成されれば業務が成立するかを利用時品質と一体で記述する。併せて、画像とテキストの双方で表現できる共通領域と、片方のモダリティに依存する固有領域を切り分け、前者は照応の評価対象、後者はモダリティ固有の判定に委ねる前提を置く。

AI に行わせたいことは、想定ユースケースごとに具体化する。例として、画像の内容と整合する説明文の生成、画像に対する質問応答、テキスト条件に合う画像の検索、コンテンツの適法性判定などがある。各タスクについて、必要な照応水準（たとえば、説明文生成なら Level 2 と Level 3、検索なら Level 1 と Level 2 など）と、出力の様式（短文か段落か、根拠の提示が必要か、不確実性の表現が必要か）をあらかじめ固定し、評価時に合否を判定できるようにする。表現空間の対応付けはこれらの達成可能性に影響する内部機構であるが、最終の合否はあくまで照応に対して行う。

AI にしてほしくないことは、照応の失敗モードとして列挙し、合否の禁止条件として明示する。典型的には、画像に存在する対象の見落とし（Level 1）、属性の取り違え（Level 2）、関係の誤認（Level 3）、抽象概念の早合点（Level 4）、画像と無関係な内容の生成や断定的記述（根拠欠如）、テキストと画像の不整合判断、出力フィルタのすり抜けなどがある。これらは評価データと試験手順に反映し、合否基準の中で禁止事項として扱う。

出力の様式は、利用時品質と照応水準の双方から決める。根拠提示が必要な場面では、文または文節単位で画像に裏付けがある主張とない主張を区別できる出力を求める。不

確実性が重要な場面では、推定や可能性の表現を許容し、断定を抑制する。追加の知識が不可欠な場面では RAG を用いて出典を明示し、出力フィルタで不適切な内容を抑制する。これらの設計は評価可能性と再現性を意識して定義し、照応の合否判定に直接接続する。

5.4 モデルアーキテクチャによる違い

マルチモーダル AI のアーキテクチャは、外部品質の達成可能性と、改善のしやすさや再設計コストに直結する。Native は複数モダリティを前提に内部表現を一体で学習する構成であり、Modular は大規模言語モデルを中核に画像エンコーダと写像層を接続して表現空間の対応付けを行う構成であり、Hybrid はその中間に位置づく。本節では、利用時品質から割り付けた外部品質目標を達成するうえでの影響と、選定時に確認すべき事項を整理する。

5.4.1 Native

Native は最初から複数モダリティを同一のモデル構造で扱うため、モデル内部の表現が多モーダル前提で形成される統合性の高さが期待できる。他方で、機能追加や挙動修正には全体の再学習が必要になる場合が多く、外部調達モデルを用いる場合は学習過程や学習資源に関与できないという制約が強い。したがって、Native を採用する場合は、システム側の設計と周辺ソフトウェアの制御で品質を担保し、モデル自体は受入れ評価と運用監視を中心に扱う前提が適合する。

5.4.2 Modular

Modular は、大規模言語モデルを中核として画像エンコーダを接続し、写像層によって表現空間の対応付けを行う構成であるため、外部品質のうち照応性能、とりわけ Level 2（属性の帰属）と Level 3（事象間関係）の達成度が、この対応付けの設計と学習に強く左右される。写像層の再学習や、必要に応じた画像エンコーダの追加学習により段階的な改善が可能である一方、変更の影響は局所に見えても照応性能に直結するため、受入れ評価と変更管理では、表現空間の対応付けの良否を測る補助評価（画像からテキスト／テキストから画像の検索整合など）と、外部品質の合否を判定する照応評価（属性・関係の一致など）を併用して回帰試験を設計する。調達と選定では、写像層の

構造、画像エンコーダの種類と凍結範囲、学習に用いたデータの性質、損失設計、ならびに表現空間の対応付けと照応に関する評価結果の開示を重視し、開示内容に基づいて改善計画と検証手順をあらかじめ確定する。

5.4.3 Hybrid

Hybrid は一部のサブモジュールの微調整を許容するため、Native ほどの再学習負荷を負わずに改修余地を確保できる。ただし、改修可能な範囲と影響範囲は構成に依存するため、改修単位と回帰評価の境界を事前に規定し、変更時の評価計画に反映する。

5.4.4 選定原則

アーキテクチャの選定は、必要なクロスモーダル照応能力の水準、改善の頻度と許容停止時間、保守時の検証コスト、調達形態における関与可能性という四点を起点に行う。外部調達を前提にする場合は、モデルの素性に関する開示が十分であることを優先条件とし、システム側と周辺ソフトウェアの設計で利用時品質を達成する戦略を明示する。
開示要求（選定時に確認する事項の例）

- アーキテクチャ類型および改修可能範囲の明示
- 写像層の構造、画像エンコーダの種類と凍結範囲、追加学習の可否
- 学習に用いたデータの性質と来歴、データの分割と再現性の管理方法
- 損失関数、学習設定、正則化や制約の方針
- 表現空間の対応付けの評価結果（画像からテキスト、テキストから画像の検索整合など）、属性と関係の一致に関する評価、合否基準の設定根拠
- 既知の失敗モードと回避策、推奨用途と非推奨用途、評価レポートの提供方法

これらの開示は、表現空間の対応付けが外部品質を左右する Modular の場合に特に重要であり、Native や Hybrid でも受入れ評価と運用上のリスク管理に不可欠である。

5.4.5 変更管理と回帰評価

Modular では写像層や画像エンコーダの入替や再学習によって局所的に改修できる一方、その変更は Level 2 と Level 3 に影響するため、表現空間の対応付け評価と回帰試験を変更単位で必ず計画する。Native は変更が全体再学習に波及しやすく、回帰評価の範囲が広がるため、導入前の受入れ評価と運用中の監視を強化し、リリース計画に十

