

SWE-benchを用いた AIモデルルーティング手法の評価

オープンソースソフトウェア工学研究室
S2520048 陸井優希

背景

多様なAIモデル

- 性質の異なるAIモデルが多数発表されている

モデルサイズの違い, 専門特化の有無, 学習戦略の違い など

モデルサイズの違い



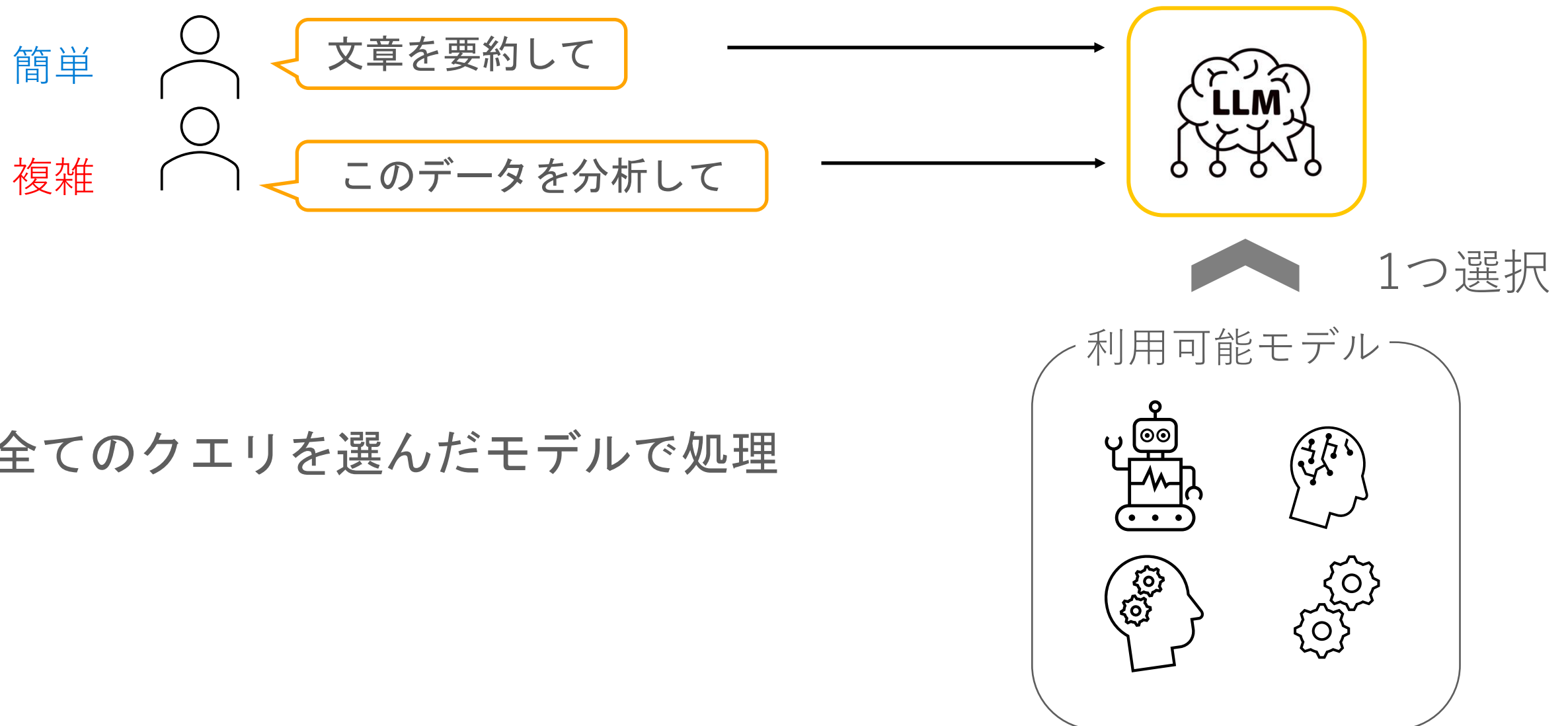
コストと精度はトレードオフ

[1] <https://openai.com/ja-JP/api/pricing/>

モデル利用の現状と課題

モデルの使われ方

- 一つモデルを決めて使うのが一般的^[4]

