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

研究コース5「人工知能とソフトウェア品質」

Q4AIチーム

研究員	：長田　章良	（ヤンマーホールディングス株式会社）
	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）

目次

1．研究の背景

- マルチモーダルLLMとは？
  生成AIの一種であり、テキスト、画像、音声等の複数形式の情報をまとめて入出力できる大規模言語モデル（LLM）
  → 従来のLLM以上に用途が多様化し、実利用が進みつつある

マルチモーダルLLMの使用例

画像キャプション

「子供と犬が遊んでいます」

図付き文書読み取り

「正解は、Bのゾウです」

音声認識

「次はこのパーツを右側に取り付けてください」

視覚支援

「2m進むと左に階段があります」

画像入力タスク

1．研究の背景

- 画像入力タスクでは、利用者による出力の正誤判定が困難
  → 出力内容の追加検証や調査を行い、作業量がむしろ増える場合も
- 画像を入力する代表的な実務例を想定

① 縦書き・図表・段組み等を含む文書の読解

利用者：文書画像を扱う作業者

特徴：専門性は低いが、画像の識別・理解が困難

② 画像に関連する情報の取得補助

利用者：画像の解析者

特徴：専門性が高く、画像外の情報と照合が必要

- 既存の生成AIの評価手法では限界
  - 固定的なデータセットや単一の自動評価指標による評価
  - 人手評価手法

LLMの出力が「良い回答」であるかどうか、目安となる指標が必要

2．本研究の目的

利用者の意思決定を支援するために、
「良い回答」を測る指標（信頼度スコア）を算出する「枠組み」を提案する

- 本研究で提案する信頼度スコア

- 本研究のポイント①　：　「枠組み」としてロバストな設計手法を提案する
- 本研究のポイント②　：　人手評価手法を補完し得るものを目指す

3．研究課題

本研究で提案する信頼度スコアは、何を要求されるか？

- **明確な品質観点・評価基準**
  - 「良い回答」の定義がぶれないようにする
  - 観点ごとに人の感覚と近い基準で点数化できる
- **低コストな自動評価**
  - 人手評価のように作業時間も手間もかからない
- **複数の品質観点の適切な統合**
  - 観点ごとの重みの違いを考慮したスコアを示せる

本研究における課題

- **上記の要求を満たす信頼度スコア算出の「枠組み」を設計すること**
- **具体的な事例を用いて人手評価と比較し、「枠組み」の有効性を示すこと**

4．信頼度スコア算出の「枠組み」

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）

- **手順①：品質観点・評価基準の明確化**
  - 品質観点：QA4AIガイドライン※より重要な観点を抽出
  - 評価基準：人手評価の実施者同士で話し合い決定
- **手順②：LLM-as-a-Judgeの実行**
  - LLMの回答を複数LLMにより評価する手法
    → 人手評価のコスト削減
    → 複数のLLMを用いることでばらつきを削減
- **手順③：複数品質観点の統合・スコア化**
  - 適切な重みづけにより利用者の重視度の違いを反映
  - 複数の評価基準の統合が可能

※ AI品質評価に関する国内ガイドライン
（https://raw.githubusercontent.com/qa4ai/Guidelines/refs/heads/main/QA4AI.Guidelines.202504.pdf）を指す

5．実験設定概要①

JRVF手順①：品質観点・評価基準の明確化

品質観点：QA4AIの品質観点から、以下の理由で大項目4種・小項目10種を採用

- 正解ラベルを用いず自動評価が可能
- 商用モデルで既に確立していない

[C] 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

[A] 正確性 (Accuracy)

出力内容がどの程度事実と合致しているか

小項目：

- 情報の真実性
- 質問への関連性・適合性
- 論理的な一貫性

[R] 再現性 (Reproducibility)

同様の質問に対して出力が一貫しているか

小項目：

- 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

[E] 説明性・妥当性 (Explainability)

根拠や説明が結論のために必要十分か

小項目：

- 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

CARE スコアと定義 = [C]onfidence + [A]ccuracy + [R]eproducibility + [E]xplainability

5．実験設定概要①

JRVF手順①：品質観点・評価基準の明確化

評価基準：品質観点の小項目10種類ごとに、人手評価のばらつき（解釈のブレ）が生じにくい粒度として5段階評価を採用し、各段階の明確な基準を協議により決定

例) 質問と評価点別の評価基準（大項目：正確性 / 小項目：情報の真実性)

質問 / スコア	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	すべ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。