分な検証時間を確保する。Hybrid は構成要素ごとに両者の中間的な扱いを定義し、改修可能な境界と試験の責任範囲を調達時点で合意しておく。

5.4.6 データに関する含意

Modular では、画像とテキストの対に基づく表現空間の対応付け重視のデータを反復的に整備し、文単位の根拠付けなどの注記を評価データに付与する設計が有効である。Native では、多様で広範な対データを通じて多モーダルな内部表現を形成する前提で評価計画を組み立てる。いずれの場合も、利用時品質で要求する Level 1 から Level 4 の到達水準と共通領域と固有領域の切り分けを、データ設計と評価設計に一貫して反映する。

5.4.7 周辺ソフトウェアとの境界

どのアーキテクチャでも、入力の整形、外部知識の参照、出力の検査や抑制などの周辺ソフトウェアが品質達成の主手段となる。したがって、モデル側の改修方針と周辺ソフトウェアの責務分担を、インタフェース仕様、出典と根拠の記録方式、版管理と監査手順とともに前提設計へ織り込む。

5.5 データセットの設計と開発

マルチモーダル AI のデータ設計は、利用時品質から割り付けた外部品質目標を確実に達成するための内部品質として位置づけ、照応性能の水準（Level 1, 2, 3, 4）と共通領域と固有領域の境界を最初に定めてから、学習用と評価用の双方を計画する。Level 1 では対象の有無を安定に扱うため、クラス分布と多様性を管理した分類データを用意する。Level 2 では同一画像内の複数対象に対する属性の帰属を正しく扱う必要があるため、対象ごとの位置情報や属性注記を明確に付与する。Level 3 では対象間の関係を誤らないことが焦点となるため、画像とテキストを対にし、関係の種類と向きを特定できる記述を整える。Level 4 は抽象的判断を含むため、可能であれば抽象概念に直接対応する対データを増やし、困難な場合は抽象を具象の組合せに分解して扱えるよう、具象所見を列挙したテキストと抽象的結論の対応を設計に反映する。これらはいずれも、文や文節ごとに画像に根拠があるかどうかの注記を付与して、根拠に基づく生成や説明の一貫性を後段で評価できるようにしておくことが望ましい。

アーキテクチャに応じた重心の置き方も重要である。Modular 構成では、表現空間の対応付けや写像の設計と学習を段階的に改善できる余地があるため、画像とテキストの対に基づくデータを反復的に整備し、対応付けの質が照応性能（特に Level 2 と Level 3）に与える影響を確認できるように、難例や取り違えが生じやすい対を意図的に含める。Native 構成では、モデル内部の表現が多モーダル前提で形成されることを期待し、対象分野の広がりと表現の多様性を重視した対データを計画する。どちらの構成でも、データの来歴、分割方法、再現性、秘匿要件を記録し、評価セットは学習と独立させたうえで、照応水準ごとに目的変数と合否基準を明瞭にする。

評価用データは、照応の水準別に意図された誤りを検出できるよう設計する。Level 1 では対象の有無に対する偽陰性と偽陽性を測り、Level 2 では属性の取り違えを顕在化させるために似通った対象や紛らわしい属性の対を含める。Level 3 では関係の向きや相対位置を取り違えやすい構図を用い、正例と近傍の負例をセットで配置する。Level 4 では抽象判断の成否を直接問う試験と、具象所見の列挙から抽象的結論へ到達できるかを問う試験を分ける。補助的には、表現空間の対応付けの状態を早期に把握するために、画像からテキスト、およびテキストから画像の検索整合のような補助評価を併用し、結果を照応の合否と併せて解釈する。

実務上は、外部調達の訓練済みモデルを用いる前提が強いため、事業者が設計に直接関与できるのは主に評価データと周辺ソフトウェアに供する参照データである。したがって、用途に固有の語彙、禁止事項や保留条件、参照情報の出典を明記したコーパスを整備し、プロンプト構築や RAG や出力フィルタで再利用できる形で管理する。これにより、モデル内部への介入が難しい状況でも、システムとしての外部品質を達成するための手段が確保される。

5.6 モデルの品質マネジメント

本節は、利用時品質から割り付けた外部品質目標を達成するために、モデルに関わる内部品質をどのように設計し評価し改善するかを述べる。まず、マルチモーダル AI に特有の前提として、画像とテキストの照応性能は、モデル内部の表現空間の対応付けに強く依存するが、最終的な合否はあくまで外部品質としての照応に対して判定することを原則とする。ここでいう内部品質には、問題領域分析とデータ設計、モデルの正確性

と安定性、プログラムの信頼性、運用時の維持性が含まれ、これは既存の機械学習品質マネジメントの枠組みに準拠して扱う。

Modular 構成では、写像層や場合によっては画像エンコーダの追加学習により表現空間の対応付けを段階的に改善できる一方、変更は属性の帰属や事象間関係といった Level 2 や Level 3 の照応性能に直結しやすい。そのため、受入れ評価と変更管理では、表現空間の対応付けの状態を把握する補助評価（画像からテキストおよびテキストから画像の検索整合など）と、照応を直接測る本評価（Level 1〜Level 3 の有無・属性・関係の一致、必要に応じて文や文節の根拠確認）を併用し、変更単位で回帰試験を設計する。Native 構成では、機能追加や挙動修正が全体再学習に波及しやすいため、モデル内部への介入は限定的であることを前提に、導入前の受入れ評価と運用中の監視の比重を高め、合否判定は照応評価を中心に組み立てる。Hybrid 構成では、改修可能な境界と影響範囲を事前に規定し、改修単位ごとの回帰試験範囲を固定する。

評価の設計では、外部品質の合否を最終判定としつつ、内部メカニズムの把握に有用な補助指標を明確に位置づける。具体的には、表現空間の対応付けに関する補助評価を早期の品質ゲートとして用い、合格後に照応評価の本試験へ移行する二段構成とする。照応評価は、Level 1〜Level 3 に加えて、抽象的判断を含む場合には Level 4 を必要最小限で適用し、抽象を具象の組合せへ分解する設計によって代替達成を図る。これにより、内部品質の改善が外部品質に与える効果を段階的に確認しながら、過剰な再学習や不要な仕様変更を抑制できる。

