EMBEDLLM: LEARNING COMPACT REPRESENTATIONS OF LARGE LANGUAGE MODELS

Richard Zhuang, Tianhao Wu , Zhaojin Wen, Andrew Li, Jiantao Jiao, Kannan Ramchandran, 16 Oct 2024.

ICIR

背景

LLMの爆発的な増加と新たな課題

現在Hugging Faceには数十万のLLMが存在し，日々増え続けている.
その中でも以下の２つが課題として挙げられる.

- LLMの効率的な比較
  どのモデルがどのタスクに最適かどう判断するのか？

- 膨大なコスト問題
  “全モデルを各タスクで評価“は，時間とお金がかかりすぎる.

従来の手法

従来の手法と問題点

ベンチマーキング

各モデルを固定のテストセットでスコアを算出.

➤新規テストセットの登場のたびに１つずつ評価するため，多大な時間とコストがかかる.

モデルルーティング

タスクに応じて最適なモデルを選択するシステムを構築.

➤各タスクに特化したモデルを学習するため，汎用性が無く応用が効かない.

EmbedLLM

LLM自体を意味のあるベクトルで表現できないか？

特徴を抽出しているベクトルにすれば，それを様々な下流タスクに再利用も可能となるため.

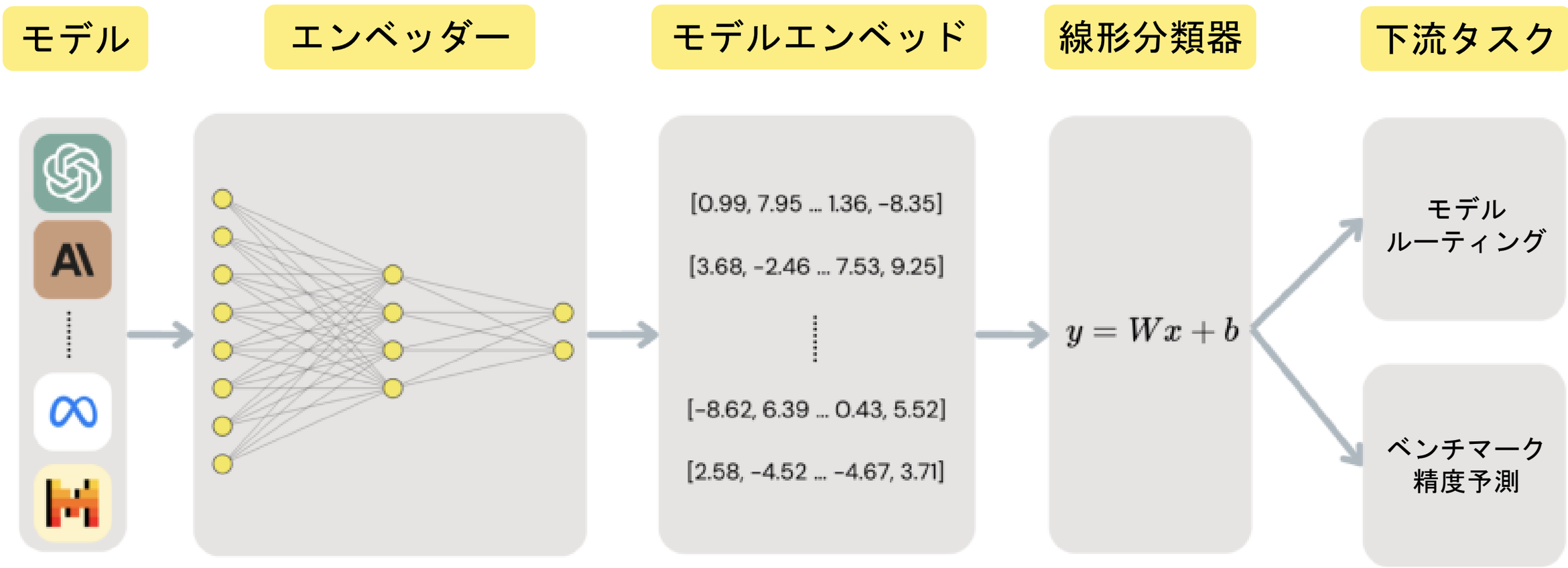
例：性能予測、モデルルーティング、ベンチマーク制度の評価など

➤各LLMをコンパクトなベクトルに変換するフレームワークであるEmbedLLMを製作した.

