Evaluating Language Models as Synthetic Data Generators

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang,
Kiril Gashteovski, Carolin Lawrence, Sean Welleck, Graham Neubig

ACL 2025

[論文リンク]

ACL読み会2025

名古屋大学 笹野研究室

D1 大鹿 雅史

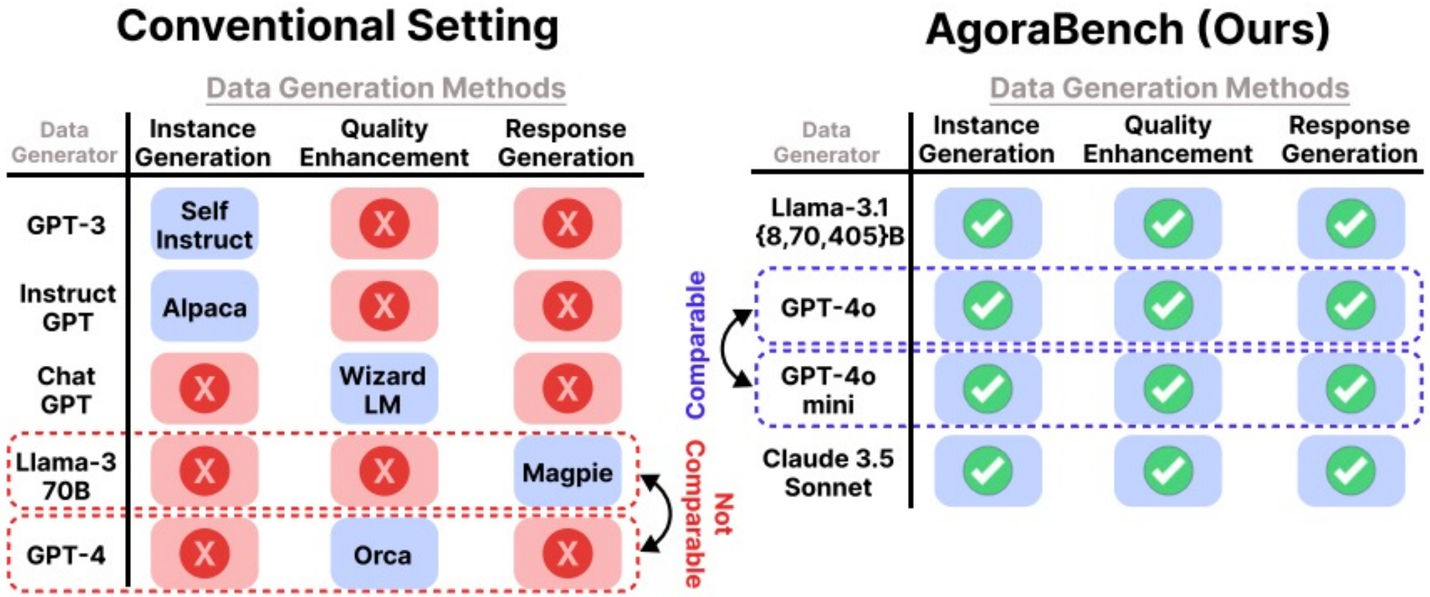
概要

- LLMを利用した合成データの作成はよく行われている
  - LLMの合成データ生成能力の比較は十分にされていない
- LLMのデータ生成能力を検証するAGORABENCHを構築し検証
  - 9つのデータ生成の設定で６つのLLMの生成能力を評価
  - LLMによって得意な領域や不得意な領域が存在することを示した

- 選定理由
  - LLMの学習への合成データの利用は広まっている
  - 数多くのLLMからどのモデルを利用するかの足がかりになる

背景

- 事後学習はLLMのタスクの解決能力を高めるために有望
  - 人手アノテーションは品質の保証はされるがスケール化が困難
    →合成データを利用することでスケール化が可能
- 数多く公開されるLLMのデータ生成能力を測るには統一的な実験が必要
  - 既存研究は実験設定がバラバラ→設定を揃えモデルのデータ生成能力を比較



