

研究発表

文献

背景

大規模言語モデル(LLM)の利活用が一般化している

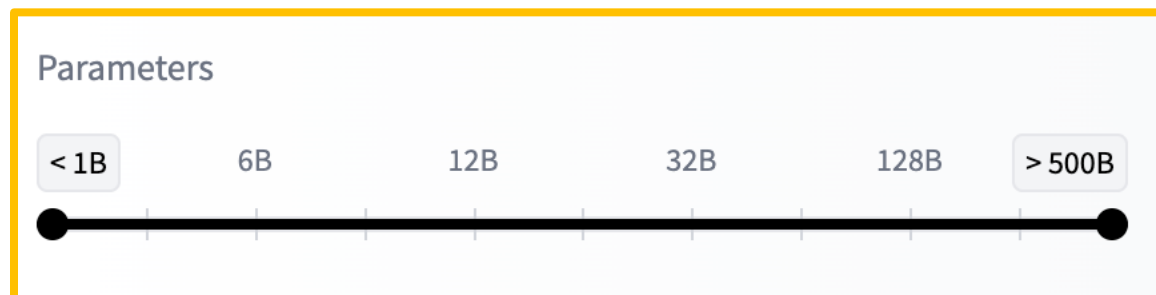
- Chat GPT, Claude, DeepSeek...

モデルの数も増加し続けており, 種類も豊富にある

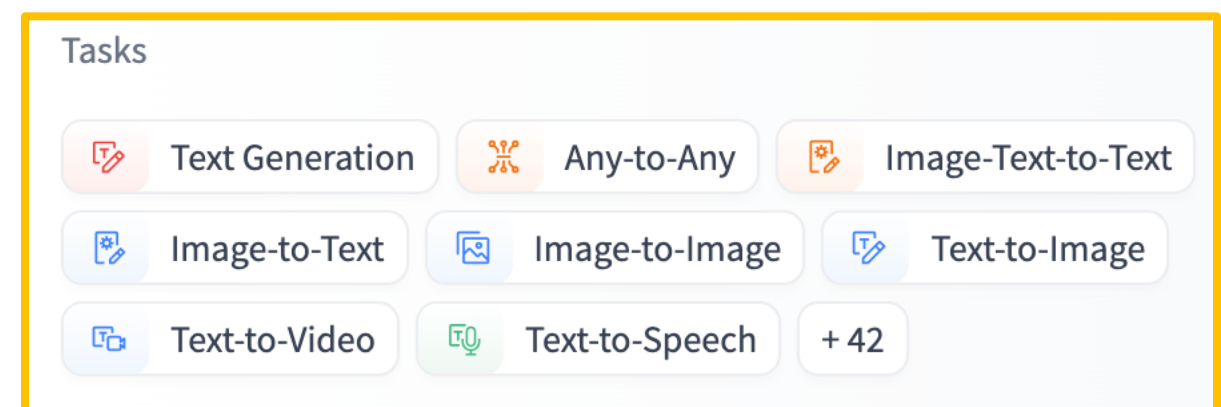
- Hugging Face : オープンソースモデル共有プラットフォーム

Models 1,902,343

2025/07/29時点



広範なパラメータ数



豊富なタスクの種類

背景

モデルのサイズと性能



大規模LLM

様々なタスクにおいて、高精度な結果
計算にかかるリソースが**高い**。

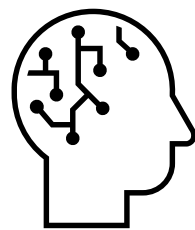


小規模LLM

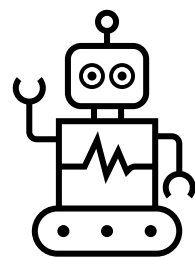
特定のタスクにおいて特化した言語モデル
汎化性能は低いものの、計算にかかるリソースが**低い**。

