大規模言語モデルによる失語症の単語産出シミュレーション

森田早織 [1]　原田宥都 [1]　直江大河 [2]
沖村宰 [2]　大関洋平 [1]
[1] 東京大学大学院総合文研究科 [2] 昭和大学発達障害医療研究所
{msaori6012, harada-yuto, oseki}@g.ecc.u-tokyo.ac.jp
{naoe_taiga, tokimura.psy}@med.showa-u.ac.jp

概要

本研究では、日本語生成モデル llm-jp-3-13B-instruct に構造化 Pruning を適用し、失語症の言語症状の再現可能性を検証した。結果、高頻度語が低頻度語よりも復唱タスクにおいて高い正答率を示し、失語症者に見られる高頻度語保持の傾向と一致する可能性が示唆された。また、層ごとの Pruning が出力に与える影響を目標語と出力語のコサイン類似度にて評価した結果、一部のモデルで目標語との意味的類似性が観察され、意味性錯語に近い現象が確認された。これにより、LLM の Pruning が失語症者の発話特徴の再現や回復過程のシミュレーションに応用できる可能性が示された。本研究は、失語症の病態の機序解明やリハビリ手法の開発など、臨床応用できる可能性を示した点で意義深いと考えられる。

1　はじめに

1.1　失語症とは

脳卒中などの後天的な疾患や事故によって発症するとされる失語症は、「大脳の損傷に由来する、一旦獲得された言語記号の操作能力の低下ないし消失」と定義づけられている [24]。脳のどの部位を損傷するかによって現れる症状は異なり、それに基づいた失語症タイプの分類も行われている。代表的なものとしては、脳の前方領域を損傷した際に発症するケースが多く、理解はある程度保たれるものの流暢な発話が困難になるブローカ失語 [2] や、脳のやや後方領域を損傷した際に発症するケースが多く、発話の流暢性は保たれるものの、言語理解が困難になることから表現選択の誤りに気づけないウェルニッケ失語 [17] などがある。また、単語の発話や理解といった側面では、錯語と呼ばれる症状がある。錯語には複数種類があるが、代表的なものとして、音の一部が誤って表現される音韻性錯語と、発話された単語が目標語と同じカテゴリに属しているが、意図された単語とは全く異なる状態を指す意味性錯語がある [23]。のちに Geschwind や Lichtheim [6, 9] によって、人間が言葉を理解したり産出したりする上で行われる情報の流れを図式化した”Wernicke-Lichtheim 図式”がうまれ、失語症には 7 つの分類タイプが生じうると提唱されていった。

しかし現代においては、1800 年代後半から 1900 年代後半にかけて確立されてきた古典的な失語分類では位置づけが困難な症状も見られ [22]、言語リハビリによる効果や回復経過にも個人差がある [23]。

患者によって多彩な症状が見られる一方で、ある程度の傾向が見られるという報告もある。特に失語症者の単語処理について、Silver と Halpern は、高頻度語は、失語症患者にとって低頻度語よりも語彙検索が容易であることが確認されており、特に日常生活でよく使われる語が保持されやすいと述べている [15]。また、特に名詞は動詞と比べても頻度の影響を受ける傾向にあるという報告もある [1]。

失語症者の言語機能はどのように回復していくかについては、発症後の経過時期によって言語処理に関わる脳部位やメカニズムが異なることが推察されている。また当事者を取り巻く環境による影響も大きい [23]。失語症者における言語機能回復の機序やそれに基づいた効果的なリハビリ方法について、今なお議論を深めることが求められている [21]。

近年、臨床応用分野の研究においては、失語症の診断やリハビリへの応用研究に対する自然言語処理分野の介入は進展しつつある。例として、事前学習済みの大規模言語モデル (LLM) を用いて、失語症状の有無・重症度・サブタイプを検出し、特にファインチューニングなしでの性能を評価し、臨床的自動診断の可能性と課題を議論する研究 [3] や、失語症者が発話した断片的な文章を LLM によって補完し、

