

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 付録 研究コース 5 (Q4A1)

1. はじめに

本文書は、「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」の詳細情報・付加情報（付録）を記載する。なお、章番号を本文と合致させている。

2. 研究背景

LLM-as-a-Judge による評価手法は、経済的コストも懸念される。

3. 品質観点の定義

画像入力タスクにおける回答品質やその信頼度には「画像認識精度」が大きく関わると考えた。

4. 提案内容

① 品質基準の明確化

品質観点ごとに 5 段階評価基準を評価者間で決定した。以下にその内容を示す。

大項目	中項目	評価用 LLM に入力する質問および評価基準の定義
正確性	情報の 真実性	回答に含まれる主要なデータや事実は、客観的に正確であり、誤情報や虚偽を含んでいませんか？
		5 点 全ての主張が完全に正確で、誤情報や虚偽はありません。
		4 点 主要な主張は正確ですが、本筋に関係ない細部は確認が必要な場合があります。
		3 点 概ね正しい情報ですが、重要な前提やデータの一部が欠落または不足しています。
		2 点 事実誤りや矛盾が一部含まれています。主要部分の信頼性は低いです。
		1 点 決定的な誤情報や虚偽を含み、回答全体が信頼できません。
	質問への関 連性・適合 性	回答は、質問で求められた内容に的確に答えていますか？ 不必要な冗長性や、中心的なテーマからの逸脱はありませんか？
		5 点 回答が質問の意図と条件を完全に満たしています。
		4 点 質問の意図は正しく捉えていますが、条件にわずかな不一致があります。
		3 点 主要な意図には沿っていますが、一部の指示を満たしていない、または不要な情報を含んでいます。
		2 点 質問の意図を部分的に誤解しており、回答の関連性が低いです。
		1 点 質問の意図を完全に誤解し、回答が的外れです。
	論理的な一 貫性	回答の根拠と結論の関係は論理的で分かりやすいですか？ 回答全体で話の流れや整合性が保たれていますか？
		5 点 論理展開は完全に合理的で、飛躍や矛盾はありません。
		4 点 論理の流れは明確ですが、一部に説明不足があります。
		3 点 結論は概ね妥当ですが、論理の飛躍や根拠不足があります。
		2 点 推論過程に矛盾や破綻が一部あり、結論の信頼性は低いです。
		1 点 論理が完全に破綻しており、結論は導かれていません。
説明性 ・ 妥当性	根拠の 提示と 説明性	回答に主張を支える根拠や理由が十分に含まれていますか？ 質問の趣旨から外れる説明や不要な情報は含まれていませんか？
		5 点 全ての回答に対して根拠が提示され、根拠と回答の因果関係が論理的に説明されています。
		4 点 全ての回答に対して根拠が提示され、概ね論理的な説明がされているが、一部に不足があります。
		3 点 回答に対して、一部の根拠は提示されるが、説明が不十分または抽象的な箇所が含まれます。
		2 点 根拠の提示が部分的であり、全体的に主張との論理的つながりが不明瞭です。
		1 点 根拠が提示されていない、または根拠が主張を説明していません。
	前提・ 適用範囲 ・限界 の明示	回答に必要な前提条件、適用範囲、限界の説明が十分に含まれていますか？ 不要な条件は含まれていませんか？
		5 点 必要な前提条件、適用範囲、限界のすべてが具体的かつ明確に記述されています。
		4 点 必要な前提条件、適用範囲、限界が記述されているが、一部に説明の不足や不要な条件が含まれています。
		3 点 必要な前提条件、適用範囲、限界の記載に誤りはないですが、いずれかが欠落しています。
		2 点 必要な前提条件、適用範囲、限界の記述があいまいで、一部誤った記載があります。
		1 点 必要な前提条件、適用範囲、限界が明示されていない、または誤った条件が記載されています。

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 付録
 研究コース 5 (Q4A1)

大項目	中項目	評価用 LLM に入力する質問および評価基準の定義
説明性 ・ 妥当性	追跡可能性	この回答が、どの根拠に基づき、どのような推論過程と判断基準を経て導かれ、情報の妥当性がどのように確保されているかを、一貫して説明できていますか？
		5点 入力から出力までの推論プロセスが詳細に説明され、各段階の判断基準が明確です。
		4点 主要な推論プロセスが示され、概ねプロセスを追跡できます。
		3点 推論プロセスの概要は示されているが、詳細は不明瞭です。
		2点 推論プロセスの一部のみが示され、全体像が把握しにくいです。
		1点 推論プロセスがブラックボックスで追跡不可能です。
自信度	自信度 (過信度)	提示された確信度は、実際の回答品質に対して妥当なスコアですか？
		5点 自身の回答内容の不足部分を確信度として正確に捉え、過信・謙虚のどちらにも寄っていません。
		4点 回答の主旨は確信度と一致しているが、若干の過小評価・過大評価があります。
		3点 回答の一部について確信度と回答品質に乖離があり、過小評価・過大評価があります。
		2点 回答の全体について確信度と回答品質に乖離があり、大きな過小評価・過大評価があります。
		1点 完全に誤った回答に対して確信度「高」、もしくは完全に正しい回答に対して確信度「低」を示しています。
再現性	同一質問再現性	同じ質問（同一トークン列）を繰り返したとき、出力の結論や重要な判断が安定していますか？
		5点 繰り返しの全試行で結論・方針・重要条件が一致しています。表現の微かな差（語順・言い回し）を除けば同一の回答です。
		4点 結論・方針が全試行で一致しています。説明の順序や補足は変わる場合もありますが、主旨は変わりません。
		3点 大筋の結論は同じですが、必要な条件が明示されていない場合があります。
		2点 繰り返しの中で一部試行が異なる結論や方針になります。過半数は同じです。
		1点 試行ごとに結論・方針がばらばらです。
	パラフレーズ耐性	質問の表現を変えた場合に、意味が同じである限りにおいて出力の結論や重要な判断が一貫していますか？
		5点 繰り返しの全試行で結論・方針・重要条件が一致しています。表現の微かな差（語順・言い回し）を除けば同一の回答です。
		4点 結論・方針が全試行で一致しています。説明の順序や補足は変わる場合もありますが、主旨は変わりません。
		3点 大筋の結論は同じですが、必要な条件が明示されていない場合があります。
		2点 繰り返しの中で一部試行が異なる結論や方針になります。過半数は同じです。
		1点 試行ごとに結論・方針がばらばらです。
	根拠・判断基準の再現性	同じまたはパラフレーズされた質問群に対し、結論を導く根拠や判断プロセスが安定していますか？
		5点 全ての回答で主要な根拠・判断基準が一致しています。
		4点 根拠の主旨は一致しています。補助的な説明や例示が追加・省略される場合もありますが、核心的判断基準に変化はありません。
		3点 結論と最重要根拠は概ね一致しますが、判断プロセスがばらつくことがあります。
		2点 回答ごとに使う根拠が異なり、同じ根拠で説明できない場合があります。
		1点 回答同士で矛盾しています。