背景

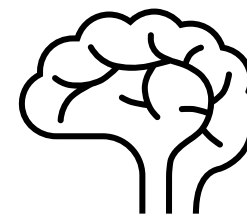
モデルにも得意・不得意がある



翻訳が得意なモデル



計算が得意なモデル



コーディングが得意なモデル

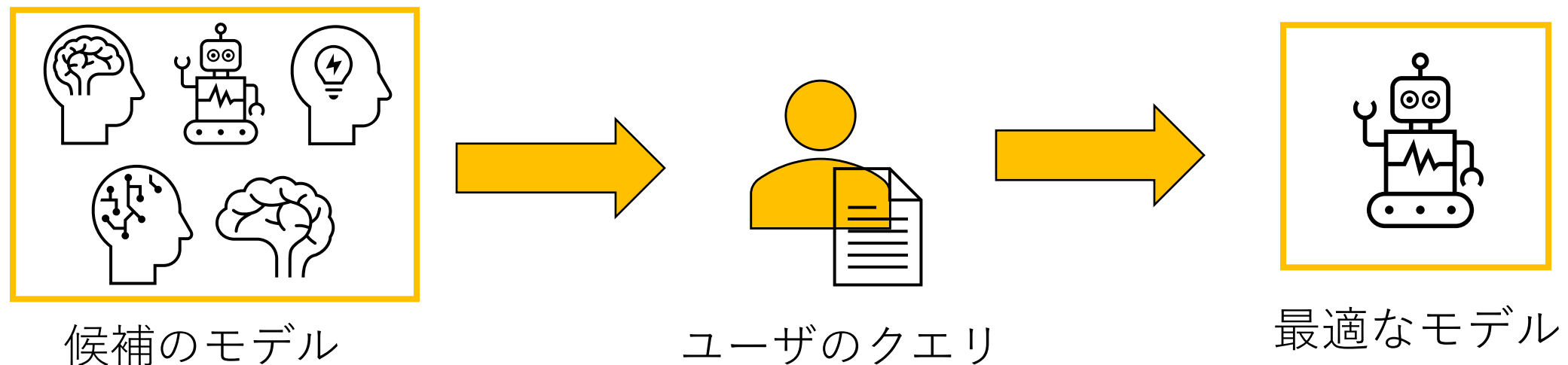


モデルのサイズやタスクに応じて
適切なモデルを選択する必要がある

先行研究

モデルルーティング

- 複数のモデルの中からコストや精度を考慮し、最もユーザーのクエリに最適なLLMを選定すること.



問題点

- 候補のモデルを手で選んでこななければならない
⇒最新のモデルに対応できない.

