マルチモーダル LLM の画像入力タスクにおける信頼性評価指標設計手法の提案

A Design Methodology for Reliable Evaluation Metrics for Image-Input Tasks in Multimodal Large Language Models

リ ー ダ ー ：長田　章良（ヤンマーホールディングス株式会社）
研　　究　　員：菊武　裕輔（アドビ株式会社）
　　中村　将大（株式会社日立ソリューションズ・クリエイト）
　　伊藤　稜　（三菱電機ソフトウエア株式会社）
主　　査：石川　冬樹（国立情報学研究所）
副　主　査：徳本　晋　（富士通株式会社）
アドバイザー：栗田　太郎（フリー株式会社）

研究概要
近年，大規模言語モデル（LLM）の実利用が進む一方で，その出力の信頼性を利用者視点で評価する指標は十分に確立されていない．特にマルチモーダル LLM を利用した画像入力を伴うタスクでは,利用者が回答の正誤や妥当性を出力文のみから判断することが難しく，また人手評価は高コストかつ主観的ばらつきを伴うという課題がある．本研究では画像入力タスクにおける LLM の出力品質を対象に，LLM-as-a-Judge を活用した回答の信頼性評価指標を算出する枠組み（JRVF）を提案する．本提案に自信度・正確性・再現性・説明性の４観点を統合した評価指標（CARE スコア）を導入し，人手評価との比較実験を実施した．その結果，CARE スコアと人手評価との間に相関を確認し，本提案の有効性を示した．

1. はじめに

近年，大規模言語モデル（LLM）はテキストの入力・生成にとどまらず，画像などの非言語情報を入出力として扱うマルチモーダル LLM へと拡張され，実利用が進みつつある．しかし，マルチモーダル化に伴い入出力の構造や評価対象が多様化した結果，固定的なデータセットや単一の自動評価指標に基づく評価では，利用者が重視する品質を十分に捉えることが難しい．一方，人手評価は入力形式に対して柔軟なデータセットを設計できる利点を有するが，評価に要する時間的・金銭的コストが大きく，さらに評価者による主観的判断のばらつきも避けることが困難である．

加えて，実利用の観点からは，利用者が LLM の出力情報のみからその正誤を判定できるとは限らない．とりわけマルチモーダル LLM では，回答が画像内の細部や複数箇所に分散した情報，あるいは同時に入力するテキストと画像の対応関係に依存する場合が多く，利用者が入力全体を精査して真偽を確認することは大きな負担になり得る．その結果，回答文のみから「出力が入力と整合しているか」「説明の根拠が十分であるか」「見落としや誤りが含まれていないか」を適切に判断することが難しくなる．そこで，出力を利用者が採用する前段階において，その信頼度の目安を提示できれば，利用者は追加的な検証の要否をより効率的に判断できることが期待される．

そこで本研究では，①各品質観点に対する評価基準の明確化，②LLM-as-a-Judge による評価の実行，③統合スコアとして算出，までを行う信頼性評価指標算出の枠組みを提案する．具体的な題材（タスク）を用いた，同一基準に基づく人手評価との比較実験を通じて，本枠組みが人手評価を補完し,利用者への信頼性評価指標として有効であるかを検証する．

2. 研究背景

画像を含むマルチモーダル LLM の評価では，入力と出力の対応関係が複雑化するため，テキストのみを対象とした従来の評価手法をそのまま適用することは困難である．本章では，マルチモーダル LLM を対象とした既存の自動評価手法，LLM-as-a-Judge による評価，および品質観点に関する先行研究を整理する．

画像を含むタスクに対する従来の自動評価手法として，参照キャプションや正解ラベルとの一致度を基準とする指標（例：BLEU[1]など）や，画像とテキストの埋め込み類似度に基づく視覚言語モデルベースの指標（例：CLIPScore[2]など）が発展してきた．しかし，入出力が多様化する現状では，あらゆる妥当な出力を網羅する参照データを作成することは現実的ではなく，ベンチマークの整備が追い付かない問題が生じている．