受入れ評価と運用においては、モデル単体の合否だけでなく、周辺ソフトウェアと組み合わせたシステムとしての合否を重視する。とくに、プロンプト構築や RAG や出力フィルタは外部品質達成の主要手段であり、モデル側の変更が困難な場合でも、これらの設計と運用によって合否を満たす経路を確保する。運用開始後は、入力分布の変化に対する監視、誤りの傾向分析、再評価のトリガ条件、版管理とロールバック手順をあらかじめ定め、評価結果と変更履歴、根拠や出典、モデルのバージョンを監査可能な形で記録する。

最後に、調達と選定に際しては、表現空間の対応付けや写像に関わる設計情報と学習に用いたデータの性質、損失設計、補助評価と照応評価の結果などの開示を重視し、開示内容に基づいて改善計画と回帰試験の手順を先に確定する。これにより、モデルに対する関与可能性の制約を踏まえつつ、外部品質の達成と継続的な維持を実務的に担保できる。

5.7 周辺ソフトウェア

周辺ソフトウェアは、マルチモーダル AI における外部品質を実運用で達成するための主要な手段である。訓練済みモデルの調達や改修に制約がある前提を踏まえ、システム設計の中で周辺ソフトウェアを品質担保の手段として計画する。その役割は、入力の正規化と安全確認、プロンプトの構築、RAG の実行、出力フィルタ、ポリシー適合の確認、確信度に応じた保留や人へのエスカレーション、外部ツールの呼び出しなどであり、いずれも利用時品質から割り付けた外部品質目標にひも付けて設計する。

基本設計では、周辺ソフトウェアの各機能を外部品質の到達基準と対応づけ、インタフェースと証跡を明示する。証跡には、入力資産の識別子、前処理の手順、プロンプトとそのバージョン、RAG で参照した出典、出力フィルタで適用した規則、モデルのバージョン、最終判断の根拠を含める。これにより、合否判定の再現性と監査可能性を確保し、モデル内部に介入できない場合でもシステムとしての品質を説明可能にする。

運用設計では、推論前後のゲートを段階化し、前段で入力の安全と形式の健全性を確認し、中段でプロンプト構築や検索拡張を適用し、後段で出力検査とポリシー適合の確認を行う。確信度のしきい値を設定し、灰色領域は人に委ねる。評価は、合否に直結する照応評価に加え、トリアージの有効性や審査負荷の削減といった業務指標を併せて観測し、影響が大きい変更はシャドー運用や段階導入で安全に展開する。

マルチモーダル固有の留意点として、画像側の前処理と安全確認、画像とテキストの整合点検、出典と根拠の提示を体系化する。入力画像については、画質と形式の確認、メタデータの抽出、必要に応じた文字認識を実施し、テキスト側の主張と一致しているかをルールベースで点検する。RAG は権威のある出典に限定し、提示時には該当箇所の根拠を示す。これらの処理は表現空間の対応付けの設計とは独立に機能し、照応性能を外側から補強する。

保守とガバナンスでは、規則群と辞書、プロンプト様式、検索拡張の索引、出力検査の基準を構成管理し、改定時には回帰試験とロールバック計画を伴わせる。運用では、入力分布の変化や誤判定の傾向を監視し、再評価と改定のトリガを明文化する。調達面では、モデルの素性が開示されていることを優先しつつ、周辺ソフトウェア側の設計と証跡によって最終的な合否をシステムとして保証する。

5.8 運用・保守・継続的改善

運用・保守・継続的改善は、AI システムとしての利用時品質を維持しながら外部品質を安定的に達成するための実務であり、計画、導入、監視、変更管理、再評価の各段階を一体に設計する。基本方針は、最終的な合否判定を照応などの外部品質に置きつつ、内部品質の評価結果をエビデンスとして連結し、計画的な再評価と改善により品質を維持することである。運用設計は、事前に定めた利用シナリオと合否基準に基づき、手順、責任、判断権限、監査に耐える証跡の最小集合を明確にする。

導入前の準備では、実運用データを用いたシャドー運用を実施し、照応評価を中心とする本評価と、表現空間の対応付けや写像に関する補助評価の双方で基準に到達していることを確認する。その後は、限定的なユーザー群またはトラフィックを対象とした段階導入を行い、サービス影響を最小化しながら想定外の失敗モードを抽出し、基準への適合を再確認する。これらの過程では、投入前ゲート、段階導入ゲート、全面展開ゲートという三段階の承認点を設け、各段階で必要な評価と承認権限を整理する。

運用監視では、照応性能を水準別（Level 1, 2, 3, 必要に応じて Level 4）に計測し、偽陰性と偽陽性、属性の取り違え、関係の誤認、抽象判断の不整合といった誤りの内訳を継続的に記録する。マルチモーダル特有の観点として、入力分布の変化（画像のカテゴリ構成、撮影条件、圧縮や解像度、テキストの語彙や文体）と、その変化が外部品質に与える影響を監視し、基準を超える偏りや性能劣化を検知した場合は自動で再評価プロセスを起動できるようにする。業務指標としては、審査負荷、Human in the loop の介入率、エスカレーション遅延などを併せて追跡し、品質と運用効率を同時に最適化する。

変更管理では、版管理、ロールバック、回帰試験の範囲を事前に定義する。Modular 構成の変更（写像層や場合により画像エンコーダの再学習）は局所的に見えても照応性能、とくに Level 2 と Level 3 に直結しやすいため、変更単位で補助評価と照応評価を併用する回帰試験を設計する。Native 構成では変更の影響範囲が広くなるため、受入れ評価と段階導入の重みを高め、A/B 試験やカナリアリリースで影響を限定しながら確証を得る。いずれの構成でも、変更理由、対象、期待効果、評価結果、承認記録、展開計画、ロールバック条件を一件ごとに証跡化する。

継続的改善では、運用データと評価結果を用いて誤りの再発を抑止する。評価に用いたデータセットと基準は定期点検し、分布や難易度の偏りを是正する。モデル側の改修が難しい場合は、周辺ソフトウェアの改定（プロンプト様式、RAG の索引、出力フィル

タ規則、禁止や保留の条件）を優先して調整し、外部品質の合否に効く箇所から改善を重ねる。改善の都度、合否基準、評価手順、運用手順に反映し、品質の計画、実行、点検、改善の循環を保つ。