失語症者のコミュニケーションのサポートを図る研究 [12, 16] などがある。

1.2 症状シミュレーション研究の背景

1.1 でも述べたように、臨床診断やリハビリのサポートに対する LLM の介入は近年進みつつあるものの、失語症の回復機序へのさらなる理解を深める手法として LLM を応用するケースは少ない。一方で、精神医学の観点では、特定の疾患から現れる言語に関わる症状の発現メカニズムを迫るために、ニューラルネットワークや LLM を応用した研究が報告されている。Hoffman ら [7] は、統合失調症における幻聴の要因を明らかにするため、音声認識に特化したニューラルネットワークの特定のシナプスを過剰に刈り込むというアプローチで探索した。その結果、特定のニューロンが一層活性化することで誤出力が発生し、統合失調症による幻聴の発症要因に対する仮説を裏づける一因となった。また、Fradkin ら [4] は GPT-2 にパラメータ調整を施すことで統合失調症者の話題逸脱的な発話のシミュレーションを行っている。さらに、直江ら [20] は日本語の LLM に Pruning とパラメータ調整を行った障害モデルを作成したところ、中間層を損傷したモデルに、発話の逸脱や脱線に加えて、失語症特有の錯語[1)]が観察されたことを示している。これにより、LLM への Pruning という手法が失語症状を再現

1.3 本研究の目的・内容

本研究では、精神医学で実際に応用されてきた計算論的なアプローチ [5, 14] を失語症の回復機序研究に取り入れ、失語症による多彩な言語特徴を機械上で再現し、さらに効果的なリハビリ方法を探るべく、回復までのシミュレーションの実現を目指すという背景がある。その一段階的な研究として、直江らの研究 [20] を踏まえ、日本語の単語や文を生成する LLM を Pruning によって障害したモデルを作成し、症状の再現が可能か否か探索した。失語症の症状にも種類があるが、今回は 1 単語を復唱させるタスクによる意味のエラーの有無と傾向について着目し、LLM の事前学習に使われたコーパスの中で特に高頻度な単語と低頻度な単語を復唱させ、得られた出力を観察した。また、障害したモデルによるエラーが、失語症で見られる意味性錯語のように、類似した意味を持つものか調べるために、プロンプトに含まれる目標語と出力した単語の意味距離を計測した。1.1 でも言及したような、単語の使用頻度と症状の現れ方との関係性が、LLM を障害した場合でも見られるか否か評価する。

[1)] 1.1 を参照

2 実験

2.1 実験設定

本研究では、5 種類の日本語コーパスを含む計 14 種のコーパスで事前学習された llm-jp-3-13B-instruct に Pruning を行った。Pruning には LLM-Pruner を用い [10]、40 層からなるモデルを 5 層ずつ区分して重みを削除した。ただし 40 層目は LLM-Pruner の仕様上、Pruning の対象外であったため、実際は 1 層目から 39 層目を Pruning している (1-5 層 6-10 層 11-15 層 16-20 層 21-25 層 26-30 層 31-35 層 35-39 層)。今回はニューロン単位での構造化 Pruning を行っており、どの重みを削除するか決定する指標には L1 ノルムを用いた。また、LLM のネットワーク全体で 7.36% の重みを削除して実験を行った。

2.2 プロンプトの作成

今回は、Bastiaanse による動詞よりも名詞の方がより頻度の影響を受けやすいという主張に基づき [1]、障害したモデルによるタスクの結果と単語の頻度の関係性を調べるためのプロンプトを作成した。llm-jp-3-13B-instruct が事前学習に使用している日本語コーパスの一つである wikipedia-ja-20230720 [8] から、高頻度語上位 100 個と、低頻度語下位 100 個を取り出し、合計 200 個の単語をそのまま繰り返して出力させた。コーパスから高頻度語と低頻度語を抽出するにあたって、Janome [13] を使った形態素解析を行い、名詞のみに絞った。また、低頻度語に関しては、その後の評価の際に意味の類似性を正確に計測するために、1-9999 回の頻度を示した単語は低頻度語プロンプトの採用対象から除外し、10000 回以上の頻度のものから 100 語を抽出した。

