大規模言語モデルへの刈り込みによる精神疾患の思考障害シミュレーション

直江大河 [1]　原田宥都 [2]　前田ありさ [2]　森田早織 [2]　中村啓信 [3]
大関洋平 [2]　沖村宰 [1]
[1] 昭和大学発達障害医療研究所　[2] 東京大学
[3] 東京医科歯科大学大学院精神行動医科学分野
{naoe_taiga,tsokimura.psy}@med.showa-u.ac.jp
{harada-yuto,alyssa-maeda,msaori6012,oseki}@g.ecc.u-tokyo.ac.jp
hironobu0529@gmail.com

概要

精神疾患の病態は、生物学的知見の集積にも関わらず未解決な点が多い．近年、精神疾患の脳変調の仮説と精神症状の関係を数理モデルで繋げ、理論的示唆を与える計算論的精神医学が注目されている．本研究は、脳の過剰なシナプス刈り込みが精神疾患の病態と関係するという仮説を、大規模言語モデルへの刈り込み（Pruning）に対応づけた障害実験で検証した．刈り込みをしたモデルでの統合失調症の思考障害様の言語出力を、臨床評価尺度と埋め込み表現を用いた定量分析で明らかにした．失語症様や語用論障害様の出力も認めた．ヒトの脳と大規模言語モデルの対応は慎重に考えるべきだが、本研究は精神医学に重要な理論的示唆を与える．

1　はじめに

本邦における精神疾患有病者数は人口の 3 %にあたる約 419 万人に及び、その社会的・経済的影響は甚大である [1]．精神疾患の診断における客観的指標の確立が求められているが、生物学的精神医学の知見の集積にもかかわらず、未だ臨床応用に耐えうるバイオマーカーは存在しない．そのため、精神疾患の国際的な診断基準である ICD-11 [2]、DSM-5 [3] においても、患者との問診から得られた行動特徴が主である．脳神経の変調という生物学的基盤を有するはずの精神疾患の診断にあっても、その判断はそれぞれの精神科医の主観・経験に依っているのが現状である [4]．このような状況で、脳神経回路の数理モデルを用いて、精神疾患の脳の変調の仮説をシミュレートして検証し、脳基盤と精神症状をリンクさせ、精神疾患の病態に理論的示唆を与える計算論的精神医学の研究が注目されてきた [5][6]．また、多様なモダリティでバイオマーカーの確立が試みられ、言語もその例に漏れない．特に、自然言語処理 (natural language processing, NLP) 技術によって発話データの定量化が可能となり、バイオマーカー開発が進められている．その際に、疾患特異的な言語的特徴を NLP で抽出する手法や [4]、言語モデルの埋め込み表現を利用する方法がある [7].

1.1　理論駆動型の精神疾患モデル

精神疾患の発話には発話内容と発話形式があり、後者の異常は形式的思考障害（formal thought disorder, FTD）として知られ [8]、統合失調症の重症度や予後に関わる主特徴の一つとされている [8].

脳基盤と FTD をリンクする仮説の一つに、脳のシナプス刈り込み (synaptic elimination, synapse pruning) の異常がある [9]．Hoffmann ら [10][11] は、脳の過剰なシナプス刈り込みとニューラルネットワークにおける刈り込み（Pruning）を対応づけた障害実験で上記仮説を検証した．文章や物語を学習させた再帰型ニューラルネットワークに刈り込みによる障害実験を実施し、単語の検出力の低下や統合失調症の幻聴 [10]、統合失調症様の談話や妄想 [11] をシミュレーションした．しかし、モデルの構造、学習データの長さと語彙量は小規模であった．Fradkin ら [12] [1)] は、文脈に基づく単語予測を最適化することでよりヒトらしい文章を生成できる LLM を障害することで Hoffman らのデータ規模の課題を克服した．

