「数」に着目したLLMの多言語能力の検証

羽根田 賢和 [1,2]　岸波 洋介 [1]　藤井 諒 [1]　森下 睦 [1]
[1] フューチャー株式会社　[2] 東北大学
{k.haneda.k3, y.kishinami.rh, r.fujii.6d, m.morishita.pi}@future.co.jp

概要

近年の自然言語処理技術の発展に伴い，大規模言語モデル (LLM) は様々な言語を扱うことが可能となっている．一方でこれらのモデルを用いた推論・生成では，英語で指示を与える際に，他の言語で指示を与える場合よりも性能が高くなる現象が確認されている．本研究では，誤差の大小の解釈が容易な「数」に着目したタスク設定を行い，指示言語によって生じるLLMの性能差のより精緻な検証を試みた．結果として指示言語の特徴により生成結果の誤差の生じ方に差が生じ，言語的特徴が性能に影響を及ぼしていることの示唆を得た．

1　はじめに

近年，様々な言語を扱うことのできる大規模言語モデル (LLM) が数多く登場している．近年の性能向上に伴い，これらのマルチリンガルなLLMは翻訳タスクに限らず，要約や雑談など，多様なタスクでの活用が可能となっている．一方でマルチリンガルなLLMを用いた推論・生成においては，英語で指示を与える際，他の言語で指示を与える場合と比較して性能が高くなる現象が報告されている [1, 2, 3]．

しかし，これらの先行研究において比較対象となっている翻訳などのタスクでは正解がただ一つに限定されない．そのため定量的な分析を行う際には，正解となるゴールドデータを適切に選定する必要があり，恣意性を完全に排除することが難しい．

また感情分析などのタスクでは正解率を0か1で測定することは容易だが，正答に対する誤りの深刻さの程度を定量的に表現することは容易ではない．そのため，例えば指示言語における英語と日本語の性能差と，英語とドイツ語の性能差を意味のある数値で表現し，分析することは困難である．

そこで本研究では性能や誤差のより精緻な定量評価のために「数」に着目したタスク設定を行った．具体的には，LLMに単語や文の文字数を数えさせるカウントタスクと，LLMに文字数を指定して文を生成させる生成長制御タスクの2つを設定し，検証を行った．GPT-4 [4] をはじめとするLLMは，文字数を正確に数えることができないということがShinら [5] やDaveら [6] によって報告されている．また，Jieら [7] の研究からも見られるように，厳密な生成長制御能力に関しても不完全であり，エラー分析の対象として適しているといえる．これら2つのタスクでは，文字数という絶対的な基準により精度を定量的に評価することが可能である．また誤りの程度を正答の文字数との差で表現することにより，性能差を意味のある数値で表現することができる．

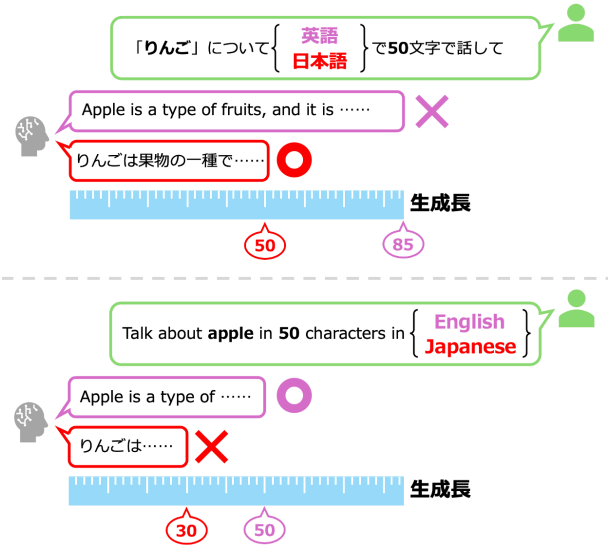

図1　生成長制御タスクの概要図．カウントタスクにおいても指示とカウント対象言語をそれぞれ変更する．

本研究では，文字列のカウントタスクと生成長の指示を伴う生成長制御タスクにおいて，指示言語による性能の差を調査した．この際，指示言語とカウント・生成対象の言語が同じ設定のみを考慮すると，厳密な対照実験とはならず，指示言語の影響を他の要因と切り離して考えることが困難である．そのため本研究では図1に示したように，指示言語とカウント・生成対象の言語をそれぞれ独立に変更し

