# 難解な入力単語を用いた日本語 CommonGen タスクによる LLM の文生成能力評価

鈴木雅人 [1]， 新納浩幸 [2]
[1] 茨城大学大学院理工学研究科情報工学専攻, [2] 茨城大学大学院理工学研究科情報科学領域
{23NM726L, hiroyuki.shinnou.0828}@vc.ibaraki.ac.jp

## 概要

大規模言語モデル（LLM）の出現により、機械による文生成の能力は大きく向上した．それに伴い、LLM の出現以前に利用されていた文生成能力を測るタスクは現在の LLM の能力を測る指標としては適切なものでなくなってしまったものが多い．その一つに常識推論能力を測る CommonGen がある．本研究では、文生成タスクの CommonGen における入力単語を特定の分野に絞った専門用語にすることで、さらに高い文生成能力を求められるタスクにする試みを行った．更に試験的に作成したデータセットを用いて、LLM による生成文の評価と人手評価を行い、LLM の文生成能力と自動評価の可能性について調査した．

## 1 はじめに

自然言語処理には常識推論と呼ばれる問題があり，我々人間が持つ常識と呼ばれるような巨大な背景知識に矛盾しないように推論できるかという重要な問題である．そのような常識推論能力を測るデータセットはいくつも考案されており，[1][2][3][4][5][6] そのなかで Lin らは CommonGen という新しい常識推論のタスクを提案した [7]．CommonGen は，数個のキーワードを入力し，それらキーワードを用いて妥当な文を生成するという制約付き文生成のタスクである．しかし，近年では，高性能な LLM の出現により CommonGen 形式のデータセットは LLM にとって難しいタスクではないことが示されている．[8]

本研究では， CommonGen タスクを LLM においても有効なものとするため，入力キーワードを専門用語に限定し，タスクの難易度を高める試みを行った．結果，近年の LLM においても，意味が通っていない文を作成するということを確認し，タスクの難易度を上げることに成功した．また機械による自動評価では，人手評価と相関が確認できたが，大きく異なる場合も見られ各評価の違いについて今後の課題を残す結果となった．

## 2 関連研究

自然言語処理の分野では QA や対話などで，深い意味理解が必要となる場面で常識推論が使われ，様々なタスクが考案されている．

CommonsenseQA [1] は元となる言葉に関係する 3 つの言葉を用意し，関係する元の言葉を含みながら，3 つの言葉の内それぞれひとつずつのみに当てはまる質問を用意する．またこの際に，追加で元となる言葉に関係する 2 つの誤答用の言葉が用意され，3 つの質問は 2 つの誤答用の言葉には当てはまらないようにする．そうして用意された 3 つの質問に対して 5 つの選択肢となる言葉からそれぞれ，どの言葉が当てはまるかを選択するタスクである．SocialIQA [2] は社会的に一般的な状況を提示し，その状況下で取る行動やその状況下での心情を尋ねる質問と 3 つの選択肢が与えられ，適切な回答を選ぶタスクである．WinoGrande [3] は 2 つの文章とその文章中にある代名詞に対して 2 つの単語の選択肢が与えられ，各文章に代名詞の指す正しい単語を選択するタスクであり，これは Winograd Schema Challenge からバイアスを除去し，クラウドソージングの手続きを改善したものである．KUCI [4] は日本語において中断されている文章とそれに続く蓋然的な関係を持つ文章を 4 つの選択肢から 1 つ選択するタスクである．SWAG [5] は「ある場面でのビデオキャプション」と 4 つの「次の場面でのキャプション」の選択肢となる文章を提示し，選択肢から本物の次の場面でのキャプションを選択するタスクである．HellaSwag [6] は SWAG を元としてさらに難しい不正解の選択肢を導入する Adversarial Filtering

や元となるビデオキャプションの厳選などを行い，SWAG を改善したものである．

これらタスクは基本的には選択式の問題である．Lin らが発表した CommonGen[7] は制約付き文生成のタスクであり，選択式の問題では扱えない問題を扱っていると考えられる．

CommonGen は数個の単語を入力として，それらの単語を含む常識的な文を作る制約付き文生成タスクとなっている．例えば「彼，犬，フリスビー」という単語を与えたなら「彼が犬にフリスビーを投げる」などといった文を生成するのが目標である．しかし論文内では，「犬が彼にフリスビーを投げる」という文法的には正しくても常識的にはおかしな文が生成されることがあるとして，常識推論能力を測る CommonGen タスクを考案し，英語のデータセットを公開した．