目的：モデル評価・活用の効率化とコスト削減

EmbedLLM

モデルの全体像



○エンベッドとは

単語や文，画像，AIモデルなどをベクトルに変換すること.

EmbedLLM

★エンコーダ・デコーダによる再構成タスクを行う

データ準備

“どのモデルがどの質問に正しく答えられたか”を記録した巨大な正解/不正解行列を用意する.

行：モデル，列：質問，値：0か1

エンコーダ（エンベッダー）

モデルが各LLMの特性，各質問をベクトルに圧縮する.

デコーダ（線形分類器）

圧縮されたモデル埋め込みと質問埋め込みから，元の正解/不正解行列を予測する.

EmbedLLM

使用するデータセット

LLMモデル

汎用・特化型を含む112個のオープンソースLLM

例：Llama， Mistral, CodeLlamaなど

ベンチマーク(データセット)

10種類の主要ベンチマークから質問を36054個

例：MMLU， GSM8K， MathQAなど

訓練データ

上記の全モデル×全ベンチマークでの正解/不正解ラベルを構築した.

EmbedLLM

今回焦点を当てる下流タスク

- 正解率予測

  未知の質問に対して各モデルが正解できるかを予測.

- モデルルーティング

  各モデルの埋め込みベクトルを利用して質問に最も適するモデルを効率的に選択できるか.

- ベンチマーク制度予測

  あるLLMを特定のベンチマークで実際に実行せずに，そのモデルがどの程度の精度を出すかを予測.

結果１：正解率予測

未知の質問に対しモデルが正解できるか.

EmbedLLMの手法（Matrix Factorization）とKNN分類器を比較

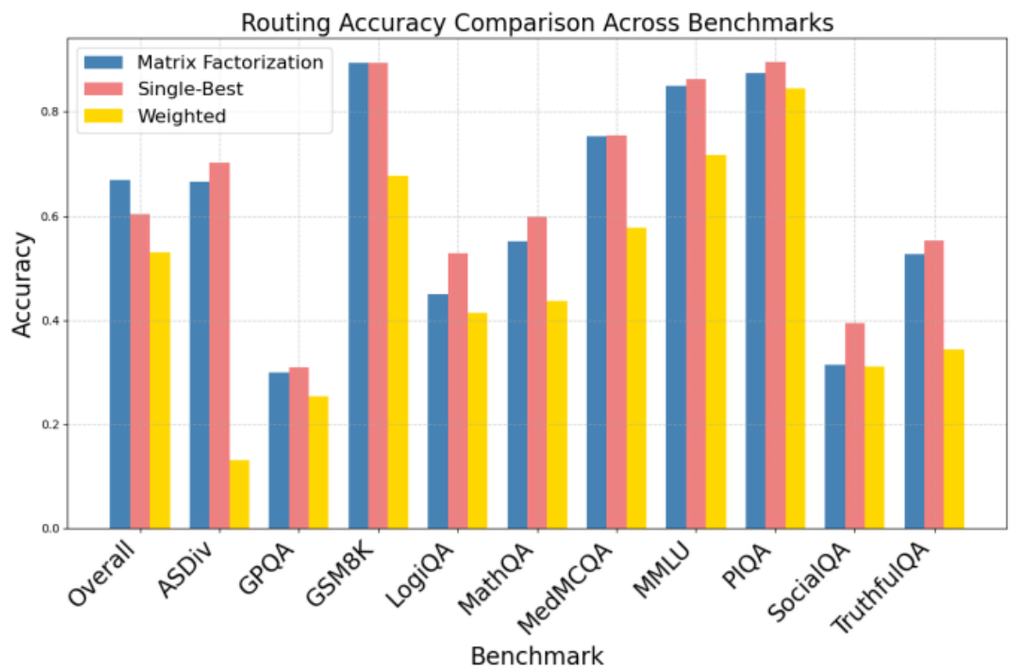
アルゴリズム	訓練時のデータセットのサイズ						
	1K	5K	10K	15K	20K	25K	Full(29K)
KNN	0.6372	0.7078	0.7107	0.7128	0.7143	0.7146	0.7152
MF	0.6443	0.7331	0.7362	0.7378	0.7390	0.7394	0.7409

➤訓練データサイズが大きいほど性能が向上する.

〇EmbedLLMの方が未知の質問への正答率が高い.