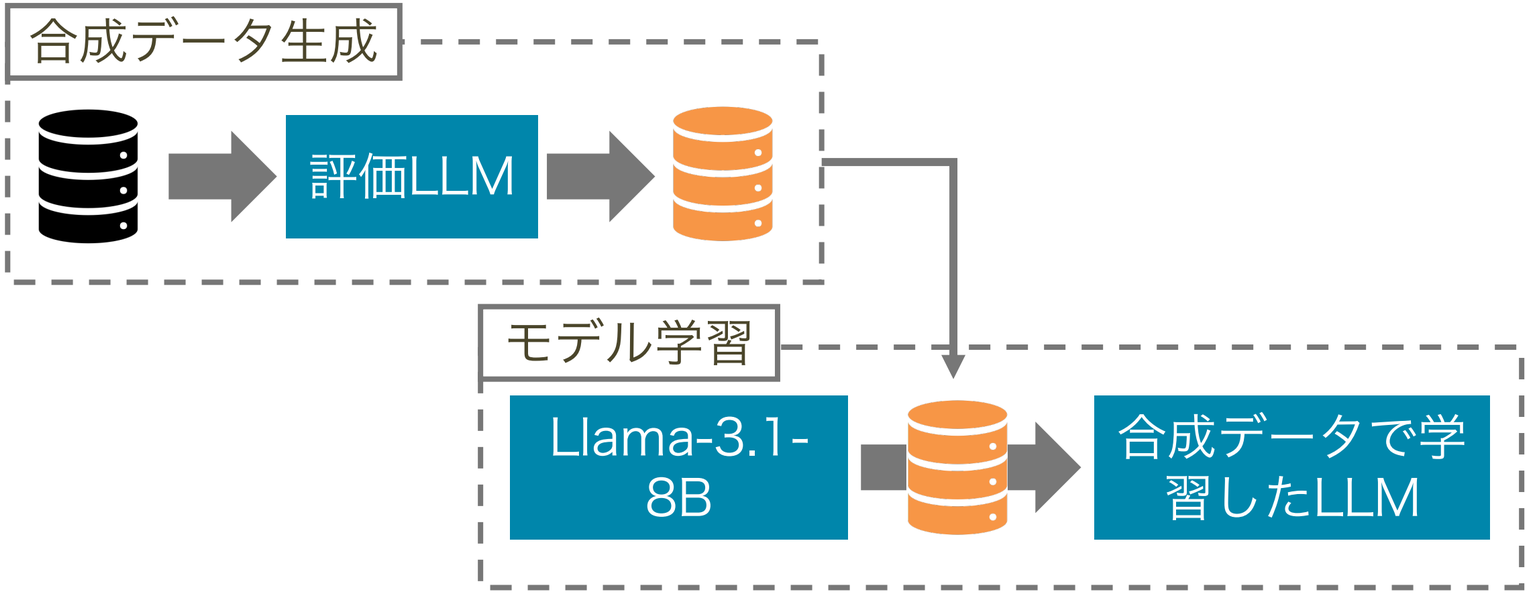
AGORABENCH

データ生成能力の評価方法

1. 既存のデータセットをもとにLLMを利用して合成データの生成
2. 事前学習済のモデルをSupervised Fine-Tuning
3. Instructionモデルと性能の比較を行う