また鈴木らが日本語での CommonGen タスクに対して ChatGPT を用いて文を生成させる論文 [8] が発表されている．研究内では，鈴木らが作成した日本語版の CommonGen のデータセットを用いて，ChatGPT-3.5 と ChatGPT-4 で文を生成させ，T5 を用いて生成した文と比較を行い，ChatGPT-4 では，ChatGPT-3.5 や T5 で生成が上手くいかなかったテストデータにおいて，ほとんどの場合で自然な生成をできており，論文内で日本語 CommonGen タスクは既に解決可能な問題であると結論付けている．

# 3 実験

## 3.1 テストデータの作成

本研究では，文生成タスクである CommonGen を現在の LLM に見合ったタスクにするために入力単語を一般的な単語から専門的な単語に変更する．また生成能力を測るために，テストデータの難易度が段階的になるよう設定する．専門な単語の分野は，情報，哲学，経済分野の単語として，表 1 のような組み合わせで設定した．表 1 中の番号が若いものほど広い分野の専門用語が出現し，難易度が高いことが分かる．各パターン 10 個のテストデータ，計 30 個のテストデータを作成した．

## 3.2 LLM による文生成

データセットが現在の LLM においても有効か確認するため claude 3.5 sonet，gemini 1.5，command R+ の三つの LLM を用いて各 LLM において 3.1 で作成したテストデータを入力として生成を行った．生成に用いたプロンプトを表 2 に示す．

**表 1** 分野別の単語組み合わせ

| | 単語 1 | 単語 2 | 単語 3 |
|---|---|---|---|
| パターン 1 | 情報 | 哲学 | 経済 |
| パターン 2 | 情報 | 情報 | 経済 |
| パターン 3 | 情報 | 情報 | 情報 |

**表 2** LLM による文生成のためのプロンプト例

以下の 3 単語を含む一般的な文を作成してください

## 単語
セキュリティ, 道徳, 買収
## ルール
- 3 単語を必ず含む
- なるべく簡潔な文
- 意味が分かり、かつ常識的な文

# 4 生成文の評価

3.2 節で生成した文を人手と ChatGPT-4o により評価した．

## 4.1 人手による評価

各 LLM が生成した文を人手によって比較評価した．評価方法は同一の入力から得られた各 LLM の出力をランダムに並び替えたものを，1 から 3 まで順番を付けるよう指示した．これを計 30 個，16 名に行ってもらった．この評価の結果を図 1 に示す．

全体の評価結果を見ると，各 LLM の性能差はあまり大きくは出ていないことが分かる．特徴的と言える点は，Claude 3.5 sonet の 2 位評価と Commnad R+ の 3 位評価が多い点が挙げられる．

## 4.2 ChatGPT-4o による評価

ChatGPT-4o を用いて各 LLM が生成した文の評価を行った．

まず専門用語の組み合わせについて検証を行うため， ChatGPT-4o に Claude 3.5 sonet が生成した文を 10 点満点で評価を行わせ，表 1 で示した専門用語の組み合わせにより難易度の変化が起きているかを確認した．プロンプトを 3 に，評価結果を表 4 に示す．評価結果から情報分野のみで統一された場合では，評価の平均点が高く情報，哲学，経済の組み合わせと情報 2 単語，経済 1 単語の組み合わせでは，あまり差がが出ていないことが分かる．