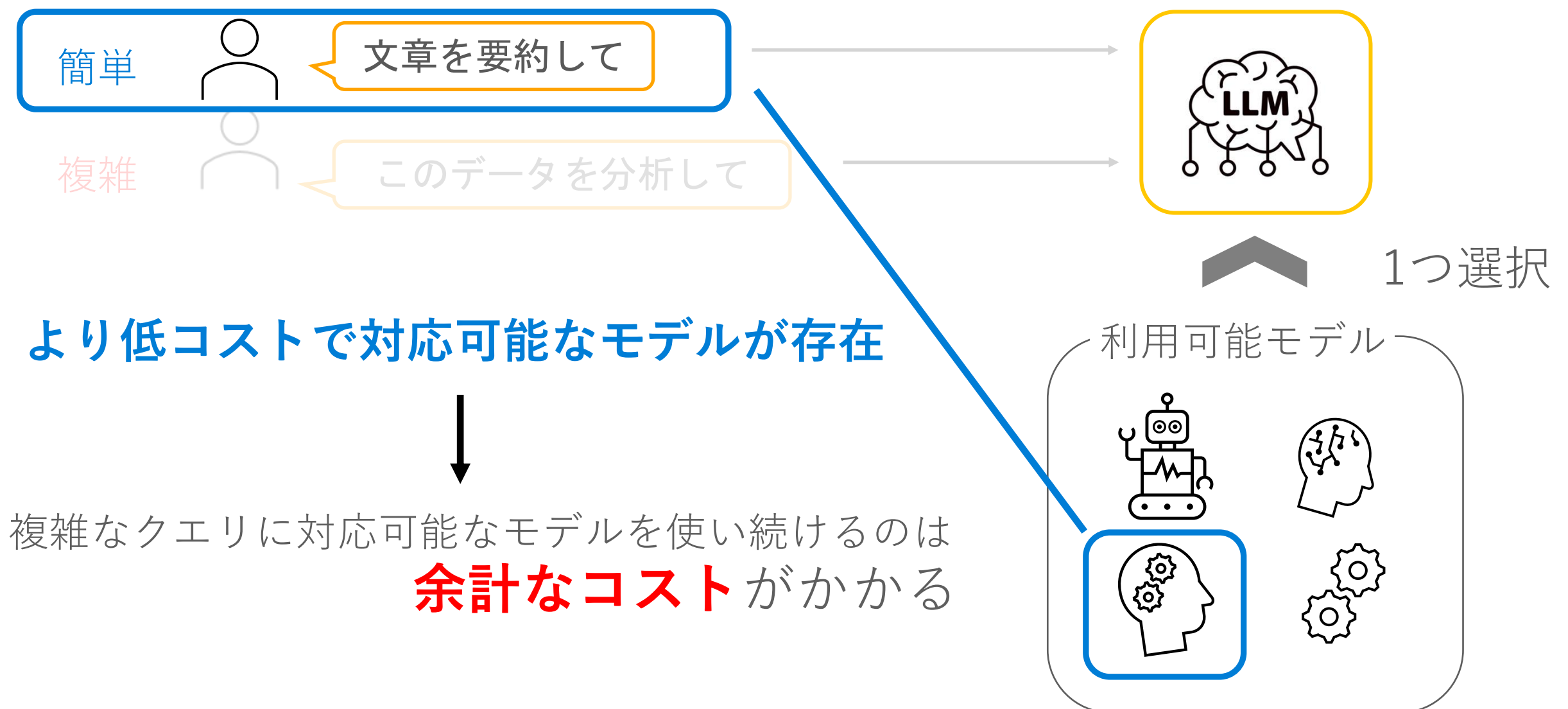


[4] Doing More with Less: A Survey on Routing Strategies for Resource Optimisation in Large Language Model-Based Systems
Clovis Varangot-Reille and Christophe Bouvard and Antoine Gourru and Mathieu Ciancone and Marion Schaeffer and François Jacquenet