結果２：モデルルーティング

質問に最も適したモデルを効率的に選択できるか.

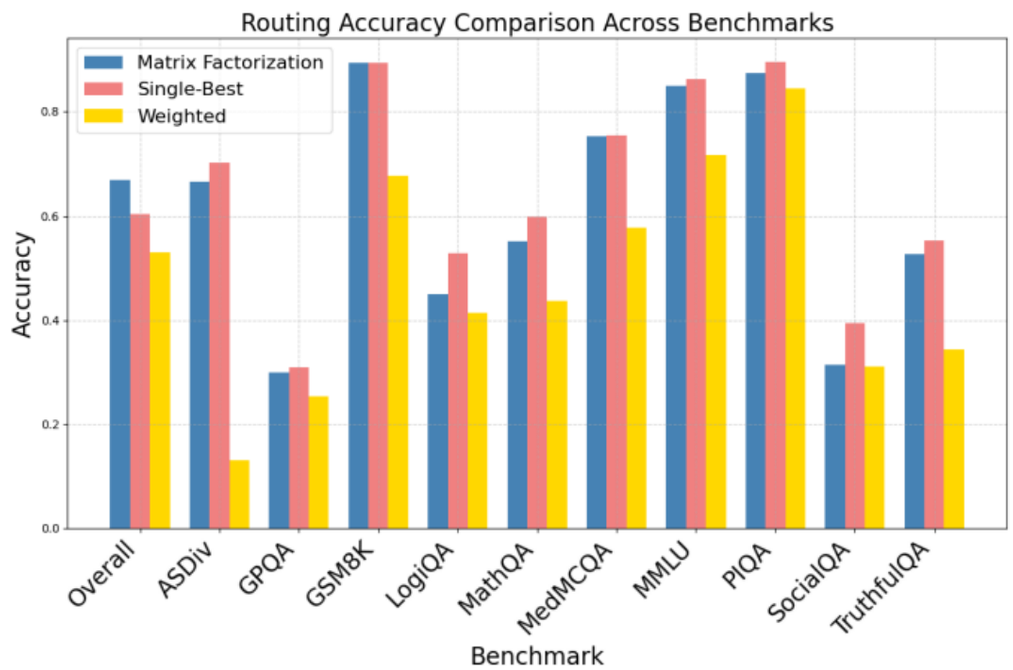


以下の3つでモデルルーティングの性能を比較した.

- Matrix Factorization：各モデルの埋め込みから正解できる確率を予測し選択.
- 単一最良モデル：常に一番性能が良いモデルを使い続ける手法.
- 加重ランダムモデル：MFルータと同じ選択比率でモデルをランダムに選択.

結果２：モデルルーティング

- 質問に最も適したモデルを効率的に選択できるか.



- 結果：MFルータが最も適切にモデルを割り当てた.
  - 圧倒的な速度

    3000の質問を3.8秒でルーティング（既存ルータの15倍以上高速）
  - 驚異的な低コスト

    訓練に必要なメモリは1GB未満（既存ルータの1/60の使用量）

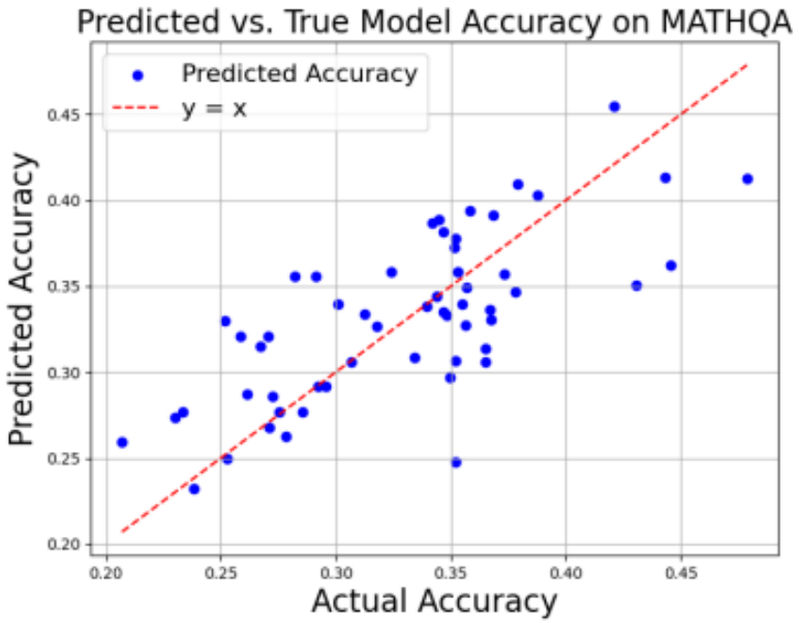
結果３：ベンチマーク精度予測

モデル埋め込みの性能予測は信用できるのか.

