

高品質な合成データに基づく ドメイン特化LLM開発フレームワーク

早稲田大学¹ 株式会社フィックスターズ²

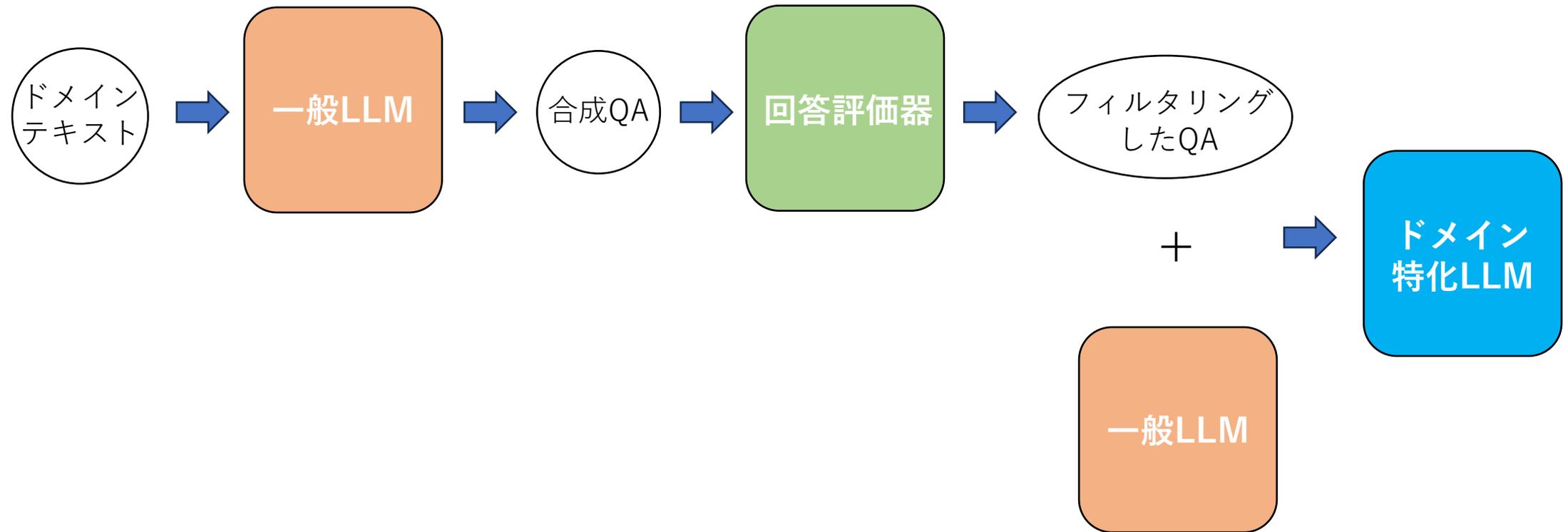
中村友亮¹ 堀尾海斗¹ 飯塚康太² 小田桐海翔² 吉藤尚生² 河原大輔¹

本研究にはFixstars AI Boosterを利用しました

目的

- 一般的な大規模言語モデル(LLM)は、特定のドメインにフォーカスして訓練されていない
- 本研究では、特定ドメインのテキストが存在するときに、そのドメインに特化したLLMを開発するためのフレームワークを提案
- 提案フレームワークでは、高品質な合成データでLLMをチューニング (SFT)することで性能向上を図る