て検証を行った．具体的には 5 つの言語を指示言語として使用し，指示言語とカウント・生成対象の言語をそれぞれ変更することで，詳細なエラー分析を試みた．結果として，生成長制御タスクにおいては，指示した長さに対応するトークン数が指示言語の種類によって決定され，生成長に影響を及ぼしていることの示唆が得られた．

2 関連研究

英語プロンプトの優位性 Ahuja ら [1] は LLM の多言語能力を調査し，英語などの高リソース言語と，低リソース言語との間に存在する性能差を指摘している．また，Bang ら [2] は，ChatGPT を用いた感情分析と言語識別タスクにおいて，英語での性能がジャワ語などの低リソース言語での性能よりも高いことを報告している．Hendy ら [3] は GPT モデル [8] を用いて翻訳タスクにおける高リソース言語の優位性を示すとともに，算術タスクにおける英語プロンプトの優位性も報告している．

LLM によるカウント能力 Shin ら [5] は LLM の文字に対する理解能力が低いことを示し，その中で文字のカウント能力が低いことに言及している．GPT-4 [4] をはじめとした各種 LLM におけるカウントタスクの正解率は 50%程度にとどまり，トークン単位での学習，推論を行う現行のモデルアーキテクチャの限界を指摘している．この論文では，英語の他に中国語や日本語などで同様の検証をしているが，指示言語とカウント対象言語は同一であり，指示言語による性能差に着目している本研究とはスコープが異なる．また Dave ら [6] は，計算理論の視点をもとに，LLM が一般の系列に対して足し算や掛け算，記号の数え上げといったタスクが可能であるかを検証し，いずれにおいてもタスクの難易度の上昇とともに精度の低下を確認している．しかしながら，モデルへの指示は英語でのみ行われており，この点において本研究とは趣旨が異なる．

厳密な生成長制御 LLM の生成長の厳密な制御は，現在の自然言語処理分野における研究課題の一つである．Jie ら [7] は，ユーザーから与えられる多様な生成長制御プロンプトを標準化する機構と強化学習を組み合わせた生成長制御手法を提案している．結果として最大文字数を指定する指示においては一定の効果を示した一方，指定長ちょうどでの出力を得る指示に対しては未だ課題が残ることが報告されている．

3 実験

文字列の文字数を LLM に数えさせるカウントタスクと，指定文字数の文章を LLM に生成させる生成長制御タスクの二種類の実験を行った．実験では，英語 (en)，スペイン語 (es)，ポルトガル語 (pt)，簡体中国語 (zh)，日本語 (ja) を使用し，モデルは GPT-4 と Claude 3.5 Sonnet v2 [9] を用いた[1)]．

3.1 文字列カウントタスク

カウントを LLM に指示するプロンプトと，カウント対象となる文字列をそれぞれの言語で用意した．カウント対象の文字列は各言語の「単語」と「文」であり，それぞれ 100 種類を使用した．

意味などの文字列そのものの難易度差によって，各言語間のタスク難易度に差が生じないよう，単語は英語の Age-of-acquisition [10] から無作為に抽出し，英語ではそれらを翻訳したものを対象とした．この際，特定の言語に特有な概念を指す単語が選択されることを避けるために，平均獲得年齢が 5〜10 歳の単語のみを対象とした．またスペイン語やポルトガル語に翻訳した際にスペースが入り複数の単語となってしまう場合や，明らかに有害表現となる単語が選ばれた場合には再抽選を行った．文に対しても，各言語間での意味による差が生じないよう対訳コーパスである FLORES-200 [11] を用いて対象文字列を選出した．実験対象の 5 つの言語に対し，テストセットから対訳文を無作為に 100 件抽出し，カウントの対象文字列とした．

言語 A での指示に対してカウント対象となるのは，言語 A を含む 5 つの言語の単語と文である．例えば単語に対する日本語のプロンプトは「次の単語の文字数を数えて，数字だけを答えてください．単語:[xxxx]」であり，カウント対象の xxxx には，「スパイス」などの日本語だけではなく「spice」や「香料」といった各言語の単語が代入される．