5. 実験設定

5.1 題材とするタスクの説明

本文ではタスクの概要のみ説明したが、各タスクの目的まで含めた詳細を記載する。

5.1.1 PDF 画像タスク

近年、PDF ドキュメントに含まれるテキスト、図表、スキャン画像などを直接解析するマルチモーダル LLM の活用が進んでいる。しかし、欧文文書と異なり、日本語文書では縦組み・横組みの併用や、漢字・ひらがな・カタカナ・アルファベットが混在するなどレイアウトが多様であり、PDF 由来の画像では文字認識の揺らぎやレイアウト解釈の誤りが生じやすい。そのため、マルチモーダル LLM がこれらの文書画像を安定して解釈できるかは必ずしも明らかではなく、また、利用者が文書の全ページを精査することは大きな負担となるため、モデルの回答をどこまで信頼でき

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 付録 研究コース 5 (Q4A1)

るかを把握することが重要である。本タスクでは、日本語特有の縦書き・ルビ・割注を含むテキスト、表組、図表など、実際の PDF 文書で一般的に見られる多様な要素を網羅したデータを対象に、マルチモーダル LLM による質問回答を評価する。さらに、各回答に CARE スコアを付与し、利用者が回答の信頼性を判断できる手がかりとなるかを検証する。

5.1.2 衛星画像タスク

近年、地球観測衛星の高分解能化および観測頻度の向上により、衛星画像を活用した広域かつ継続的な環境モニタリングが実用段階に入りつつある。これに伴い、衛星画像の解析・運用を支えるシステム開発が進展しているが、その過程において正解情報（実際の撮影箇所に関する補足情報）の収集が大きなコスト要因の一つとなっている。マルチモーダル LLM の活用により作業効率の向上が期待される一方で、その推論精度や適用可能な課題（質問）範囲は必ずしも明確ではない。そこで本タスクでは、無償で広く利用されている光学衛星である Sentinel-2 画像を対象とし、質問への回答に CARE スコアを併記することで、マルチモーダル LLM から得られる情報の信頼性を担保し得るかを検証する。その際、LLM が直接扱える画像形式ではないため、画像を 3 バンドずつに分割し、計 4 枚の PNG ファイルに変換して同時に入力した。さらに付加情報として、国立研究開発法人宇宙航空研究開発機構 (JAXA) が公開している高解像度土地利用・土地被覆図（以下、土地被覆図）を用いた。

5.2 実験設定

5.2.1 回答用 LLM・評価用 LLM の設定

評価に用いたプロンプトの一覧を示す。

<タスク回答用プロンプト (PDF 画像タスク) >

```
=== SYSTEM ===
あなたは画像分析の専門家です。
提供された画像を分析し、質問に回答してください。
## 回答の構成
以下の 3 点を含めて、自然な文章で簡潔に記述してください：
1. **回答内容**：質問への答え
2. **根拠と推論**：
  - どのページのどこに何が書かれているか
  - その情報からどう判断したか
3. **信頼性と限界**：
  - この回答の確信度 (高/中/低) と理由
  - 回答できない点や不確実な点
## 注意事項
- 画像内の情報のみに基づいて回答すること
- 推測や憶測は避け、確認できる事実のみを記述すること
- 不明な点は明確に「不明」と述べること
=== HUMAN ===
{各質問と、質問に使う画像を動的に投入}
```

<タスク回答用プロンプト (衛星画像タスク) >

```
=== SYSTEM ===
あなたは Sentinel-2 衛星画像の解析者です。
## 入力データの仕様
### 衛星画像データ
10m/pix の衛星画像の raw データを 3 バンドずつ 16bit PNG 化した画像を入力します。
**大気反射率への変換方法**：
- 1000 を引いてから 10000 で割った値を使用してください
- 計算式：`反射率 = (ピクセル値 - 1000) / 10000`
**バンド構成**：
- **1 番目の画像**：バンド 2, 3, 4 (青、緑、赤)
- **2 番目の画像**：バンド 8, 1, 5 (近赤外、Coastal aerosol、植生レッドエッジ 1)
- **3 番目の画像**：バンド 6, 7, 8a (植生レッドエッジ 2, 3、狭帯域近赤外)
- **4 番目の画像**：バンド 11, 12, 9 (短波長赤外 1, 2、水蒸気)
```

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 付録
 研究コース 5 (Q4A1)

```

### 土地被覆図
**5 番目の画像**が土地被覆図です。
**ラベル定義**:
- **黒色**: 不明
- **白色**: 雲域
- **青色**: 水域
- **緑色**: 森林
- **黄色**: 草地
- **肌色**: 耕作地
- **水色**: 水田
- **茶色**: 裸地
- **赤色**: 人工構造物
- **紫色**: ソーラーパネル
- **灰色**: 岩礁・干潟
## 回答の構成
以下の3点を含めて、自然な文章で簡潔に記述してください:
1. **回答内容**: 質問への答え
2. **根拠と推論**:
    - どのページのどこに何が書かれているか
    - その情報からどう判断したか
3. **信頼性と限界**:
    - この回答の確信度 (高/中/低) と理由
    - 回答できない点や不確実な点
## 注意事項
- 画像内の情報のみに基づいて回答すること
- バンド情報を適切に解釈すること (特に植生指数、水域判定など)
- 推測や憶測は避け、確認できる事実のみを記述すること
- 不明な点は明確に「不明」と述べること
- 土地被覆図のラベルを正確に参照すること
=== HUMAN ===
# 位置情報
衛星画像の位置情報は以下の通りです:
- **緯度の範囲**: {個別情報を動的に投入}
- **経度の範囲**: {個別情報を動的に投入}
- **座標系**: {個別情報を動的に投入}
- **撮影日**: {個別情報を動的に投入}
---
# 質問
{個別情報を動的に投入}
---
# 回答指示
上記の衛星画像、土地被覆図、位置情報、撮影日を踏まえて、質問に回答してください。
--- 画像 1: バンド 2,3,4 (青、緑、赤) ---
{個別画像を動的に投入}
--- 画像 2: バンド 8,1,5 (近赤外、Coastal aerosol、植生レッドエッジ 1) ---
{個別画像を動的に投入}
--- 画像 3: バンド 6,7,8a (植生レッドエッジ 2,3、狭帯域近赤外) ---
{個別画像を動的に投入}
--- 画像 4: バンド 11,12,9 (短波長赤外 1,2、水蒸気) ---
{個別画像を動的に投入}
--- 画像 5: 土地被覆図 ---
{個別画像を動的に投入}
    
```

<パラフレーズ生成用プロンプト>

```

=== SYSTEM ===
あなたは質問を言い換える専門家です。
与えられた質問を、意味を保ちながら異なる表現で言い換えてください。
## 言い換えの原則
1. **意味の保持**: 元の質問の意図を正確に保つ
2. **表現の多様性**: 異なる語彙や文構造を使用
3. **自然さ**: 自然で理解しやすい日本語
4. **質問の本質**: 求めている情報は変えない
## 言い換えるの種類
- **語彙変更**: 同義語や類義語を使用した言い換え
- **構文変更**: 文の構造を変更した言い換え
- **詳細化**: より具体的な表現に変更
- **簡潔化**: より簡潔な表現に変更
- **丁寧度変更**: 丁寧さのレベルを変更
## 出力形式
以下の JSON 形式で出力してください:
`json`
{
  "paraphrased_questions": [
    {
      "paraphrased_question": "<言い換えた質問>",
      "paraphrase_type": "<言い換えるの種類>"
    }
  ]
}
    
```

