LLM-jp を用いて生成したデータセットによる Embedding モデルの Fine-tuning

○ チームS（高木純平、 村中勇輝 、小林伸次、浦上暉允、河野聖）

株式会社セック
SEC Systems Engineering Consultants Co.,LTD.

課題：RAGの精度と速度のトレードオフ

汎用 Embedding モデルの限界

専門分野のベクトル表現が
困難なため、**検索精度が低下**

Rerank の検索速度

毎回 Rerank 処理を行うため、
検索速度が低下

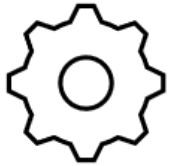
解決策：Embedding モデルの Fine-tuning

生データ

自社文書
専門ドキュメント

LLM-jp

高品質な
[問い] / [答え]
データセット



Fine-tuning

データセットで
Embedding モデルを
Fine-tuning

専門文書に特化した Embedding モデルにチューニング

Embedding Search Demo

all-MiniLM-L6-v2(FT) | all-MiniLM-L6-v2×Rerank | all-MiniLM-L6-v2(ベース)

2024年度の科学技術振興費(A)は2023年度と比べてどのくらい増えましたか？

検索

all-MiniLM-L6-v2(FT)

正解ランク
1
/ 100

検索速度
0.035s

RAG 回答

2024年度の科学技術振興費(A)は14,092億円で、2023年度の13,942億円と比べて150億円増加しました。増加率は約1.1%です。

all-MiniLM-L6-v2×Rerank

正解ランク
1
/ 100

検索速度
10.895s

RAG 回答

2024年度の科学技術振興費(A)は、2023年度の13,942億円から14,092億円に増加し、約150億円（約1.1%）の増加となりました。

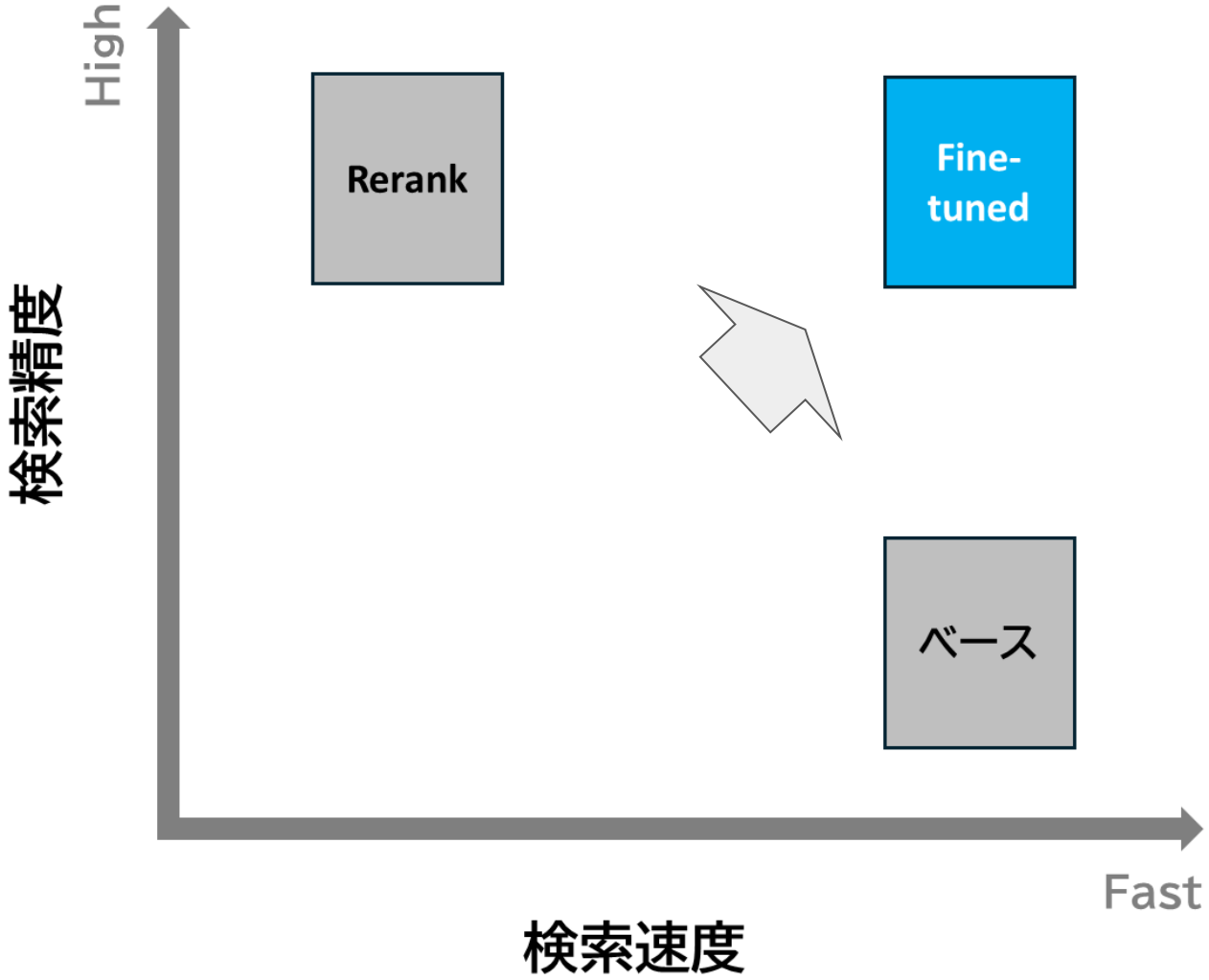
all-MiniLM-L6-v2(ベース)

正解ランク
32
/ 100

検索速度
0.031s

RAG 回答

提供された参考情報には、2024年度の科学技術振興費(A)と2023年度の比較に関する具体的な金額や増加率の記載がありません。そのため、参考情報から「2024年度の科学技術振興費(A)は2023年度と比べてどのくらい増えましたか？」という質問に対しては回答できません。



高品質データセットで
高い精度と高い速度を
両立可能