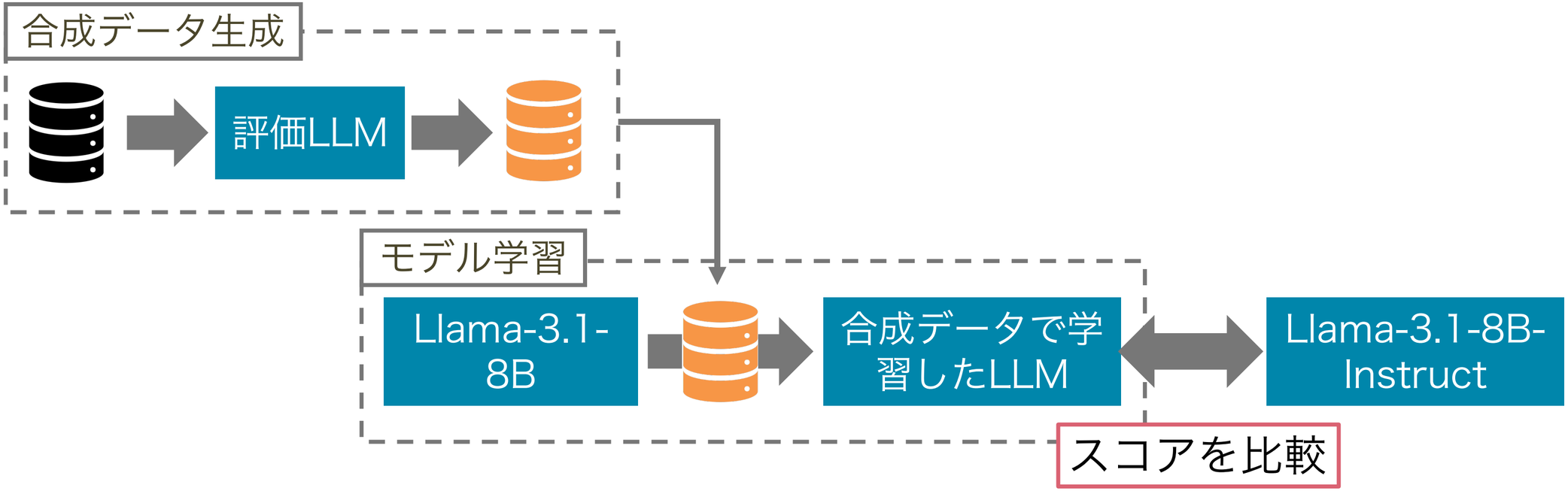
データ生成能力の評価方法

1. 既存のデータセットをもとにLLMを利用して合成データの生成
2. 事前学習済のモデルをSupervised Fine-Tuning
3. Instructionモデルと性能の比較を行う



データ生成能力の評価方法

1. 既存のデータセットをもとにLLMを利用して合成データの生成
2. 事前学習済のモデルをSupervised Fine-Tuning
3. Instructionモデルと性能の比較を行う



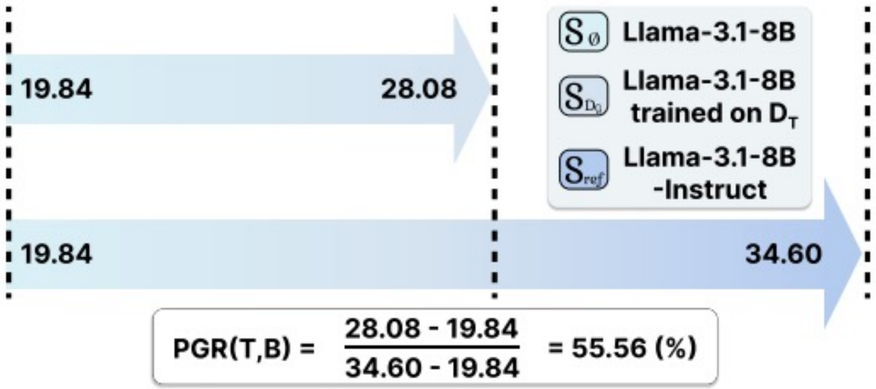
評価指標

Performance Gap Recovered (PGR)

- 合成データで学習したモデルがInstructionモデルのスコアをどれくらい再現するか

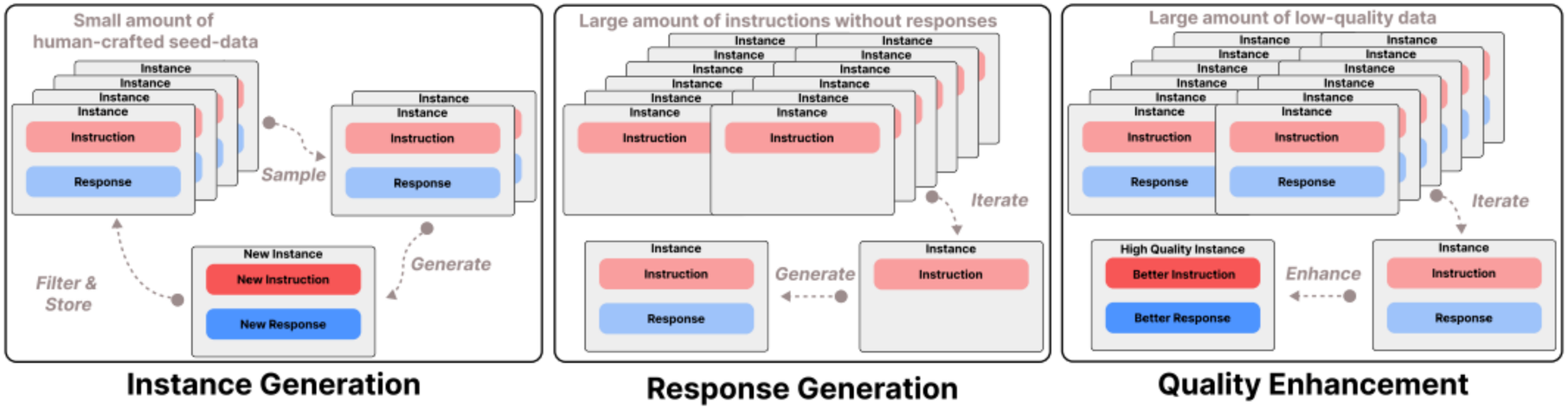
$$PGR(G,B) = \frac{score_B(S_{D_G}) - score_B(S_\emptyset)}{score_B(S_{ref}) - score_B(S_\emptyset)} \times 100$$