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 付録 研究コース 5 (Q4A1)

```
### フィールドの説明
- **paraphrased_questions**：言い換えた質問の配列
- **paraphrased_question**：言い換えた質問文（文字列）
- **paraphrase_type**：言い換えの種類（上記の種類から選択）
## 重要な注意事項
- 純粋な JSON のみを出力（追加のテキストや説明は不要）
- コードブロックマーカー（` `json` `）は使用しない
- 質問の本質的な意味を変えない
- 各言い換えは明確に異なる表現にする
- すべてのフィールドは必須
応答は必ず{で始まり}で終わる、パース可能な JSON でなければなりません。
=== HUMAN ===
以下の質問を 3 通りに言い換えてください：
**元の質問**： {質問内容を動的に投入}
**要求事項**：
- 3 通りの異なる表現で言い換える
- 各言い換えは意味を保ちながら表現を変える
- 言い換えの種類を明示する
上記の質問を 3 通りに言い換え、JSON 形式で出力してください。
```

<評価用プロンプト（再現性以外の評価項目）>

```
=== SYSTEM ===
あなたは文書分析の品質評価者です。
以下の回答を{評価観点名を動的に投入}の観点から評価してください。
## 評価の観点
{評価観点に関する個別質問を動的に投入}
## 採点基準（5段階評価）
**5点**： {採点基準を動的に投入}
**4点**： {採点基準を動的に投入}
**3点**： {採点基準を動的に投入}
**2点**： {採点基準を動的に投入}
**1点**： {採点基準を動的に投入}
## 出力形式
以下の JSON 形式で評価結果を出力してください：
` `
{
  "score": 0,
  "feedback": "<評価の詳細な説明>"
},
### 出力ルール
1. **score**： 1~5 の整数（上記の採点基準に従う）
2. **feedback**：
- なぜそのスコアをつけたのか、具体的な根拠を示す
- 回答の良い点と改善点を明記する
- 採点基準のどの項目に該当するか説明する
## 重要な注意事項
- 純粋な JSON のみを出力（追加のテキストや説明は不要）
- コードブロックマーカー（` `json` `）は使用しない
- score は必ず 1~5 の整数
- feedback は具体的に建設的な内容にする
応答は必ず{で始まり}で終わる、パース可能な JSON でなければなりません。
=== HUMAN ===
# 評価対象
## 質問
{質問内容を動的に投入}
## 回答
{生成された回答を動的に投入。再現性評価の場合、繰り返し回答・パラフレーズ質問および各回答を投入。}
## 元文書
[画像は下に提供されます]
---
# 評価指示
上記の回答を以下の観点から評価してください：
{評価観点に関する個別質問を動的に投入}
## 採点基準
- **5点**： {採点基準を動的に投入}
- **4点**： {採点基準を動的に投入}
- **3点**： {採点基準を動的に投入}
- **2点**： {採点基準を動的に投入}
- **1点**： {採点基準を動的に投入}
## 出力形式
` `
{
  "score": 0,
  "feedback": "<評価の詳細な説明>"
},
**今すぐ上記の形式で JSON 評価結果を出力してください**
```

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 付録
研究コース 5 (Q4AI)

5.2.2 CARE スコアの算出・各評価観点の重みの設定

人手評価後に実施した AHP を実施し、得られた重み (倍率) を示す。

大項目	大項目満点	中項目	中項目満点	評価結果に対する倍率
正確性	61.0	情報の真実性	46.2	9.2
		質問への関連性・適合性	7.5	1.5
		論理的な一貫性	7.3	1.5
説明性	7.7	根拠の提示と説明性	4.3	0.9
		前提・適用範囲・限界の明示	1.6	0.3
		追跡可能性	1.8	0.4
自信度	6.6	自信度	6.6	1.3
再現性	24.7	同一質問再現性	12.9	2.6
		パラフレーズ耐性	9.3	1.9
		根拠・判断基準の再現性	2.4	0.5

6. 結果と考察

6.1 結果

6.1.1 PDF 画像タスク

PDF 画像タスクの画像データ一覧を以下に示す。

項目	要約中間結果	格納結果
要約中間結果	8,800	8,312
要約中間結果	1,991	1,991
要約中間結果	2,703	2,703
要約中間結果	1,372	1,372
要約中間結果	1,838	1,838
要約中間結果	1,659	1,659

図 1 入力画像 1~4

年次	販売台数	割合
2022年	8,800	100.0%
2023年	8,312	94.4%

年次	トキインホテルズ	西日本サンライズホテル
2021.3	28.5	56.4
2022.3	50.8	72.7
2023.3	73.8	85.5
2024.3	76.5	85.5
2025.3	76.2	84.3

図 2 入力画像 5~8

図 3 入力画像 9~10

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 付録 研究コース 5 (Q4A1)

次に、PDF 画像タスクの回答用 LLM 向けの質問文の一覧を示す。

番号	質問文
1-1	パリ五輪の男子ストリートで金メダルを取ったと推測できる選手は誰ですか？
1-2	ミシェル・ド・ビュルヌ夫人のサロンで踊ったエキゾチックな女性ダンサーたちが後に出演した劇場のうち、“異国趣味のショーで知られていた” 劇場はどれですか？
1-3	腰を抜かして動けなかった期間はどれくらいだったと推測できますか？
1-4	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？
1-5	「日中・中日機械翻訳システム」の提供が開始されたのはいつ頃ですか？
1-6	連結販売台数について、日本の 2025.4-2026.3 の今回見通しは、前期実績と比較してどのぐらい増減がありましたか？
1-7	2021年3月から2025年3月まで西日本サンライズホテル稼働率は何パーセント増減しましたか？
1-8	フォームをオーダー後に最初にアクションを取る担当は誰で、どのような処理を行うと推測できますか？
1-9	作業課・射出課・ブロー課・原料課がすべて同じ部の配下にあることから、これらの課が属している部はどこになりますか？
1-10	毎月賃料と管理費に合計いくら払いますか？

<GPT-4.1 mini の回答例 (質問 1-7 : 高スコア例) >

<p>1. **回答内容**: 西日本サンライズホテルの稼働率は、2021年3月の56.4%から2025年3月の84.3%へ27.9ポイント増加しました。</p> <p>2. **根拠と推論**: - ページ1の右下中央の「ホテル稼働率 (%)」グラフに、西日本サンライズホテルの稼働率の推移が示されています。 - 2021年3月の稼働率は56.4%、2025年3月は84.3%と記載されています。 - 差分を計算すると、84.3% - 56.4% = 27.9ポイント増加となります。</p> <p>3. **信頼性と限界**: - 確信度は高いです。グラフに明確な数値が記載されており、視覚的にも読み取りやすい内容です。 - ただし、細かい数値の読み取り誤差はわずかにあり得るものの、全体的な増減の傾向は明確です。 - 他の期間や月単位の詳細増減は不明です。</p>
--