モデル利用の現状と課題

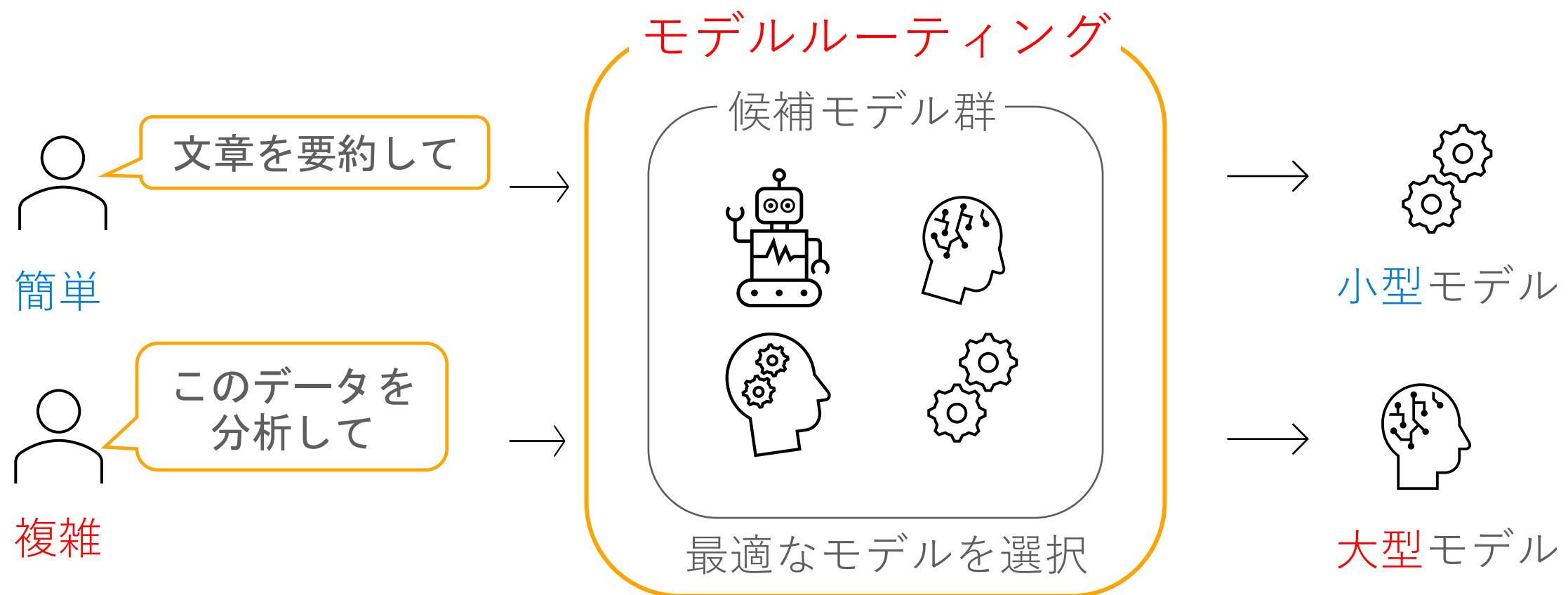
モデルの使われ方

- 一つモデルを決めて使うのが一般的¹⁾



AIモデルルーティング

クエリごとに最適なモデルを自動的に選択する技術



NLPタスクで複数の手法が提案・検証されている[2][3]

クエリごとに最適なモデルを選択することで**コスト削減**が可能

[2] MetaLLM: A High-performant and Cost-efficient Dynamic Framework for Wrapping LLMs Quang H. Nguyen and Thinh Dao and Duy C. Hoang and Juliette Decugis and Saurav Manchanda and Nitesh V. Chawla and Khoa D. Doan <https://arxiv.org/abs/2407.10834>

[3] RouterDC: query-based router by dual contrastive learning for assembling large language models Chen, Shuhao and Jiang, Weisen and Lin, Baijiong and Kwok, James T. and Zhang, Yu Proceedings of the 38th International Conference on Neural Information Processing Systems

SEタスクへの適用

NLPタスクとSEタスクでは入力異なる

- NLPタスクの入力

短いテキスト（質問・文章）

例）感情分析 “This movie was great!”