5．実験設定概要②

JRVF手順②：LLM-as-a-Judge の実行

特定のモデルやタスク特性による評価のばらつきを防ぐため、ベースモデルや回答生成過程の違いを考慮し4種のLLMを採用し、回答と評価で異なる役割を設定し評価

回答用 LLM

- Claude Haiku 4.5
- GPT-4.1 mini

評価用 LLM

- Claude Haiku 4.5
- GPT-4.1 mini
- Claude Haiku 4.5 (拡張思考モードON)*
- GPT-5 mini *

*よりロバストな評価を実現するため、評価側には高度な思考プロセスを持つ「推論モデル」を追加

5．実験設定概要③

JRVF手順③：複数品質観点の統合・スコア化

「LLM の回答を判断する際、何を重視すべきか？」という主観を定量化するため、AHP (階層化意思決定法) で求めた重みを用いた重みづけを採用

大項目同士、小項目同士でどちらがより重要かを定性的に評価
人手評価の実施後に重みを決定し、利用者の重視度をより反映

大項目	非常に重視する	重視する	やや重視する	どちらかというと重視する	同等	どちらかというと重視する	やや重視する	重視する	非常に重視する	
正確性	○									説明性
正確性			○							自信度
正確性					○					再現性
説明性								○		自信度
説明性									○	再現性
自信度							○			再現性

大項目	正確性	説明性	自信度	再現性	幾何平均
正確性	1	9	5	1	2.59
説明性	0.1	1	0.14	0.11	0.20
自信度	0.2	7	1	0.2	0.73
再現性	1	9	5	1	2.59

重み（大項目）

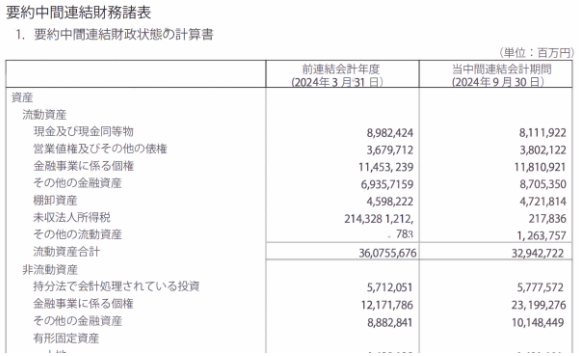
5 .実験設定概要④

実利用を想定した2種類の画像タスク (各タスクで画像と質問のセットを10個ずつ) を設定し、質問への回答に対し、JRVFと人手評価*の結果を比較

PDF 画像タスク

特徴：複雑な文書・図表の迅速な把握が困難

画像と質問のセット例



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）

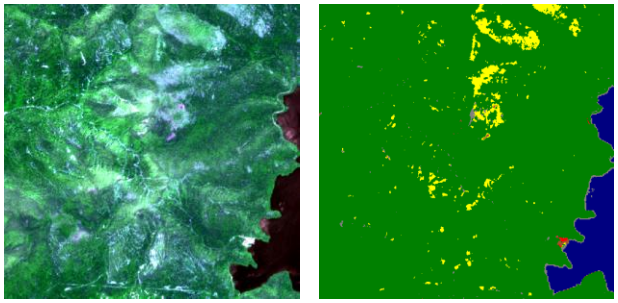
	前連結会計年度 (2024年３月31日)	当中間連結会計期間 (2024年９月30日)
資産		
流動資産		
現金及び現金同等物	8,982,424	8,111,922
営業債権及びその他の債権	3,679,712	3,802,122
金融事業に係る債権	11,453, 239	11,810,921
その他の金融資産	6,935,7159	8,705,350
棚卸資産	4,598,222	4,721,814
未収法人所得税	214,328 1,212,	217,836
その他の流動資産	. 783	1, 263,757
流動資産合計	36,0755,676	32,942,722
非流動資産		
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		

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？

衛星画像タスク

特徴：画像が不確実で正解情報の収集が困難

画像と質問のセット例



Contains modified Copernicus Sentinel data 2024

提供：高解像度土地利用土地被覆図(JAXA)

