

最先端ソフトウェア工学ゼミ 成果報告

2024年7月11日

氏名 家村 康佑
所属 富士通



発表内容

1. 設定テーマと選定理由
 - ・プロンプトエンジニアリングの有用性評価
2. プロンプトエンジニアリング
3. LLMの評価手法
4. 調査内容
5. 調査結果
6. 考察
7. 所感
8. まとめ



1. 設定テーマと選定理由

■ 設定テーマ

最新LLMにおけるプロンプトエンジニアリングの有用性評価

■ 選定理由

- 最新のLLMでは回答の精度が高くなっており、従来のプロンプトエンジニアリングの必要性に疑問
- モデルの変化による、プロンプトの改善の必要性を確認し、今後のLLM活用に反映したい。
- プロンプトエンジニアリングが有用な、ジャンルやタスクを明確にし、LLM活用時の効率化の目安とする。

2. プロンプトエンジニアリング

- プロンプトを工夫して、LLMの精度向上を行う手法として多くの手法が提案・発表されている。

代表的なプロンプトエンジニアリング手法例

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

• Few-shot:

いくつかの回答方法を例示

• Few-shot-CoT:

例示の際に段階的に思考の手順を例示

• Zero-shot:

問題だけを与える

• Zero-shot-CoT:

問題に加え段階的に考える指示を追加

3. LLMの評価手法

■ 様々な評価手法

- 様々なタスク(文章分類、類似性判定)用に作成された、評価用データセットを用いて、LLMの出力を評価する手法
- LLMの出力を人間や、最新の別のLLMによって評価する手法

代表的な評価用データセット例

Dataset	タスク概要	Release
VQAv2	画像含めた問題	2017-04
GLUE	文章分類、類似性、関係性判断等 9種類で構成	2018-04
<u>MMLU</u>	自然言語のQA。数学、天文学等57のジャンル 4択問題で構成される1万6千問	2020-09
MATH	数学問題	2021-03
GSM8K	数学問題	2021-10

- 今回、プロンプトエンジニアリングを比較・評価する手法として、様々なジャンル、改変の容易性などからMMLUを用いた評価を実施



4. 調査内容

- Chain-of-Thought(CoT)でのMMLUの評価環境を改版して評価
- モデル: GPT3.5Turbo, GPT4Turbo
- 改版内容

- ① Few-Shot-CoT
- ② ①に「Step-by-Step」を追加
- ③ Few-shot
- ④ ③に「Step-by-Step」を追加

①, ②: 最後にStep-by-Stepを追加することによる効果を比較
 ③: 単純な回答を期待するFew-Shot
 ④: Zero-shot-CoTに近い形

① Few-Shot-CoT

Q: {Question1}
 A: CoT {Ans1}
 Q: {Question2}
 A: CoT {Ans2}

Q: {Question}
 A:

② ①+Step-by-Step

Q: {Question1}
 A: CoT {Ans1}
 Q: {Question2}
 A: CoT {Ans2}

Q: {Question}
 A: step by step

③ Few-Shot

Q: {Question1}
 A: {Ans1}
 Q: {Question2}
 A: {Ans2}

Q: {Question}
 A:

④ ③ + Step-by-step

Q: {Question1}
 A: {Ans1}
 Q: {Question2}
 A: {Ans2}

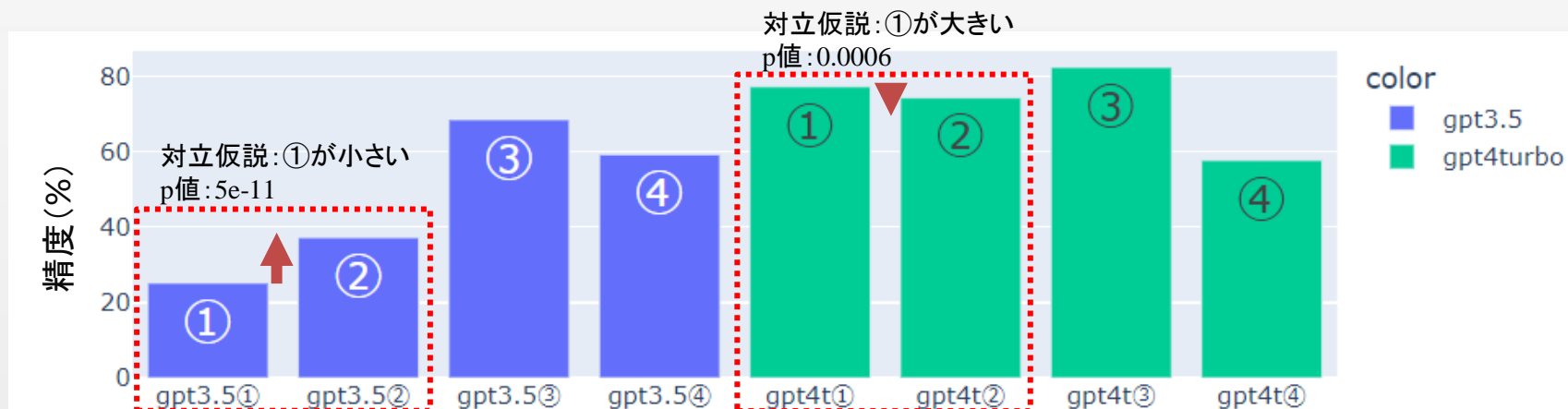
Q: {Question}
 Step by step
 A:

- Few-shot: 回答フォーマットを指定することにも使用
- 「Let's think Step-by-Step」を与えることによる回答の変化を確認



5. 調査結果

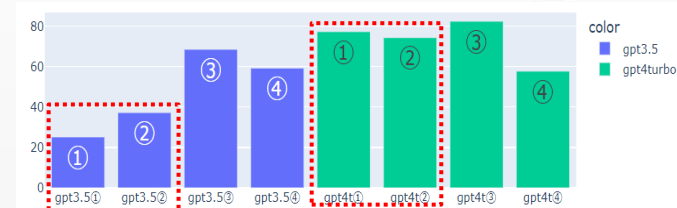
- ① Few-Shot-CoT
- ② ①に「Step-by-Step」を追加
- ③ Few-shot
- ④ ③に「Step-by-Step」を追加



- ①, ②の結果では、Step-by-Stepを追加することでGPT3.5Turboは精度が上がり、GPT4Turboは精度が下がる結果
- 元々の回答精度が低いものについて効果的
- ④で精度が下がる部分は、回答結果の漏れ(ご参考情報)

5. 調査結果

- ① Few-Shot-CoT
- ② ①に「Step-by-Step」を追加



57種のジャンルごとの①と②の精度差分(pt)



- GPT4 はベースラインが概ね高いが、「college_mathematics」、「college_computer_science」はさらなる精度向上が見られた。



6. 考察

- モデルの違いによるプロンプトエンジニアリングの効果
 - Step-by-Step により、GPT3.5Turboでは精度上昇、GPT4Turboでは精度低下
 - 元々の回答の精度が高い場合には、段階的な回答により回答の複雑性が増し、誤る場合が存在
 - 最新のLLMでは簡易なプロンプトでも十分な精度
- Step-by-Stepの効果的なジャンルは、段階的に考えることが必要な高度な課題（数学・コンピュータサイエンス等）

使用するモデルや、質問の難易度、期待する回答に応じて、適切にプロンプトの工夫をする必要がある





7. 所感

- LLMの評価方法の技術の習得
 - LLMの実際の評価手法の環境の構築
 - 評価方法や評価環境を確認し、実態・課題を認識
- 進化が早く、短期間の間に状況が変化
 - 最新のLLMや手法の習得、評価手法の改善が必要
- LLMの評価の難しさ
 - 自然言語の回答から、回答の抽出が困難
⇒ 回答結果をLLMで抽出するなど改善余地
 - ランダム性による回答の変動
 - 時間当たりの制限による回答の制限
 - 評価するための絶対的な時間
(ご参考: 20秒/問 × 16k問 ÷ 4日間)





