MIRU2026 若手プログラム 「わかりやすい，読みやすい」を深掘る

ベンチマーク論文をベンチマークする

田上 鈴奈（中京大学） 趙 在瀛（東京大学） 石山 遼（九州大学）
加藤 駿（慶應義塾大学） 堀田 南（早稲田大学） 千葉 倫太郎（慶應義塾大学）

Project page

ベンチマークとは：モデルの性能を評価するために用いられる，研究コミュニティにおける「**物差し**」的な存在

ベンチマークとデータセットの違い

- **データセット論文**：**データそのもの**が主役
- **ベンチマーク論文**：**評価の仕組み**や**品質**が主役（データは要素の1つ）

ベンチマークの4つのモチベーション

領域開拓型	欠落補完型	更新・難化型	評価再設計型
新しい物差し自体を作る	物差しで測れる範囲を広げる	物差しのレパートリーを増やす	物差しの妥当性を見直す

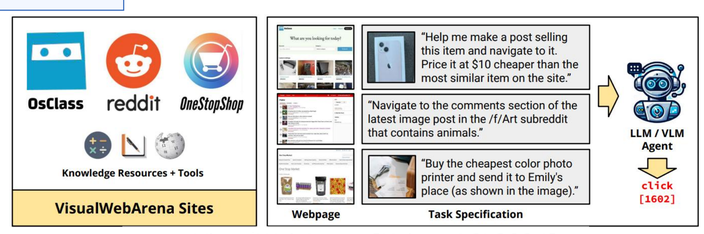
セクション別に見るベンチマーク論文の「わかりやすい，読みやすい」

Introduction

新しいタスク・評価対象 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

図１で多くを語る（読者が概要を早期につかめる）

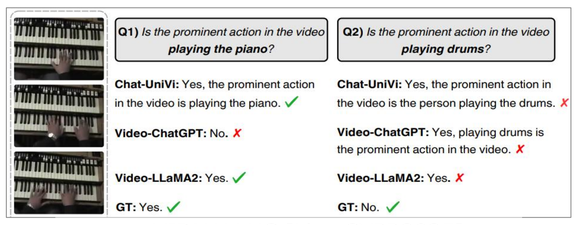
- 既存ベンチのタスクとの違いを記述・図示する
- 既存ベンチでは出せないSOTA手法の失敗例を主張する
- 評価の際の入出力を詳細に記述・図示する



VisualWebArena [Koh+, ACL2024]

既存ベンチマークとの差分 〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

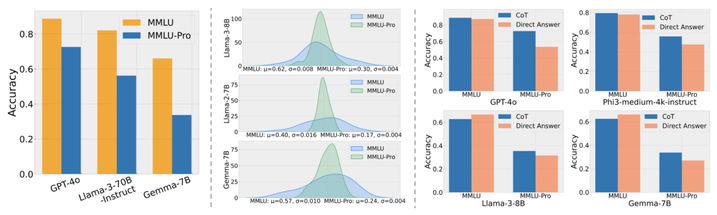
- 既存ベンチとのデータ数・カテゴリ数などの差や，提案ベンチによってSOTA手法ができなくなったタスクを一目でわかるようにする
- カバレッジ表などで提案ベンチが多数の従来ベンチよりも網羅性あることを示す



VIDHALLUC [Li+, CVPR2025]

結果等の要約 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

- シンプルなグラフで既存ベンチとの比較結果を示す
- Key findingsで，SOTA手法に対する改善の余地，手法間の性能差，手法に関する新たな知見などを端的に述べる

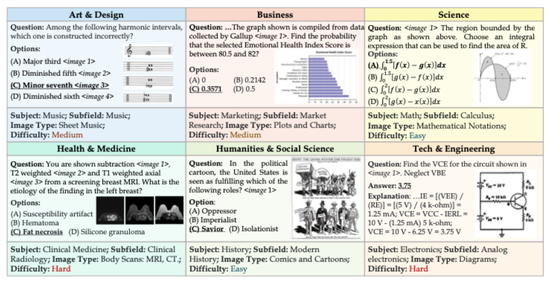


MMLU-Pro [Wang+, NeurIPS2024]