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。

＊ 人手評価は各タスク2名（専門家・非専門家）で実施

実験結果：CARESコア①

CARESコア（平均値一覧）

JRVF（4種のモデル）、及び人手評価(4名)によって算出されたCARESコアの平均値を示す。

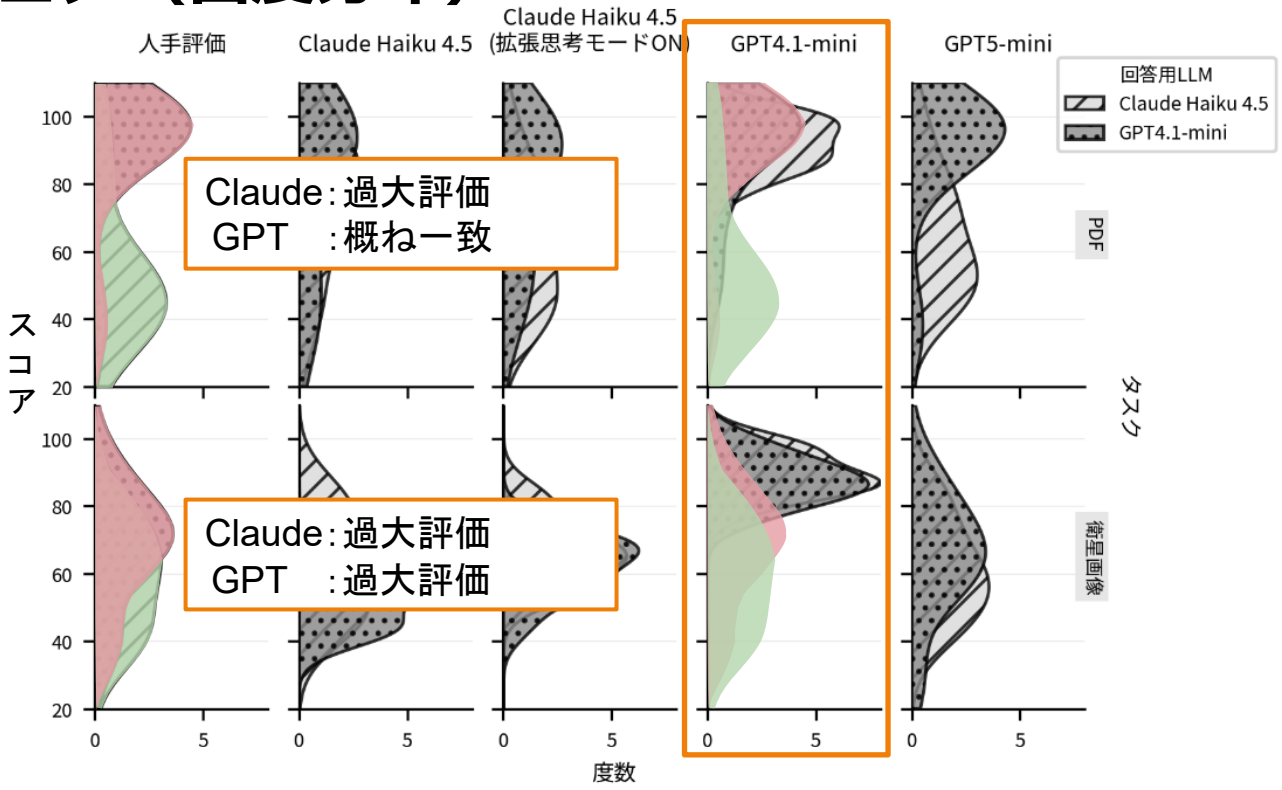
タスク ＼ 回答用LLM	Claude Haiku 4.5		GPT-4.1 mini	
	JRVF	人手	JRVF	人手
PDF画像タスク	73.2	54.5	84.3	91.0
衛星画像タスク	69.9	58.2	67.1	67.9

- JRVFによるPDFスコアが相対的に高かった
  → 衛星画像タスクの専門性が高かった（緯度経度情報、数値計算、etc.）ため
- 両タスクに共通してGPT-4.1 miniのスコアが高かった
  → **Claude系モデルは画像が読み取れていなかったため**