8. まとめ

- プロンプトエンジニアリング (CoT) の効果を、LLM のモデルの新旧で評価を実施
- モデル・課題によりプロンプトエンジニアリングの効果は限定的
- 現時点では高度な課題の場合にStep-by-Stepにより精度が向上する傾向
- LLMは急速に進化を続けており、モデルに対するプロンプトエンジニアリングの技術や、評価の仕組みなどは継続的な習得、追求が必要

【ご参考】評価用データセットの改版

■ MMLU-Pro データセット

- Release: 2024-06
- MMLUで精度が頭打ちになり、評価に差がつかないことに言及
- 難易度を向上させた10択問題で対処
- MMLUは、CoTよりもDirect Answerが精度が高いことにも言及

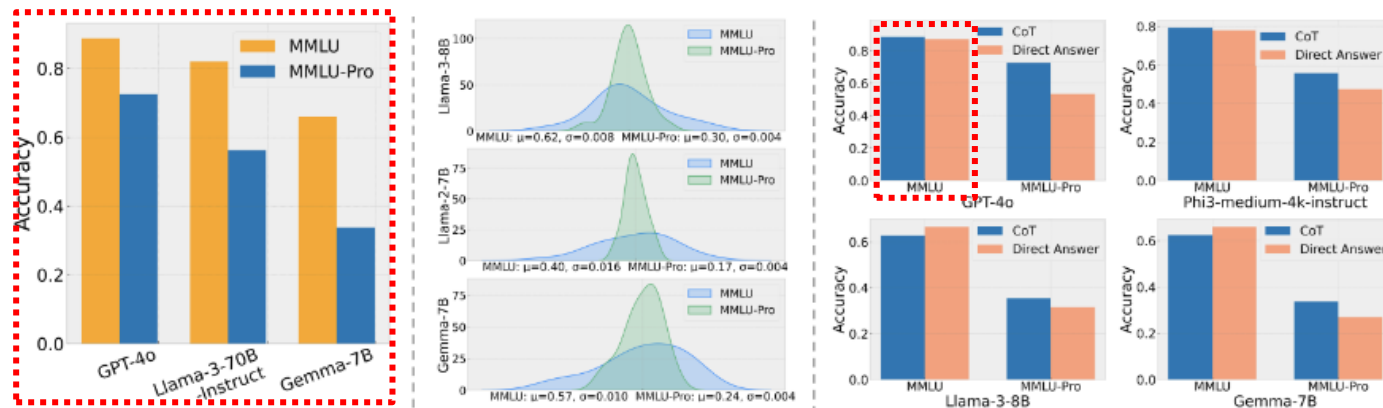


Figure 1: Comparing between MMLU and MMLU-Pro: (Left) Performance gap; (Center) Accuracy distributions affected by 24 prompts, with taller and thinner profiles indicating more stability and shorter and wider profiles indicating greater fluctuations; (Right) Performance using CoT vs. Direct.

⇒ モデルの能力と問題の難しさに対する調査・比較は重要



ありがとうございました。

【ご参考】使用したコンテキスト例

①, ②のFew-shot-CoT

回答例1

4択

CoT

Q: How did Donald Trump attack globalization in the 2016 campaign?
 (A) Globalization had made men like him too rich (B) Globalization only benefited certain American states, such as New York (C) Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it (D) Globalization encouraged damaging trade wars
 A: Let's think step by step. We refer to Wikipedia articles on us foreign policy for help. Trump attacked globalization because he believed ordinary Americans lost jobs due to it, and so he wanted to blame liberals who had encouraged it. The answer is (C).

回答例2

4択

CoT

Q: How did NSC-68 change U.S. strategy?
 (A) It globalized containment. (B) It militarized containment. (C) It called for the development of the hydrogen bomb. (D) All of the above
 A: Let's think step by step. We refer to Wikipedia articles on us foreign policy for help. NSC-68 outlined a variety of courses of action, including globalization of containment, militarization of containment, and the development of the hydrogen bomb. The answer is (D).