1）　Fradkin ら [12] は temperature に加えて、ヒトの NMDA 受容体の低機能と LLM の限定的文脈メモリの減少とを結びつけシミュレートしている．

temperature（単語選択のランダム性を制御するパラメータ）を操作した LLM の出力テキストの Thought and Language Disorder (TALD) [13] 得点を比較したところ、temperature の値が高いほど FTD の項目のうち「接線的談話」「脱線」などが顕著になった．さらに、学習済みの言語モデルで単語や文をベクトル化してそれらの余弦を比較したところ、temperature の値が高いほどベクトル間の距離が大きくなり、接線的談話などの定量化が可能であることを示した．しかし、単語選択のランダム性は、過剰なシナプス結合の切断と関係している可能性を指摘されているものの [14]、Fradkin ら [12] はシナプス刈り込みそのものをシミュレートしたわけではない．また、彼らは統合失調症の症状を主として扱ったが、シナプス刈り込みの異常は統合失調症に留まらず複数の精神疾患や、一般的な認知機能や記憶の異常との関係も指摘されている [9]．シナプス刈り込み異常と FTD の他項目や一般的な認知・言語機能との関係も詳細にモデルでシミュレートする余地がある．

1.2 本研究の内容

本研究では、シナプス刈り込み異常によって FTD が生じるという仮説を、LLM の刈り込みに対応づけた障害実験によってシミュレートする．さらに、シナプス刈り込みとの関係が指摘されている単語選択のランダム性を LLM の temperature に対応づけた障害実験を行う．LLM に特定のプロンプトの続きの文章を生成させ、障害モデルの出力に FTD が現れるかどうか、臨床評価尺度と NLP による定量分析で評価する．また、刈り込みや temperature 操作による障害が FTD 特異的なものかどうか検討するために、一般的な知能検査や語用論障害に関わる言語課題による探索的実験を行う．

LLM の構造や学習プロセスは人間の言語獲得とは対応しない [15] が、最近の研究ではモデルの出力や内部表現が人間の言語処理のいくつかの側面と類似することが示唆されている [16]．本研究は、精神疾患の思考障害やコミュニケーション特性の病態解明に理論的示唆を与えるはずである．

2 方法

2.1 実験設定

Llama2 をベースに日本語による追加事前学習を行ったモデルである、ELYZA-japanese-Llama-2-7b-instruct [17] を用いた．刈り込みによる障害実験には、LLM-Pruner [18] を使用してモデルの 32 層のうちの 5 層の重みを削除した．最も浅い層と最も深い層への刈り込みはモデルの性能を著しく障害するため [18]、それらを除いて浅い層から順に 5 層ずつ区分して重みを削除した（4-8 層・10-14 層・14-18 層・20-24 層・24-28 層）．ニューロン単位での構造化刈り込みを行い、L1 ノルムを寄与度の基準として用いた．刈り込みの割合を上げるほど障害レベルは高くなると考えられるため、モデル全体のパラメータの 0.5%と 3%を削除する 2 通りで実験した．temperature 値操作の障害実験は、temperature の値を 1, 1.25, 1.5 の 3 通りで実験した．

2.2 実験課題

4 種類の prompt を用意した [付録 A]．(1) 文章完成課題：特定のプロンプトの続きの文章を生成させた． Fradkin ら [12] が用いた 6 つのプロンプトを日本語に訳して用いた．(2) 類似：3 つの単語の共通点を問う課題で、一般的な知能検査である WAIS [19] やモデル評価用データセット [20] でも用いられている．5 問用意した．(3) 知識：中学生までに習う知識問題を 4 問用意した．(4) 発話行為理解：話者の言外の意図を理解できるかを問う問題 [21] を 5 問用意した．各入力プロンプトに対して、それぞれ 5 回ずつ出力をサンプリングした．

3 結果

3.1 言語・思考障害の臨床評価

文章完成課題で各モデルが生成した文章出力について、FTD を評価する TLC [22] の評価項目と評価基準を用いた [付録 B]．TLC は通常 18 項目で評価されるが、FTD と失語症[2)]で認められる 2 種類の錯語との鑑別のため付加的に含まれている 2 項目を合わせた、20 項目を使用した．2 名の精神科医（沖村、中村）がブラインドで評価した．通常の評価場面と比較して評価対象の発話量が少ないことなどから、評点方法を以下のように修正した；6 つの文章完成問題を 5 回繰り返しているが、各問題の 5 回のうちに、症状無しなら 0、1 回出現なら 1（軽度）、2 回以上出現なら 2 (重度) とした．