指示言語とカウント対象言語を 5 言語ずつ組み合わせ，計 25 パターンのカウントタスクを，単語と文のそれぞれを対象に行った．

3.2 生成長制御タスク

生成のテーマとなるテーマ単語を用意し，それぞれの単語に対して 50 文字の説明文を LLM に生成さ

1) Llama 3.1 70B を用いた検証も行ったが複雑なタスクであったためか分析に値する生成結果が得られなかった．

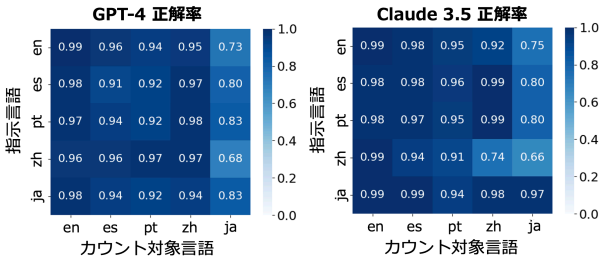

図 2　各モデルの単語カウントタスク正解率

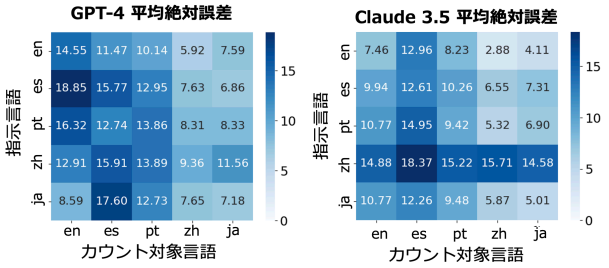

図 3　各モデルの文カウントタスクの平均絶対誤差

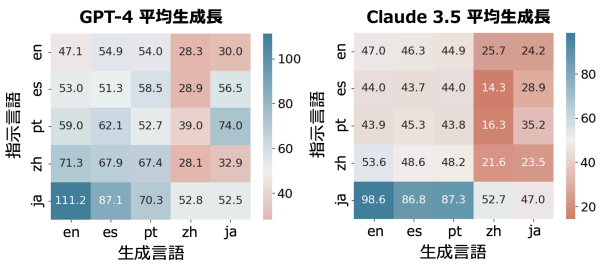

図 4　各モデルの平均生成長

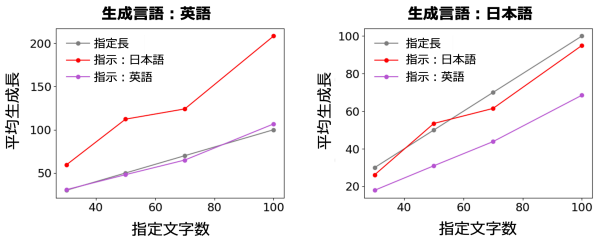

図 5　指定文字数を変更した際の平均生成長 (GPT-4). Claude に対しても同様の結果が得られた.

せた．語彙の難易度が生成難易度に影響を与えないよう，テーマ単語はカウントタスクで選出した単語を用いた．プロンプトで用いる言語 A に対し，生成対象言語となるのは，言語 A を含む 5 つの言語全てである．例えば指示言語が日本語，生成対象言語が英語の場合，英語で 50 文字の文を生成するように，日本語で指示を与えることになる．具体的には「「スパイス」について英語で 50 文字で説明してください」のようなプロンプトを与えることになる．

指示言語と生成言語を 5 言語ずつ組み合わせ，計 25 パターンで検証を行った．また指定した文字数による影響を調査するため，日本語と英語においては生成長を 30，70，100 に変更し実験を行った．

4　結果

4.1　文字列カウントタスク

図 2 に各モデルの単語カウントタスクにおける正解率を示す．単語のカウントは，GPT-4，Claude 3.5 Sonnet v2 の双方のモデルにおいて，どの言語の組み合わせでも比較的高い精度を持つことが確認された．一方で日本語をカウント対象とした際，精度が若干低下する場合が見られた．また Claude では，中国語で指示した際，中国語を含むアジア言語に対する精度が大きく低下した．

図 3 に各モデルの文カウントタスクにおける平均絶対誤差を示す．単語のカウントでの高い精度と比べ，文のカウントでは全体として精度が低下していた[2)]．また Claude では，中国語で指示した際，いずれのカウント対象言語においても誤差が大きくなる現象が確認された．

4.2　生成長制御タスク

図 4 に生成長制御タスクの実験結果を示す．GPT-4，Claude 3.5 Sonnet v2 ともに，ヨーロッパ言語はヨーロッパ言語の指示で生成させる方が，アジア言語はアジア言語の指示で生成させる方が精度が高い傾向があった．またどちらのモデルでも，中国語と日本語を生成させた場合，生成結果の平均長は指定した 50 文字よりも短くなる傾向があった．さらに特筆すべき結果として，日本語で指示を与えた際，英語，スペイン語，ポルトガル語の生成結果が著しく長くなる現象が確認された．また，図 5 に日本語と英語において指定文字数を変更した際の結果を示す．指定文字数を変更した場合でも一貫して 50 文字の際と同様の傾向を持つ結果となった．

5　考察

5.1　文字列カウントタスク

単語のカウントは全体的に高精度であったのに対し，文のカウントでは著しく精度が低下していた．これは，カウント対象の系列長が長くなったことが一つの要因であると考えられる[3)]．LLM は対象の系

2)　各モデルの正解率は Appendix A に記した.
3)　カウント対象とした単語，文の平均文字数は Appendix D に記した.

