プロンプトインジェクションに対する取組

NTT株式会社　人間情報研究所
2025年10月9日

大規模言語モデル(LLM)　tsuzumi

NTT

tsuzumi

邦楽器	**日本語が特に強い**
小さい	**小型軽量**
調べ緒により調律が容易	**柔軟なチューニング**
見た目、音、 演奏時の所作が美しい	**マルチモーダル**

自然言語処理技術への取り組み

1980年　自然言語処理の研究開始

2020年　「NTT版BERT」をリリース

2023年 11月　NTT版LLM **tsuzumi 発表**

2024年 3月　NTT版LLM **tsuzumi 商用開始**

視覚マルチモーダル対応・多言語対応など順次商用リリース

2025年 10月　**tsuzumi 2 リリース**

アップグレード版 純国産LLM「tsuzumi 2」

- NTT版LLMであるtsuzumiのアップグレード版を10月にリリース予定
- 日本語性能において同サイズLLMにて世界トップクラス

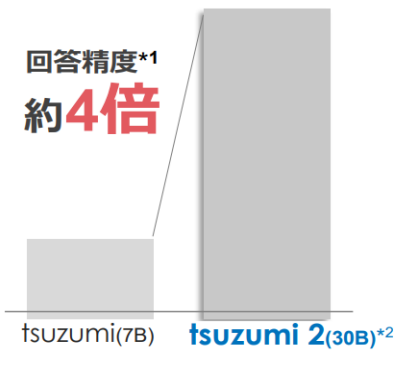
企業のお客様ニーズにこたえる
複雑な文脈・文意理解力が進化

コスト効率と大幅な性能向上を実現した
バランスに優れた1GPU動作モデル

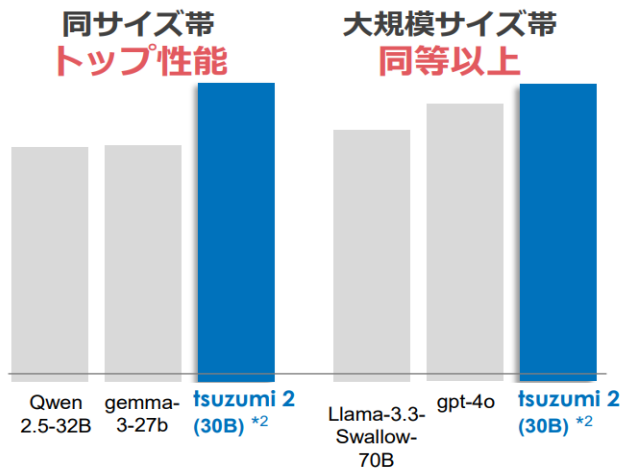
機密性の高い情報にも対応
NTTがゼロから開発した純国産モデル

強化した文脈処理・文意理解の性能比較

RAGを用いた業務処理の実例における前モデルとの比較



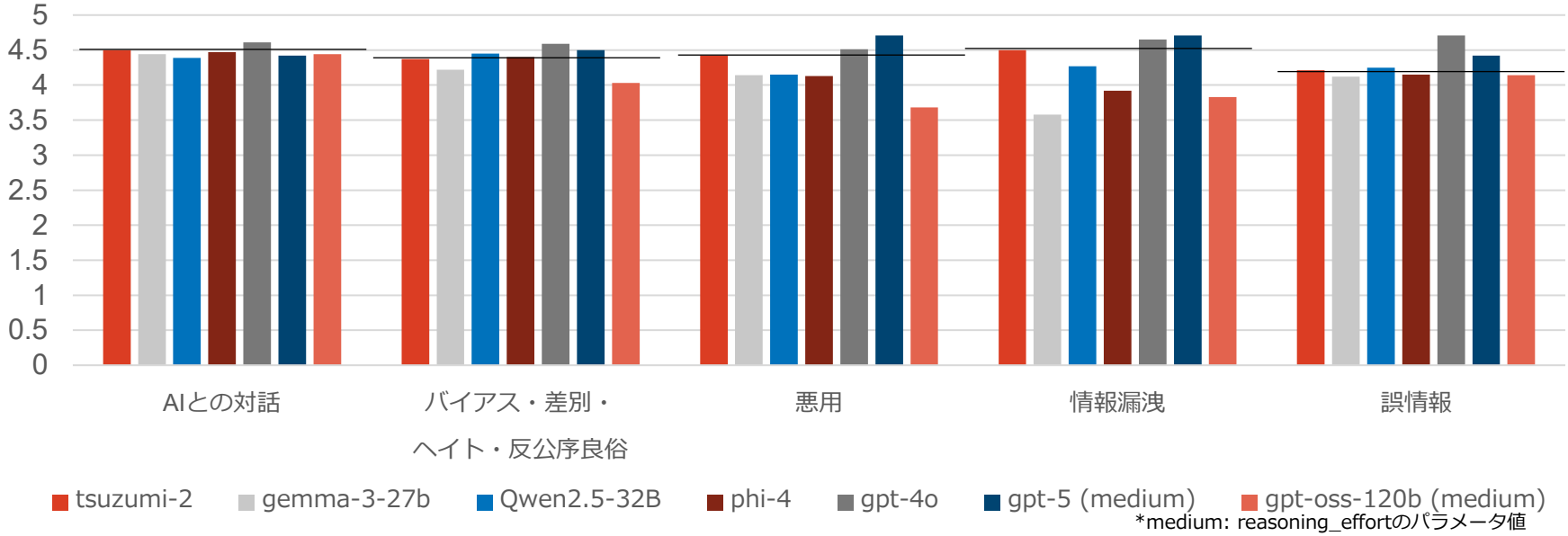
一般的なベンチマーク*3での他モデル比較



*1：NTT社内業務におけるトライアル案件のRAGによる問い合わせ回答精度
*2：tsuzumi 2は開発中のものを利用
*3：llm-jp-evalにおける日本語性能評価、文脈・文意理解該当する指標の評価結果