- $S_\emptyset$：Llama-3.1-8B
- S_{ref}：Llama-3.1-8B-Instruct（25M以上の合成データと公開データで事後学習）
- S_{D_G}：合成データで事後学習したLlama-3.1-8B



AGORABENCH

- ３つのドメイン及び３つの合成手法に関するデータセット
  - ３つのドメイン
    - 数学, コード, Instruction following
  - ３つの合成手法
    - Instance Generation：既存の事例をコンテキストとして新しい事例を生成する
    - Response Generation：instractionのみのデータに応答を生成する
    - Quality Enhancement：指示-応答ペアを入力し品質改善する



AGORABENCH

- 以下のデータセットを元にしてデータ合成
- 各設定で10,000件データを生成し生徒モデルの学習に利用

Domain	Data Generation Method	Seed Data	Seed Data Size
Math	**Instance Generation**	GSM8K, MATH (train set)	14,856
	Response Generation	Magpie-Reasoning (math)	10,000
	Quality Enhancement	WebInstruct (math)	10,000
Code	**Instance Generation**	MBPP (train set), xP3x	874
	Response Generation	Magpie-Reasoning (code)	10,000
	Quality Enhancement	CoNaLa	10,000
Inst. Follow	**Instance Generation**	LIMA	503
	Response Generation	Magpie-Pro	10,000
	Quality Enhancement	WebInstruct (code)	10,000

実験

実験設定

- 評価LLM
  - GPT-4o, GPT-4o-mini, Claude-3.5-Sonnet, Llama-3.1-8B/70B/405B-Instruct
- 生徒モデル
  - Llama-3.1-8B
- 評価ベンチマーク
  - Math：GSM8K, MATH
  - Code：MBPP, HumanEval
  - Instruction Following：AlpacaEval 2.0, Arena-Hard

$$PGR(G,B) = \frac{score_B(S_{D_G}) - score_B(S_\emptyset)}{score_B(S_{ref}) - score_B(S_\emptyset)} \times 100$$

実験結果

- GPT-4oは多くの設定で高い性能を示す
  - Quality EnhancementのCodeにおいては低性能
- ClaudeがQuality Enhancementで高い性能を示す
- LLaMAがcodeのドメインで強い結果を示す

Data Generator	Instance Generation				Response Generation				Quality Enhancement			
	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg
GPT-4o	**20.6**	**73.6**	**46.1**	**46.8**	46.7	28.5	**30.3**	**35.2**	**21.9**	-8.8	7.1	6.7
GPT-4o-mini	16.1	41.9	18.0	25.3	**48.1**	18.9	13.7	26.9	17.8	-11.2	9.9	5.5
Claude-3.5-Sonnet	8.9	23.4	40.1	24.1	29.0	**44.5**	12.7	28.8	15.7	**16.1**	**21.8**	**17.9**
Llama-3.1-405B	10.4	12.6	7.4	10.1	31.7	35.4	4.9	24.0	-11.8	7.5	3.6	-0.2
Llama-3.1-70B	9.6	58.7	6.5	24.9	23.0	37.1	4.5	21.5	-21.8	6.9	2.7	-4.1
Llama-3.1-8B	6.5	55.7	6.2	22.8	27.6	25.8	5.0	19.4	-1.7	15.4	3.0	5.6

実験結果

- GPT-4oは多くの設定で高い性能を示す
  - Quality EnhancementのCodeにおいては低性能
- ClaudeがQuality Enhancementで高い性能を示す
- LLaMAがcodeのドメインで強い結果を示す

Data Generator	Instance Generation				Response Generation				Quality Enhancement			
	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg
GPT-4o	**20.6**	**73.6**	**46.1**	**46.8**	46.7	28.5	**30.3**	**35.2**	**21.9**	-8.8	7.1	6.7
GPT-4o-mini	16.1	41.9	18.0	25.3	**48.1**	18.9	13.7	26.9	17.8	-11.2	9.9	5.5
Claude-3.5-Sonnet	8.9	23.4	40.1	24.1	29.0	**44.5**	12.7	28.8	15.7	**16.1**	**21.8**	**17.9**
Llama-3.1-405B	10.4	12.6	7.4	10.1	31.7	35.4	4.9	24.0	-11.8	7.5	3.6	-0.2
Llama-3.1-70B	9.6	58.7	6.5	24.9	23.0	37.1	4.5	21.5	-21.8	6.9	2.7	-4.1
Llama-3.1-8B	6.5	55.7	6.2	22.8	27.6	25.8	5.0	19.4	-1.7	15.4	3.0	5.6