3 実験結果と評価

3.1 モデルごとのタスク正答率

Pruning によって作成した 9 つのモデルとオリジナルモデルに対して、出力する語の数を指定するパラメータ (max new token) を 1 に設定し、200 語の単語を復唱するようにプロンプトを与えた。その結

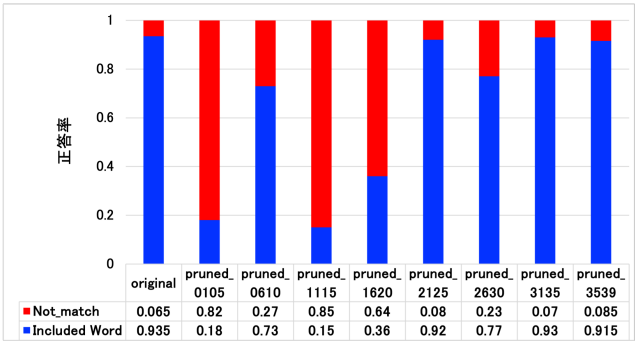

図 1　復唱タスクの成功率

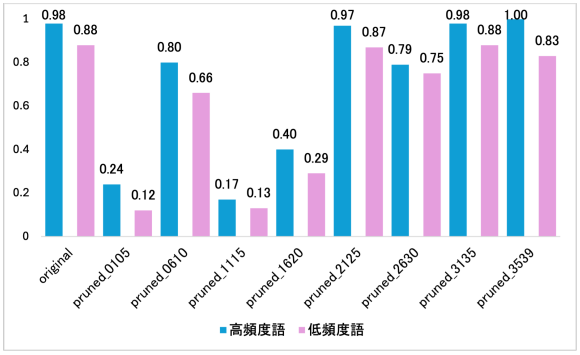

図 2　高頻度語と低頻度語の復唱タスクの正答率

果、図 1 のような正答率の傾向が見られた。LLM の最も浅い層に対する Pruning はモデルの性能を著しく損傷することがいわれている [10] が、本実験でも、1 から 5 層と 11 層から 15 層と浅い層での Pruning 障害モデルでの正答率が顕著に障害された。また、中間層である 16 層から 20 層の Pruning 障害モデルも、その他の障害モデルと比べて正答率が低い傾向が見られた。一方で、深い層での Pruning 障害モデルでの正答率はオリジナルモデルとの正答率は変わらない傾向があった。つまり、全体の層の中でも浅い層から中間層にかけての損傷にて、1 単語のみの復唱でも正確に出力できなくなることが示唆された。

3.2　高/低頻度語の正答率の差

高頻度な単語および低頻度な単語それぞれの復唱タスクにおける正答率についても分析した。図 2 で示している数値は、高頻度語 100 語、低頻度語 100 語に対する正答率を示している。つまり、例えば 35 層から 39 層を Pruning したモデルの高頻度語復唱タスクにおける正答率は 1.00 と示されているため、100 語全て正確に復唱することができていたことを示している。max new token を 1 に設定した上で、高頻度語と低頻度語を復唱させると、全てのモデルで高頻度語の方が低頻度語より高い正答率を示していることが読み取れる。

3.3　意味距離による分析

200 語の単語を復唱したモデルによって得られた結果と、プロンプトに含まれる目標語が意味の類似性が見られるかどうか、どれほど一致しているかを調べるために、両者間の単語の意味距離を計測することで、意味性錯語らしさの有無を評価した。目標語および得られた出力のベクトル変換およびコサイン類似度の測定には、日本語版 Wikipedia の本文全てから学習している、日本語 Wikipedia エンティティベクトル [19] を用いた。単語ベクトルの学習には Word2vec [11] が使用されている。

