

LLMを用いたコード検索のコスト削減手法の分析

S2310157 谷口

1. 研究背景

コード検索

データベースからコードを取得

- ・ 文章や単語を入力し，関連するコードを取得するタスク
- ・ 企業や組織，プロジェクトが専用のデータベースで行う
- ・ 安全性が検証済みであるコード資産の再利用



コード検索の課題

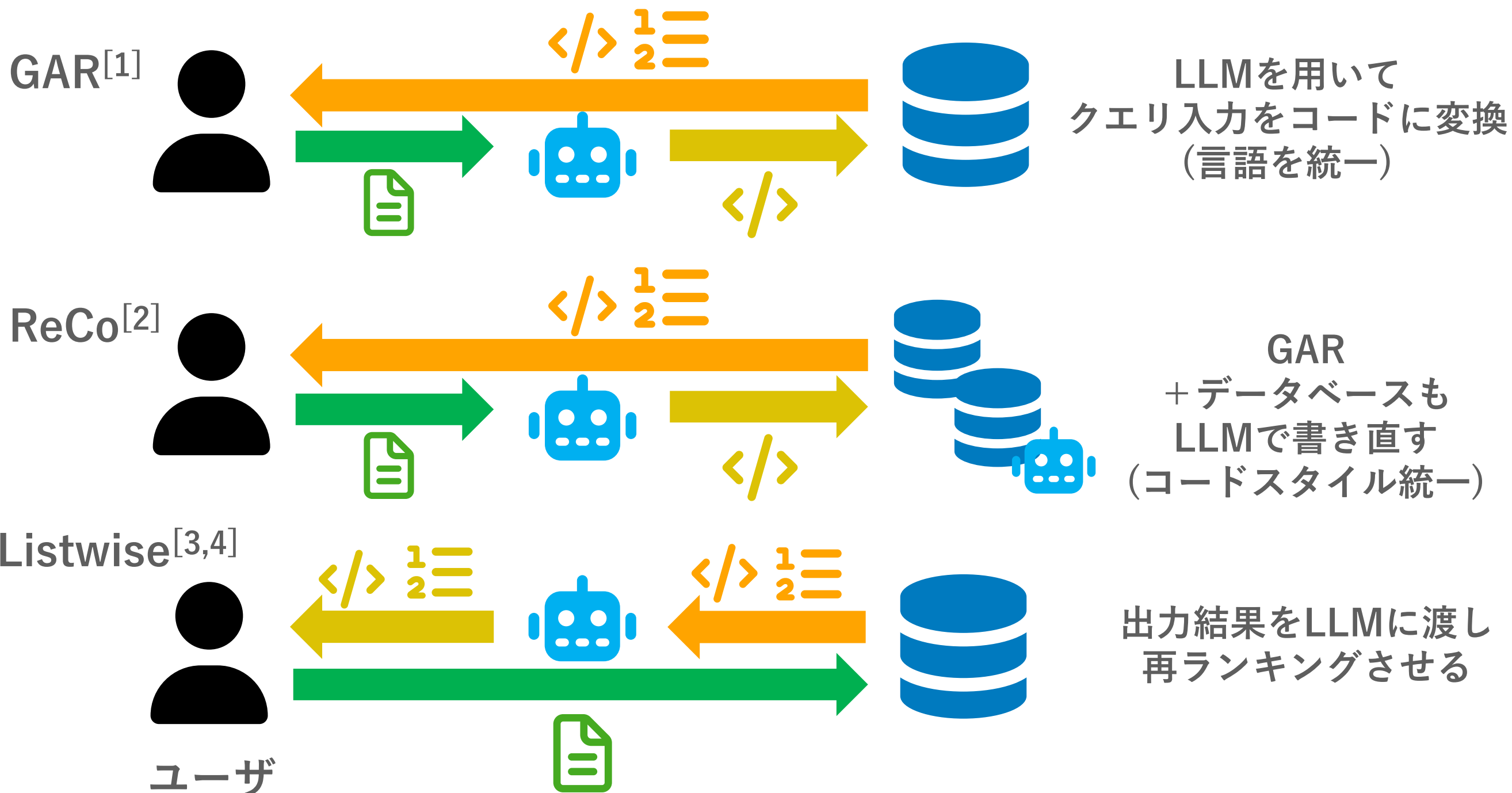
自然言語とコードの表現形式の差

- ・ 自然言語とコードは語彙と文法が大きく異なる
- ・ キーワード検索でも機械学習モデルでも精度向上を阻害



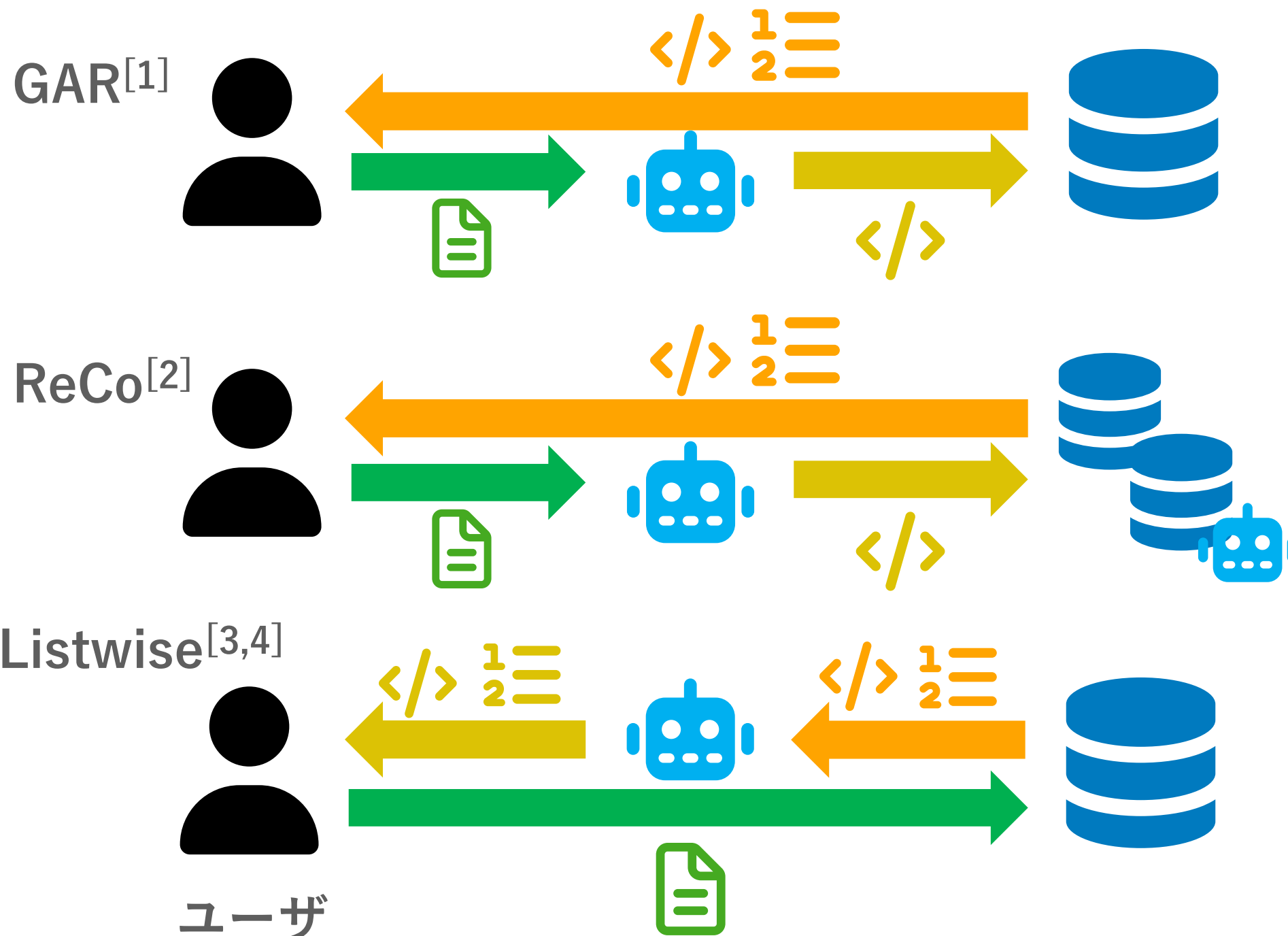
LLMを用いたアプローチ

課題に対する代表的なアプローチ



LLMを用いたアプローチ

課題に対する代表的なアプローチ

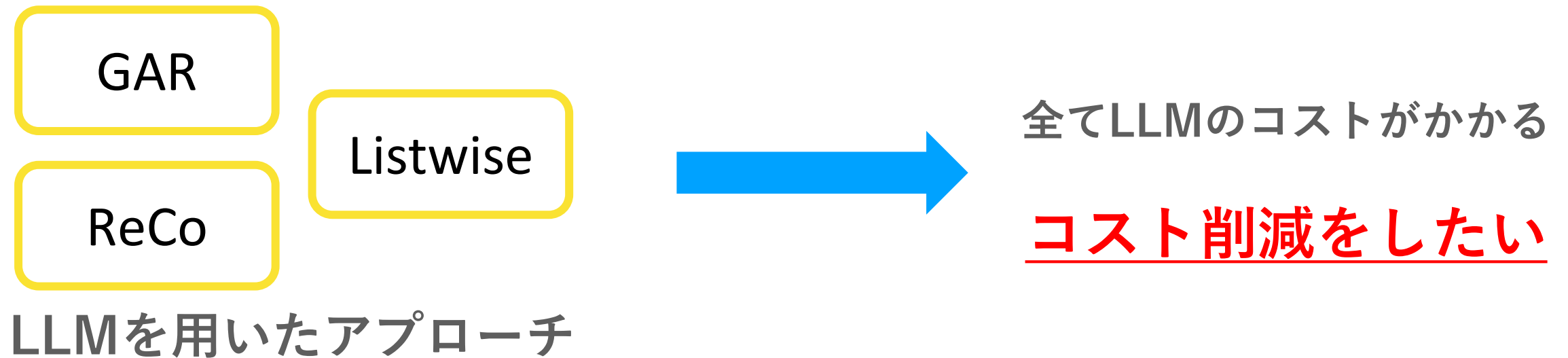


どの手法も
LLMの推論コストが
必要

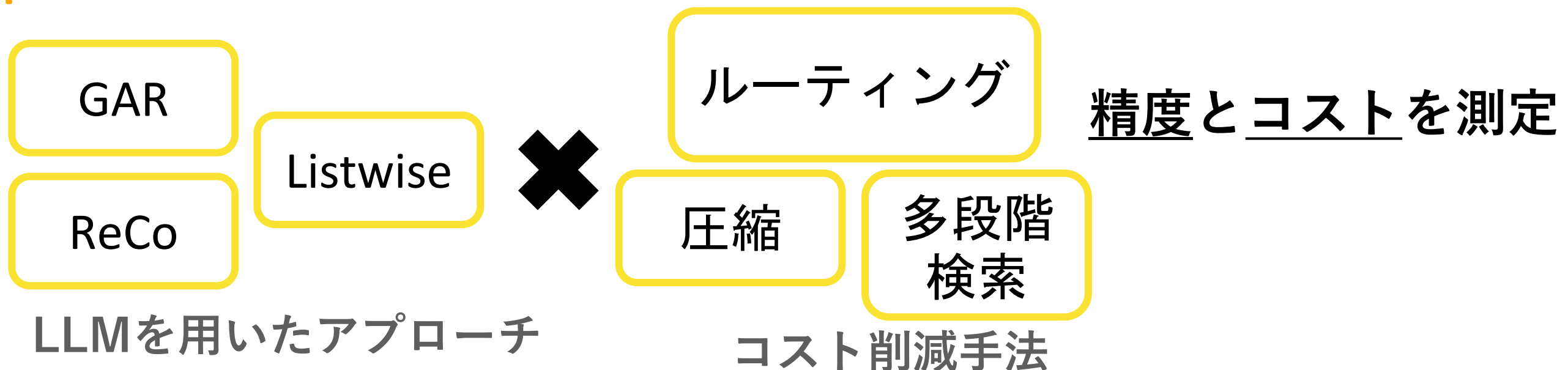
ReCoは高精度だが
一番コストが高い

研究内容

目的：コード検索のLLMコストの削減



アプローチ：既存のコスト削減手法を適用し比較



2. 関連研究

2.1 LLMによるクエリ拡張

入力に対してのアプローチ：クエリ拡張

- **GAR(生成拡張検索)**：自然言語クエリをコードクエリに変換
- **ReCo**：GAR + データベースもLLMに書き直させる

出力に対してのアプローチ：Listwise

- RankGPT：検索結果全件をLLMに与え再ランキング
 - **RLT**：件数を絞った再ランキング
 - CodeRankLLM^[5]：コード検索特化の再ランキング