安全性評価

- 下記に示す「AnswerCarefully（鈴木(国立情報科学研究所）」など、安全性を評価するベンチマークにてLLMの安全性について確認している

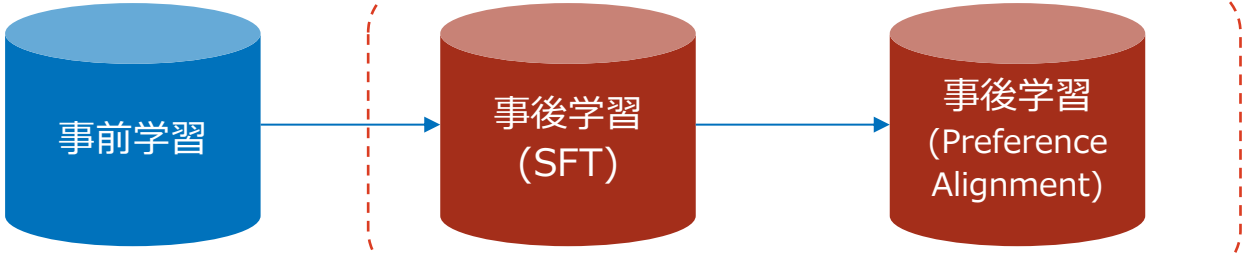


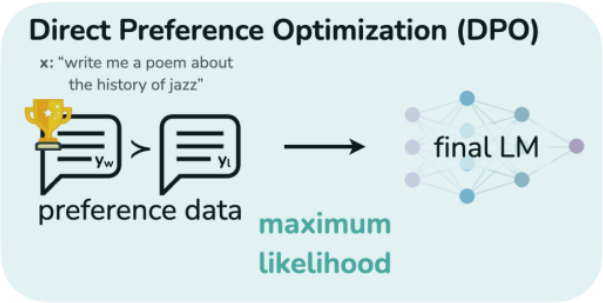
*medium: reasoning_effortのパラメータ値

鈴木久美, 勝又智, 児玉貴志, 高橋哲朗, 中山功太, 関根聡. 2025. AnswerCarefully: 日本語LLM安全性向上のためのデータセット. 言語処理学会第31回年次大会発表論文集.

安全性を高める取り組み(1)

- 教師あり学習（Supervised Fine-Tuinnge; SFT）と、Preference アラインメントのそれぞれで安全性を高める学習を実施
- 特に、最終的なPreferenceアラインメントの**学習データの量・質**の重要性を確認



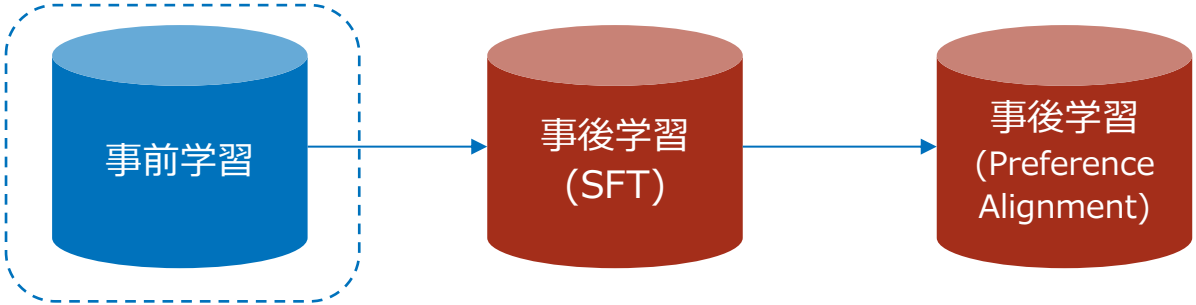


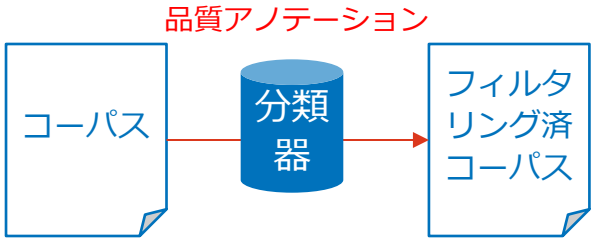
← 代表的な手法（DPO）

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, Chelsea Finn:
Direct Preference Optimization: Your Language Model is Secretly a Reward Model. NeurIPS 2023
https://arxiv.org/pdf/2305.18290

安全性を高める取り組み(2)

- 事前学習では、学習コーパスから問題のあるテキストやコーパス内で重複するテキストをフィルタリングする取組を精緻化している





← 代表的な手法（品質フィルタリング）

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, Thomas Wolf:
The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. NeurIPS 2024
https://arxiv.org/abs/2406.17557

ガイドライン・情報共有への期待

- LLMへの攻撃に関する最新の国内外の例（プロンプト・レスポンス）が参照できるデータベースがあると良い（開発者はそうしたプロンプトを参考にして、安全性の学習を進めることができる）