3.4　類似度範囲と意味性錯語

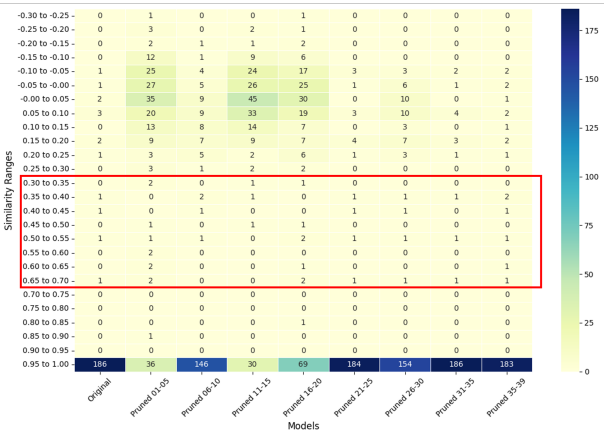

図 3　モデルごとの類似度範囲分布

目標語とモデルによる出力で得られた単語の類似度を計測し、各類似度範囲における組み合わせ数の分布を図 3 に示した。最も正答率が低かった 11 層から 15 層を Pruning したモデルは、語彙よりも括弧やシャープなどの記号、あるいは「はい」や「次」などといった、モデルが前置きに発する文章の一部をを出力しているケースが多かった。そのような出力は-0.1 から 0.1 の範囲に集中した。また、1 層目から 5 層目を Pruning したモデルでも同じような傾向が見られた。

全てのモデルに共通して、記号や前置きの一部が出力するケースと、正確に出力するケースのどちらかに大きく偏っていたが、同時にどのモデルでも 0.3 から 0.7 あたりの類似度を示す組み合わせも観察された。特に 0.3 から 0.7 の範囲に含まれる目標語と出力の組み合わせによっては、意味的に関連する誤りを含んでいる可能性があると考えた。

3.5 錯語との類似性

図 3 上の赤色枠で示した、類似性が 0.3 から 0.7 の範囲に含まれている目標語と出力の単語の組み合わせを見ると、目標語の 2 文字目あるいは 3 文字目が欠落した形で出力されている例が特に多く見られた。しかし、最も正答率が低かった 11 層から 15 層を Pruning したモデルに匹敵する正答率を示した、1 層から 5 層を Pruning したモデルと、3 番目に正答率が低かった 16 層から 20 層を Pruning したモデルで、目標語と類似性が高い出力や、1 文字異なる漢字で類義語を出力する例、すなわち、意味性錯語が見られた (図 4)。

Models	入力	出力	類似度
pruned 0105	政治	経済	0.64604837
pruned 0105	戦争	平和	0.4884842
pruned 0105	降板	脱退	0.55192029
pruned 0105	満州	チベット	0.52424937
pruned 0105	信者	神	0.31631511
pruned 0105	奪取	獲得	0.65162629
pruned 0105	あまり	もちろん	0.3127194
pruned 0105	除去	排除	0.60733002
pruned 1115	以上	以下	0.47289106
pruned 1620	監督	映画	0.4916172
pruned 1620	会社	株式会社	0.50617486
pruned 1620	自身	自分	0.67557859
pruned 1620	技師	技術	0.32089338
pruned 1620	得票	選挙	0.60004735
pruned 2630	警告	了解	0.35495299

表 1 類似度が 0.3-0.7 を示した単語

4 考察

本研究では、日本語の文や単語を生成する LLM である llm-jp-3-13B-instruct に構造化 Pruning を行い、以下の 2 点に着目して検証した。

- 失語症者は単語の使用頻度が症状や回復過程に影響するが、Pruning を行ったモデルでも類似した傾向が見られるか
- Pruning を行ったモデルによる出力に、失語症の症状の一つである意味性錯語が見られるか