多肢選択QA “What is the capital of France? A) London B) Paris”

- SEタスクの入力

Issue文・コード・自然言語の指示

例）SWE-bench (Issue文 + コード)

“Modeling’s separability_matrix does not compute separability correctly...”

+ 数百行のコード

ルーティングの判断材料である入力異なる

→ 既存手法がSEタスクに有効かどうかは不明

目的・アプローチ

本研究の目的

- SEタスクでのモデルルーティングの効果を検証する

アプローチ

- 1, 既存手法を再現
- 2, SWE-benchでの効果を確認
精度・コストを確認

実験設定

データセット

- SWE-bench (Verified)

GitHub Issueからの実際のバグ修正タスク

近年のLLMの標準的な評価ベンチマーク

5分割交差検証

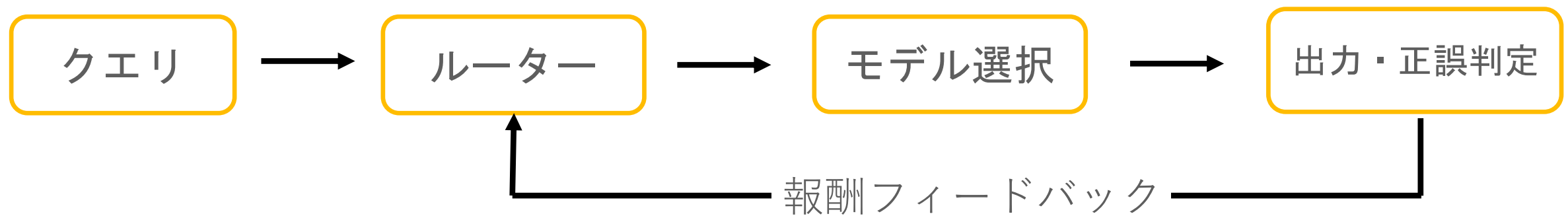
- 限られたデータを最大限利用し訓練・評価するため

	fold 1	fold 2	fold 3	fold 4	fold 5
1 ~ 100	[訓練]	[訓練]	[訓練]	[訓練]	<u>[評価]</u>
101 ~ 200	[訓練]	[訓練]	[訓練]	<u>[評価]</u>	[訓練]
201 ~ 300	[訓練]	[訓練]	<u>[評価]</u>	[訓練]	[訓練]
301 ~ 400	[訓練]	<u>[評価]</u>	[訓練]	[訓練]	[訓練]
401 ~ 500	<u>[評価]</u>	[訓練]	[訓練]	[訓練]	[訓練]