実験結果

- GPT-4oは多くの設定で高い性能を示す
  - Quality EnhancementのCodeにおいては低性能
- ClaudeがQuality Enhancementで高い性能を示す
- LLaMAがcodeのドメインで強い結果を示す

Data Generator	Instance Generation				Response Generation				Quality Enhancement			
	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg
GPT-4o	**20.6**	**73.6**	**46.1**	**46.8**	46.7	28.5	**30.3**	**35.2**	**21.9**	-8.8	7.1	6.7
GPT-4o-mini	16.1	41.9	18.0	25.3	**48.1**	18.9	13.7	26.9	17.8	-11.2	9.9	5.5
Claude-3.5-Sonnet	8.9	23.4	40.1	24.1	29.0	**44.5**	12.7	28.8	15.7	**16.1**	**21.8**	**17.9**
Llama-3.1-405B	10.4	12.6	7.4	10.1	31.7	35.4	4.9	24.0	-11.8	7.5	3.6	-0.2
Llama-3.1-70B	9.6	58.7	6.5	24.9	23.0	37.1	4.5	21.5	-21.8	6.9	2.7	-4.1
Llama-3.1-8B	6.5	55.7	6.2	22.8	27.6	25.8	5.0	19.4	-1.7	15.4	3.0	5.6

実験結果

- GPT-4oは多くの設定で高い性能を示す
  - Quality EnhancementのCodeにおいては低性能
- ClaudeがQuality Enhancementで高い性能を示す
- LLaMAがcodeのドメインで強い結果を示す

Data Generator	Instance Generation				Response Generation				Quality Enhancement			
	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg	Math	Code	Inst.	Avg
GPT-4o	**20.6**	**73.6**	**46.1**	**46.8**	46.7	28.5	**30.3**	**35.2**	**21.9**	-8.8	7.1	6.7
GPT-4o-mini	16.1	41.9	18.0	25.3	**48.1**	18.9	13.7	26.9	17.8	-11.2	9.9	5.5
Claude-3.5-Sonnet	8.9	23.4	40.1	24.1	29.0	**44.5**	12.7	28.8	15.7	**16.1**	**21.8**	**17.9**
Llama-3.1-405B	10.4	12.6	7.4	10.1	31.7	35.4	4.9	24.0	-11.8	7.5	3.6	-0.2
Llama-3.1-70B	9.6	58.7	6.5	24.9	23.0	37.1	4.5	21.5	-21.8	6.9	2.7	-4.1
Llama-3.1-8B	6.5	55.7	6.2	22.8	27.6	25.8	5.0	19.4	-1.7	15.4	3.0	5.6

分析

良いデータ生成LLMの要素は何か？

1. タスク性能とデータ生成能力の関係
2. 生成データから改善可能性を予想できるか？

タスク性能とデータ生成能力の関係

- 6つのベンチマークの平均性能とタスク性能を比較
- 実験結果
  - タスクを解く性能とデータ生成能力は必ずしも一致しない
  - コスト面とのバランスではGPT-4o, GPT-4o-mini, Llama3.1-8Bが高性能

Data Generator	API Cost		Prob. Solv.	Data Gen.
	Input	Output	Avg	Agora Bench
GPT-4o	$2.50	$10.00	80.9	29.5%
GPT-4o-mini	$0.15	$0.60	75.4	19.2%
Claude-3.5-Sonnet	$3.00	$15.00	80.5	23.6%
Llama-3.1-405B	$1.79	$1.79	75.0	11.3%
Llama-3.1-70B	$0.35	$0.40	69.6	14.1%
Llama-3.1-8B	$0.055	$0.055	50.2	15.9%

タスク性能とデータ生成能力の関係

- 6つのベンチマークの平均性能とタスク性能を比較
- 実験結果
  - タスクを解く性能とデータ生成能力は必ずしも一致しない
  - コスト面とのバランスではGPT-4o, GPT-4o-mini, Llama3.1-8Bが高性能

Data Generator	API Cost		Prob. Solv.	Data Gen.
	Input	Output	Avg	Agora Bench
GPT-4o	$2.50	$10.00	80.9	29.5%
GPT-4o-mini	$0.15	$0.60	75.4	19.2%
Claude-3.5-Sonnet	$3.00	$15.00	80.5	23.6%
Llama-3.1-405B	$1.79	$1.79	75.0	11.3%
Llama-3.1-70B	$0.35	$0.40	69.6	14.1%
Llama-3.1-8B	$0.055	$0.055	50.2	15.9%