あるモデルを特定のベンチマークで実際にテストせずに，そのスコアを予測できるのか検証した.

ある特定のベンチマークのデータを除外して，モデル埋め込みに学習させ，それを特徴量として線形回帰モデルを訓練し，除外スコアを予測させる

Benchmark	Significance
MathQA	100
LogiQA	100
MedMCQA	100
PIQA	98
TruthfulQA	96
MMLU	94
GSM8K	93
GPQA	10
ASDiv	6
SocialQA	3



結果３：ベンチマーク精度予測

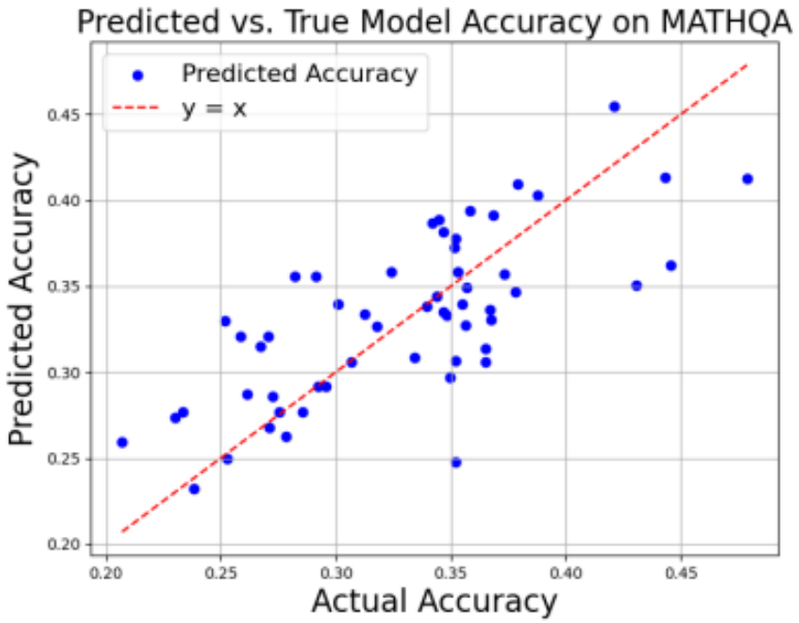
結果

評価した10個のベンチマークのうち，７個で統計的に有意な相関（正しく順序付けできている関係）が見られた.

特にMMLUのような大規模ベンチマーク（14000問）に対しても他のベンチマークのデータから精度を予測可能.

➤モデル埋め込みが未知のベンチマークに対する性能を予測するための有効な情報を含んでいることを示す.

Benchmark	Significance
MathQA	100
LogiQA	100
MedMCQA	100
PIQA	98
TruthfulQA	96
MMLU	94
GSM8K	93
GPQA	10
ASDiv	6
SocialQA	3



モデル埋め込みに含まれる情報とは？

モデル埋め込みはどのような意味のある情報を捉えているのかを探るため，２つのプロービング実験を行う.

実験１：類似性に対する検証

学習された埋め込みが，基本的な直観に合っているかを検証.

例：似た特性を持つモデルはベクトル空間でも近くに配置されるべき

実験２：学習データ特性に対する検証

埋め込みはそれを学習するために使用したデータの特性を反映しているかを検証.

例：特定分野のベンチマークを学習させるとその分野の能力が向上しているのか.