監査とガバナンスでは、入力識別子、前処理、プロンプトまたは入力構造、RAG の出典、モデルのバージョン、確信度、出力とその検査結果、最終判断の理由、運用者の介入記録を、再現可能で秘匿要件に適合する形で保存する。調達先や委託先が関与する場合は、開示情報の更新時期、再評価の義務、重大インシデント時の報告と共同調査の手順を契約上も明確にし、再評価や入替の判断に必要な情報が継続的に入手できる体制を確保する。

5.9 評価フレームワーク

評価フレームワークは、AI システムの利用時品質を最終判定対象としつつ、機械学習要素と周辺ソフトウェアの外部品質を合否に結び付け、内部品質の評価結果をエビデンスとして連結する枠組みである。最終判定は照応（Level 1, 2, 3, 必要に応じて Level 4）に対して行い、表現空間の対応付けに関する評価は、照応の達成可能性を早期に見極める補助的手段として位置づける。これにより、内部メカニズムの改善と外部品質の到達を因果的に対応づけたうえで、計画的な再評価と変更管理に移行できる。

評価は、システム全体と構成要素の二層で設計する。システム層では、想定する利用シナリオに対する合否を決め、要素層では、機械学習要素と周辺ソフトウェアの外部品質がシステム合否に寄与するかを検証する。導入プロセスでは、受入れ評価、段階導入、全面展開の各ゲートで必要な証跡と合否基準を定め、運用中は定期再評価と回帰試験を変更単位で実施する。

照応評価は、Level 1 では対象の有無、Level 2 では属性の帰属、Level 3 では事象間の関係を、それぞれ画像とテキストの対応として判定する。Level 4 は抽象的判断が業務上不可欠な場合に限って最小限で適用し、それ以外では抽象を具象の組合せに分解して検証可能性を確保する。文または文節の単位で、画像に根拠が存在する主張と存在しない主張を区別できるように評価手順を整備し、根拠に基づく生成や説明の一貫性も観測する。

表現空間の対応付けに関する評価は、照応に先行する品質ゲートとして用いる。画像からテキスト、テキストから画像の検索整合（R@K など）を補助指標とし、Modular 構

成での写像層や画像エンコーダの変更が、照応の上位課題（とくに Level 2 と Level 3）へ波及し得るかを早期に把握する。補助指標は合否の参考値にとどめ、最終判定は照応評価で行う。

評価用データは、学習と独立させ、来歴、分割方法、再現性、秘匿要件を記録する。マルチモーダルの特性を踏まえ、共通領域と固有領域のタグ付けを行い、照応で扱うべき範囲と、モダリティ固有の判定に委ねる範囲を分離する。試験計画では、代表性と難例の両方を確保し、取り違えや関係誤認を顕在化させる近傍負例を適切に配置する。

合否基準は、利用時品質の終点に合致する外部品質指標に対して数値化し、必要に応じて信頼区間や許容誤差を併記する。安全や公平性、プライバシー、セキュリティなど、機械学習品質マネジメントで定義される外部品質特性を評価観点に含め、業務上の重要度に応じて重み付けを行う。

運用評価では、照応の主要指標に加えて、トリアージや審査負荷の削減などの業務指標を追跡する。RAG の到達率と出典の信頼度、出力フィルタの適合率と再現率、確信度しきい値の調整による灰色領域の人手審査率、段階導入時の逸脱検知率などを併記し、モデルの変更が難しい場合でも周辺ソフトウェアの改定で外部品質を満たせるかを検証する。

変更管理に伴う回帰試験は、写像層や画像エンコーダの変更（Modular）、全体再学習または入替（Native）、限定的微調整（Hybrid）など、アーキテクチャごとの変更単位に対応させる。補助評価で表現空間の対応付けや写像の変化を確認した後、照応評価で外部品質の合否を判定し、合格時のみ次工程へ進む。

評価の証跡は、入力識別子、前処理の手順、プロンプトや入力構造、RAG で参照した出典、モデルのバージョン、確信度、出力と出力フィルタの判定、最終判断の理由、運用者の介入記録を、再現可能な形で保存する。調達先や委託先が関与する場合は、開示情報の更新時期、再評価の義務、重大インシデント時の報告と共同調査の手順を明記し、評価で必要となる情報が継続的に入手できる体制を維持する。

5.10 ガバナンスと調達

ガバナンスと調達は、AI システムとしての利用時品質を持続的に達成するための統治と契約の実務であり、責任分界、方針、証跡、開示要求、受入れ評価、再評価の義務を体系化する。まず、組織として、AI システム、機械学習要素、周辺ソフトウェアの責

務と権限を明記し、合否判定の対象を外部品質に置いたうえで、内部品質の評価結果をその根拠として連結する。評価と監査に必要な最小限の証跡は、入力識別子、前処理の手順、プロンプトまたは入力構造、RAG で参照した出典、モデルのバージョン、確信度、出力と出力フィルタの判定、最終判断の理由、運用者の介入記録とし、再現性と秘匿要件を同時に満たす保管計画を定める。調達では、外部調達を前提として、開示が十分なモデルと構成要素を優先する。とくに、アーキテクチャ類型、表現空間の対応付けや写像の設計（写像層の構造、画像エンコーダの種類と凍結範囲）、学習に用いたデータの性質と分割、損失設計、受入れ評価に資する評価結果、既知の失敗モード、推奨用途と非推奨用途の開示を重視し、入手した開示に基づいて受入れ評価、回帰試験、段階導入の手順をあらかじめ確定する。契約では、開示情報の更新時期、再評価の義務、重大インシデント時の報告と共同調査の手順、モデルのバージョン更新や置換の通知と影響評価、修補や改修の責任分界を条項として明文化する。調達後の統治では、方針、標準、テンプレート、ガイドを整備し、合否基準、承認ゲート、ロールバック条件、回帰試験の範囲を一貫した書式で管理し、変更単位での影響範囲と記録の保全を徹底する。規制や業界指針の更新が品質に与える影響を定期的に点検し、必要に応じて再評価と契約更新を行う。これらの運用は、機械学習品質マネジメントで確立された利用時品質、外部品質、内部品質の連鎖を基盤に据え、生成やマルチモーダルの特性に応じた開示と受入れ評価を重ねることで、外部調達の制約下でもガバナンスの実効性を確保する。