各モデルにはスコアのみならず
点数の理由も併記させたために判明

実験結果：CARESコア②

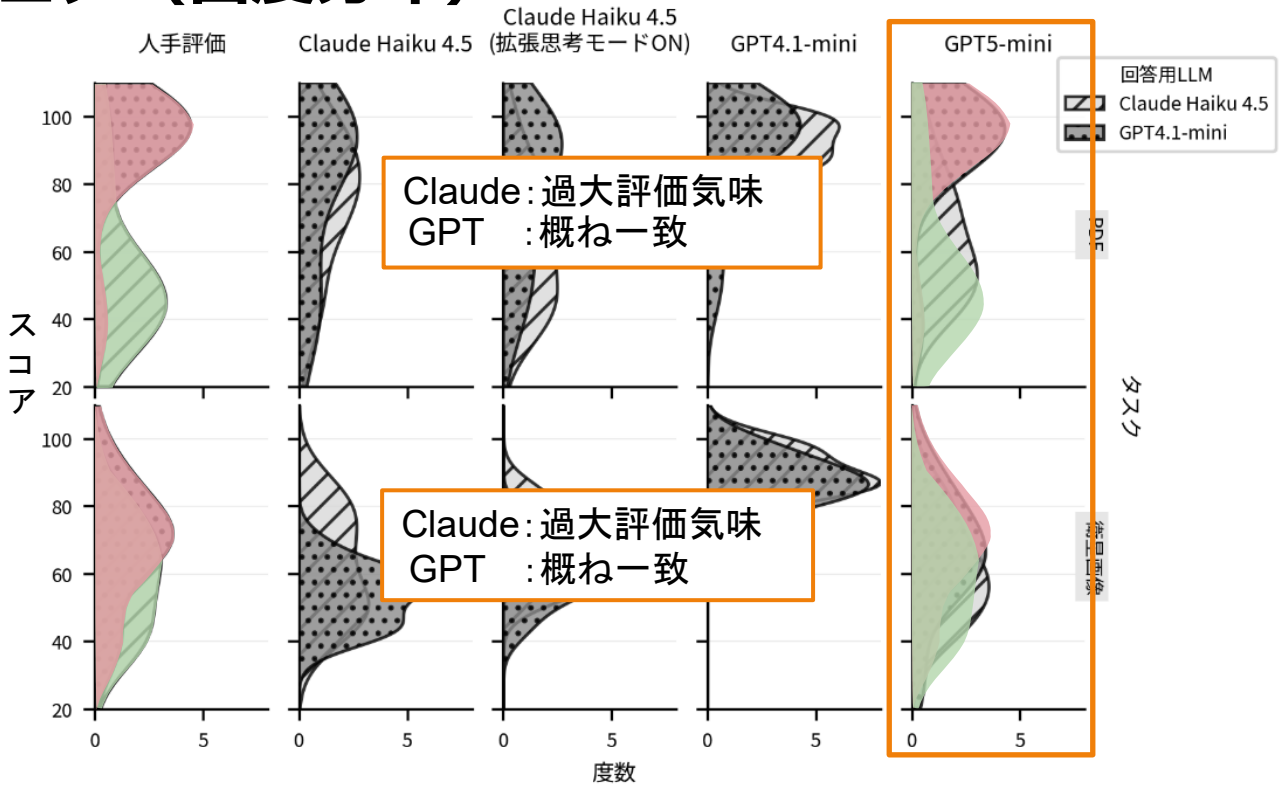
CARESコア（密度分布）



- タスク差や回答品質差に応じて分布のピーク位置やその大きさの違いを確認

実験結果：CAREスコア②

CAREスコア（密度分布）



- タスク差や回答品質差に応じて分布のピーク位置やその大きさの違いを確認

→ 推論モデルのほうがより人手評価に近しい

アプローチの妥当性考察①

JRVF手順①：品質観点・評価基準の明確化

品質観点（大項目）別に人手評価とLLM評価（平均値）との相関係数を示す。

	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**

衛星画像タスクでは、
専門家と非専門家で
相関係数に差が見られた
⇒ 専門性の高さが影響？

- **再現性が最も相関が高く、比較的評価基準の認識が一致した**
- 説明性・妥当性はいずれのタスクでも相関が低い

→ LLMと人間（或いは人間同士）の評価基準認識がばらついた可能性

  - 説明性・妥当性の評価基準のさらなる具体化・詳細化の検討
  - 採用した4種類以外の観点の必要性を検討

アプローチの妥当性考察②

JRVF手順②：LLM-as-a-Judgeの実行

1問あたりのJRVFの処理時間と費用を以下に示す。
なお、人手評価にはいずれのタスクも1問当たり50分程度かかった。

タスク名	処理時間	LLM処理回数	1処理あたりの処理時間	発生費用
PDF画像タスク	21.8[分]	96	**13.6[秒]**	$0.89
衛星画像タスク	24.8[分]	96	**15.5[秒]**	$0.67