このような制約を背景として，LLM-as-a-Judge 等の正解ラベルに依存しない評価手法が注目されている．LLM-as-a-Judge は，事前に定義された基準に基づき LLM が出力を評価する手法である．LLM-as-a-Judge を評価に用いるメリットとして，正解ラベル不要で任意形式の出力を柔軟に評価でき，大量出力を短時間で処理できる点が挙げられる．同様のメリットを持つ手段として，対数確率など内部確率情報を用いる評価も提案されている[3]が，多くの商用 LLM では取得が困難である．一方デメリットとして，評価の信頼性が挙げられる．先行研究[4]では，信頼性向上策として「プロンプト設計最適化」,「モデル改善」,「出力後最適化」が述べられている．特にプロンプト設計最適化は，タスクの概要や評価観点を明確に定義し，LLM が的確に解釈・運用できるような形式化が重要と述べている．つまり，「何をもって良い出力とみなすか」という評価観点の設計が依然として重要である．

LLM の品質観点については AI プロダクトの品質保証コンソーシアムである QA4AI (Quality Assurance for AI)が公開するガイドライン[5]に記載がある．LLM の品質特性の大分類として「回答性能」,「事実性・誠実性」,「倫理性・アラインメント」,「頑健性」,「AI セキュリティ」の 5 つを提示しており，各分類で具体的評価項目を整理している．本研究ではこの考え方を受け，実務的かつ再現可能な評価体系に集約して扱う．また既存研究では，評価の自動化や個別観点の定義に焦点が当てられてきたが，利用者の意思決定を支援することを目的とした,複数の品質観点を同一基準で評価･統合する試みは限定的である．

3. 品質観点の定義

本研究では,LLM-as-a-Judge を用いた LLM の品質評価観点として,「自信度」,「正確性」,「再現性」,「説明性」の 4 つを採用する．これら 4 観点は，正解ラベルを用いずに自動評価が可能であり，回答の妥当性だけでなく，一貫性や根拠提示まで含めて多面的に品質を捉えられる．さらに，QA4AI のガイドラインにおいて品質保証に関わる観点として整理されており，画像入力タスクの回答信頼性を評価する上で適切であると判断した．

一方で，同ガイドラインに示される「頑健性」や「倫理性」は，特に主要な商用モデルで既に一定程度確立しており，利用者がリアルタイムに解釈・判断が難しいことから，本研究の評価観点からは除外した．以下に，採用した 4 つの観点（大項目）の詳細を示す．

3.1 自信度 (Confidence)

出力モデル自身に自信度を出力させ，回答内容とどの程度一致しているかを評価する．例えば自信度が高いと言った場合に，実際に正解率も高い場合に最高評価となる．

3.2 正確性 (Accuracy)

出力内容がどの程度事実と合致しているかを評価する．以下の 3 つの小項目から成る．

- 情報の真実性：虚偽を含んでいるかどうか.
- 質問への関連性・適合性：質問の意図に沿った回答かどうか.
- 論理的な一貫性：根拠と結論の関係がわかりやすく，説明に整合性があるか.

3.3 再現性 (Reproducibility)

同様の質問に対して,出力が一貫しているかを評価する.以下の 3 つの小項目から成る.

・ 同一質問再現性：同じ質問を繰り返した場合に，結論や判断が一貫しているか.
・ パラフレーズ耐性：表現の異なる質問をした場合に，結論や判断が一貫しているか.
・ 根拠・判断基準の再現性：同様の質問をした場合に，判断プロセスが安定しているか.

3.4 説明性・妥当性（Explainability）

根拠や説明が結論のために必要十分なものかを評価する.以下の 3 つの小項目から成る.
・ 根拠の提示と説明性：根拠が回答に対して必要十分か.
・ 前提・適用範囲・限界の明示：前提・適用範囲・限界の説明が必要十分か.
・ 追跡可能性：説明のプロセスが明確か.