<Claude Haiku 4.5 の回答例 (質問 1-1 : 低スコア例) >

<p># パリ五輪男子ストリートの金メダル選手について</p> <p>## 回答内容</p> <p>画像に掲載されている情報からは、**パリ五輪の男子ストリート (スケートボード) で金メダルを取った選手を特定することはできません**。</p> <p>## 根拠と推論</p> <p>提供いただいた画像 (全 11 ページ) は「2024 年パリオリンピック ハイライト」という朝日小学生新聞の特集号です。これらのページを確認すると:</p> <ul style="list-style-type: none"> - **ページ 2**: スケートボード女子ストリートについて「吉沢恋選手 遠くなかった金」と記載されており、これは**女子種目**の記事です - **ページ 10**: スケートボード女子パーク開場選手が銀について記載されています <p>男子ストリート種目に関する記事は、提供していただいた画像内には見当たりません。</p> <p>## 信頼性と限界</p> <p>**信頼度: 低**</p> <ul style="list-style-type: none"> - 理由: ページに明記されていない情報についての推測を避けるため、確認できない内容については回答できません - 限界: 本特集号が全競技種目をすべて網羅していない可能性があり、提供いただいたページのみでは男子ストリート金メダリストの情報不足しています <p>より詳しい情報が必要な場合は、別途資料をご確認いただくことをお勧めします。</p>

6.1.2 衛星画像タスク

衛星画像タスクの画像データ一覧を以下に示す. 左は衛星画像 (RGB), 右は土地被覆図である.

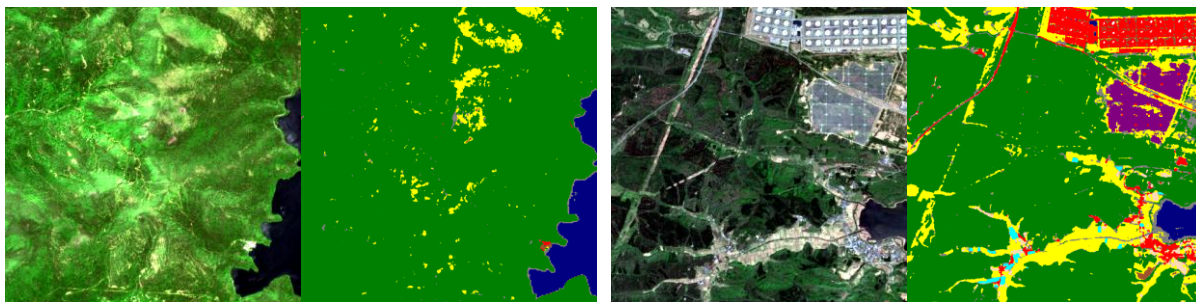


図 3 入力画像例 (左 : 北海道、右 : 青森県)



図 4 入力画像例 (左 : 茨城県、右 : 東京都)

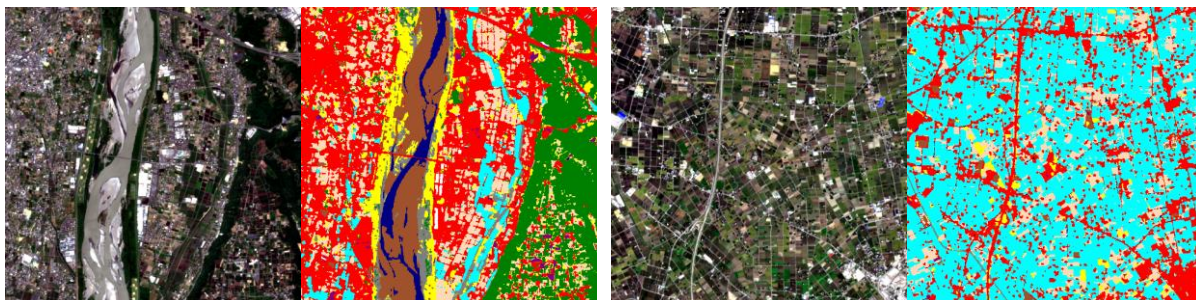


図 5 入力画像例 (左 : 静岡県、右 : 石川県)

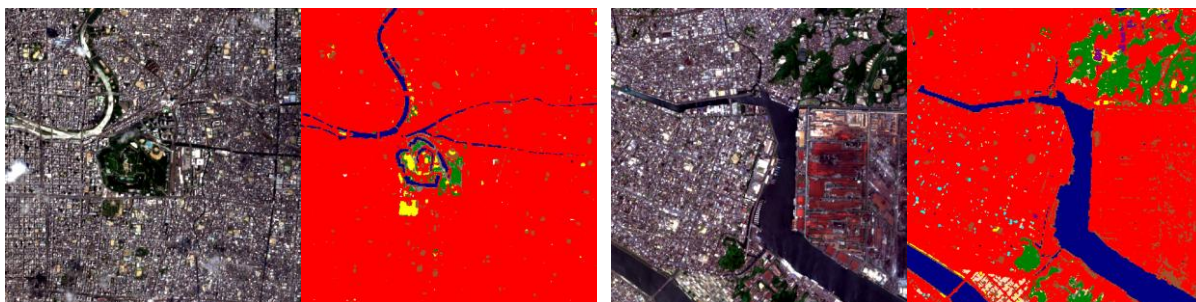


図 6 入力画像例 (左 : 大阪府、右 : 岡山県)

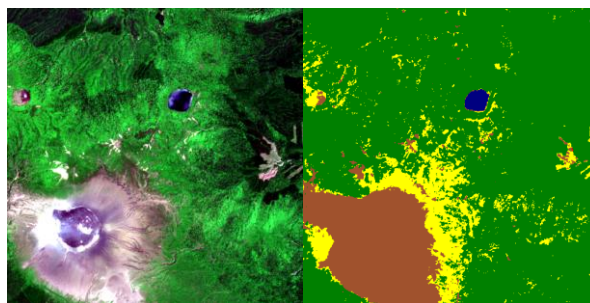


図 7 入力画像例 (鹿児島県)

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 付録 研究コース 5 (Q4A1)

次に、衛星画像タスクの回答用 LLM 向けの質問文の一覧を示す。

番号 (地域)	質問文
2-1 (北海道)	画像内にある最大の水域の名称を教えてください。 また、その判断をした根拠も明示してください。
2-2 (青森)	画像内のソーラーパネルの発電量を MW 単位で推定して下さい。 複数ある場合はその数や位置ごとに詳細に説明してください。 また、その判断をした根拠も明示してください。
2-3 (茨城)	画像内にある高速道路の本数とその名称を教えてください。 また、IC や JCT がある場合はその名称と画像における位置を教えてください。 また、その判断をした根拠も明示してください。
2-4 (東京)	画像内にある河川の名称と画像内の位置関係を全て回答してください。 また、その判断をした根拠も明示してください。
2-5 (東京)	画像の中央上部にて河川が蛇行し東向きから南向きに方向転換している箇所と線路の間に 挟まれた大きな人工構造物は何ですか。 また、その判断をした根拠も明示してください。
2-6 (静岡)	画像内の水域及び裸地の土地被覆ラベルは、衛星画像とあっていないように見えます。 面積としてそれぞれどの程度異なるか割合 [%] を示してください。 また、その判断をした根拠も明示してください。
2-7 (石川)	画像内の水田の面積 [ha] と、水稻の収量 [kg/10a] を推定して下さい。 また、その判断をした根拠も明示してください。
2-8 (大阪)	画像の中央付近にある公園の名称を教えてください。 また、その面積 [ha] を推定して下さい。その判断をした根拠も明示してください。
2-9 (岡山)	画像内に含まれる埋め立て地の位置とその面積 [ha] を教えてください。 また、その判断をした根拠も明示してください。
2-10 (鹿児島)	画像内の左下あたりに火口の見える山の名称を教えてください。 また、その火口の直径 [m] を推定して下さい。その判断をした根拠も明示してください。