タスク性能とデータ生成能力の関係

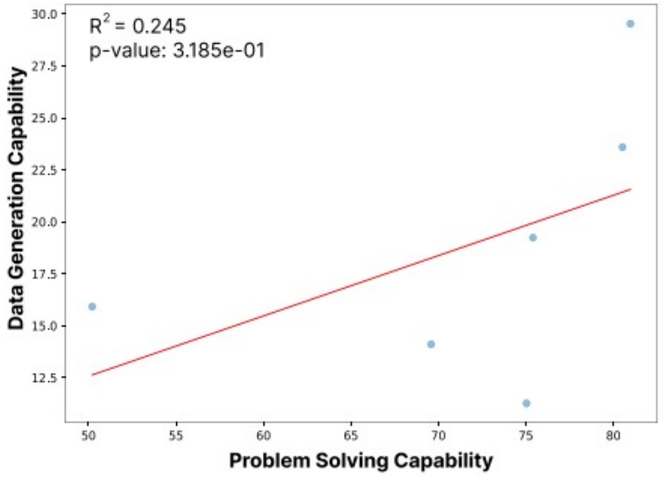
- 6つのベンチマークの平均性能とタスク性能を比較
- 実験結果
  - タスクを解く性能とデータ生成能力は必ずしも一致しない
  - コスト面とのバランスではGPT-4o, GPT-4o-mini, Llama3.1-8Bが高性能

Data Generator	API Cost		Prob. Solv.	Data Gen.
	Input	Output	Avg	Agora Bench
GPT-4o	$2.50	$10.00	80.9	29.5%
GPT-4o-mini	$0.15	$0.60	75.4	19.2%
Claude-3.5-Sonnet	$3.00	$15.00	80.5	23.6%
Llama-3.1-405B	$1.79	$1.79	75.0	11.3%
Llama-3.1-70B	$0.35	$0.40	69.6	14.1%
Llama-3.1-8B	$0.055	$0.055	50.2	15.9%

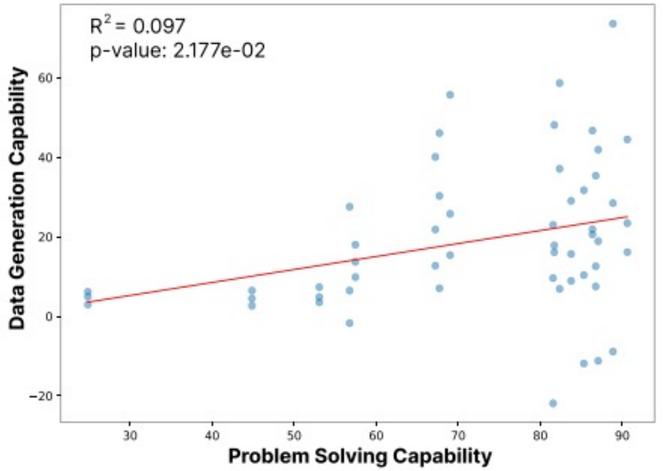
タスクを解けるLLMは良いデータ生成器なのか？

- 2つの設定で線形回帰分析を実施
  1. 6つのLLMのベンチマークの平均スコアとPGRの関係性
  2. 6つのLLMのベンチマークの個別のスコアとPGRの関係性

- 実験結果
  - ●強い相関は見られない→タスク性能からデータ生成能力を測ることはできない



Coarse-grained Analysis:
Average Scores across all Settings (n=6)



Fine-grained Analysis: Individual Scores
per Domain and Generation Setting (n=54)

生成データから改善可能性を予想できるか？

- 生成データのどの特性がPGRの改善に貢献するかを分析
- 生成データの評価指標
  - 応答品質
    - LLM-as-a-Judge：1～5のスコアを生成させ評価
    - 報酬モデル：品質のスカラー値を出力するモデルを利用
  - 指示の複雑さ
    - LLM-as-a-judge： 1～5のスコアを生成させ評価
  - 応答のパープレキシティ
    - 指示をLlama-3.1-8Bに入力した時の応答のパープレキシティを計測
  - 事例の多様性
    - 生成した事例の埋め込みのコサイン類似度を算出し多様性を定義