4. 提案内容

本研究では，LLM-as-a-Judge を活用して各評価観点を一つの指標に統合するための手順・枠組みとして Judge 主導型 LLM 信頼性可視化フレームワーク（Judge-driven LLM Reliability Visualization Framework，JRVF）を提案する．JRVF により算出された信頼性評価指標は，品質観点ごとに詳細な評価内容を提示した場合と比べて評価過程の詳細を把握しにくくなるという側面を有する．一方で，複数の評価結果を集約することで，利用者が評価結果を直感的に理解し，迅速に意思決定を行えるという利点がある.

JRVF は，LLM-as-a-Judge による評価結果を，品質観点の定義から重み付け統合までを含めた一連の手順として体系化し，評価観点の解釈のばらつきや主観性を構造的に整理可能とする枠組みである．従来は個別観点ごとの評価や経験的な統合が行われることが多かったが,本研究では評価構造そのものを明示化し,再現可能な統合プロセスとして提示する.

LLM 回答の実用的な信頼性評価指標として機能させるために，以下の手順で統合スコアを算出する.

① 評価基準の明確化
第 3 章で示したように重要な品質観点を定義し，それぞれの観点からの評価を促す質問および評価基準を明確化する.本研究では評価対象の違いを明確に区別しつつ,人手評価のばらつきが生じにくい 5 段階評価を採用した.具体的な実施結果は付録に示す.

② 複数 LLM を利用した LLM-as-a-Judge の実行
複数の LLM で個別に評価を実行する．単一の LLM に回答の評価を一任しないことで，LLM の持つバイアスやタスクごとの得意・不得意等，モデルやタスクの特性に起因する評価スコアの不正確さを排除する.

③ 各品質観点への重み付け
品質観点それぞれ 1〜5 点で評価された結果に対し，品質観点ごとに重みを付けた加重和を算出する．重みの算出方法の一例として，本研究では AHP（Analytic Hierarchy Process：階層化意思決定法）を採用し,「LLM の回答の良し悪しを判断するとき，何を重視すべきか？」という主観的な判断を定量的に整理し「重み」とした.

JRVF を用いる上で最適な評価基準やモデルの組み合わせは，社会状況・技術の進展によって変化する．そこで本研究では，提案する JRVF に第 3 章で定義した品質観点を適用し，上記①〜③の手順で算出した評価指標「CARE スコア（Confidence / Accuracy / Reproducibility / Explainability の頭字語)」を定義する．この CARE スコアの有効性を検証し，JRVF が信頼性評価指標を算出するための枠組みとして機能することを示す.

5. 実験設定

5.1 題材とするタスクの説明

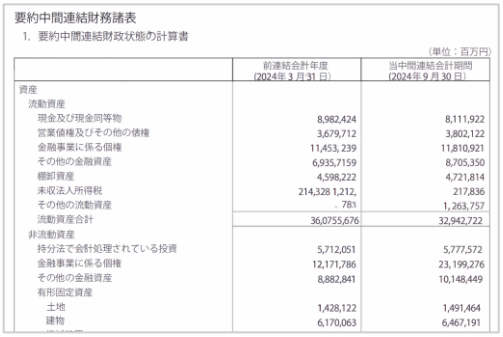
本実験では，想定した利用者及び利用ケースに対応する 2 種類のタスクを設定した．専門外領域における知識不足，複雑な文書構造，図表データの迅速な傾向把握の困難さにより短時間での正確な判断が困難な状況を PDF 画像タスクとし,知識ギャップ・読解コスト・構造的複雑性の軽減を目的とした利用場面を想定した．また，衛星画像の解析において，