<GPT-4.1 mini の回答例 (質問 2-8 : 高スコア例) >

<p>1. 回答内容: 画像中央付近にある公園の名称は「大阪城公園」と推定します。推定される面積は約 100 ヘクタール (ha) です。</p> <p>2. 根拠と推論: - 土地被覆図(画像 5)の中央付近には、赤色 (人工構造物) が主要部分ですが、その周囲に緑色 (森林) と黄色 (草地) の区画がまとまって存在しており、公園の緑地帯として識別できます。 - 中央の人工構造物の形状は、大阪城の特徴的な五重の堀と緑地帯のパターンに合致しています。 - 位置情報 (緯度約 34.67~34.71、経度約 135.50~135.55、EPSG:32653 の日本大阪一帯) からも、大阪城公園の位置と一致しています。 - 衛星画像での形状と土地被覆パターンから、このエリアが「大阪城公園」だと判断しました。 - 面積の推定は、画像が 10m/pix であることから、土地被覆図の緑および草地領域のピクセル数を概算すると約 1 万ピクセル前後となり、1 ピクセル=10m×10m=100 m²なので、約 100ha (1ha=10,000 m²) と推定しました。</p> <p>3. 信頼性と限界: - 確信度は中です。大阪城公園は日本の有名な緑地公園であり、位置・形状・土地被覆の特徴から強く示唆されていますが、画像のみの情報で名称を正確に断定することはできません。 - 公園の正式名称は地形情報や行政の資料、現地確認なしには確実に特定できません。 - 面積推定は画像解像度と土地被覆マスクに基づく概算であり、正確な境界線や建物配置の詳細はわかりません。</p>
--

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 付録 研究コース 5 (Q4A1)

<Claude Haiku 4.5 の回答例 (質問 1-1 : 低スコア例) >

```
# 回答

## 1. 水田面積と水稲収量の推定

**水田面積**：推定不可 (**0 ha**)

**水稲収量**：推定不可

---

## 2. 根拠と推論

### 土地被覆図の分析
提供いただいた土地被覆図 (画像 5) を詳細に検査したところ、以下の構成が確認できます：

- **水色ピクセル (水田)**：ほぼ検出されない
- **青色ピクセル (一般水域)**：わずかに存在
- **赤色ピクセル (人工構造物)**：大量に分布
- **肌色ピクセル (耕作地)**：わずかに分布
- **黄色ピクセル (草地)**：ほぼ検出されない
- **その他**：大部分が青～水色で統一 (ノイズまたは背景)

### 画像データとの照合
衛星バンド画像 (1~4) を検証した結果：
- **バンド 2, 3, 4 (RGB)**：ほぼ真っ黒 (信号が非常に弱い)
- **バンド 8, 1, 5 (NIR)**：近赤外を含むにもかかわらず、ほぼ真っ黒
- **バンド 6, 7, 8a**：暗い背景にわずかな構造が見える程度
- **バンド 11, 12, 9 (SWIR)**：散発的な赤色信号のみ

### 結論
画像全体の約 95%以上が水色で表示されており、これは：
1. **データの不完全性**：土地被覆図のラベル付けが未完了の可能性
2. **画像品質の問題**：衛星バンド画像の信号が極めて弱く、反射率計算が困難
3. **空間的不適合**：土地被覆図と衛星スペクトルデータが適切に対応していない

---

## 3. 信頼性と限界

### 確信度：**低**

**根拠**：
- 土地被覆図の大部分が同一色で占められており、適切な分類がなされていない
- 衛星バンド画像がほぼ黒色で、大気反射率への変換に必要な十分な信号がない
- 10m 解像度で水田を正確に識別するための照度情報が不足

### 不確実な点：
1. **水田の正確な面積**：土地被覆図の水色ピクセル (判定困難) からは面積計算不可
2. **水稲の生育ステージ**：5月14日時点での北陸地方 (推定位置) の水稲は田植え～初期成長段階だが、画像信号が弱く植生指数 (NDVI) を算出できない
3. **収量推定の根拠**：現在の画像では生育状態の判別が不可能であり、収量推定に必要な分光情報が欠落

**推奨**：より高品質な衛星画像またはドローン画像による検証が必要です。
```

以下、衛星画像タスク固有の考察を示す。

衛星画像のファイル形式がマルチバンドの TIFF 形式であり、実験に使用した API にサポートされておらず、TIFF そのままの入力ができなかった。そのため、情報の欠落がない形で PNG に変換して入力を行った。変換の結果、全体的に暗めな画像となり LLM の回答の中には「暗くて認識ができない」といった回答も見られた。回答用 LLM に入力するプロンプトに「自動で MinMax スケーリングをして相対的な画素値 (輝度) を見てください」等を追記する等して LLM の回答性能の向上が見られるのかはさらに検討が必要と考えられる。

画像とともに緯度・経度情報を入力した場合でも、当該地点の近傍に、より知名度が高い (インターネット上の情報量が多い) 建物や湖沼、河川などが存在すると、それらに注目した回答が生成される事例が確認された。また、「〇〇西部」を「〇〇東部」と誤認するなど、位置関係を取り違えるケースも見られた。一方で、大阪城公園や新燃岳のように、特定の地物を明示的に指す質問については、比較的高い正解率が得られた。以上より、位置情報の解釈は必ずしも厳密ではなく、回答生成においては、周辺の情報量や知名度の高い対象に強く影響される傾向があると考えられる。