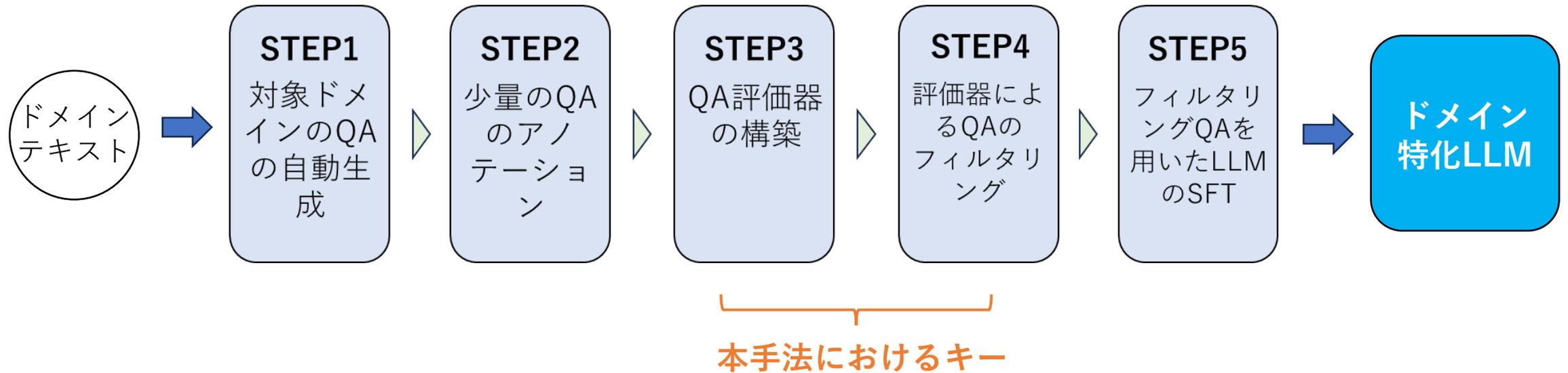
フレームワーク概観



関連研究

- SFTによるLLMへの知識注入
 - LLMに知識を与える方法についての研究は数多く存在しており、質問応答タスクにおける活用法の一つとして、SFTによる知識注入アプローチを取ったもの [[Mecklenburg+ 2024](#)]などが存在
- Retrieval-Augmented Generation (RAG) [[Lewis+ 2020](#)]
 - 質問応答タスクにおける手法の一つ
 - 質問応答のたびに参照テキストを入力として与えることで、確実な知識の補強を行うことが可能
 - RAGとSFTを比較した研究結果 [[Ovadia+ 2024](#), [Balaguer+ 2024](#)] が報告されているが、どちらを選ぶことがより適切かは状況によって異なる
- 合成データによるモデルの訓練
 - モデルのSFTなどに用いる訓練データを収集する方法として、LLMによる合成データを用いる手法がある [[Puri+ 2020](#), [Shakeri+ 2020](#), [Takahashi+ 2023](#)]
 - 合成データを用いることで人手で一から構築する場合よりも、比較的容易に大量のデータを生成することが可能

提案手法



1. 対象ドメインのQAの自動生成



1. 対象ドメインのQAの自動生成

- 対象のドメインにおける文章を一定の大きさ（参照文書）に切り分け、そのまとまりごとにLLMを用いて質問、回答を生成
- 似通ったものを削除するため、生成したQAに対して質問同士および回答同士の類似度（BERTScore）を計算し、両方の類似度が0.8を超えるものは1つ残してそれ以外を削除