専門的な知識不足，画像の持つ情報の不確実性により正解情報の収集が困難な状況を衛星画像タスクとし，専門的な説明および作業負担の削減を目的とした利用場面を想定した．なお，前者と比べ後者のタスクの回答品質が低い（難しい）ことが想定される．

5.1.1 PDF 画像タスク

日本語 PDF を画像解析する場合，文字認識やレイアウト解釈の揺らぎが生じやすい．本タスクでは，日本語独自の文字組・表組など多様な要素を含む日本語 PDF 文書を対象に，マルチモーダル LLM による質問応答を行い，回答に CARE スコアを併記する設定とした．

質問例：要約中間連結財務諸表にて，流動資産合計が増加していますが，この増加はどの項目の増加による影響が最も大きいと推測できますか？


要約中間連結財務諸表

1. 要約中間連結財政状態の計算書

（単位：百万円）

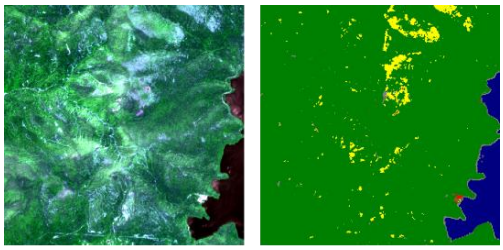
	前連結会計年度 (2024年3月31日)	当中間連結会計期間 (2024年9月30日)
資産		
流動資産		
現金及び現金同等物	8,982,424	8,111,922
営業債権及びその他の債権	3,679,712	3,802,122
金融事業に係る債権	11,453, 239	11,810,921
その他の金融資産	6,935,7159	8,705,350
棚卸資産	4,598,222	4,721,814
未収法人所得税	214,328 1,212,	217,836
その他の流動資産	: 781	1,263,757
流動資産合計	36,0755,676	32,942,722
非流動資産		
持分法で会計処理されている投資	5,712,051	5,777,572
金融事業に係る債権	12,171,786	23, 199,276
その他の金融資産	8,882,841	10,148,449
有形固定資産		
土地	1,428,122	1,491,464
建物	6,170,063	6,467,191

図 1 PDF 画像例

5.1.2 衛星画像タスク

衛星画像の解析・運用に必要な正解情報の収集は依然として大きなコスト要因である．本タスクでは，光学衛星 Sentinel-2 画像と土地被覆図を対象とし，マルチモーダル LLM による質問応答を行い，回答に CARE スコアを併記する設定とした．

質問例:画像内にある最大の水域の名称を答えてください．また，その判断した根拠も明示してください．



Contains modified Copernicus Sentinel data 2024

提供：高解像度土地利用土地被覆図（JAXA）

図 2 衛星画像例

5.2 実験設定

5.2.1 回答用 LLM・評価用 LLM の設定

本研究では「複数 LLM を利用した LLM-as-a-Judge の実行」を提案している．ベースモデルの多様性，回答生成過程の違い（推論モデルと汎用モデル）の観点から以下に示す LLM モデルを実験に用いた．

- 回答用 LLM　：【汎用】Claude Haiku 4.5, GPT-4.1 mini
- 評価用 LLM　：【汎用】Claude Haiku 4.5, GPT-4.1 mini
  【推論】Claude Haiku 4.5（拡張思考モード ON), GPT-5 mini

5.2.2 CARE スコアの算出・各評価観点の重みの設定

第 4 章で定義した評価基準を評価用 LLM に入力するプロンプトに明記し，回答用 LLM が出力した回答に対して LLM-as-a-Judge と人手評価でそれぞれ 5 段階評価を実行した．

その後，人手評価の実施者（評価者）により AHP による各品質観点の重み付けを段階的に実施した．まず大項目（自信度，正確性，再現性，説明性）間でどちらを重視するかを 5 段階で総当たり評価し，その後大項目ごとに小項目間で同様に総当たり評価を行った．最後に評価者 4 人がそれぞれ算出した評価結果の幾何平均を取り重みとして設定し，加重和を合計 100 点で算出した．重みの詳細は付録に示す．