実験設定 | 再現した手法

MetaLLM

- 強化学習ベースの手法



- 選定理由

NLPタスクで実績がある

モデル選択時に複数回コード生成を行わない（コストが低い）

→ コスト削減効果がより期待できる

- 報告されている効果

精度：±0.5% コスト：57～12%削減

実験設定 | ルーティング候補モデル

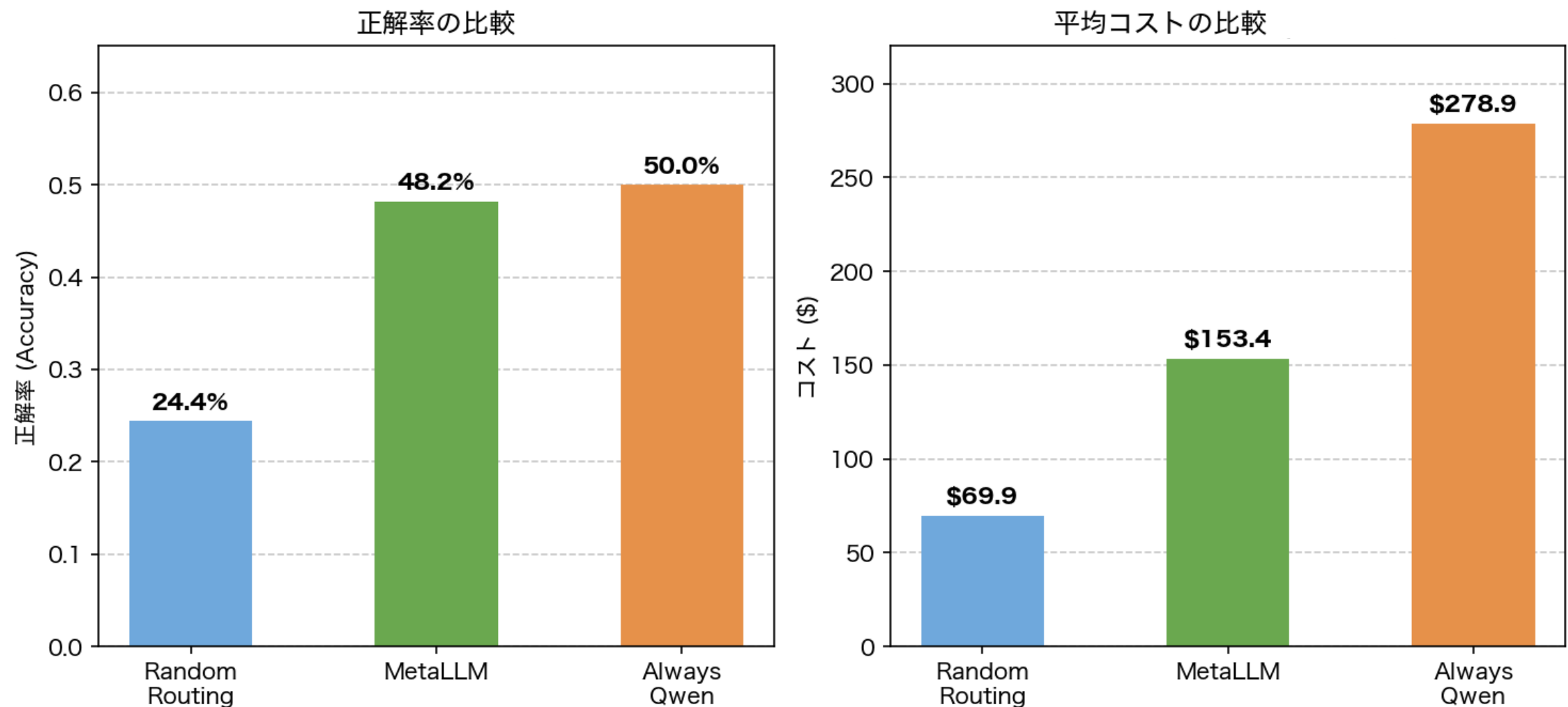
5 種類のオープンソースモデル

- 幅広い価格・性能のモデル

SWE-bench (Verified) 実行結果

モデル	正解率 (500問中)	コスト
Qwen3-coder-480B	250 (50%)	\$1,394
DeepSeek-V3.1	241 (48.2%)	\$271
gpt-oss-120b	103 (20.6%)	\$ 28
gpt-oss-20b	13 (2.6%)	\$ 8
gemma-3n-E4B-it	4 (0.8%)	\$ 5