1. 対象ドメインのQAの自動生成

- 対象のドメインにおける文書を一定の大きさ（参照文書）に切り

参照文書	質問	回答
<p>LJP は国外において盛んに研究されている。欧州人権裁判所のデータを用いた研究 \$[1,2,3,4]\$ はその代表例である。米国においても Katz ら [5] による米国最高裁判所におけるデータを用いた例、Semo ら [6] によるクラスアクション訴訟を取り扱った例が挙げられる。中国においては刑事事件向け \$¥operatorname{LJP}[7,8,9,10,11]\$ の研究が盛んである。また、Chalkidis ら [12] は複数のタスクから構成される英語圏の法律分野向けのベンチマークデータセット LexGLUE を提案している。国内においては、機械学習を用いた実験に適する規模を持つ実事件データを元にした LJP データセットは存在しない。法に関するデータを用いた Competition on Legal Information Extraction/Entailment (COLIEE) [13] が存在するものの、COLIEE では日本の司法試験短答式問題を用いている点で狙いが異なる。言語・法制度は国毎に異なるため、LJP 研究において日本独自の LJP データセットを構築することは重要である。そこで日本の民事事件判決書を用いた大規模なデータセットとして Japanese Tort-case Dataset (JTD) を構築する。</p>	<p>LJP用のデータセットの状況について説明してください。</p>	<p>LJP用のデータセットの状況は、国別に異なります。国際的には、欧州人権裁判所、米国最高裁判所、クラスアクション訴訟などのデータを用いた研究が行われています。国内においては、COLIEEが存在し、法に関するデータを用いた研究が行われていますが、民事事件判決書を用いた大規模なデータセット Japanese Tort-case Dataset (JTD) を構築することになっています。</p>
	<p>LJP が国外でどのように研究されているのでしょうか？</p>	<p>LJP は国外で盛んに研究されています。欧州人権裁判所のデータを用いた研究や、米国最高裁判所におけるデータを用いた研究、クラスアクション訴訟を取り扱った研究などが挙げられます。また、中国においては刑事事件向けの LJP の研究も盛んです。</p>

質問ペア・回答ペアそれぞれで BERTScore を計算

2. QAの人手アノテーション



2. QAの人手アノテーション

- 生成したQAの一部をQA評価器の訓練データ用に切り出し、人手でアノテーションを実施
- 質問のアノテーション
 - 参照文書を読み、以下の基準で質問を評価し、適切かどうかの二値のラベルを付与
 - 文書の情報のみで回答できる
 - 文書の内容に関係している
 - 日本語である
- 回答のアノテーション
 - 質問が適切であるものに対して、以下の基準で回答を評価し、適切かどうかの二値のラベルを付与
 - 文書の情報のみで回答している
 - 文章として破綻していない
 - 回答が正しい

アノテーションの例

評価	参照文書	質問
悪い	<p>【Sports Watch】星野監督、開幕投手を決めた経緯を告白「今バラしますがけど……」##いよいよ来週12日に開幕するプロ野球。見所の一つは、田中将大、岩隈久志の二大エースを擁し、メジャー帰りの岩村明憲、松井稼頭央が新戦力として加わった東北楽天ゴールデンイーグルスの戦いだろう。####とくに田中は岩隈から開幕投手の座を奪うと宣言していた。しかし、東北震災により、その開幕が延期になると、千葉（QVCマリン）での開幕戦は岩隈が投げ、田中は甲子園が舞台となるホーム開幕、15日の対オリックス戦に登場することになった。</p>	田中将大はメジャーリーグでどのような実績を残しているか？
良い	<p>■標準と同じサイズで70mAhも大容量##「ソフトバンク AQUOS PHONE 102SH 対応スタンダード大容量バッテリー」は、文字通りにAQUOS PHONE 102SH用に開発した大容量バッテリーだ。####標準バッテリー（1520mAh）と比べて70mAh多い1600mAhと大容量だが、バッテリーサイズが標準バッテリーと同じだ。今まで使っているケースをそのままのかたちで使い続けることができる。#### ## ####・ソフトバンク AQUOS PHONE 102SH対応スタンダード大容量バッテリー ####■標準の約2.1倍の大容量##「ソフトバンク AQUOS PHONE 102SH対応大型大容量バッテリー」は、より長時間使いたい人のために開発された大型大容量バッテリーだ。</p>	Aquos Phone 102SH用の標準的なバッテリー容量は何ですか？

3. 回答評価器の構築



3. 回答評価器の構築

- 人手アノテーションの結果、質問が適切と評価されたものののみを用いて訓練データを作成
- 訓練データでエンコーダーLLMをファインチューニングし、回答評価器を構築
- 回答評価器は参照文書、質問、回答を入力とし、回答の二値の評価ラベルを出力とする
 - 学習は二値分類として早大RoBERTaをファインチューニング