5.2.3 人手評価方法の設定

本研究における評価基準（評価用の質問，点数の基準）の設計は複数の評価者による合議により実施した．これにより評価者間の認識を統一し，人手評価によるばらつき軽減を図った．また画像を入力とするタスクでは画像解釈の曖昧性が評価に影響を及ぼす可能性があるため，本研究では各タスクに専門家・非専門家を一人ずつ配置し，CARE スコアとの比較を行った．ここで専門家とは，PDF 画像タスクでは PDF の技術的構造（レイアウト，読み順，表構造等）および文字認識（OCR）の特性の評価に長けた者を指し，衛星画像タスクでは衛星画像判読・分析を業務として扱い，画像からの情報収集に長けた者を指す．

6. 結果と考察

6.1 結果

評価用 LLM の評価結果と，人手評価結果（評価者 2 人の平均点）との相関係数を表 1 に示す．PDF 画像タスクでは人間の評価結果と高い相関がみられた一方，衛星画像タスクでは比較的相関が低い結果となった．また PDF 画像タスクに絞って比較すると，Claude 系・GPT 系ともに推論モデルが汎用モデルよりも人間の評価結果と高い相関を示した．

表 1　人手評価の CARE スコアと評価用 LLM の CARE スコアの相関係数一覧

回答用 LLM / 評価用 LLM	PDF 画像タスク		衛星画像タスク	
	Claude Haiku 4.5	GPT-4.1 mini	Claude Haiku 4.5	GPT-4.1 mini
Claude Haiku 4.5	0.52	0.61	-0.12	0.44
Claude Haiku 4.5 (拡張思考モード ON)	0.70	0.71	0.33	0.29
GPT-4.1 mini	0.09	0.83	0.19	0.44
GPT-5 mini	0.59	0.99	0.13	0.13

また各タスクに対する人手評価の CARE スコア（評価者 2 人の平均点）を表 2 に示す．今回の実験においては下記の傾向があることが分かった．

- 2 つのタスク間で難易度に違いがある
  衛星画像タスクは，緯度経度情報の把握や数値計算など比較的高い専門性が求められる質問が多く，CARE スコアの回答用 LLM 間平均値が低くなった．
- 2 つのモデル間で回答品質に差異がある
  両タスクで共通して Claude Haiku 4.5 が GPT-4.1 mini と比べ低いスコアとなった．CARE スコアの値が 55 前後と近しいため，共通の要因として Claude では画像が読み取れなかったことが挙げられる．(評価用 LLM の Claude 系モデルも認識不可な画像である前提で評価した事例が多かった)特に PDF 画像タスクにおいてその差は顕著だった．

表 2　各タスクにおける人手評価結果

タスク / 回答用 LLM	Claude Haiku 4.5	GPT-4.1 mini
PDF 画像タスク	54.5	91.0
衛星画像タスク	58.2	67.9

また各評価用 LLM の評価結果を分析すると，上記に挙げた今回の実験における特徴を捉えて評価できているモデル（GPT-5 mini 等）があるのに対して，特徴を捉えられていないモデル（GPT-4.1 mini 等）もあることが分かった．特に GPT-4.1 mini はタスクの違い・回答品質の違いに関わらず高いスコアを付ける傾向にあった．詳細分析結果は付録に示す．

6.2 考察

本研究の目的である「LLM 出力の信頼性指標を提示する」という観点から，CARE スコアの有効性を検証するにあたって特に重視して検討すべき課題を以下に示す．

課題①：評価コストと一貫性（LLM-as-a-Judge におけるバイアスの有無，実行コスト）
課題②：品質観点，評価基準の定義の妥当性
課題③：入力画像解釈の曖昧性