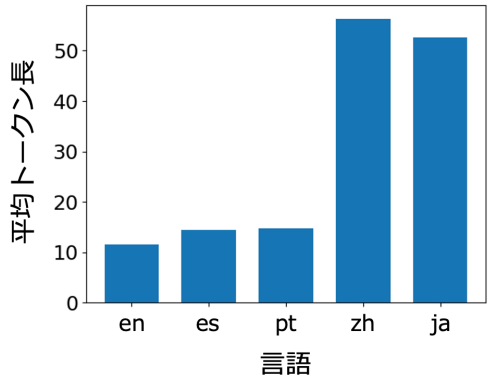

図 6 各言語の 50 文字あたりの平均トークン長．FLORES-200 の開発セットを 50 文字ずつ GPT-4 でトークナイズした．

列長が長いほどカウントに失敗しやすい傾向を持ち[4]，このことが単語と文における精度の差を生んだ最も大きな要因であると考えられる．

5.2 生成長制御タスク

日本語で指示を与えた際に，英語，スペイン語，ポルトガル語で生成結果が著しく長くなる現象は，複数の要因によるものであると推測される．

一つの大きな要因として考えられるのは，漢字によって生じる文字数とトークン数との関係の違いである．英語・スペイン語・ポルトガル語では，ほとんどの場合で複数文字に対して 1 トークンが割り当てられ，極端な場合であっても，基本的に 1 文字は 1 トークン以上で表されることはない．しかし漢字は 1 文字が 2〜3 トークンで表現されることも多く[5]，漢字を用いる言語とそうでない言語では文字数とトークン数の関係は大きく異なる．実際に，図 6 に示した通り，日本語における 50 文字の文のトークン数は英語よりも平均的に長くなっている．そのため，日本語で 50 文字を指定した際にモデル内部で想定されるトークン数は英語で 50 文字分にあたるトークン数よりも多くなってしまう．その結果として英語などでの生成結果が 50 文字よりも長くなると推測される．

一方，ヨーロッパ言語での指示がアジア言語を短く生成させるという結果に関しても，同様にして考えることが可能である．英語における 50 文字の文の平均トークン数は日本語や中国語におけるそれよりも短い．そのためモデル内部で想定されるトークン数をもとに生成が行われると，日本語や中国語では 50 文字に満たない生成になると考えられる．

したがって生成長制御においては，指示言語によって生成すべき長さがトークン基準で決定され，それが生成言語を問わず適応されることで，文字数に影響を及ぼしていると考えることが可能である．

また，この考察を支える結果として，生成長制御タスクではヨーロッパ言語どうし，アジア言語どうしでの精度が高い傾向があったことが挙げられる．これらの似通った言語間では 50 文字あたりのトークン数も近しい値をとっていることが図 6 から確認できる．そのため，ヨーロッパ言語どうし，アジア言語どうしでは比較的高い精度を保ったままでの生成が可能であったと考えられる．

しかしながら，以上の点だけではこれらの現象を説明するには不十分である．例えば Claude では中国語を生成した際の文字数の短さが顕著であり，中国語自身で指示を与えた際にも同様であった．文のカウントにおいても，Claude のみ中国語を指示言語にした際に著しい性能の低下が確認されている．これらの結果から Claude は中国語を苦手としていると推察され，このことが生成長制御におけるズレを引き起こすことは十分に考えられる．他にも英語などの言語は単語がスペースで分割されており，文字単位ではなく単語単位での文長調整が一般的である可能性が高い．今回の結果を説明しきるためには，より精緻な分析が必要であると考えられる．

6 おわりに

本研究では，文字列カウントタスクと生成長制御タスクにおいて，指示言語による性能の違いを調査した．結果として生成長制御タスクにおいては指示した長さに対応するトークン数が指示言語の種類によって決定され，生成長に影響を及ぼしていることの示唆が得られた．