結果として 9 種類の Pruning モデル全てで、1 単語の復唱をする場合には、事前学習コーパスの中で高頻度に使用されている単語の方がより正確に復唱可能であることが明らかとなった。これは、高頻度な名詞がより保持されやすい傾向にある失語症者の特徴と類似性が高いと考えられる。また、正答率が特に低かった 1 層から 5 層、11 層から 15 層、16 層から 20 層の Pruning モデルは、わずかではあるが目標語との類似性が見られる出力も観察された。

失語症者の中で、意味性錯語は脳の後方領域に損傷が見られる患者に顕著に現れるという報告があり、これは喚語すべき単語のカテゴリーを選択することはある程度できているものの、厳密に単語を想起させることに障害をきたしているからであるという解釈がなされている [18]。今後 LLM を損傷させるというアプローチで、脳の後方領域への損傷による失語症者の発話特徴をより細かく再現できる可能性がある。

また、LLM-Pruner は、Pruning によるモデルの性能低下を補うため、追加のトレーニングデータを用いてモデルを再調整する Recovery Stage が構成の中に含まれている [10]。しかし本研究では、この Recovery Stage まで行っておらず、重みを削除したモデルを使用した復唱タスクを行う実験にとどまっている。今後、LLM-Pruner を用いた Pruning による症状の再現を行った上で、Recovery Stage でのモデルに対する再調整を行うというアプローチから、これまで予測が困難とされてきている失語症者の回復過程を再現することにも結びつき、より効果的な言語リハビリテーションの探索を実現することも期待される。

ただ LLM の場合は、特に正答率が低かったモデルは語彙の出力に至ることができておらず、また全体的に前置きの文章を出力するケースも多かった。また、本研究では現時点で、実際の生理学的基盤に基づいたデータおよび失語症者の発話データを用いた分析にまで至れていない。また、失語症の多彩な症状を幅広く再現していくために、Pruning 以外の手法でも実現可能か引き続き検討していく必要がある。今後、より正確に失語症者の発話特徴を LLM 上で再現した上で、治療やリハビリテーションの LLM 研究ならでは可能性として、障害モデルでの他のパラメータ調整、層別のファインチューニング、障害モデルのままで大規模かつ多様なプロンプト入力による出力語の改善などを検討している。さらに、Pruning するネットワークの割合を調整したモデルを新たに作成した上で再実験を行うことや、より臨床医療との繋がりを深められるような評価をしていく必要があると考える。

謝辞

本研究は JST さきがけ JPMJPR21C2、JSPS 科研費 24H00087/23H05493 および 22K18480 の助成を受けたものです.

参考文献

[1] M. Wieling R. Bastiaanse and N. Wolthuis. The role of frequency in the retrieval of nouns and verbs in aphasia. **Aphasiology**, 30(11):1221–1239, 2016.

[2] P. P. Broca. Perte de la parole, ramollissement chronique et destructionpartielle du lobe antérieur gauche du cerveau. **Bulletin de la Société Anthropologique**, 2:235–238, 1861.

[3] Y. Cong, J. Lee, and A. LaCroix. Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Danielle Bitterman, editors, **Proceedings of the 6th Clinical Natural Language Processing Workshop**, pages 238–245, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] I. Fradkin, M. M. Nour, and R. J. Dolan. Theory-driven analysis of natural language processing measures of thought disorder using generative language modeling. **Biological Psychiatry: Cognitive Neuroscience and Neuroimaging**, 8(10):1013–1023, 2023.

[5] K. J. Friston, K. E. Stephan, R. Montague, and R. J. Dolan. Computational psychiatry: the brain as a phantastic organ. **The Lancet Psychiatry**, 1(2):148–158, 2014.

[6] N. Geschwind. Disconnexion syndromes in animals and man. i. **Brain: a journal of neurology**, 88(2):237–294, 1965.

[7] R. E. Hoffman and T. H. McGlashan. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated 'voices' in schizophrenia. **The American Journal of Psychiatry**, 154(12):1683–1689, 1997.

[8] Izumi Lab. Wikipedia japanese dataset 2023-07-20.