実験結果 | 精度とコストの比較

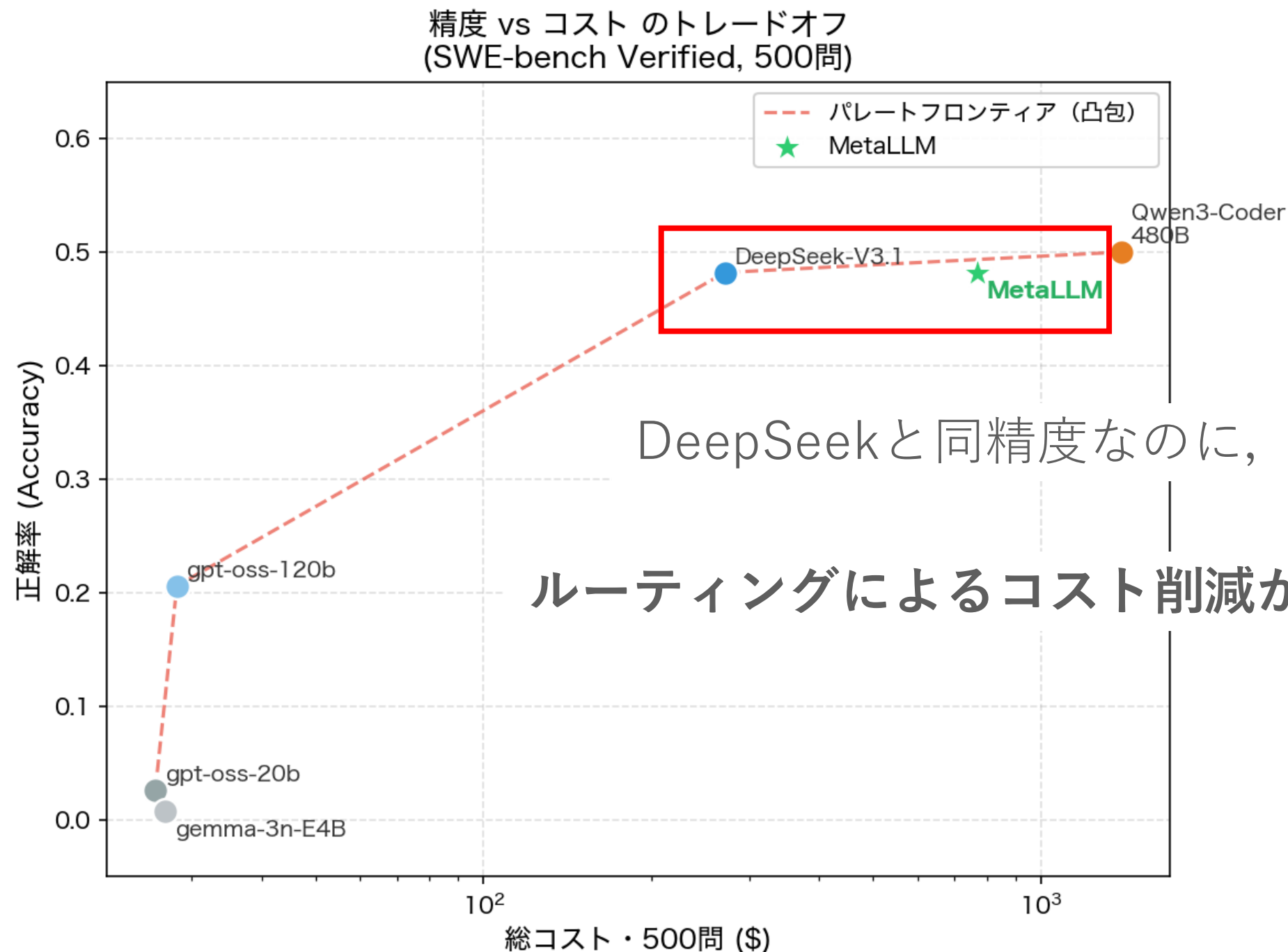


最良のモデル (Qwen3-coder) を使い続けた場合と比較して

精度：1.8%減

コスト：45%減

実験結果 | トレードオフ分析



詳細分析

| 各問題のルーティング結果を分類

- 適切な選択

正解かつ、より安いモデルでは解けなかった