(エージェントタスク， 訓練必須)

どの手法も検索精度を向上させるがコストがかかる

2.2 既存の推論コスト削減手法

入力の短縮

- 重要度の低いトークンを除くプロンプト圧縮
 - LLMによるプロンプト圧縮を軽量モデルに蒸留(LLMLingua^[6])
 - 各検索手法にどのように適用でき、
どのように効果が出るかが明らかでない

検索工程の効率化

- 異なる粒度/精度の検索を組み合わせるカスケード(FrugalGPT^[7])
 - 一部検索手法に適用可能
- クエリごとに使用モデルを選択するルータ(RouteLLM^[8])
 - 各検索手法で等しく効果が出る

各削減手法を適用し、コスト削減効果や精度変化を比較する

2.3 コスト比較の研究

検索コスト比較

- EnrichIndex^[9]
 - 一般文書検索
 - LLMリランク vs インデックス強化 (Listwise vs ReCoに近い)

やること

関連研究(2章)の整理

- この研究の方向性や目的を確定させる
- 実験計画の骨格を作る
- 無秩序な追加実験を防ぐ

実験の詳細の決定

- 使用するデータセットは？
 - データリークの有無は？
- 比較するLLM検索手法は？
- 各手法にどの削減手法が適用できるか？
- 得られる知見が重複した実験はないか？

- [1] Dong Li, Yelong Shen, Ruoming Jin, Yi Mao, Kuan Wang, and Weizhu Chen, “Generation-Augmented Query Expansion for Code Retrieval,” arXiv preprint arXiv:2212.10692, 2022.