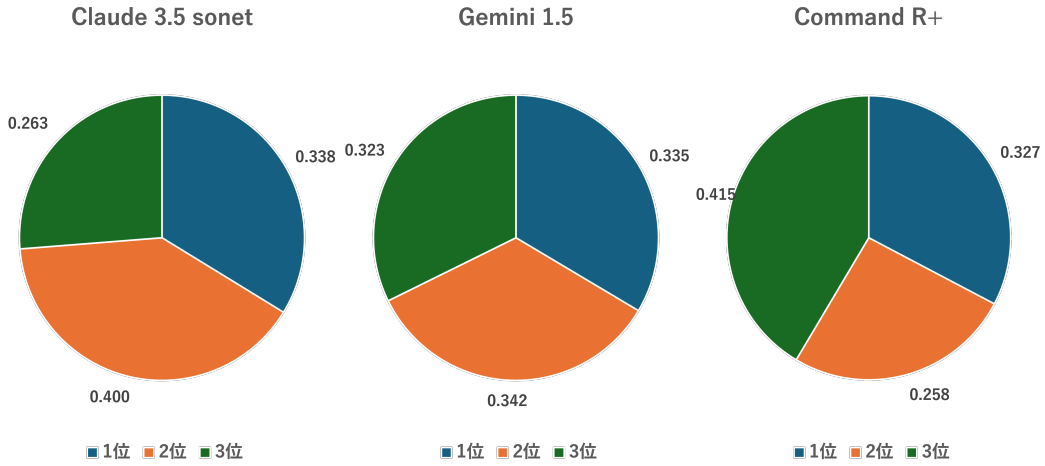


**図 1**　人手による評価で各 LLM が 1 位から 3 位と評価された割合

**表 3**　ChatGPT-4o による文の妥当性の評価のためのプロンプト例

| |
|---|
| 文の自然さについて 10 点満点中何点か<br>評価してください<br>また下の 3 つの点に注意してください<br>・文全体で意味が通っているか<br>・常識的におかしな点がないか<br>・単語の意味に誤りがないか<br><br>企業買収の際には、セキュリティと道徳の両面を考慮することが重要だ。 |

**表 4**　単語組み合わせ別の評価値

| 専門用語組み合わせ | 評価平均点 |
|---|---|
| 情報，哲学，経済 | 8.4 |
| 情報，情報，経済 | 8.5 |
| 情報，情報，情報 | 10 |

また評価を行うために同一の入力を与えた際の各 LLM が生成した 3 つの文を入力し，順位をつけさせた．プロンプトの例を表 5 に示す．なお，プロンプトに入力される生成文はランダムに並び替えている．実験結果を表 6 に示す．

実験結果は人手評価と比較してはっきりと性能差をつけており， ChatGPT-4o は Claude 3.5 sonet の出力が最もよく，ついで Gemini 1.5，最後に Command R+ という順で評価したことが分かる．

**表 5**　ChatGPT-4o による文評価のためのプロンプト例

| |
|---|
| 以下の 1 から 3 の文を評価して順番を付けてください<br><br>評価の基準は自然な文かつ意味の分かる文であるかどうかです<br><br>1. 非同期的に実行される MMT システムは、誤謬推理を検出して修正するメカニズムを組み込むことで、より強固で正確なものとなる。<br><br>2.MMT の批評家は、非同期的な経済効果を考慮しないことを誤謬推理の一例として指摘する。<br><br>3.MMT 理論は、非同期的な経済政策を通じて、外生的な経済ショックに対応することを提案している。 |

**表 6**　ChatGPT4-o による生成文の比較結果

| | 1 番 | 2 番 | 3 番 |
|---|---|---|---|
| claude 3.5 sonet | 16 | 12 | 2 |
| gemini 1.5 | 8 | 16 | 2 |
| command R+ | 2 | 6 | 22 |

# 5　考察

## 5.1　専門用語に限定する手法の効果の考察

本研究にて提案した入力単語を専門的な用語に制限することにより，文生成の難易度を上げる手法の効果について考察する．生成文の妥当性に関しては，主観による評価となってしまうことを注意が必要である．各 LLM による生成文は，一見するともっともらしい文が多い．しかし，専門用語を多用する関係で文全体を通してみると意味がかなりあいまいになっており，間違っている文ではないが，的外れな文や無理に単語を使用しているような文

**表 7**　各 LLM の生成文と人手評価，機械評価例

| LLM | 生成文 | 1 位　2 位　3 位（人手評価の割合） | 順位（ChatGPT-4o） |
|---|---|---|---|
| claude 3.5 sonet | 企業買収の際には、セキュリティと道徳の両面を考慮することが重要だ。 | 0.375　0.500　0.125 | 2 |
| gemini 1.5 | 企業買収においては、セキュリティと倫理に配慮することが重要である。 | 0.438　0.375　0.188 | 1 |
| command R+ | セキュリティ企業の買収には、道徳的な問題がつきまとう。 | 0.188　0.125　0.688 | 3 |
| claude 3.5 sonet | IR チームは、機密情報の処理に関するジレンマに直面することがある。 | 0.125　0.500　0.375 | 3 |
| gemini 1.5 | IR 担当者は、情報処理における倫理的ジレンマに直面することが多い。 | 0.563　0.375　0.063 | 2 |
| command R+ | IR 処理におけるジレンマは、ノイズ除去と特徴抽出のバランスにある。 | 0.313　0.125　0.563 | 1 |