以降の考察ではこれらの課題を重点的に検討した．6.2.4 項ではまとめた考察を示す．

6.2.1 評価コストと一貫性

まず課題①に関する考察を行う．ここでは，先行研究[4]で示された LLM-as-a-Judge の持つバイアス（先入観や偏見）の内，個々のタスクや回答内容の違いに直接起因しない，よりタスク非依存的なバイアスに焦点を当てる．ただし，実際の評価過程ではタスクに依存したバイアスも混入し得ることに留意されたい．本研究で想定する LLM モデルに起因する

タスク非依存的バイアスは，主に以下の 2 点である．

a) モデル種別によるバイアス：GPT 系モデルと Claude 系モデルの間の評価傾向の差

b) 自己強化バイアス：同一系統のモデル同士の場合に相対的に過大評価する傾向

6.1 より，Claude 系モデルは画像を読み取れず，当該タスクにおける回答品質が GPT 系モデルと比較して相対的に低いことが分かっている．そのため，本研究の設定では，モデル間の回答品質の差と a)のバイアスとを統計的に切り分けることは困難である．そこで，本節では b)の自己強化バイアスの有無の検証に焦点を絞る．評価には Wilcoxon 符号付順位検定を用い，モデルが生成した評価が，異なる系統のモデルの評価分布と同じ範囲に含まれる確率を算出し，評価項目ごとに表 3 に示した．本研究では，0.05（5%）未満の項目は自己強化バイアスが統計的に有意であると判断し，該当箇所に色を付けて示した．

表 3 自己強化バイアスの可視化

大項目	小項目 ＼ 回答 LLM モデル	Claude Haiku 4.5		GPT-4.1 mini	
		PDF 画像タスク	衛星画像タスク	PDF 画像タスク	衛星画像タスク
自信度（過信度）		0.91	0.14	0.06	0.06
正確性	情報の真実性	0.15	0.15	0.06	0.03
	質問への関連性・適合性	0.89	0.89	0.16	0.05
	論理的な一貫性	0.48	0.02	0.27	0.41
再現性	同一質問再現性	0.89	0.75	0.06	0.18
	パラフレーズ耐性	0.66	0.78	0.03	0.69
	根拠・判断基準の再現性	0.95	1.00	0.00	0.00
説明性	根拠の提示と説明性	0.91	0.22	0.22	0.75
	前提・適用範囲・限界の明示	0.72	0.13	0.10	1.00
	追跡可能性	1.00	0.13	0.22	0.59
合計スコア		0.31	0.25	0.01	0.01

表 3 に示すとおり，GPT 系モデルでは，合計スコアを含む一部の評価項目において 0.05 未満となり，自己強化バイアスが存在する可能性が示唆された．一方で，Claude 系モデルについては，多くの項目で統計的に有意な自己強化バイアスは確認されなかった．

ただし，いずれのタスクにおいても Claude 系モデルが入力画像を正しく認識できていない事例が多数観測されており，このことが評価値を相対的に低下させ，結果として確率に影響を与えた可能性がある．この点については，モデル固有のマルチモーダル処理能力の差異を考慮した追加検証の必要がある．

また，LLM を用いた評価に PDF 画像タスクは＄8.8（3.5 時間），衛星画像タスクは＄6.6（4 時間）かかった．なお，人手評価はいずれのタスクも 9 人時程度かかる作業であった．

6.2.2 品質観点，評価基準の定義の妥当性

表 4 品質観点別の人手評価との相関係数

	PDF 画像タスク	衛星画像タスク
正確性	0.62	0.06
説明性・妥当性	0.27	-0.05
自信度	0.60	0.45
再現性	0.83	0.70

続いて，課題②に関する考察を行う．第 3 章で定義した 4 種の品質観点について，LLM(GPT-5 mini)と人手評価との相関および人手評価者間一致度を比較した結果を表 4 に示す．両タスクにおいて再現性は相関・一致度ともに相対的に高かった一方，説明性・妥当性はいずれも低い値を示した．これは，根拠をどの程度具体的に示せば「十分」とみなすかという閾値や，明記されていない前提・適用範囲・限界をどこまで補って解釈するかといった評価者ごとの基準の違いに強く依存するためである．実際，PDF 画像タスクでは結論が誤っていても論理展開の丁寧さを評価する者がいたことや，衛星画像タスクでは専門家と非専門家で暗黙の前提の扱いが異なることが，説明