質問

4択

②④で追加

Q: How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour?
 (A) Defensive realists place greater emphasis on the role of international institutions (B) Defensive realists place less emphasis on geographical factors (C) Offensive realists give more priority to the national interest than Defensive realists. (D) Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers
 A: Lets think step-by-step

③, ④のFew-shot

Q: How did Donald Trump attack globalization in the 2016 campaign?
 (A) Globalization had made men like him too rich (B) Globalization only benefited certain American states, such as New York (C) Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it (D) Globalization encouraged damaging trade wars
 A: The answer is (C).

Q: How did NSC-68 change U.S. strategy?
 (A) It globalized containment. (B) It militarized containment. (C) It called for the development of the hydrogen bomb. (D) All of the above
 A: The answer is (D).

Q: How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour?
 (A) Defensive realists place greater emphasis on the role of international institutions (B) Defensive realists place less emphasis on geographical factors (C) Offensive realists give more priority to the national interest than Defensive realists. (D) Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers
 Lets think step-by-step
 A:

【ご参考】③④の回答の違い

Q: You want to cluster 7 points into 3 clusters using the k-Means Clustering algorithm. Suppose after the first iteration, clusters C1, C2 and C3 contain the following two-dimensional points: C1 contains the 2 points: $\{(0,6), (6,0)\}$ C2 contains the 3 points: $\{(2,2), (4,4), (6,6)\}$ C3 contains the 2 points: $\{(5,5), (7,7)\}$ What are the cluster centers computed for these 3 clusters? (A) C1: (3,3), C2: (4,4), C3: (6,6) (B) C1: (3,3), C2: (6,6), C3: (12,12) (C) C1: (6,6), C2: (12,12), C3: (12,12) (D) C1: (0,0), C2: (48,48), C3: (35,35)

A: The answer is (A).

Q: You want to cluster 7 points into 3 clusters using the k-Means Clustering algorithm. Suppose after the first iteration, clusters C1, C2 and C3 contain the following two-dimensional points: C1 contains the 2 points: $\{(0,6), (6,0)\}$ C2 contains the 3 points: $\{(2,2), (4,4), (6,6)\}$ C3 contains the 2 points: $\{(5,5), (7,7)\}$ What are the cluster centers computed for these 3 clusters? (A) C1: (3,3), C2: (4,4), C3: (6,6) (B) C1: (3,3), C2: (6,6), C3: (12,12) (C) C1: (6,6), C2: (12,12), C3: (12,12) (D) C1: (0,0), C2: (48,48), C3: (35,35)

Let's think step by step.

A: To find the cluster centers for each cluster, we calculate the mean of the points in each cluster.

For cluster C1:

- Points: (0,6) and (6,0)
- Mean: $((0+6)/2, (6+0)/2) = (3,3)$

For cluster C2:

- Points: (2,2), (4,4), (6,6)
- Mean: $((2+4+6)/3, (2+4+6)/3) = (4,4)$

For cluster C3:

- Points: (5,5) and (7,7)
- Mean: $((5+7)/2, (5+7)/2) = (6,6)$

Thus, the cluster centers are:

- C1: (3,3)
- C2: (4,4)
- C3: (6,6)

The correct answer is (A) C1: (3,3), C2: (4,4), C3: (6,6).



【ご参考】④の場合の回答の抜きの例

■ 「the correct answer is:」で回答が終わり、最終的な解が不明な例

From laboratory measurements we know that a particular spectral line formed by hydrogen appears at a wavelength of 486.1 nanometers (nm). The spectrum of a particular star shows the same hydrogen line appearing at a wavelength of 485.9 nm. What can we conclude?

(A) The star is getting hotter. (B) The star is getting colder. (C) The star is moving toward us. (D) The star is moving away from us.

Let's think step by step.