- **人手評価作業と比べて作業時間が50%未満**
  → 並列化によるさらなる作業高速化も期待される。
- LLMの進化に伴い、回答品質の高い（期待値を十分に満たす）モデルが安価で使用可能になることも期待される

アプローチの妥当性考察③

JRVF手順③：複数品質観点の統合・スコア化

AHPによる重みづけの有無による、CAREスコアの人手評価とLLM評価の平均値の相関係数をそれぞれ示す。

また、AHPの重みを閾値でフィルタリングした結果も併せて示す。

AHP利用方法 \ 回答用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
AHP重みづけなし	0.75	0.91	0.46	0.63
AHP重みづけあり	0.73	0.88	0.41	0.57
AHPフィルタリング	**0.80**	**0.93**	**0.60**	**0.74**

- AHPの有無によって結果に大差は見られなかった
  → 品質観点ごとの重みを考慮した評価基準であったため
- **AHPの重みでフィルタリングをすると相関係数が良化した**
  → 比較的相関の低い観点（説明性・妥当性）を除外できたため

まとめ

- 本研究では、マルチモーダルLLMの画像入力タスクにおいて、利用者が「良い回答」を判断するための「信頼度の目安」を提示する評価枠組み（JRVF）を提案した。
- JRVFにより算出されるCAREスコアの有効性を検証した。
- 特にPDFタスクにおいて、人手評価との間に高い相関（相関係数 0.88）が確認された。

LLMが継続的に進化している現状を踏まえると,
JRVFによる評価の信頼性は今後さらに向上する可能性がある。
JRVFとの相性がより良い品質観点・LLMモデル・統合手法を模索していきたい！

今後の展望

課題解決に向け、今後検討の必要がある項目を課題別に示す。

- **説明性・妥当性の評価基準の曖昧さ**
  - CAREスコアに用いた品質観点が必要十分であるか検証する
- **評価LLM・スコア統合手法の一般化**
  - LLMモデル別に評価した結果の統合手法（重みづけの有無など）

謝辞

本論文の執筆に際し、以下の方々にご指導賜りました。
厚く御礼申し上げます。

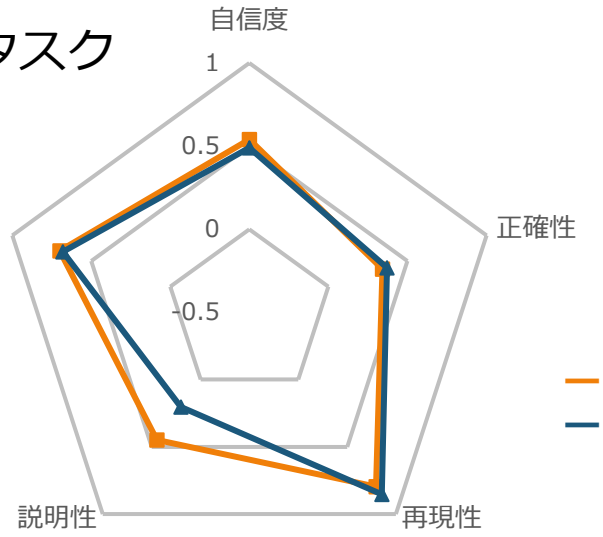
- 主査　　：石川　冬樹（国立情報学研究所）
- 副主査　　：徳本　晋　　（富士通株式会社）
- アドバイザー：栗田　太郎（フリー株式会社）