- 正解だが非適切

正解したが、もっと安いモデルで解けた

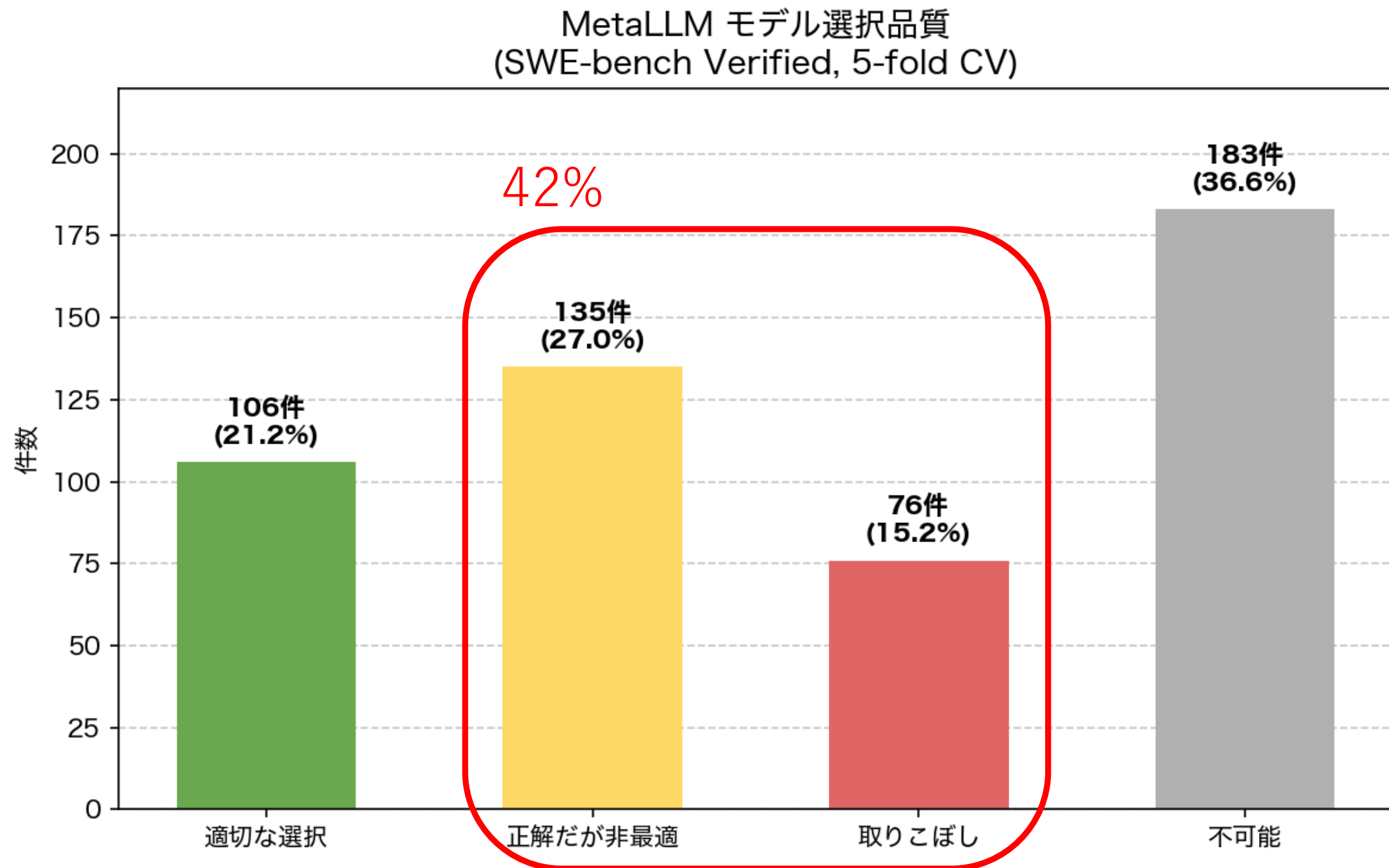
- 取りこぼし

不正解だが、別のモデルなら解けた

- 不可能

どのモデルでも解けなかった

分類結果



ルータの役割を果たせてないケースが多い

まとめ・今後の予定

| SWE-benchにおけるルーティング手法の効果を検証