目的

モデルの候補を機械的に選定する手法を提案

- Hugging Faceに登録されているモデルから，モデルの候補を機械的に選定

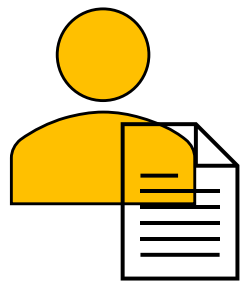


人手で選定するより，精度の高い選定を目指す

アプローチ

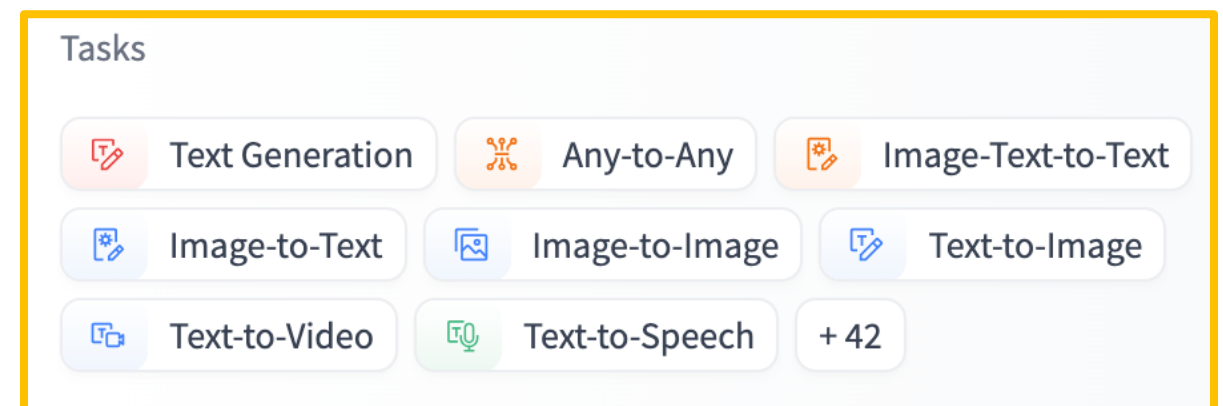
1. クエリを識別

- ユーザが入力したクエリをHuggnig Faceに登録されている50のタスクに分類.



ユーザのクエリ

few-shot学習された
LLMで分類



Hugging Faceに登録されているタスクの例

アプローチ

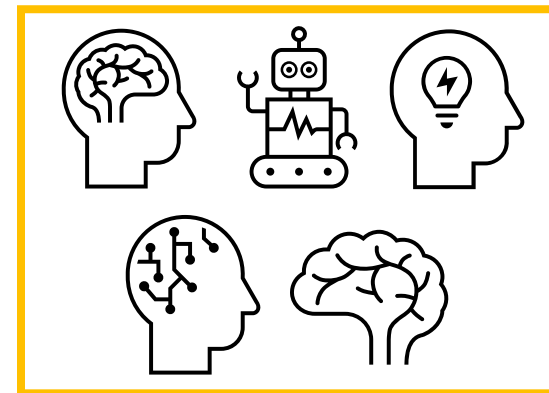
2. ダウンロード数やトレンドが上位のモデルを抽出

- 専門家がLLMをダウンロードする際に、最も重視しているのは
ダウンロード数, Like数
- 最新モデルに対応するため、トレンドや最近アップロードされたものも考慮



Text Generation

識別されたタスク

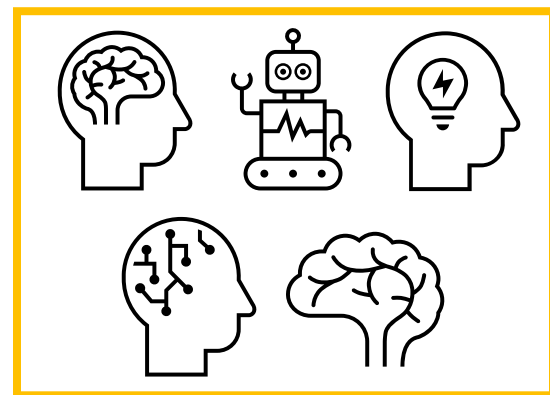


ダウンロード数や
トレンドが上位の
モデル

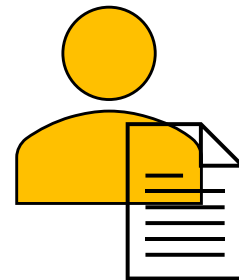
先行研究

モデルルーティング

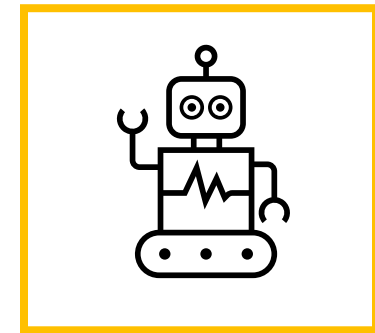
- 複数のモデルの中からコストや精度を考慮し、最もユーザーのクエリに最適なLLMを選定すること.