モデル埋め込みに含まれる情報とは？

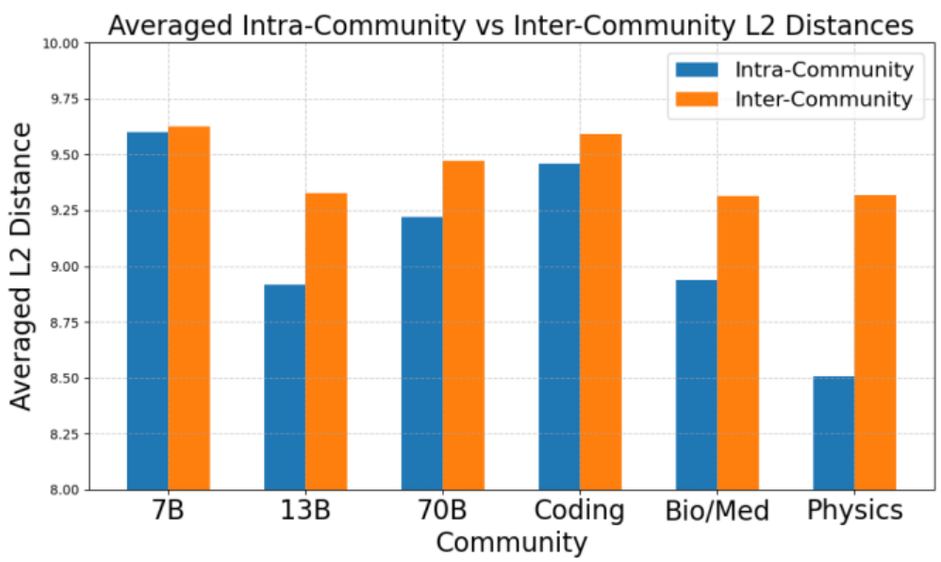
実験１：類似性に対する検証

モデルをその名称に含まれるキーワードに基づき６つのコミュニティに分類.
コミュニティ内とコミュニティ外のモデル間の平均距離を比較する.

結果

コミュニティ内の平均距離の方が短かった.

➤モデル埋め込みが基本的な特性を正しく捉えていることが分かる.

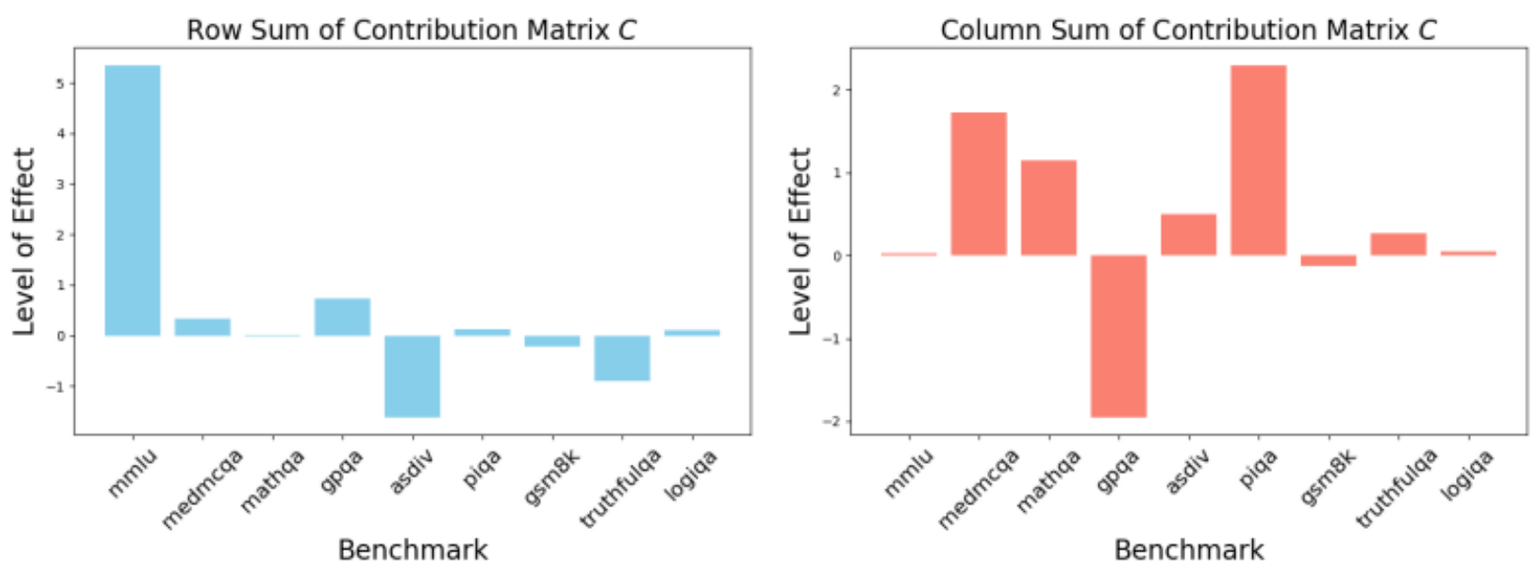


モデル埋め込みに含まれる情報とは？

実験２：学習データ特性に対する検証

ベンチマークの性能を予測する際に，訓練データから別のベンチマークを除外した場合と含めた場合で，予測精度がどう変化するかを測定する.