- [2] Haochen Li, Xin Zhou, and Zhiqi Shen, “Rewriting the Code: A Simple Method for Large Language Model Augmented Code Search,” Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pp. 1371–1389, 2024.
- [3] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren, “Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents,” Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 14918–14937, 2023.
- [4] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke, “Ranked List Truncation for Large Language Model-based Re-Ranking,” Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 141–151, 2024.
- [5] Tarun Suresh, Revanth Gangi Reddy, Yifei Xu, Zach Nussbaum, Andriy Mulyar, Brandon Duderstadt, and Heng Ji, “CoRNStack: High-Quality Contrastive Data for Better Code Ranking,” Proceedings of the 13th International Conference on Learning Representations, 2025.
- [6] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu, “LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models,” Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 13358–13376, 2023.
- [7] Lingjiao Chen, Matei Zaharia, and James Zou, “FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance,” Transactions on Machine Learning Research, 2024.
- [8] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica, “RouteLLM: Learning to Route LLMs with Preference Data,” Proceedings of the 13th International Conference on Learning Representations, 2025.
- [9] Peter Baile Chen, Tomer Wolfson, Michael Cafarella, and Dan Roth, “EnrichIndex: Using LLMs to Enrich Retrieval Indices Offline,” Proceedings of the 2nd Conference on Language Modeling, 2025.