[9] L. Lichtheim. On aphasia. **Brain**, 7(4):433–484, January 1885.

[10] X. Ma, G. Fang, and X. Wang. LLM-Pruner: On the structural pruning of large language models. **arXiv preprint arXiv:2305.11627**, 2023. Accepted at NeurIPS 2023.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[12] R. Misra, S. S Mishra, and T. K. Gandhi. Assistive completion of agrammatic aphasic sentences: A transfer learning approach using neurolinguistics-based synthetic dataset, 2022.

[13] Mocobeta. Janome - 日本語形態素解析ツール.

[14] A. D. Redish and J. A. Gordon, editors. **Computational Psychiatry: New Perspectives on Mental Illness**. The MIT Press, Cambridge, MA, 2016.

[15] L. S. Silver and H. Halpern. Word-finding abilities of three types of aphasic subjects. **Journal of Psycholinguistic Research**, 21(5):317–348, 1992.

[16] S. van Vaals, Y. Matusevych, and F. Tsiwah. Generating completions for fragmented broca's aphasic sentences using large language models, 2024.

[17] C. Wernicke. **Der aphasische Symptomen-complex. Eine psychologische Studie auf anatomischer Basis**. M. Cohn und Weigert, Breslau, 1874.

[18] 大槻 美佳. 教育講演 3　失語症の診療－最近の進歩－. **臨床神経学**, 48(11):853–856, 2008.

[19] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. In **言語処理学会第 22 回年次大会 (NLP2016)**, March 2016.

[20] 直江大河, 原田宥都, 前田ありさ, 森田早織, 中村啓信, 大関洋平, 沖村宰計算論的精神医学としての大規模言語モデルへの pruning による精神疾患の思考障害シミュレーション. In **第 30 回自然言語処理学会年次大会 (NLP2024) 論文集**, 2024.

[21] 中川 良尚, 佐野 洋子, 北條 具仁, 木嶋 幸子, 加藤 正弘. 失語症の超長期的経過. **高次脳機能研究 （旧 失語症研究）**, 31(4):373–383, 2011.

[22] 東川 麻里, 波多野 和夫 再帰性発話と反響言語の合併についてー症例報告ー. **高次脳機能研究 （旧 失語症研究）**, 23(4):281–288, 2003.

[23] 藤田一郎, 立石雅巳, 菅野正裕. **標準言語聴覚障害学　失語症学　第 3 版**. 医学書院, 2021.

[24] 山鳥 重. **神経心理学入門**. 医学書院, 1985.

A　付録

A.1　実験プロンプト

復唱タスクは、以下のプロンプトで行った。
次の単語を一度だけ繰り返してください。「目標語」

また、復唱タスクに含まれていた目標語は以下の表 1 に含まれる高頻度語 100 個、低頻度語 100 個の合計 200 個である。janome による形態素解析の際、固有名詞は省かずに抽出している。

高頻度語	低頻度語
日本,現在,時代,放送, 世界,作品,場合,昭和, 使用,東京,活動,大学, 映画,学校,当時,研究, 番組,存在,代表,概要, 同年,可能,発表,出身, 以下,開発,地域,発売, 選手,試合,出場,年代, 以降,平成,歴史,記録, 開始,関係,大会,開催, 一部,参加,時間,中心, 登場,利用,出演,所属, 一般,鉄道,以上,監督, 音楽,明治,事業,人物, 委員,国際,問題,必要, 優勝,世紀,会社,卒業, 影響,結果,教育,計画, 社会,設置,変更,位置, 都市,中国,担当,自身, 選挙,政府,獲得,設立, 最初,これら,地方,女性, 事件,情報,年間,建設, 政治,機関,契約,当初, 就任,構成,制作,中央, 大阪,収録,人口,戦争	農村,会合,流用,債務, 知的,保全,技師,青春, 浮上,元素,東芝,制約, 国軍,焦点,照明,検定, 各社,上場,分間,西方, 室内,寄与,採択,他社, 弾薬,摂取,再会,警告, 調節,先住民,流出,反射, 充実,全部,収益,徒歩, 強度,過剰,大都市,三菱, 神聖,野村,土壌,葬儀, 正当,幹事,重大,島根, 近畿,白色,同型,艦艇, はるか,阪急,降板,職務, 幹線,適応,得票,推奨, 季節,満州,倫理,開局, 電池,本国,最適,倶楽部, 排出,満足,壊滅,損失, 信者,染色,奪取,乗務, 転用,女神,通路,歩行, 休日,役者,標的,喪失, 広大,貴重,客員,看板, 大尉,運輸,撃墜,代数, 代行,体育館,牧師,一時期,遭遇,あまり,庶民, 除去