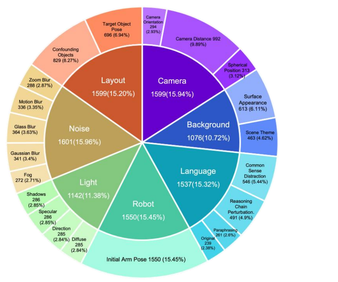
Method / Benchmark Curation

データの設計 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉

- 評価したい能力に対応するタスク・カテゴリを設計する
  - カテゴリ/ドメイン/難易度の分類を可視化し，どの能力範囲を測るのかを明確にする
  - 何を含め，何を含めないかの設計判断を説明する



MMMU [Yue+, CVPR2024]

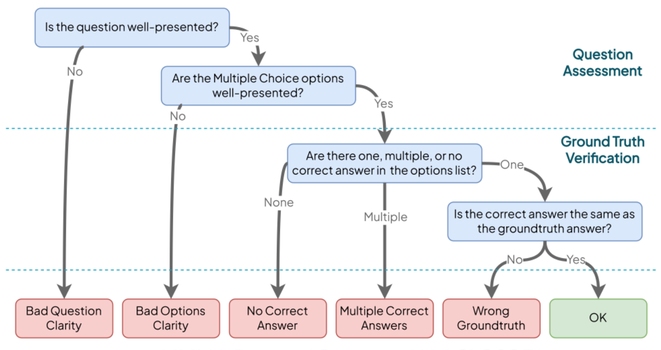


LIBERO-Plus [Fei+, CVPR2026]

品質の保証 〈更新・難化型〉〈評価再設計型〉

- データ作成/アノテーション/フィルタリングの流れを透明化する
- 人間や専門家がどの段階で何を確認したかを明示する
- エラー分類や除外基準を明確にし，評価結果がデータ品質に左右されないことを保証する

HLE [Phan+, Nature2026]

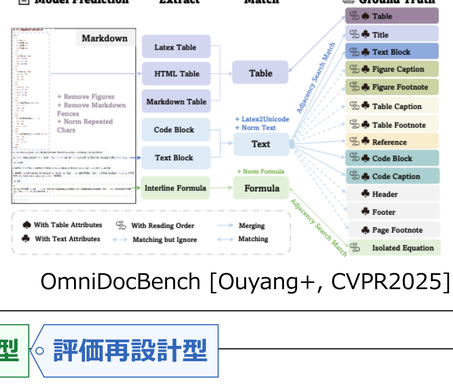


MMLU-Redux [Gema+, NAACL2025]

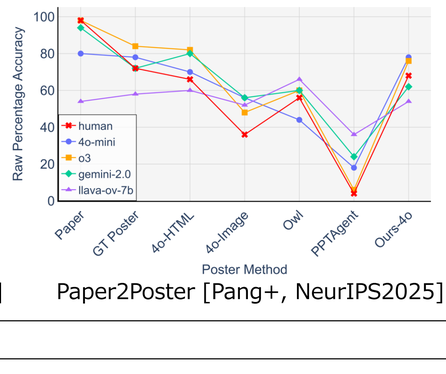
Experiments

評価指標とその妥当性 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

- データの入力形式，出力形式，評価方法までの評価プロトコルを明瞭に示す
- 評価指標の設計とその計算方法について詳述する
- 評価指標の妥当性と測定結果を示す（例：人間との整合性，カバレッジ，頑健性）



OmniDocBench [Ouyang+, CVPR2025]

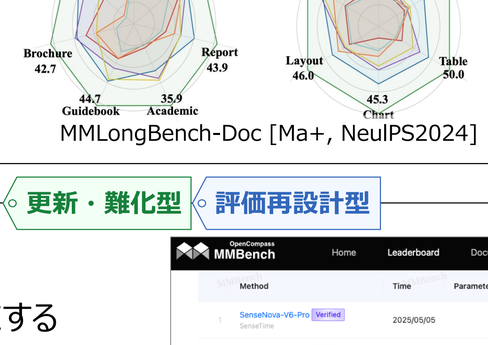


Paper2Poster [Pang+, NeurIPS2025]