| ルーティングの効果はなかった

→ SEタスク特有の工夫が必要（？）

| 今後

- 失敗した原因を調査
- コスト削減効果を改善する方法を模索

補足資料

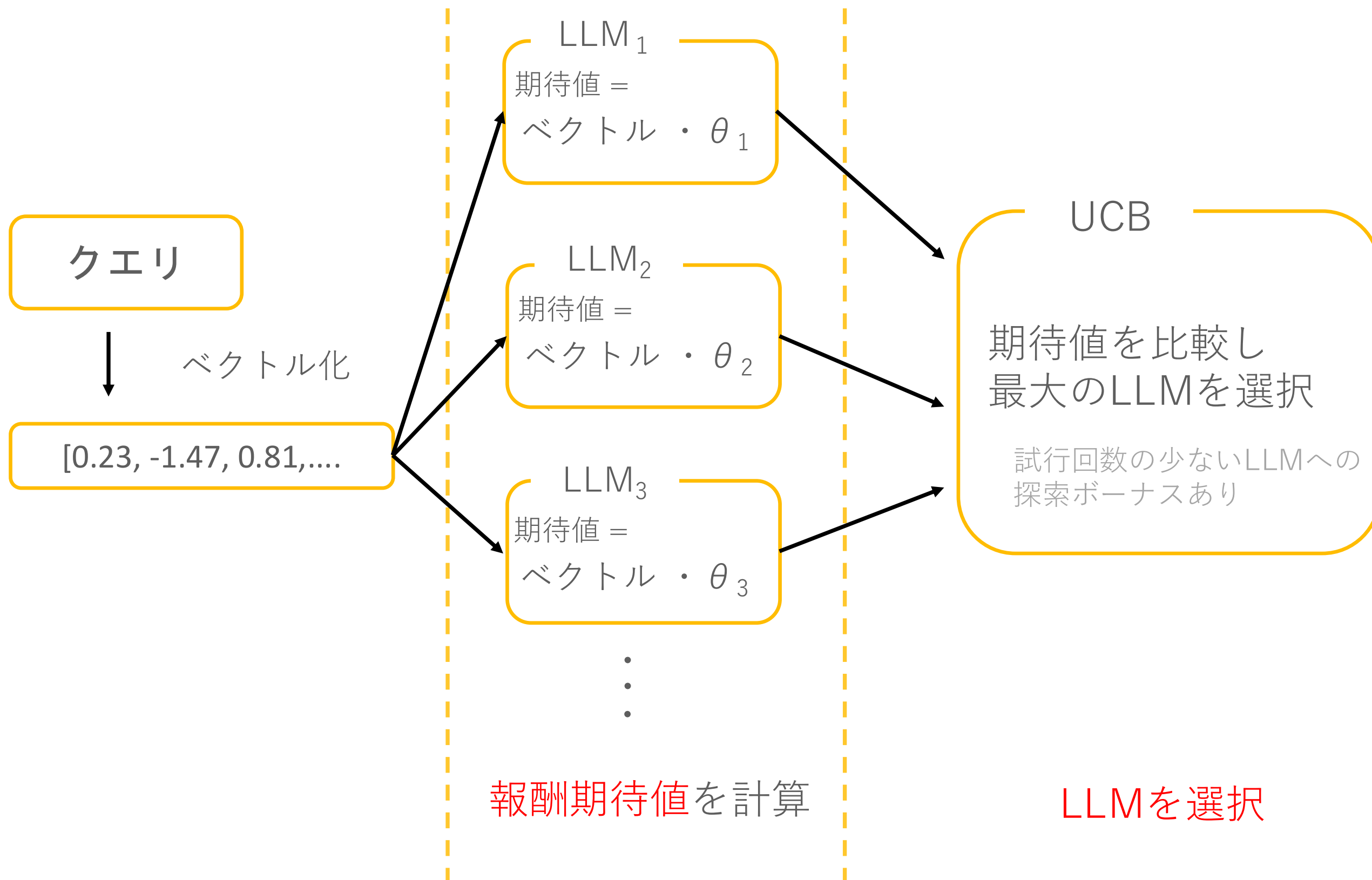
MetaLLM 概略

全体像：

LLMを選択



後処理



MetaLLM 概略

全体像：

LLMを選択



後処理