表 2　プロンプト作成に使用した単語

A.2　max new token=30 にした際の結果

最終的に行った実験では、max new token=1 のみならず、30 の場合でも復唱タスクをモデルに行った。max new token が 30 の場合だと、低頻度語の正答率の方が全体的に向上した。しかし、全体的に正答率が低く、Pruning による影響が大きかったと思われる、1 層から 5 層を Pruning したモデルと 11 層から 15 層を Pruning したモデルは高頻度語のほうが正答率が高い。

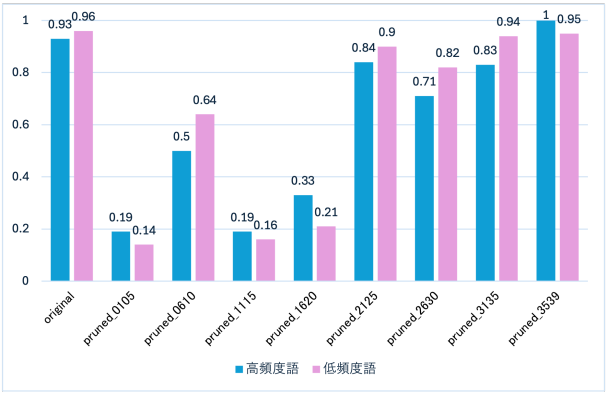

図 4　max new token=30 の際の高/低頻度語の正答率

A.3　類似度評価の結果

4.3 に関連して、0.3 から 0.7 の類似度の範囲内に含まれていた目標語と出力の組み合わせは、以下の表のようなものも観察された。目標語の 2 文字目あるいは 3 文字目が欠落するような出力であったが、このような出力になる単語の種類はモデルごとでばらつきが目立つ様子ではなく、どのモデルでも同じような傾向が見られた。単語の後半部分が欠落するさまは、音韻性錯語的と捉えられる可能性もある。

Models	入力	出力	類似度
original	国軍	国	0.41214275
original	西方	西	0.66252738
original	神聖	神	0.54082328
original	広大	広	0.37564707
pruned_0105	年間	年	0.58770233
pruned_0105	西方	西	0.66252738
pruned_0610	国軍	国	0.41214275
pruned_0610	分間	分	0.38444334
pruned_0610	神聖	神	0.54082328
pruned_0610	広大	広	0.37564707
pruned_1115	広大	広	0.37564707
pruned_1620	西方	西	0.66252738
pruned_1620	神聖	神	0.54082328
pruned_2125	国軍	国	0.41214275
pruned_2125	西方	西	0.6625274
pruned_2125	神聖	神	0.5408233
pruned_2125	広大	広	0.37564707
pruned_2630	国軍	国	0.41214275
pruned_2630	西方	西	0.66252738
pruned_2630	神聖	神	0.54082328
pruned_3135	西方	西	0.6625274
pruned_3135	神聖	神	0.5408233
pruned_3135	広大	広	0.37564707
pruned_3539	国軍	国	0.41214275
pruned_3539	分間	分	0.38444334
pruned_3539	西方	西	0.6625274
pruned_3539	神聖	神	0.5408233
pruned_3539	広大	広	0.37564707
pruned_3539	一時期	一時	0.60175

表 3　プロンプト作成に使用した単語