性スコアのばらつきを拡大したと考えられる．このことから，説明性・妥当性についてはルーブリックの具体化・詳細化，あるいは代替となる品質観点の検討が必要である．

6.2.3 入力画像解釈の曖昧性

課題③に関し画像解釈の曖昧性に関する考察を示す．利用者の質問内容および入力画像に対する理解度の違いにより，質問ごとに要求される「正解」の粒度は異なる．このような評価基準の違いが LLM の評価とどう関係しているかを検討するため，各 LLM による評価結果と，専門家および非専門家による評価結果との相関係数を算出し，比較を行った．なお，各タスクは専門家 1 人と非専門家 1 人によって評価作業を実施している．

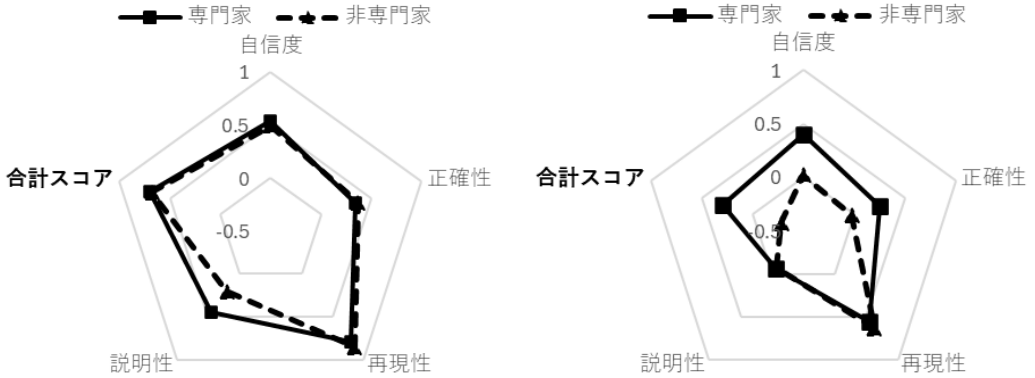


図 3 専門家と非専門家の LLM 評価との相関係数
（左：PDF 画像タスク，右：衛星画像タスク）

図 3 に示す通り，PDF 画像タスクにおいては，多くの評価項目で専門家と非専門家の間に大きな相関の差は認められなかった．一方，衛星画像タスクでは，特に合計スコアにおける相関係数が専門家と非専門家の間で 0.5 以上異なる結果となった．

この結果は，専門性の高い分野を対象としたタスクにおいて，LLM の評価傾向が非専門家よりも専門家の評価に近づく可能性を示唆している．すなわち，タスクの専門性が高まるほど，LLM は一般的な理解に基づく評価よりも，専門的観点に基づく評価基準を内在的に反映する傾向があると考えられる．

6.2.4 CARE スコアの有効性まとめ

これまで示した実験結果・考察内容から，CARE スコアの有効性について考察する．6.1 節で示した結果からタスクの難易度・回答品質の違いに関わらず同一な評価を行う LLM が存在することが示され，また 6.2.1 項では自己強化バイアスが存在することが示された．これらの結果は，単一モデルの評価ではなく複数モデルの評価結果を統合する提案は，指標の安定性を高める手段として有効であることを示唆している．