5.11 文書化と導入準備

文書化と導入準備は、AI システムの利用時品質を最終判定の対象とし、そこから要素の外部品質目標と内部品質の評価項目へと連結する証跡の連鎖を確立する実務である。まず、利用時品質、外部品質、内部品質の関係を章内の用語に従って定義し、合否判定は照応に対して行い、表現空間の対応付けや写像に関する評価は合否の根拠を補強する位置づけとして証跡に組み込む。これにより、評価と運用で得られる事実を、導入前の受入れ評価、段階導入、全面展開、定期再評価の各ゲートにおける承認と説明責任に結び付けられるようにする。

証跡は最小集合を定義したうえで拡張可能に設計し、再現性と秘匿要件を同時に満たす保管方針を定める。最小集合には次を含める。

- 入力識別子と前処理手順（画像の品質確認、メタデータ抽出など）の記録

- プロンプトとその版、適用時の入力構造
- RAG で参照した出典と該当箇所の情報
- モデルのバージョン、推論設定、しきい値などの運用パラメータ
- 出力と出力フィルタの判定結果、適用規則の版
- 照応評価と補助評価の結果、合否基準に対する判定
- 最終判断の理由と運用者の介入記録、承認・却下の決裁履歴

この最小集合は、学習データへの関与が限定される状況でも、システムとしての品質達成と監査可能性を維持するための根幹となる。

導入準備では、試験計画、手順書、責任と権限、承認ゲート、ロールバック条件をひとつの書式で統一し、段階導入の対象範囲、逸脱検知基準、エスカレーション手順、人の介入条件を実行可能な形で文書化する。受入れ評価の完了報告、段階導入計画、回帰試験計画、運用手順、インシデント対応計画、容量と性能の想定、教育と訓練計画をそろえ、想定する利用シナリオに対して外部品質の合否を判定できる状態を確認する。

変更管理に関しては、要求（変更理由、期待効果、影響範囲）から評価（補助評価と照応評価の手順、近傍負例の再現方法）と承認（責任者、承認条件）を経て展開（段階導入、カナリア、ロールバック）に至る一連の記録様式を定義し、モデルのバージョンと周辺ソフトウェアの版を一体で管理する。Modular では写像層や場合によっては画像エンコーダの変更が外部品質、特に Level 2 と Level 3 に直結するため、変更単位で回帰試験の範囲と合否基準を先に固定してから実施する。Native では全体再学習に波及しやすい前提を織り込み、受入れ評価と段階導入の重みを高めて展開判断を行う。

調達とガバナンスに連動する文書として、ベンダから入手した開示情報（アーキテクチャ類型、表現空間の対応付けや写像の設計、学習に用いたデータの性質と分割、損失設計、既知の失敗モード、推奨用途と非推奨用途、評価結果）を整備し、更新時期、再評価の義務、重大インシデント時の報告と共同調査の手順を契約に反映する。開示内容は受入れ評価と回帰試験の手順に直接結び付け、段階導入と全面展開の承認に必要な根拠として保全する。

運用に向けた準備として、評価用データと業務上の指標を保守対象に含め、分布や難易度の偏りを定期点検し、再評価のトリガを明文化する。人が関与する判断（Human in the loop）のための訓練資料、判断基準、エスカレーションの連絡体制を整備し、RAG の出典選択基準や出力フィルタの規則群の改定手順を他の改修と同等に版管理する。これらを通じて、モデル内部への介入が制約される場合でも、周辺ソフトウェアの改定と評価の再実施を通じて、外部品質の合否を安定して満たす体制を維持する。

参考文献

[MLQM 2023] 機械学習品質マネジメント検討委員会、機械学習品質マネジメントガイドライン第 4 版、2023 https://www.digiarc.aist.go.jp/publication/aiqm/guideline-rev4.html

[MLQM 2025] 機械学習品質マネジメント検討委員会、生成 AI 品質マネジメントガイドライン第 1 版、2025 https://www.digiarc.aist.go.jp/publication/aiqm/genaiqm-guidelines-v1.html

[Shucor 2025] Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, Alaaeldin El-Nouby, *Scaling Laws for Native Multimodal Models*, 2025 arXiv:2504.07951 [cs.CV]

付録1 参考：アプリケーション事例の検討

A1.1　はじめに

本章は、前章で示した品質マネジメントの手法を具体の場面に移し替える道筋を示すことを目的とする。要求の定義、設計上の要点、データと評価、運用とガバナンスという枠組みを、実務の流れに沿って検証できるようにし、事例ごとの差異はあっても考え方の移植性が保てることを確認する。

取り上げる三つの事例は、いずれも架空のものであるが、マルチモーダルの難易度と意思決定の性質が異なる代表例である。EC 商品説明生成は、画像で確認できる具体的事実を正確に言語化し、画像から確認できない事項の断定を避けるという要件が中心となる、比較的素朴で具体的な事例である。ここでは、人が最終確認と修正を行う運用に適合する下書きの生成が狙いとなる。インフラ点検支援は、想定する利用者が新任点検員か主任技術者かによって必要な照応の深さが変わり、ときに抽象的判断を含む Level 4 の扱いが鍵になる。 SNS コンテンツモデレーションは、単一モダリティで完結する判定、モダリティ間の照応が必要な判定、文脈に依存して適法に転ぶ判定という層を区別し、AI と人の役割分担を前提に運用設計を行う複雑な事例である。

A1.2　EC 商品説明生成

商品画像で確認できる具体的な事実に基づいて説明文の草案を作成し、画像からは確認できない事項の断定を避けるという位置づけの事例である。生成物は完成原稿ではなく編集しやすい出発点として設計し、画像と文章の一致を品質の第一基準とする。