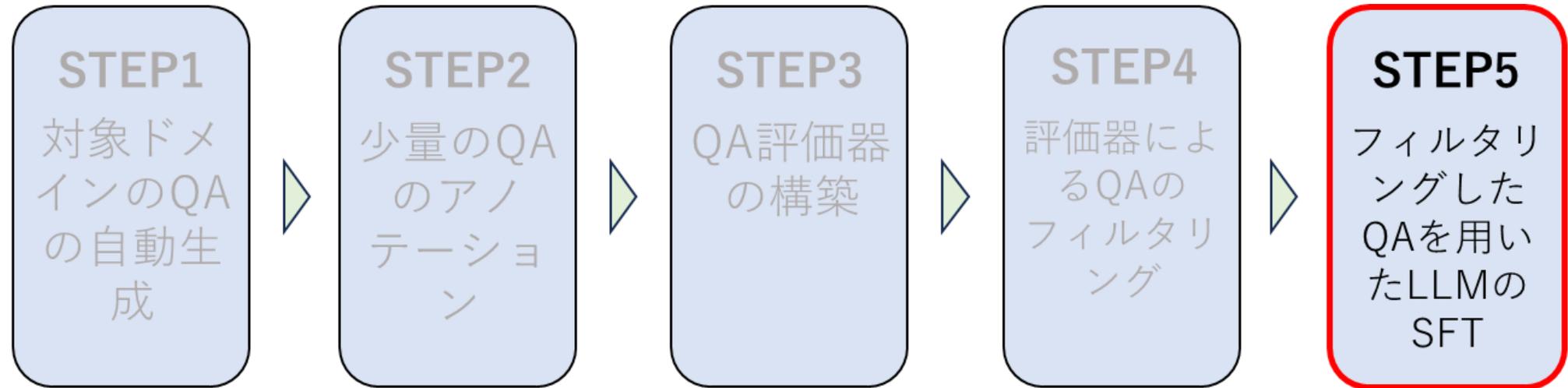
4. 評価器によるQAのフィルタリング



4. 評価器によるQAのフィルタリング

- 生成したQAの一部をSFTに用いる訓練データとして切り出し、構築した回答評価器でスコア付け
 - スコアは、回答評価器に参照文書、質問、回答を入力し、出力されたラベルの予測値を使用
- スコアの高い順に25, 50, 75 %の部分データを作成

5. フィルタリングQAを用いたLLMのSFT



5. フィルタリングQAを用いたLLMのSFT

- フィルタリングした QA データを訓練データとして、LLM の SFT を実施
- 入力には参照文書 + 質問を与え、回答を出力させる

実験設定：合成データの生成

- 対象ドメインは3つで、Mixtral-8x7Bを用いて質問、回答を生成

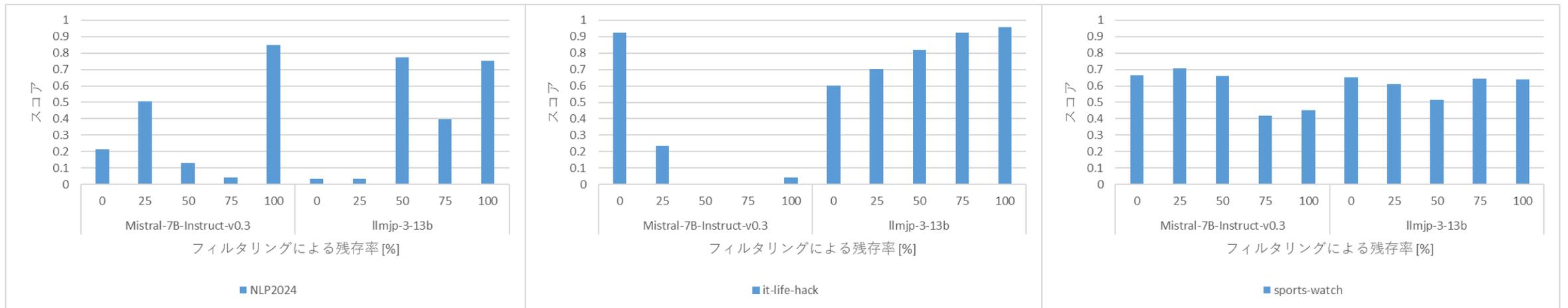
ドメイン名	ドメイン文書数	Q (すべて)	Q (○)	QA (○)	アノテーション
NLP2024	14論文	3,165	930	655	著者2人が実施
livedoor news corpus it-life-hack	100記事	6,620	3,837	2,925	クラウド ソーシング
livedoor news corpus sports-watch	100記事	3,740	2,081	1,577	クラウド ソーシング