また 6.2.2 項では品質観点ごとに人手評価と LLM 評価の一致度には差があることを示し，特に説明性・妥当性については人間の評価者間でもばらつきが大きいことも 6.2.3 項で示した．このような一致度構造のもとでは，重み配分を変更・調整したとしても，CARE スコアと人手評価との相関が向上するとは言い切れないと考えられる．表 5 に，AHP による重み付け有無による CARE スコアと人手評価の相関係数の変化を示す．AHP による加重和を計算した CARE スコアは，重み付け無しの場合に比べて人手評価との相関が低くなる結果となったが，変動は限定的であった．

一方で AHP を重み決定手法としてではなくフィルターとして活用し，重要度が低いと判断される品質観点を除外したうえで統合した場合，相関は改善する傾向がみられた．表 5 に，AHP による重みを上位から加算し，累積寄与率が 90%に達する項目までを採用して統合した結果も併記した．以上から AHP は，評価構造を可視化し，各品質観点の影響を把握・

取捨選択をするための枠組みとして有効である可能性が示された.

なお，表 5 における複数モデルの評価結果を統合する方法は平均値とし，衛星画像タスクについては専門家のスコアを採用，PDF 画像タスクについては 2 人の評価者の平均点を採用した.

表 5 人手評価との相関係数一覧（AHP の利用方法による比較）

回答用 LLM / AHP 利用方法	PDF 画像タスク		衛星画像タスク	
	Claude Haiku 4.5	GPT-4.1 mini	Claude Haiku 4.5	GPT-4.1 mini
AHP 重み付け無	0.75	0.91	0.46	0.63
AHP 重み付け有	0.73	0.88	0.41	0.57
AHP をフィルター利用	0.80	0.93	0.60	0.74

以上から，各品質観点を適切に整理しつつ，複数モデルの評価結果を統合した CARE スコアは，人手評価と一定の相関を示し信頼性指標として有効であると結論づけられる.

7. まとめ

7.1 結論

本研究では，マルチモーダル LLM の実利用において，利用者が出力の正否や信頼性を判断することが困難であるという課題に対し，LLM-as-a-Judge を活用した信頼性評価の枠組みとして JRVF を提案した．本枠組みの中で，自信度・正確性・再現性・説明性の 4 つの品質観点を定義し，複数 LLM による評価結果を AHP に基づいて統合した指標として CARE スコアを導入した．実験の結果，CARE スコアは PDF 画像・衛星画像タスクの両方において，人手評価と一定の相関を示した．特に，複数 LLM の評価結果を統合することで，タスク難易度や回答品質の違いに依らず，人手評価とロバストに整合する傾向が確認された.

以上より，JRVF より算出される CARE スコアは，LLM の出力を採用するか，あるいは確認や再質問等が必要かを判断するための実用的な信頼性指標として機能し得ることを示した.

7.2 今後の展望

第一に重要な課題として，説明性・妥当性の評価基準の曖昧さが挙げられる．本研究では，説明性に関する評価において LLM と人手評価の相関が低く，評価基準の解釈にばらつきが生じた可能性が示唆された．説明構造の形式をより具体的に定義するルーブリックの精緻化，あるいは QA4AI に記載された他の品質観点で代替・再編することにより，評価の安定性向上が期待される.

第二に，評価用 LLM およびスコア統合・提示方法の一般化が課題である．モデル更新やタスク特性の違いにより，最適な評価用 LLM や統合方法は変化し得る．今後は，複数 LLM の評価結果を前提としつつ，統合方法をタスクに応じて切り替える手法や，スコアの内訳を併記するなど，利用者が注意すべき点を把握しやすい提示方法の検討が求められる.

8. 参考文献

[1] K.Papineni et al., Bleu: a Method for Automatic Evaluation of Machine Translation, 2002
[2] J.Hessel et al., CLIPScore: A Reference-free Evaluation Metric for Image Captioning, 2021
[3] X.Liu et al., Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey, 2025
[4] J.Gu et al., A Survey on LLM-as-a-Judge, 2024
[5] AI プロダクト品質保証コンソーシアム，AI プロダクト品質保証ガイドライン 2025.04 版，2025