A1.2.1　背景と目的

EC サイトでは、商品画像と基本情報を基に、商品説明文の草案を生成し、人が最終確認と修正を行って公開する運用が一般的である。これは AI が最終的な文章を確定する役割ではなく、執筆作業の出発点を提供する役割であることを意味する。したがって、品質上の主眼は安全性の厳格な統制よりも、画像内容と一致した正確さと編集のしやすさに置かれる。

A1.2.2　スコープと役割分担

本事例が対象とする AI の出力は、公開用の完成原稿ではなく、編集可能な草案である。人は草案を点検し、必要な情報を補い、表現を整え、最終承認を行う。AI は画像から確実に読み取れる事実の抽出と、それらの事実に基づく簡潔な文の提示を担う。

A1.2.3　要求する照応レベルと領域の扱い

本事例の中心は、商品カテゴリや色や形状の正しい同定など、画像に現れる事実と説明文の一致であり、これは主として Level 1 と Level 2 に相当する。素材や機能など、画像だけでは確定できない事項は、画像の固有領域外にある情報として扱い、人の確認を前提に運用する。

A1.2.4　設計上の要点

AI の出力は二段で構成する。第一に、画像に根拠を持つ可視属性を箇条書きで提示する。第二に、可視属性に基づく短い説明文の草案を提示する。画像から直接は読み取れない情報を断定的に記述することは避け、未確定事項は人が追記できるように残す。説明文は過度に完成形の広告文に寄せず、編集容易性を優先する。

A1.2.5　データセットとアノテーション

評価と学習に用いるデータは、画像と属性の正解対応を明確に含むことが望ましい。文単位の評価を可能にするため、参照説明文の各文に対して、画像に根拠があるかどうかの区別を付与する。誤ったカテゴリ付与や色の取り違えなど、頻出の失敗例を収集し、継続的に補強する。

A1.2.6　評価設計

整合性の評価では、カテゴリの一致率や色の一致率や形状の一致率を用いる。説明文の評価では、画像に根拠を持つ文の割合を測定し、画像からは確定できない事項の断定的記述がどの程度抑制されているかを確認する。運用効果の評価では、商品の一件当たり編集時間や修正量を基準と比較し、草案が編集作業をどの程度短縮したかを測定する。

A1.2.7　運用とガバナンス

公開前の最終確認と承認は人が行う。公開後は、返品や苦情の理由と生成された説明文を突き合わせ、属性誤りや誤認誘発の指標を継続的に監視する。修正済みの最終文は、品質向上に資する範囲で学習資料に反映させる。ただし、画像からは確定できない事項の断定的表現は学習資料から除外する。

A1.2.8　失敗モードと対策

- カテゴリの取り違え
- 色や形状の取り違え
- 画像から確定できない事項の断定的記述
- 過度に完成形の広告文への偏り

これらに対しては、カテゴリを先に確定する手順、属性ごとの自信度に基づく抑制、未確定事項の保留運用、草案の書式統一による編集容易性の確保を組み合わせて対処する。

A1.2.9　派生と拡張

商品情報管理システムや在庫データベースに保存されている確かな出典を参照できる場合は、画像根拠のない事項を出典付きで候補提示し、人が採否を決める設計に拡張できる。商品種別ごとに属性スロットと文型を標準化し、再利用性と一貫性を高めることも有効である。

振り返り

下書き作成の効率化を目的関数として明示し、画像に根拠がある事実を先に確定して優先的に文章化し、根拠の無い要素は保留するという順序を徹底する。素材や機能などの非可視情報は人が補う対象であり、断定は行わない方針を規範として宣言する。商品カテゴリと属性語彙は社内の分類体系と整合させ、評価指標として属性一致率と根拠付き文の割合を用いることで、学習データと業務運用の整合を保つ。

A1.3　インフラ点検支援

本事例では、想定する利用者が新任点検員か主任技術者かによって必要となる照応の深さが変わり、特に抽象的判断を含むレベルの扱いが状況に応じて変化する。設計と評価は、利用者像を先に固定し、Level 1、Level 2、Level 3 と順に土台を築いた上で、必要に応じて抽象判断を最小限で加える流れで組み立てる。

A1.3.1　背景と目的

インフラ点検支援では、橋梁やトンネルなどの構造物の画像や動画、既往の図面や点検記録などの視覚情報を入力として、劣化所見や補修判断の材料を文章として提示する。新任点検員の支援では典型所見の見落とし防止と記録作成の平準化が中心となり、主任技術者の支援では複数所見の組み合わせや経年変化の読み取り、優先度付けの判断を支えることが重要になる。これらの違いを明確にしないまま設計を進めると、過剰に高度な機能を組み込んで現場運用を重くするか、または必要な能力が不足して維持管理の意思決定に適合しない恐れがある。

A1.3.2　スコープと役割分担

本事例は、現地点検の前後に参照できる機械生成の文章と図化を提示し、人が内容を確認し、必要に応じて追記や修正を行い、最終判断を行う運用を前提にする。新任点検

員の支援では、画像に顕在化している所見の有無と位置と名称の提示を中心に据える。主任技術者の支援では、単一所見では確定しにくい劣化の組み合わせや進展の可能性の示唆、対策の選択肢の提示、確信度に応じた慎重な言い回しを重視する。いずれの場合も、最終的な解釈と意思決定は人が担い、機械は判断材料の抽出と表現を担う。

A1.3.3　要求する照応レベルと領域の扱い

新任点検員の支援では、画像に存在する事象の有無や属性の帰属に関する整合を重視し、主として Level 1 と Level 2 を達成する。主任技術者の支援では、複数所見の関連や構造的文脈の解釈が重要となるため、Level 2 と Level 3 を基盤にしつつ、抽象的な判断の表現にも配慮する。たとえば、ひび割れや剥離などの所見から補修優先度の示唆に至る表現は、単純な視覚特徴の検出に還元できないため、所見の組み合わせと状況解釈を併せて扱う前提で設計する。共通領域は照応の評価対象とし、素材や設計背景のように画像から直接は確定できない事項は固有領域として扱い、照応とは分けて評価する。