想定質問回答：評価者によるばらつき

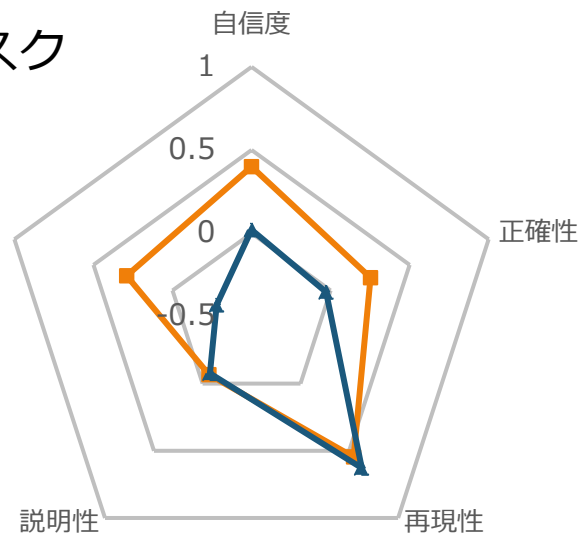
JRVF手順①：品質観点・評価基準の明確化

評価者別に人手評価とLLM評価（平均値）との相関係数の分布を示す。

PDF画像タスク



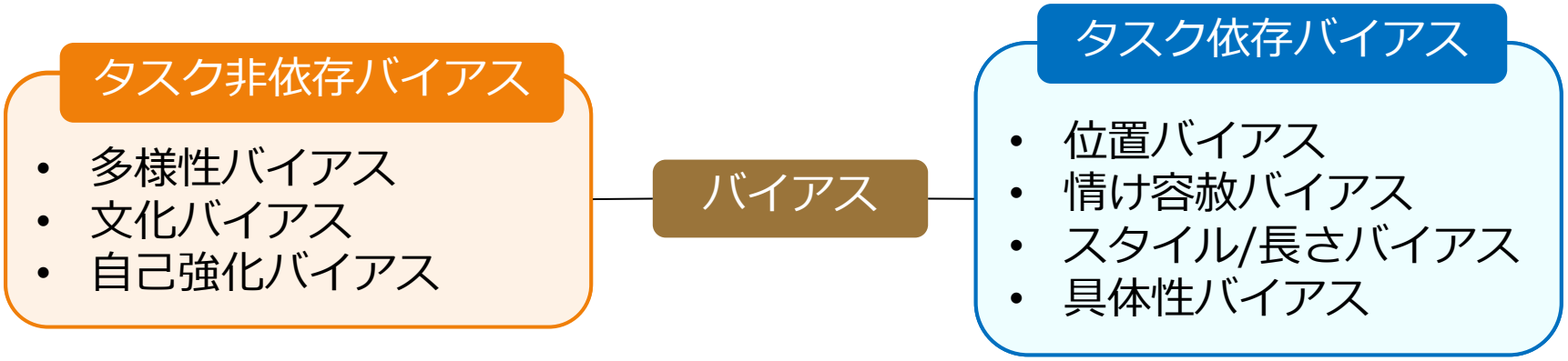
衛星画像タスク



- PDF画像タスクでは専門家と非専門家で大差ない
  → 専門性の低いタスクでは、専門性に寄らない評価基準によって評価できた。
- 衛星画像タスクでは専門家の評価との相関係数がより高い
  → **専門性の高いタスクでも、LLMは専門家により近い回答をする傾向がある。**

想定質問回答：LLM-as-a-Judgeの偏り

LLM-as-a-Judgeには、バイアス（先入観や偏見）がある※



- タスクや質問、回答の内容に依存するバイアス
  → ロバストな手法を検討しているため、詳細に分析しなかった。
- モデル種別による影響の大きいバイアス
  → ClaudeとChatGPTでは画像入力タスクに対する回答性能が大きく異なる。
- **自己強化バイアス（ 自身によって生成された回答を好む傾向）に着目**

※ A Survey on LLM-as-a-Judge(2024), J.Gu et al.

想定質問回答：LLM-as-a-Judgeの偏り

LLMの回答に対し、同系統モデルによる評価と異なる系統による評価で**Wilcoxon符号付順位検定**を実施し、出力値が0.1未満の場合は有意差があると見なした

『2つの条件で中央値に差がない』という仮定のもとで、
実際に観測された差が偶然生じる確率を求める手法

大項目	回答LLMモデル / 小項目	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
CAREスコア		0.31	0.25	0.01	0.01

- GPT系統モデルの回答に対する評価ではバイアスを確認
  → Claudeは画像が読めない等、回答性能が低いためバイアスが目立たない？
- **同系統モデルの評価を過信しないようにするため、重みづけ等を検討**

想定質問回答：評価LLM統合手法

LLMモデルによるバイアスが考えられるため、複数系統のモデルによる評価が必要
→どう統合すると人手評価との相関が上がるか？

回答用LLM / LLM統合方法	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
平均値	0.73	0.88	0.41	0.57
最大値	0.09	0.99	0.43	0.45
最小値	0.85	0.61	0.41	0.54

- 最大値：高品質の回答が多いタスクへの評価では人手評価と高い相関を示す
- 最小値：低品質の回答が多いタスクへの評価では人手評価と高い相関を示す
- 平均値：回答品質によらずロバストに人手評価と相関がみられる

複数モデルの評価結果を平均値として統合しCAREスコアは、人間の評価と相関があると言える

→ **LLMの出力を評価する信頼性指標として有効である**

- 重みづけ等を用いた最適化を検討する必要がある