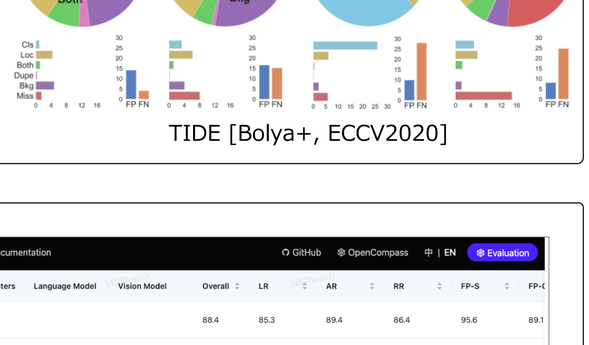
Discussion

評価結果の解釈 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

- カテゴリ/ドメイン/難易度ごとの性能差を示し，手法ごとの得意・不得意を明らかにする
  - 高スコアの要因を検証し，見かけの性能と真の能力を切り分ける
  - Failure Caseを分類し，手法の限界や失敗傾向を明らかにする



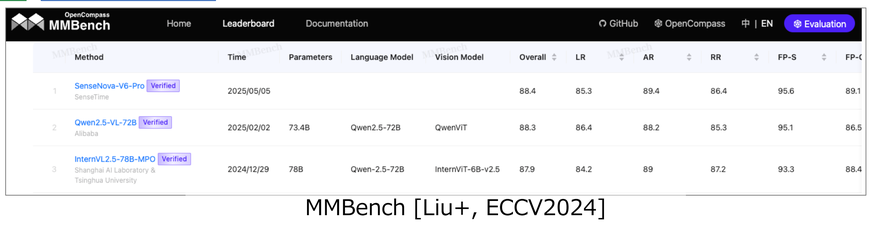
MMLongBench-Doc [Ma+, NeulPS2024]



TIDE [Bolya+, ECCV2020]

再利用性・発展性 〈領域開拓型〉〈欠落補完型〉〈更新・難化型〉〈評価再設計型〉

- リーダーボードや評価コードを公開し，同一条件下で手法間の性能を比較可能にする
- 評価設計の改善点や未評価の能力を整理し，後続研究への指針を示す



MMBench [Liu+, ECCV2024]

VLMで評価してみた

- **モデル**: GPT-5.5
- **評価方法**:
  論文をキャプチャし，モデルに入力して評価
- **評価基準**:左記の観点にしたがって，論文の「わかりやすさ」を0-5点で評価する．今回は，🤖の観点に対するコメントを抜粋する．

GPT-5.5

わかりやすさ: ･･/5.0
わかりやすさに繋がるポイント:

1. Introが･･･の構造になっている．
2. タイトルだけで主張がわかる．

･･･

5. Figure1だけで論文の概要が理解できる

Introduction

■ BLINK [Fu+, ECCV2024]

わかりやすさ: 4.2 / 5.0

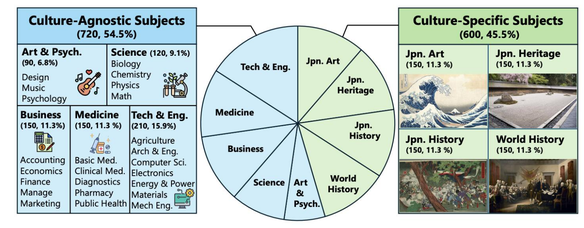
ポイントの抜粋: **Figure 1から，読者は本文で定義を読む前に，「画像中の細かな位置・形状・対応関係を直接判断させるベンチマークなのだ」と大まかに理解できます．**特に，各タスクを実際の選択問題として示している点が有効です．

Method / Benchmark Curation

■ JMMMU [Onohara+, NAACLS2025]

わかりやすさ: 4.2 / 5.0

ポイントの抜粋:著者らはMMMUの30科目を文化依存性で分類し，BiologyやChemistryなどは翻訳して継承する一方，ArtやHistoryなどは日本向けの科目に置き換えています．これにより，**言語の翻訳だけで対応できる部分と，内容自体を再設計すべき部分を区別した判断が明確に伝わります．**



Experiments

■ Paper2Poster [Pang+, NeurIPS2025]

わかりやすさ: 4.3 / 5.0

ポイントの抜粋:**見た目・文章品質・情報理解を複数の観点から評価しています．さらに，評価モデルを変えても傾向が保たれるか，人間判断と整合するかを確認**することで，評価指標の妥当性を示しています．

Discussion

■ LIBERO-Plus [Fei+, CVPR2026]

わかりやすさ: 4.5 / 5.0

ポイントの抜粋: 7種類の摂動と5段階の難易度で，VLAモデルがどこで崩れるかを明確化しています．特に**カメラ視点や初期状態に弱く，言語摂動への強さも「言語理解」ではなく，「言語を使っていない」可能性として分析**している点が有効です．