A1.3.4　設計上の要点

新任点検員の支援では、代表的所見を名称と位置の情報とともに簡潔に列挙し、必要に応じて画像上の根拠提示を伴う。主任技術者の支援では、弱い所見の組み合わせや経年比較の観点を提示し、確信度に応じた自然な言い回しで文章化し、補修や追加点検の選択肢を示す。いずれの設計でも、画像に根拠がない断定を避け、根拠と意見を分けて提示する構成とする。図面や過去の点検記録、施工情報などの参照が必要な場合は RAG を用いて出典を明示し、出力フィルタで不適切な断定を抑制する。

A1.3.5　データセットとアノテーション

新任点検員の支援に必要な学習と評価では、所見の有無に加えて、位置や寸法や名称といった属性を明確にした正解注釈が有用である。主任技術者の支援では、複数所見の複合、進展の有無に関する参照、補修要否や優先度の判断に関する参照データが望ましい。生成される文章に対しては、文ごとに画像に根拠があるかどうかを識別できる注釈を付与し、根拠に基づく表現の一貫性を評価できるようにする。時系列比較を前提にする場合は、同一部位の対応関係を明確にし、比較に用いた根拠の位置合わせを正解として記録する。

A1.3.6　評価設計

新任点検員の支援では、所見の有無と位置に関する一致や名称の一致を中心に評価する。主任技術者の支援では、確信度を伴う表現の適切さ、提案された補修や追加点検の

妥当性、提示された根拠の網羅性を評価し、点検時間や点検者間の一致の変化も併せて測定する。いずれの場合も、画像と文章の整合が損なわれていないかを文単位で確認し、過度の断定や不適切な省略がないかを点検する。経年比較を行う場合は、対応付けの正確さと差分の説明の妥当性を評価に含める。

A1.3.7　運用とガバナンス

導入時には、実運用の前段で機械生成の文章と図化を人が参照する試行運用を実施し、誤警報や見落としの傾向を把握する。運用開始後は、入力画像や動画、生成された文章、最終判断を体系的に記録し、再評価のたびに、整合のレベルや不確実性表現の適切さ、対策提案の妥当性を確認する。組織としては、どの範囲を機械の提示に委ね、どの範囲を人の判断に委ねるかを明文化し、変更が生じた場合に再評価を行う手順を整える。RAG の出典と索引の管理、出力フィルタの規則とモデルのバージョンの管理を含め、監査に耐える証跡を保持する。

A1.3.8　失敗モードと対策

- 典型所見の見落とし
- 誤警報の多発
- 位置情報や対応付けの不整合
- 画像に根拠のない断定的表現
- 補修や追加点検の過剰提案または過小提案

これらに対しては、所見の検出と位置付けを先に提示する構成、画像根拠の提示、確信度に応じた言い回しの標準化、経年比較の手順化、誤り事例の継続的な学習を組み合わせて対処する。

A1.3.9　派生と拡張

構造種別や点検基準に合わせて、重要所見の優先度や確信度のしきい値を調整する。時系列の画像やセンサ記録の比較を前提に、前回点検との差異を自動で要約する設計に拡張する。組織内で合意された語彙と表現範囲に合わせて、対策提案の標準文型を整備する。

振り返り

新任点検員支援を所見の有無や属性の提示に重点を置く設計とし、主任技術者支援を複数所見の組み合わせや対策の示唆に重点を置く設計とする。照応レベルは、新任点検員支援で Level 1 と Level 2、主任技術者支援で Level 2 と Level 3 を基盤にし、案件と目

的に応じて抽象的判断の適用を必要最小限で追加する。不確実性の表現と根拠提示を評価と運用に直結させ、導入前の試行運用と合議による検証を前提とすることで、現場の安全と有用性を両立させる。

A1.4　SNS コンテンツモデレーション

本事例では、単一モダリティで完結する判定、モダリティ間の照応が必要な判定、文脈を考慮して適法に転ぶ判定という階層を前提に、判定設計と運用を段階化する。AI は大量投稿から確認が必要な案件を抽出し、人が確定判断を行う役割分担を設計段階で明確に定義する。

A1.4.1　背景と目的

SNS やユーザー生成コンテンツを扱うプラットフォームでは、投稿が利用規約や社会的規範に違反していないかを迅速かつ大規模に確認する必要がある。マルチモーダル AI を用いるコンテンツモデレーションでは、画像だけで判断できる投稿、テキストだけで判断できる投稿、画像とテキストを組み合わせて初めて意味が確定する投稿が混在し、判断の焦点は画像とテキストの意味的な照応の理解に移る。

典型的な判定例を次に列挙する。

単一モダリティで判定可能な投稿の例

- 明白な暴力行為を映す静止画像。
- 差別語や暴力の直接的表現を含むテキスト単体。

画像とテキストの照応が必要な投稿の例

- 家事をする女性の画像と、各自の役割を強調するテキスト。
- 映画館の画像とビデオ撮影の方法を問うテキスト。

文脈により適法に転ぶ可能性のある投稿の例

- 報道目的で掲出された暴力場面の写真。
- 医療教育のための手術場面の画像や解説。

ここでの AI の役割は、人の判断を完全に置き換えることではなく、人による確認が必要な投稿を効率的に抽出することであり、設計段階で AI と人の役割の境界を明確に定義することが重要となる。

A1.4.2　スコープと役割分担

本事例は、違反の疑いがある投稿を AI が抽出し、人が最終判断を行う運用を前提にする。明白な暴力画像や明白なヘイトスピーチを含むテキストなど、単一のモダリティで判断が完結する投稿は自動判定の対象とする。画像とテキストの照応関係を理解しなければ意味が確定しない投稿は、照応判定に特化した処理段階に回送し、必要に応じて人に委ねる。画像が一見不適切であっても、報道や批評や教育目的などの文脈で適法に転ぶ可能性がある投稿は、人の判断を前提に運用する。