さらに、画像を見てと特定の土地被覆の画素数を数える、数値計算時に桁数や単位変換を間違えるなどの計算ミスも多く見られた。

6.1.3 各評価用 LLM の評価スコアと人手の評価スコアの傾向比較

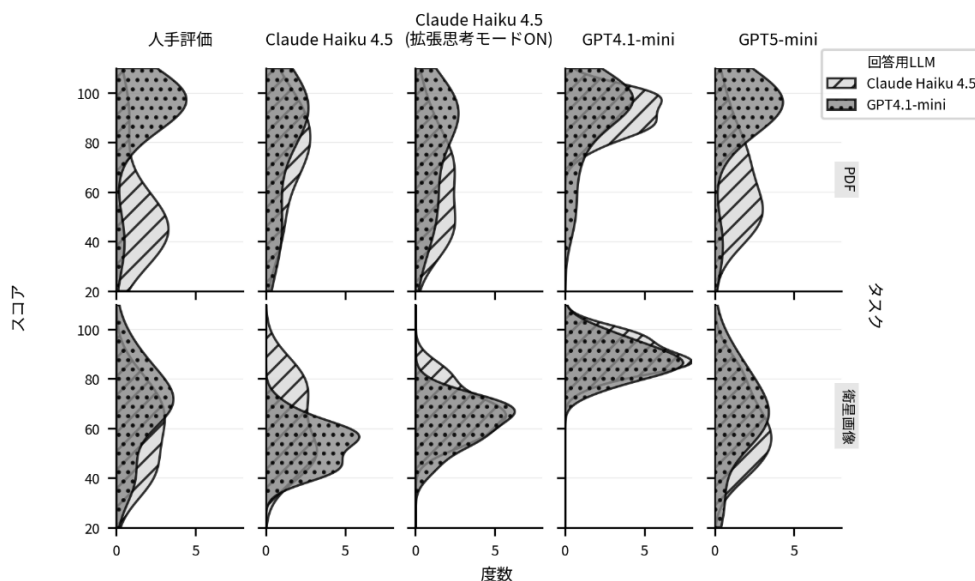


図 8 各評価手法による CARE スコアの密度分布

本実験では、タスク難易度の差（衛星画像タスクは専門性が高いタスクであり、PDF 画像タスクより CARE スコアの平均が低い）と、回答品質の差（両タスクで Claude Haiku 4.5 の方が GPT-4.1 mini より CARE スコアが低く、特に PDF 画像タスクで差が大きい）という 2 つの特徴がある。図 8 は、これらの差を評価用 LLM がスコアに反映しているかを確認するために、人手および各評価用 LLM の CARE スコアを度数密度分布として示したものである。ここで、縦軸は 100 点満点の CARE スコアを示し、横方向に太いスコア帯ほど、その点数が多く付与されたことを表す。したがって図の見方としては、(1) タスクごとに分布の山の位置が変わるか（難易度差の反映）、(2) 同一タスク内で回答用 LLM ごとに分布が分離するか（回答品質差の反映）、(3) 分布が高得点側に集中して低中得点が出にくくなっていないか（天井効果により差が潰れていないか）に着目する。図 1 より、GPT-5 mini 等は人手の傾向に近く、タスク差や回答品質差に応じて分布の山が移動し、分布の分離も一定程度確認できる。一方、GPT-4.1 mini はタスクや回答品質の違いに関わらず高スコア帯に分布が集中する傾向があった。

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 付録
 研究コース 5 (Q4AI)

6.1.4 結果詳細抜粋

Claude Haiku 4.5 と GPT4.1-mini で最も CARE スコアの差が大きかった質問について、各タスク実際の結果を示す。

ID	回答 LLM	評価者 (LLM・人間)	正確性			説明性			自信度	再現性			重み付け加重和	
			情報の真実性	質問への関連性・適合性	論理的な一貫性	根拠の提示と説明性	前提・適用範囲・限界の明示	追跡可能性	自信度 (過信度)	同一質問再現性	パラフレーズ耐性	根拠・判断基準の再現性		
PDF-7	Claude4.5-haiku	Claude4.5-haiku	5	5	5	5	4	4	2	2	1	1	78.2	
		Claude4.5-haiku (拡張思考 ON)	2	2	2	3	4	3	1	2	2	2	40.6	
		GPT4.1-mini	5	5	5	5	5	5	5	2	2	3	85.7	
		GPT5-mini	2	2	5	5	4	5	2	2	1	2	46.8	
		人間 (専門家)	1	5	5	5	5	5	2	2	2	2	44.3	
		人間 (非専門家)	1	4	5	5	2	5	1	1	1	2	36.0	
	GPT4.1-mini	Claude4.5-haiku	5	5	5	5	4	4	5	4	4	4	94.4	
		Claude4.5-haiku (拡張思考 ON)	5	5	5	5	3	4	4	4	4	4	92.7	
		GPT4.1-mini	5	5	5	5	5	5	5	5	4	5	98.1	
		GPT5-mini	5	5	5	5	4	5	5	5	5	5	99.7	
		人間 (専門家)	5	5	5	5	4	5	5	5	5	5	99.7	
		人間 (非専門家)	5	5	5	5	5	5	5	5	5	5	100.0	
	衛星画像-tokyo2	Claude4.5-haiku	Claude4.5-haiku	2	4	3	4	4	4	3	2	1	1	46.5
			Claude4.5-haiku (拡張思考 ON)	3	4	3	4	4	4	3	2	1	1	55.7
GPT4.1-mini			4	5	5	5	5	5	5	3	2	2	78.5	
GPT5-mini			4	4	3	4	4	4	5	2	1	2	68.1	
人間 (専門家)			1	5	5	5	3	5	3	2	1	2	43.1	
人間 (非専門家)			1	3	2	3	4	3	3	3	1	3	36.6	
GPT4.1-mini		Claude4.5-haiku	3	3	4	4	4	4	4	4	3	4	67.4	
		Claude4.5-haiku (拡張思考 ON)	4	3	4	4	4	4	4	3	3	3	73.6	
		GPT4.1-mini	5	5	5	5	5	5	5	5	5	5	100.0	
		GPT5-mini	5	5	4	4	4	4	5	4	4	4	92.0	
		人間 (専門家)	5	5	5	5	5	5	5	4	4	5	95.5	
		人間 (非専門家)	3	2	3	3	3	4	5	4	4	5	66.9	

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 付録 研究コース 5 (Q4A1)

6.1.5 実験コスト

今回の実験で発生するコストについて示す。ただし実行時間については LLM へ逐次投入処理を行っており、並列処理によって大幅な改善が見込まれる。

なお下表については、コスト情報取得のために単独で実行した結果を示している。LLM の特性上、本文における結果の導出プロセスで発生した実験コストとは多少の前後が生じている。

タスク名	プロセス	発生費用	LLM 処理回数	処理時間	入力トークン数	出力トークン数
PDF 画像タスク	回答	\$0.1	20	2.8 分	182,350	6,932
	評価	\$8.8	940	215 分	9,080,129	767,982
	合計	\$8.9	960	218 分	9,262,479	774,914
衛星画像タスク	回答	\$0.1	20	4.0 分	59,812	14,614
	評価	\$6.6	940	244 分	4,076,827	1,014,978
	合計	\$6.7	960	248 分	4,136,639	1,029,592

6.2 考察

課題①～③に関し、本実験結果を基に考察した結果を示す。

6.2.1 評価コストと一貫性

LaaJ において LLM モデルに起因するバイアスとして、主に以下の 2 種類が想定される。

- モデル種別によるバイアス：GPT 系モデルと Claude 系モデル間の評価傾向
- 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

本項では、a) の「モデル種別によるバイアス」が本実験において生じていたかどうかを評価した。なお実験結果より、gpt4.1-mini は過大評価傾向が強いことが確認されたため、GPT 系の評価値としては gpt5-mini の結果のみを用いた。

モデル種別によるバイアスを定量化する指標として、以下の式を定義する。

$$R_i = \frac{E_{G \rightarrow G,i} + E_{G \rightarrow C,i}}{2} - \frac{E_{C \rightarrow C,i} + E_{C \rightarrow G,i}}{2} \quad \dots \text{式(6-1)}$$