実験設定：生成評価

- SFT 後の各モデルでテストデータに対して回答を生成、評価
- 構築した回答評価器を用いて、SFT モデルの出力を評価
 - 評価器には参照文書、質問、生成した回答を結合したものを入力として与える
 - 評価器が適切と評価するようなテキストを生成出来るかを確認するため、評価器による自己評価を重視

回答評価器による生成評価の結果

- 2つのモデル (Mistral-7B-Instruct-v0.3, llmjp-3-13b) をSFTに利用
- 25%の多くのケースで0%と比べて精度が向上



まとめと今後の課題

- まとめ

- ドメインテキストとアノテーションによってドメイン特化を達成
- 特に厳しいフィルタリング (25%) を適用することによって得られる高品質なQAデータセットを用いた場合に精度向上が見られた

- 今後の課題

- 質問、回答の生成にはMixtral-8x7Bを用いたが、より日本語性能の高いモデルを利用することで精度が向上する可能性がある
- SFTモデルでQA生成し、フレームワークをループすることでさらなる改善があるか調査が必要
- より様々なドメインについて実験を行う
- 自己評価のみならず、客観評価での分析

付録. karakuri-apmによる生成評価の結果

ベースモデル	dataset	NLP2024(n=93)					it(n=383)					sp(n=208)				
llm-jp-13B instruct	dataset割合	0	25	50	75	100	0	25	50	75	100	0	25	50	75	100
	clean	28 30.10	31 33.33	7 7.53	6 6.45	7 7.53	110 28.72	51 13.32	44 11.48	30 7.83	19 4.96	32 15.38	5 2.40	10 4.81	9 4.33	11 5.29
	wrong	65 69.89	62 66.66	86 92.47	86 92.47	86 92.47	271 70.76	332 86.68	339 88.51	353 92.16	364 95.04	176 84.62	203 97.60	198 95.19	199 95.67	197 94.71
	invalid	0 0%	0 0%	0 0%	1 1.08%	0 0%	2 0.52%	0 0%								
gemma2 9B instruct	clean	36 38.71	5 5.38	0 0	0 0	0 0	112 29.24	22 5.74	14 3.66	3 0.78	4 1.04	36 17.31	2 0.96	6 2.88	2 0.96	4 1.92
	wrong	57 61.29	88 94.62	93 100	93 100	93 100	268 69.97	361 94.26	369 96.34	380 99.22	379 98.96	172 82.69	206 99.03	202 97.12	206 99.03	204 98.08
	invalid	0 0%	0 0%	0 0%	0 0%	0 0%	3 0.78%	0 0%								
mistral 7B instruct	clean	20 21.51	1 1.08	1 1.08	5 5.38	2 2.15%	56 14.62	7 1.83	1 0.26	1 0.26	0 0%	10 4.81	1 0.48	2 0.96	2 0.96	1 0.48
	wrong	73 78.49	90 96.77	89 95.70	83 89.25	89 95.70	327 85.38	376 98.17	382 99.74	380 99.21	382 99.74	198 95.19	207 99.52	206 99.04	206 99.04	207 99.52
	invalid	0 0%	2 2.15%	3 3.23%	5 5.38	2 2.15%	0 0%	0 0%	0 0%	2 0.52%	1 0.26%	0 0%	0 0%	0 0%	0 0%	0 0%