A: The shift in the wavelength of the spectral line from its original position indicates a Doppler shift. A shift to a shorter wavelength (from 486.1 nm to 485.9 nm) is known as a blue shift, which occurs when an object is moving toward the observer. Therefore, the correct answer is:

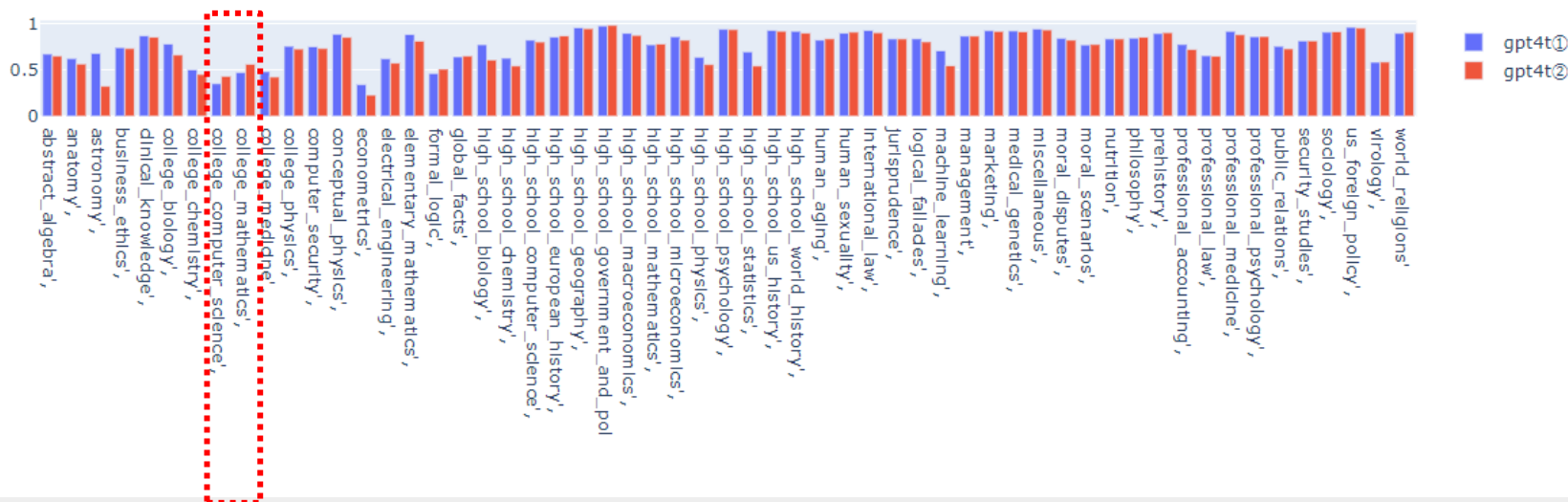
実験室での測定から、水素によって形成される特定のスペクトル線が486.1ナノメートル (nm) の波長で現れることがわかっている。特定の星のスペクトルでは、波長485.9 nmに同じ水素線が現れる。結論はどうでしょうか？

(A) その星はだんだん熱くなっている。(B) 星がだんだん寒くなっている。(C) 星がこちらに向かって動いている。(D) 星は私たちから遠ざかっている。
順を追って考えましょう。

スペクトル線の波長の元の位置からのシフトはドップラーシフトを示す。より短い波長 (486.1 nmから485.9 nm) へのシフトは青シフトとして知られており、物体が観察者に向かって移動するときに起こる。したがって、正解は次のようになります。

【ご参考】精度の絶対値

- GPT4Turbo の ①, ②で精度向上が見られたジャンルは、他のジャンルと比較したときに①の精度が、相対的に精度が低い傾向
- GPT4o を用いた追加評価で、①のパターンではさらなる精度向上を確認



	GPT3.5Turbo①	GPT4Turbo①	GPT4o①
college_computer_science	8.0%	35.0%	71.0%
college_mathematics	7.0%	47.0%	67.0%



参考文献

- MMLU:
arxiv.org/pdf/2009.03300
- MMLU-Pro:
arxiv.org/pdf/2406.01574
- Large Language Models are Zero-Shot Reasoners
arxiv.org/pdf/2205.11916
- Chain-of-Thought Hub:
[FranxYao/chain-of-thought-hub: Benchmarking large language models' complex reasoning ability with chain-of-thought prompting \(github.com\)](https://github.com/FranxYao/chain-of-thought-hub)