ここで、 $E_{G \rightarrow G,i}$ 、 $E_{G \rightarrow C,i}$ はそれぞれ GPT 系モデルの回答に対する GPT 系および Claude 系モデルの評価結果を示す。同様に、 $E_{C \rightarrow G,i}$ 、 $E_{C \rightarrow C,i}$ は Claude 系モデルの回答に対する評価結果である。この指標 R_i は、同一系統モデルによる自己強化バイアス (b) を平均化によって相殺した上で、モデル種別による評価の偏りを抽出することを目的としている。 $R_i > 0$ であれば GPT 系モデルに有利なバイアス、 $R_i < 0$ であれば Claude 系モデルに有利なバイアスが存在することを示す。評価項目別に算出した R_i を表 1 に示す。なお、絶対値が 0.5 (合計スコアは 10) を超える項目については、バイアスが顕著であると判断し黄色で示した。

表 1 モデル種別によるバイアスの可視化

大項目	小項目	PDF タスク	衛星画像タスク
	自信度 (過信度)	0.33	0.10
正確性	情報の真実性	1.03	0.00
	質問への関連性・適合性	1.10	0.28
	論理的な一貫性	0.43	0.23
再現性	同一質問再現性	1.43	-0.68
	パラフレーズ耐性	0.73	-0.13
	根拠・判断基準の再現性	0.78	-0.05
説明性	根拠の提示と説明性	0.55	0.03
	前提・適用範囲・限界の明示	0.10	-0.03
	追跡可能性	0.05	0.03
合計スコア		18.11	-1.11

PDF タスクにおいては、正確性および再現性に関する複数の評価項目で、5 段階評価においてお

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 付録
 研究コース 5 (Q4A1)

およそ 1 程度 (合計スコアで 20 程度), GPT 系モデルに有利な過大評価傾向が確認された. 一方, 衛星画像タスクでは同様の傾向はほとんど見られなかった. この要因として, GPT 系モデルの回答品質が Claude 系モデルよりも良いことを示す結果であると考えられるため, この実験結果でモデル種別のバイアスを評価することは難しいと考えられる.

6.2.2 品質観点の定義の妥当性

品質観点の評価に用いた指標値の一覧を示す.

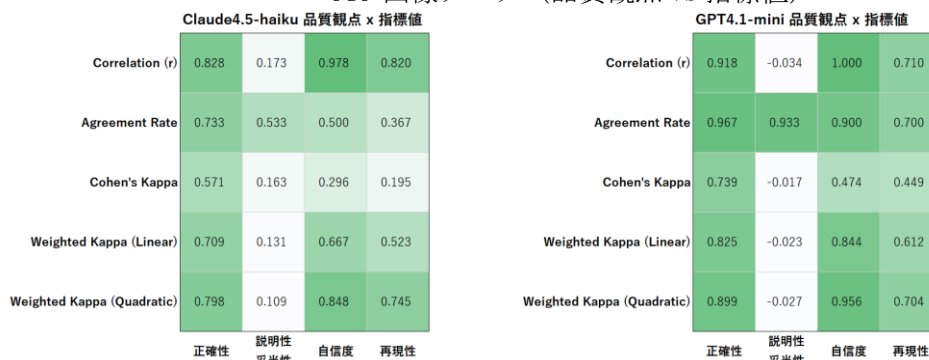
表 2 品質観点別の人手評価との相関係数と重み付きカッパ係数 (PDF 画像タスク)

		Correlation (r)	Weighted Kappa (Quadratic)
正確性	Claude Haiku 4.5	0.247	0.223
	Claude Haiku 4.5 Think	0.443	0.370
	GPT-4.1 mini	0.152	0.111
	GPT-5 mini	0.617	0.600
説明性・妥当性	Claude Haiku 4.5	0.234	0.183
	Claude Haiku 4.5 Think	0.131	0.082
	GPT-4.1 mini	0.317	0.137
	GPT-5 mini	0.269	0.222
自信度	Claude Haiku 4.5	0.434	0.394
	Claude Haiku 4.5 Think	0.614	0.525
	GPT-4.1 mini	0.412	0.196
	GPT-5 mini	0.595	0.583
再現性	Claude Haiku 4.5	0.814	0.751
	Claude Haiku 4.5 Think	0.799	0.724
	GPT-4.1 mini	0.792	0.754
	GPT-5 mini	0.832	0.809

表 3 品質観点別の人手評価との相関係数と重み付きカッパ係数 (衛星画像タスク)

		Correlation (r)	Weighted Kappa (Quadratic)
正確性	Claude Haiku 4.5	0.132	0.089
	Claude Haiku 4.5 Think	0.137	0.079
	GPT-4.1 mini	0.309	0.090
	GPT-5 mini	0.060	0.063
説明性・妥当性	Claude Haiku 4.5	-0.166	-0.117
	Claude Haiku 4.5 Think	-0.236	-0.129
	GPT-4.1 mini	0.000	0.000
	GPT-5 mini	-0.045	-0.040
自信度	Claude Haiku 4.5	0.226	0.257
	Claude Haiku 4.5 Think	0.045	0.094
	GPT-4.1 mini	0.000	0.000
	GPT-5 mini	0.445	0.420
再現性	Claude Haiku 4.5	0.585	0.299
	Claude Haiku 4.5 Think	0.569	0.288
	GPT-4.1 mini	0.464	0.418
	GPT-5 mini	0.697	0.467

PDF 画像タスク (品質観点 vs 指標値)



衛星画像タスク (品質観点 vs 指標値)

Claude4.5-haiku 品質観点 x 指標値

Correlation (r)	0.626	0.285	0.530	0.668
Agreement Rate	0.433	0.267	0.500	0.300
Cohen's Kappa	0.247	0.068	0.383	0.130
Weighted Kappa (Linear)	0.384	0.127	0.375	0.444
Weighted Kappa (Quadratic)	0.527	0.182	0.397	0.648
	正確性	説明性 妥当性	自信度	再現性

GPT4.1-mini 品質観点 x 指標値

Correlation (r)	0.350	0.111	0.376	0.550
Agreement Rate	0.333	0.267	0.500	0.367
Cohen's Kappa	0.048	-0.028	0.206	0.224
Weighted Kappa (Linear)	0.198	0.015	0.298	0.387
Weighted Kappa (Quadratic)	0.328	0.053	0.370	0.542
	正確性	説明性 妥当性	自信度	再現性

PDF 画像タスク (品質項目 vs 指標値)

Claude4.5-haiku 品質項目 x 指標値

Correlation (r)	0.515	0.163	0.558	-0.076	0.173	0.524	0.530	0.735	0.826	0.427
Agreement Rate	0.600	0.300	0.400	0.400	0.100	0.300	0.500	0.200	0.400	0.300
Cohen's Kappa	0.437	0.041	0.016	0.155	-0.047	-0.029	0.383	0.024	0.178	0.146
Weighted Kappa (Linear)	0.444	0.041	0.074	0.052	0.030	0.070	0.375	0.419	0.545	0.295
Weighted Kappa (Quadratic)	0.502	0.032	0.160	-0.019	0.087	0.180	0.397	0.686	0.811	0.365
	情報の真実性	質問への関連性・適合性	論理的な一貫性	根拠の提示と説明性	前提・適用範囲・限界の明示	追跡可能性	自信度 (過信度)	同一質問再現性	パラフレーズ耐性	根拠・判断基準の再現性
	正確性		説明性・妥当性		自信度		再現性			