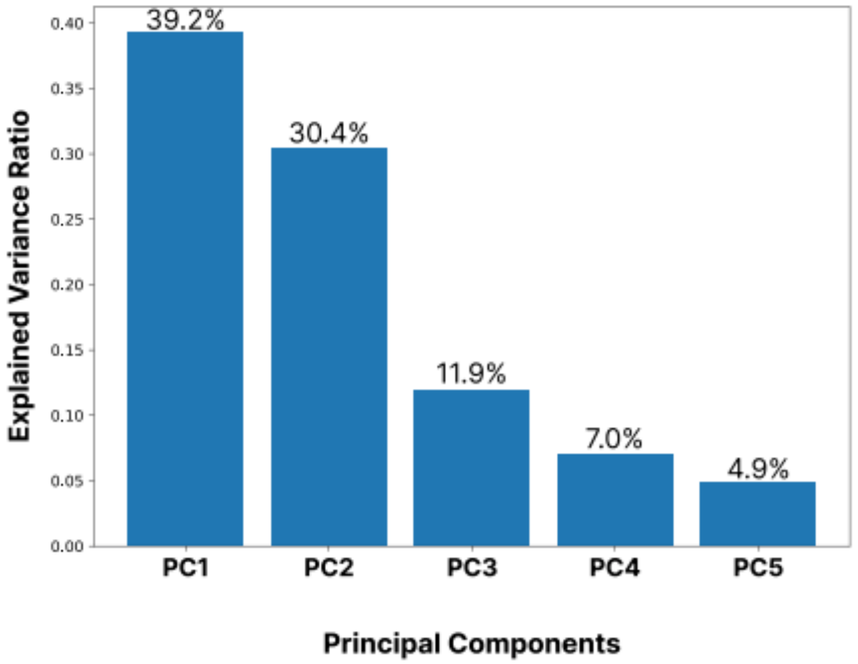
生成データから改善可能性を予想できるか？

- 生成データのどの特性がPGRの改善に貢献するかを分析
- 生成データの評価指標
  - 応答品質
    - LLM-as-a-Judge：1～5のスコアを生成させ評価
    - 報酬モデル：品質のスカラー値を出力するモデルを利用
  - 指示の複雑さ
    - LLM-as-a-judge： 1～5のスコアを生成させ評価
  - 応答のパープレキシティ
    - 指示をLlama-3.1-8Bに入力した時の応答のパープレキシティを計測
  - 事例の多様性
    - 生成した事例の埋め込みのコサイン類似度を算出し多様性を定義