が散見される．改善すべき点がある文を生成することができるという点においては，タスクとして現在の LLM に見合ったものにすることができたと考える．また，専門用語の組み合わせに関しても 4 から組み合わせによる難易度の上昇効果が確認できており，複数の専門用語の組み合わせは生成難易度の上昇に効果があると考えられる．ただ，本研究で入力とした単語には，専門用語であるため，従来の CommonGen タスクと比べ正解文の評価が人間にも難しく，本研究で提案する手法による改変を加えた CommonGen タスクでは，一般的な知識に対する文生成能力を測るタスクとしての効果は疑問点が残り，常識の範囲をどう設定するかという問題は顕在化している．

### 5.2　人手と機械による評価結果の考察

人間による評価結果から分かる特徴として，Claude 3.5 sonet の 3 位評価が少なく，Command R+ の 3 位評価が多いことが分かる．このことから，3 つの LLM の内，妥当な文の生成能力においては Claude 3.5 sonet が最も高く，Command R+ の能力が低いと考えられる．この傾向は，ChatGPT-4o による評価とも相関があり，機械においても文の自然さの評価ができているのではないかと考えられる．そのため今後は，LLM による生成文の自動評価という手法は実用的であり，自動評価手法の確立が難しい他の文生成タスクなどにおいても，有効なのではないかと考える．しかし同時に ChatGPT-4o による評価では，はっきりと差が出ているにもかかわらず人手による評価では，差が大きく見られなかった点に関しても考察を行う．人手による評価は、全体の割合では差が見られなかったものの質問ごとの比較では，ある程度割れている結果が出ており，感覚の個人差により差が出ていないというわけではないと考えられる．中には，ChatGPT-4o と人手評価で正反対の評価結果となる場合も確認されており，どのような要素が人手評価と機械による評価において優劣をつけるのかを今後分析する必要があると考える．

## 6　おわりに

本研究では，[8] で示された LLM を用いて日本語 CommonGen 形式のデータセットは文を生成する手法に入力単語を専門用語に限定する手法を提案し，提案手法をもとにテストデータを作成した．作成したテストデータを用いて，3 つの LLM に文を生成させ，それぞれの生成文を人手による評価と ChatGPT-4o による自動評価を行い，それぞれの評価を比較した．比較結果は，3 位となる評価が最も多い LLM と　3 位評価が最も少ない LLM が一致しており機械と人手とで一定の相関が見られることが分かった．しかし同時に LLM による評価でははっきりと差が見られたのに対し，人手評価には差が大きくは見られなかったことに対しては今後検証が必要である．また提案手法に関しては，主観による評価とはなるが，現在の LLM においても有効な難易度に変更できる手法ではないかと結論付ける．

## 謝辞

本研究は国立国語研究所の共同研究プロジェクト「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」及び JSPS 科研費 23K11212 の助成を受けています.

## 参考文献

[1] Talmor Alon, Herzig Jonathan, Lourie Nicholas, and Berant Jonathan. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4149–4158. Association for Computational Linguistics, 2019.

[2] Sap Maarten, Rashkin Hannah, Chen Derek, Le Bras Ronan, and Choi Yejin. Social iqa: Commonsense reasoning about social interactions. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4463–4473. Association for Computational Linguistics, 2019.

[3] Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[4] Omura Kazumasa, Kawahara Daisuke, and Kurohashi Sadao. A method for building a commonsense inference dataset based on basic events. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2450–2460. Association for Computational Linguistics, 2020.

[5] Zellers Rowan, Bisk Yonatan, Schwartz Roy, and Choi Yejin. Swag: A large-scale adversarial dataset for grounded commonsense inference. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 93–104. Association for Computational Linguistics, 2018.

[6] Zellers Rowan, Holtzman Ari, Bisk Yonatan, Farhadi Ali, and Choi Yejin. Hellaswag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4791–4800. Association for Computational Linguistics, 2019.

[7] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1823–1840. Association for Computational Linguistics, 2020.

[8] 鈴木 雅人, 新納 浩幸. 日本語 CommonGen に対する ChatGPT の性能調査. 自然言語処理研究会 (第 256 回), 2023.