一方で本研究は調査対象が API 経由で動作するモデルに限定されており，モデル内部の挙動に対する分析が困難であった．今後はよりオープンなモデルでの調査を進めていきたい．

実験を通して指示言語によって誤差の生じ方が異なることが確認されたが，この要因については，更なる検証を要する．今後の研究では，本論文で言及した文字とトークン数についての精緻な分析をはじめ，学習リソース量の差や各々の言語的特徴など，影響を及ぼしていると推測される他の要素についても検証を進めていきたい．

4） 詳細は Appendix B に記した．
5） 詳細は Appendix C に記した．

参考文献

[1] Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 4232–4267, 2023.

[2] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 675–718, 2023.

[3] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are GPT models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210, 2023.

[4] OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[5] Andrew Shin and Kunitake Kaneko. Large language models lack understanding of character composition of words. arXiv preprint arXiv:2405.11357, 2024.

[6] Neisarg Dave, Daniel Kifer, C. Lee Giles, and Ankur Mali. Investigating symbolic capabilities of large language models. In First International Workshop on Logical Foundations of Neuro-Symbolic AI (LNSAI 2024), 2024.

[7] Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. Prompt-based length controlled generation with multiple control types. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 1067–1085, 2024.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp. 1877–1901, 2020.

[9] Anthropic. Claude 3.5 sonnet, 2024.

[10] Victor Kuperman, Hans Stadthagen-González, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, Vol. 44, pp. 978–990, 2012.

[11] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. Scaling neural machine translation to 200 languages. Nature, Vol. 630, No. 8018, pp. 841–846, 2024.

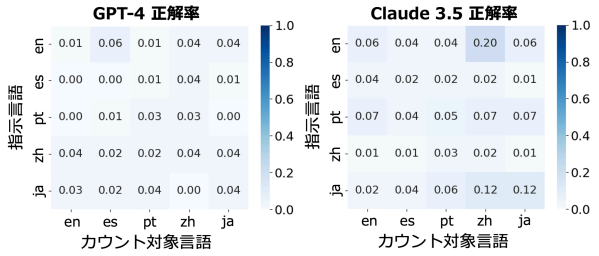

図 7　各モデルの文カウントタスク正解率

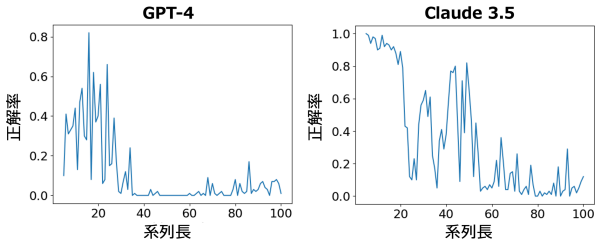

図 8　ランダムな数字列に対する各モデルのカウント精度

A　文カウントタスクの正解率

図 7 に，文カウントタスクの各モデルの正解率を示す．

B　系列長とカウント精度

図 8 に，カウント対象の長さを変更した際の正解率を示す．ここでは言語の影響を極力排すために，カウント対象はランダムな数字列としている．なお指示言語はどちらのモデルに対しても英語を用いた．

C　1 文字に対するトークン数

英語，スペイン語，ポルトガル語で用いられるアルファベット計 80 文字を，それぞれ 1 文字ずつ GPT-4 のトークナイザーでトークン化した．結果としてほとんどの文字が 1 文字 1 トークンで表され，1 文字が 2 トークンとなったのは，ポルトガル語で用いられる Ê・Ô・Õ の 3 文字のみであった．

中国語と日本語に関しては FLORES-200 のそれぞれの言語のテキストに存在する漢字のうち，CJK 統合漢字の基本漢字 20992 字に含まれる漢字に対して，同様に 1 文字ずつトークン化を行った．

結果として中国語は，2044 種の文字中，1 トークンのものが 519 字，2 トークンのものが 1292 字，3 トークンのものが 233 字であった．

日本語では，1441 種の文字中，1 トークンのものが 342 字，2 トークンのものが 885 字，3 トークンのものが 214 字であった．

表 1　カウント対象の平均文字数

	単語	文
英語	7.03	122.21
スペイン語	7.52	146.49
ポルトガル語	7.85	131.15
簡体中国語	2.14	39.27
日本語	3.06	51.72

D　カウント対象の平均文字数

表 1 に，各言語におけるカウント対象の単語，文の平均文字数を示す．