候補のモデル



ユーザーのクエリ

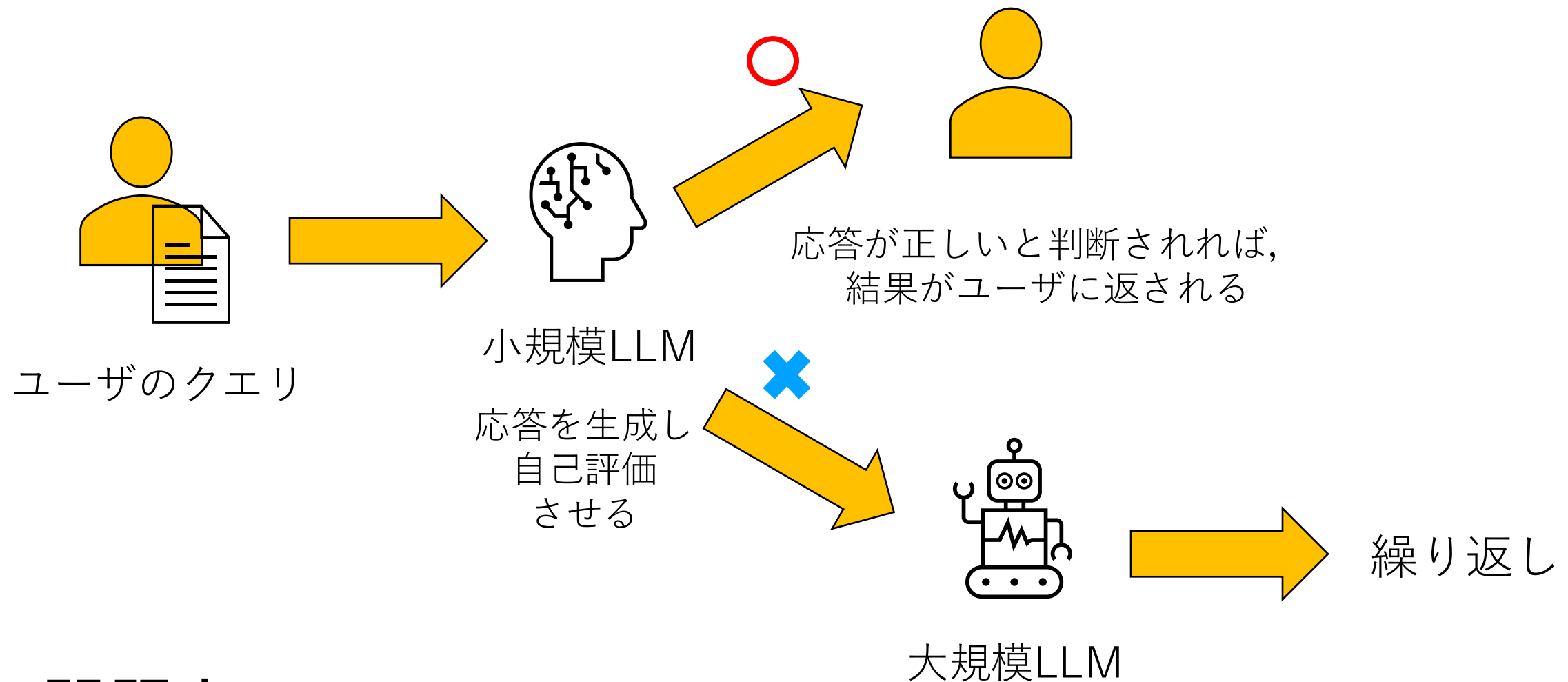


最適なモデル

これを実現することが大学院で研究する内容

先行研究

LLMカスケード



問題点

- モデルを自己評価させている

大学院での研究

現時点で考えている手法(既存研究)

- Universal Model Routing for Efficient LLM Inference
 - LLMをベクトル化し、類似度を計算する手法
- **Cost- and Latency-Constrained Routing for LLMs** (ICLR 2025)
 - コストとレイテンシの制約下で品質を最大化するルーティングを定式化し、制約を満たしながら最適なモデル選択を行う枠組みを提示