これらを特徴量
として主成分分析

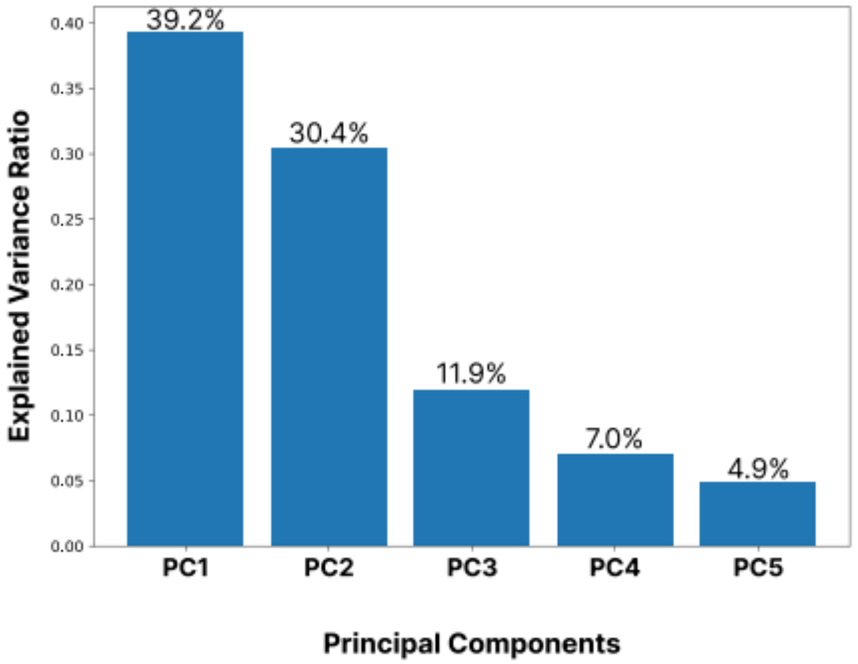
生成データから改善可能性を予想できるか？

- 上位5成分で93.4%を説明可能＝解釈可能なパターンがある
  - 指示や応答の多様性、指示の難しさなどが関係する



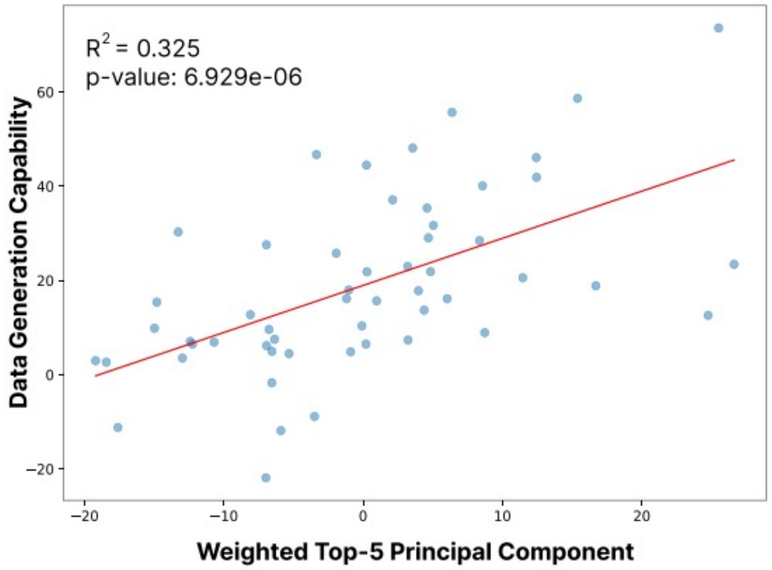
生成データから改善可能性を予想できるか？

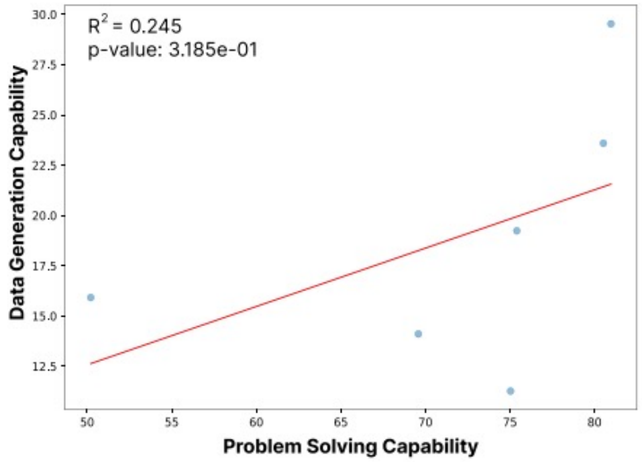
- 上位5成分で93.4%を説明可能＝解釈可能なパターンがある
  - 指示や応答の多様性、指示の難しさなどが関係する



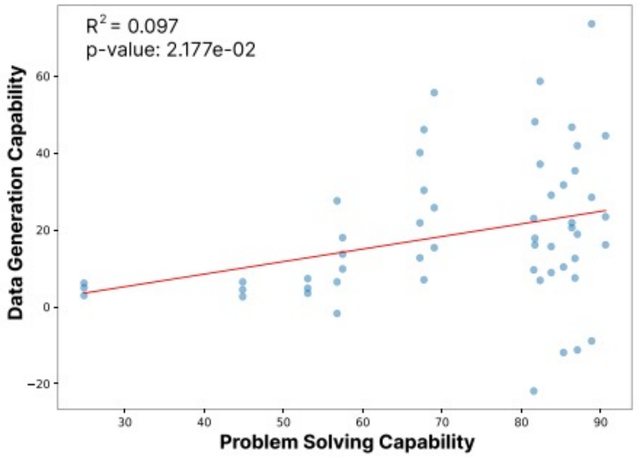
上位5成分とPGRの関係性

- 上位5成分に対して線形回帰分析を行いPGRとの関係性を考察
- LLMのタスク性能とPGRの関係性よりも説明可能性が向上
  ＝LLMのデータ生成能力を予測するには生成データを評価する必要がある





Coarse-grained Analysis:
Average Scores across all Settings (n=6)



Fine-grained Analysis: Individual Scores per Domain and Generation Setting (n=54)

まとめと疑問点

まとめ

- LLMのデータ生成能力を検証するための評価方法AGORABENCHを構築
- 6つのLLMについてデータ生成能力を評価
- 生成データのどの要素がPGRの改善に影響するかを検証

疑問点

- 複数のLLMを生徒モデルとして欲しい
  - 生徒とするLLMが異なればデータ構築に強いモデルは違いそう
- 比較がInstructモデルとの比較でいいのか？
  - 単純なSFTによる事後学習以外の処理が行われている