A1.4.3　要求する照応レベルと領域の扱い

本事例では、単一モダリティで完結する判定には Level 1 と Level 2 を適用し、対象の有無と属性の帰属を確実に行う。画像とテキストの組み合わせで意味が成立する投稿には Level 2 と Level 3 を適用し、対象と属性の整合に加えて関係の一致を確認する。文脈で適法に転ぶ可能性がある投稿には、提示目的や文脈の把握を前提に、照応の評価に加えて適法性の判断に必要な情報の抽出と提示を行う。

A1.4.4　設計上の要点

段階的な判定パイプラインを設計し、まず画像のみの判定、次にテキストのみの判定を行い、その後に画像とテキストの照応判定を行う。照応判定で疑義が残る投稿は、報道や批評や教育目的などの適法文脈の可能性を示す手掛かりの抽出を行い、人の確認に回す。自動で措置する範囲と人に委ねる範囲は、違反カテゴリと危険度と信頼度の組み合わせで定義し、危険度が高い場合は速やかな暫定措置と後審の組み合わせで運用する。

A1.4.5　データセットとアノテーション

評価用および学習用のデータは、単一モダリティで判定が完結するセット、画像とテキストの照応が必須のセット、文脈により適法に転ぶセットに分けて準備する。同一画像に対してキャプションのみを変更して適法と違反が反転するペアを含め、照応と文脈の影響を分離して検証できる構成を整える。画像とテキストの整合に関しては、対象と属性と関係の一致と不一致を注記し、文脈依存の可否を明確にする。

A1.4.6　評価設計

検出性能の評価では、カテゴリ別の適合率と再現率を測定し、危険度の高いカテゴリに関しては見逃し率の上限を設定する。照応の評価では、対象と属性と関係の整合を指標化し、画像と無関係なテキストやテキストと無関係な画像に引きずられる誤照応の発生率を監視する。文脈に関する評価では、報道や批評や教育目的などの適法文脈を考慮

して過剰なブロックを抑制できた割合を測定する。運用の評価では、人へのエスカレーションの適切さ、審査者の負荷の削減、意思決定までの時間の短縮を測定し、トリアージとしての有効性を確認する。

A1.4.7　運用とガバナンス

運用開始前に、実流量に近い条件で機械の提案のみを記録して人の判定と突き合わせる試行を行い、過剰ブロックと見逃しの傾向を把握する。運用開始後は、入力と出力と信頼度と最終措置と逆転事例の理由を監査可能な形で保存し、違反カテゴリと運用ルールの改定に合わせて再評価を行う。AI が自動で措置する範囲と人に委ねる範囲、根拠提示の要件は文書で明確に定義し、地域や言語や社会規範の差異に応じて版管理を行う。

A1.4.8　失敗モードと対策

- 過剰ブロック
- 見逃しの増加
- 照応の破綻
- 皮肉やミームへの脆弱性
- 文脈の見落とし

これらに対しては、適法文脈の可能性を示す手掛かりの提示、誤判定の再学習への反映、照応必須の評価セットの継続更新、危険度と信頼度に応じた二段階の措置を組み合わせて対処する。

A1.4.9　派生と拡張

地域や言語ごとのポリシー差分に対応するため、違反カテゴリ表の地域別プロファイルを整備し、運用上のルールセットを切り替える設計に拡張する。ハラスメントや違法性の新たな領域に対応する場合は、照応必須のデータと文脈依存のデータを優先的に追加し、試行運用で過剰ブロックと見逃しのバランスを調整する。

振り返り

三層構造を判定設計の座標系として先に定義し、単一モダリティ判定、照応判定、文脈判定を段階的に適用する。AI はふるい分けを担い、人が確定判断を担うという役割分担を宣言し、危険度が高い場合は暫定措置と後審を組み合わせる。報道や批評や教育目的などの適法文脈を見落とさない方針を位置づけとして明記し、過剰ブロックの抑制を評価項目に含める。違反カテゴリ表と運用ポリシーは版管理の対象とし、改定時には再評価を必ず行う。

機械学習品質マネジメント検討委員会メンバーリスト

2025 年度 （委員長、副委員長を除き、五十音順、敬称略）

中島　震(委員長)	産業技術総合研究所
妹尾　義樹(副委員長)	産業技術総合研究所
江川　尚志	産業技術総合研究所
越前　功	国立情報学研究所
大岩　寛	産業技術総合研究所
岡本　球夫	SHIN-JIGEN
小川　秀人	日立製作所
桑島　洋	デンソー
小林　健一	富士通
小宮山　英明	コニカミノルタ
新原　敦介	日立製作所
鈴木　知道	東京理科大学
高田　眞吾	慶應義塾大学
徳　隆弘	三菱電機
中島　裕生	NPO MedXML コンソーシアム
難波　孝彰	産業技術総合研究所
浜谷　千波	アドソル日進
福島　真太朗	トヨタ自動車
三島　浩一	DynamicRiskM
宗像　一樹	富士通
安井　裕司	本田技術研究所
山田　敦	日本 IBM
若松　直哉	日本電気

ガイドライン詳細検討タスクフォース メンバーリスト

2025 年度　（五十音順、敬称略）

磯部　祥尚	産業技術総合研究所
江川　尚志	産業技術総合研究所
大岩　寛	産業技術総合研究所
岡本　球夫	SHIN-JIGEN
小川　秀人	日立製作所
川本　裕輔	産業技術総合研究所
北村　崇師	産業技術総合研究所
北村　弘	IPA/AI セーフティ・インスティテュート
桑島　洋	デンソー
小西　弘一	産業技術総合研究所
小林　健一	富士通
小宮山　英明	コニカミノルタ
新原　敦介	日立製作所
妹尾　義樹	産業技術総合研究所
田中　哲	産業技術総合研究所
田部　尚志	日本電気
中島　震	産業技術総合研究所
中島　裕生	NPO MedXML コンソーシアム
難波　孝彰	産業技術総合研究所
浜谷　千波	アドソル日進
林谷　昌洋	日本電気
福住　伸一	産業技術総合研究所
三宅　和公	住友電気工業
三宅　武司	サイバー創研
若松　直哉	日本電気

本プロジェクトに関係して産業技術総合研究所に特定集中研究専門員として部分的に在籍しているメンバーについては、それぞれの出身母体で記載した。