精度の変化を**貢献度スコア**として定量化し，どのベンチマークが役立つのかを分析する.



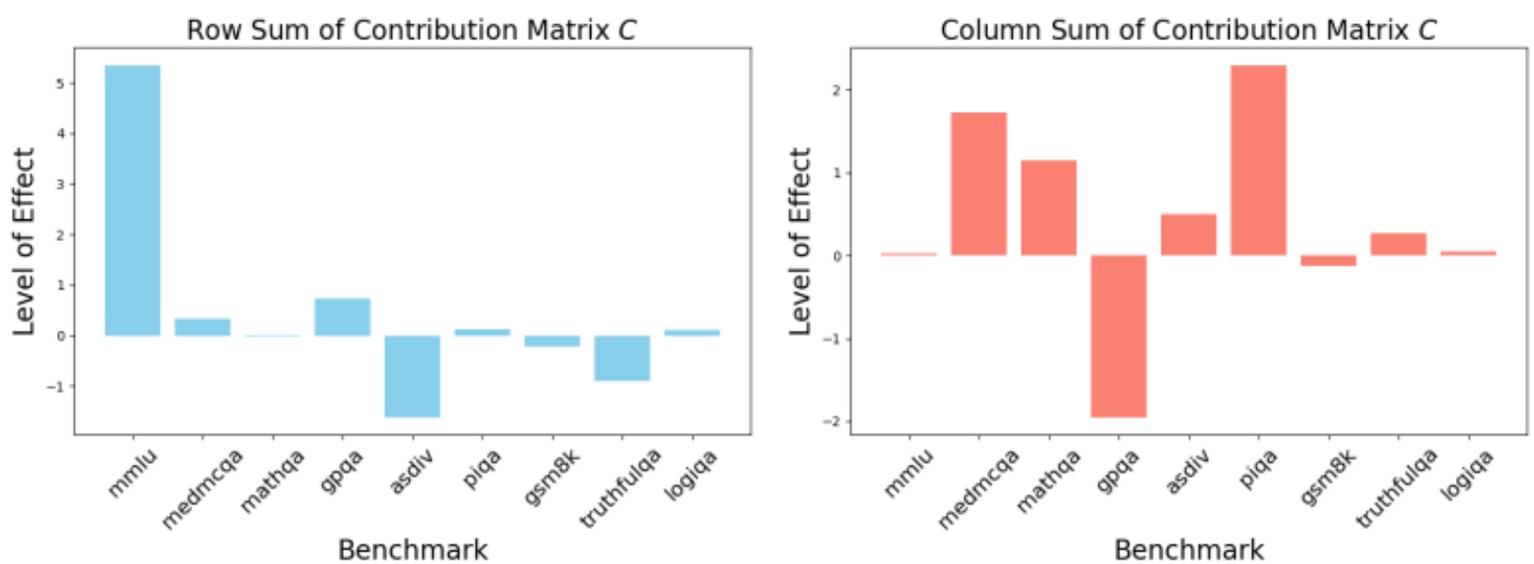
モデル埋め込みに含まれる情報とは？

結果

・MMLU（一般知識）を含めると，どのベンチマークも精度が向上した．➤MMLUは網羅的なベンチマークであることが分かる．

・GPQA（超難問）に他のベンチマークを訓練データに含めると精度が低下した．➤超難問を解く際に簡単な問題を解く能力は必要ないことが分かる．

・数学的ベンチマークは互いに精度を向上させることが分かった．
➤モデルの数学的能力の側面は埋め込みの変化により捉えている．



まとめ

LLMをコンパクトなベクトルにするEmbedLLMの提案

有効性

・正解性予測，モデルルーティング，ベンチマーク制度予測といった複数のタスクで有効性を実証.

知見

・学習されたベクトルが，モデルのサイズ，専門性，知識などの重要な特性を捉えていることを解明.

➤LLMが溢れる現代において，LLMを効率的に整理・評価・活用するための強力な基盤技術となり得る.