GPT4.1-mini 品質項目 x 指標値

Correlation (r)	1.000		1.000				1.000	1.000	0.603	0.621
Agreement Rate	1.000	1.000	0.900	0.900	0.900	1.000	0.900	1.000	0.700	0.400
Cohen's Kappa	1.000		0.474	0.000	0.000		0.474	1.000	0.500	0.077
Weighted Kappa (Linear)	1.000		0.474	0.000	0.000		0.844	1.000	0.533	0.365
Weighted Kappa (Quadratic)	1.000		0.574	0.000	0.000		0.956	1.000	0.585	0.548
	情報の真実性	質問への関連性・適合性	論理的な一貫性	根拠の提示と説明性	前提・適用範囲・限界の明示	追跡可能性	自信度 (過信度)	同一質問再現性	パラフレーズ耐性	根拠・判断基準の再現性
	正確性		説明性・妥当性		自信度		再現性			

衛星画像タスク (品質項目 vs 指標値)

Claude4.5-haiku 品質項目 x 指標値

Correlation (r)	0.515	0.163	0.558	-0.076	0.173	0.524	0.530	0.735	0.826	0.427
Agreement Rate	0.600	0.300	0.400	0.400	0.100	0.300	0.500	0.200	0.400	0.300
Cohen's Kappa	0.437	0.041	0.016	0.155	-0.047	-0.029	0.383	0.024	0.178	0.146
Weighted Kappa (Linear)	0.444	0.041	0.074	0.052	0.030	0.070	0.375	0.419	0.545	0.295
Weighted Kappa (Quadratic)	0.502	0.032	0.160	-0.019	0.087	0.180	0.397	0.686	0.811	0.365
	情報の真実性	質問への関連性・適合性	論理的な一貫性	根拠の提示と説明性	前提・適用範囲・限界の明示	追跡可能性	自信度 (過信度)	同一質問再現性	パラフレーズ耐性	根拠・判断基準の再現性
	正確性		説明性・妥当性		自信度		再現性			

GPT4.1-mini 品質項目 x 指標値

Correlation (r)	0.207	-0.199	0.742	0.241	-0.263	-0.333	0.376	0.447	0.313	0.407
Agreement Rate	0.300	0.300	0.400	0.400	0.000	0.400	0.500	0.300	0.400	0.400
Cohen's Kappa	0.157	-0.129	0.000	0.143	-0.149	-0.200	0.206	0.125	0.250	0.211
Weighted Kappa (Linear)	0.167	-0.100	0.091	0.118	-0.103	-0.200	0.298	0.273	0.265	0.259
Weighted Kappa (Quadratic)	0.204	-0.105	0.259	0.096	-0.068	-0.200	0.370	0.444	0.310	0.333
	情報の真実性	質問への関連性・適合性	論理的な一貫性	根拠の提示と説明性	前提・適用範囲・限界の明示	追跡可能性	自信度 (過信度)	同一質問再現性	パラフレーズ耐性	根拠・判断基準の再現性
	正確性		説明性・妥当性		自信度		再現性			

6.2.3 入力画像解釈の曖昧性

小項目ごとの相関係数、一致率及び 2 次重み付き Kappa 係数による比較を実施した。なお、値が正であり、かつ一方の評価者群に対する相関係数が他方の 50%未満となる項目を、傾向に顕著な差があるものとして黄色で示した。

表 4 評価者の専門度と LLM の評価結果との関係 (相関係数)

品質観点	回答 LLM モデル 品質項目	専門家		非専門家	
		PDF タスク	衛星画像 タスク	PDF タスク	衛星画像 タスク
	自信度 (過信度)	0.54	0.40	0.49	0.02
正確性	情報の真実性	0.38	0.24	0.35	-0.20
	質問への関連性・適合性	0.24	0.30	0.38	0.06
	論理的な一貫性	0.53	0.15	0.40	0.23
再現性	同一質問再現性	0.90	0.58	0.91	0.61
	パラフレーズ耐性	0.69	0.56	0.84	0.74
	根拠・判断基準の再現性	0.66	0.32	0.61	0.30
説明性	根拠の提示と説明性	0.49	-0.14	0.38	-0.06
	前提・適用範囲・限界の明示	0.34	0.16	-0.22	-0.11
	追跡可能性	0.44	-0.07	0.12	-0.08
	合計スコア	0.70	0.29	0.68	-0.28

表 5 評価者の専門度と LLM の評価結果との関係 (一致率)

品質観点	回答 LLM モデル 品質項目	専門家		非専門家	
		PDF タスク	衛星画像 タスク	PDF タスク	衛星画像 タスク
自信度 (過信度)		0.60	0.31	0.51	0.28
正確性	情報の真実性	0.48	0.20	0.44	0.19
	質問への関連性・適合性	0.61	0.30	0.59	0.26
	論理的な一貫性	0.61	0.28	0.59	0.32
再現性	同一質問再現性	0.75	0.28	0.68	0.35
	パラフレーズ耐性	0.44	0.33	0.53	0.50
	根拠・判断基準の再現性	0.36	0.19	0.44	0.22
説明性	根拠の提示と説明性	0.60	0.33	0.55	0.29
	前提・適用範囲・限界の明示	0.30	0.29	0.24	0.21
	追跡可能性	0.35	0.30	0.29	0.39
合計スコア		0.00	0.00	0.00	0.00

表 6 評価者の専門度と LLM の評価結果との関係 (2 次重み付き Kappa 係数)

品質観点	回答 LLM モデル 品質項目	専門家		非専門家	
		PDF タスク	衛星画像 タスク	PDF タスク	衛星画像 タスク
自信度 (過信度)		0.45	0.25	0.39	0.03
正確性	情報の真実性	0.34	0.14	0.31	-0.13
	質問への関連性・適合性	0.10	0.05	0.35	0.02
	論理的な一貫性	0.42	0.02	0.37	0.13
再現性	同一質問再現性	0.87	0.48	0.88	0.48
	パラフレーズ耐性	0.67	0.46	0.80	0.62
	根拠・判断基準の再現性	0.61	0.15	0.57	0.18
説明性	根拠の提示と説明性	0.32	-0.02	0.28	-0.03
	前提・適用範囲・限界の明示	0.19	0.05	-0.13	-0.04
	追跡可能性	0.24	-0.01	0.06	-0.08
合計スコア		-	-	-	-

一致率では専門家・非専門家で大きな差が見られなかった。一方で、相関係数や 2 次重み付き Kappa 係数のような一致していない評価の差も重視する指標では、PDF タスクでは説明性に、衛星画像タスクでは自信度や正確性に顕著な差が現れた。つまり、比較的評価が容易な回答に対しては LLM と一致していたが、比較的評価の難しい回答でばらつきが見られたことがわかる。