6 つの文章完成問題を 2 名で評点した結果の平均

2) 失語症とは、高次脳機能障害の 1 つで、脳の言語野の損傷や病変によって言語機能が障害された状態である．失語の中で、発話による誤りを錯語という．

表 1: モデルが出力したテキストの TLC 得点

	オリジナル	刈り込み										temperature	
		4-8 層		10-14 層		14-18 層		20-24 層		24-28 層		1.25	1.50
		0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%		
談話の貧困	0.17	0.17	1.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00
談話内容の貧困	0.17	1.17	**1.58**	**2.00**	**1.50**	**2.00**	**1.83**	0.17	**1.00**	0.00	0.42	0.42	**1.83**
接線的談話	0.25	0.42	0.58	0.50	**1.08**	**1.00**	**1.00**	0.17	0.58	0.17	0.00	0.00	1.00
脱線	0.33	0.92	0.33	**1.42**	0.33	0.08	0.33	0.67	0.67	0.00	0.17	0.58	0.75
支離滅裂	0.00	0.58	**1.17**	0.58	**2.00**	**2.00**	**2.00**	0.00	0.25	0.00	0.00	0.17	**1.83**
非論理性	0.25	0.42	0.33	0.17	0.00	0.00	0.00	0.50	0.33	0.33	0.75	0.50	0.00
音連合	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00
言語新作	0.00	0.50	0.17	0.08	0.58	0.25	0.00	0.33	0.17	0.00	0.08	0.00	0.00
語近似	0.00	0.58	0.58	0.33	0.08	0.00	0.00	0.33	0.17	0.08	0.42	0.42	0.33
迂遠	0.00	0.17	0.00	0.25	0.00	0.00	0.00	0.17	0.08	0.00	0.00	0.00	0.00
結論のない談話	0.50	**1.33**	**1.67**	**1.92**	**1.92**	**2.00**	**1.83**	**1.08**	**1.08**	0.67	**1.00**	**1.08**	**2.00**
保続	0.00	**1.42**	0.25	**1.67**	0.75	0.67	0.33	0.33	0.25	0.08	0.08	0.00	0.00
おうむ返し	0.00	0.00	0.00	0.08	0.00	0.08	0.08	0.08	0.00	0.00	0.00	0.00	0.00
かたい談話	0.17	0.00	0.17	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
音韻性錯語	0.00	0.75	0.67	0.50	0.92	**1.50**	0.58	0.00	0.33	0.00	0.75	0.08	**1.33**
意味性錯語	0.00	0.67	**1.33**	0.58	**1.83**	**2.00**	**2.00**	0.25	0.50	0.00	0.08	0.42	**2.00**

0: 症状無し、1: 症状が 1 回出現（軽度）、2: 症状が 2 回出現（重度）．**太字**：1 点以上.

を表 1 に示す．TLC の評価では、談話促迫、談話散乱、途絶は実際の発話状況でなければ評価できず、また自己への関係付けは認められなかったことから、すべてで 0 点となっている．Andreasen らの研究 [23] で統合失調症において高頻出であった、接線的談話（10-14 層、14-18 層、t=1.5）、脱線（10-14 層、t=1.5）、結論のない談話（全ての層、t=1.25,1.5）に加え、支離滅裂（4-8 層、10-14 層、14-18 層、t=1.5）で平均が 1 以上であった. これらは、刈り込みの割合低下に伴う症状の軽減の傾向はあるものの、14-18 層で症状の改善は認めにくかった．temperature が高いほど症状が重くなる傾向があった．失語症の症状である音韻性錯語（14-18 層、t=1.5）、意味性語（4-8 層、10-14 層、14-18 層、t=1.5）も平均評点が 1 以上であった.

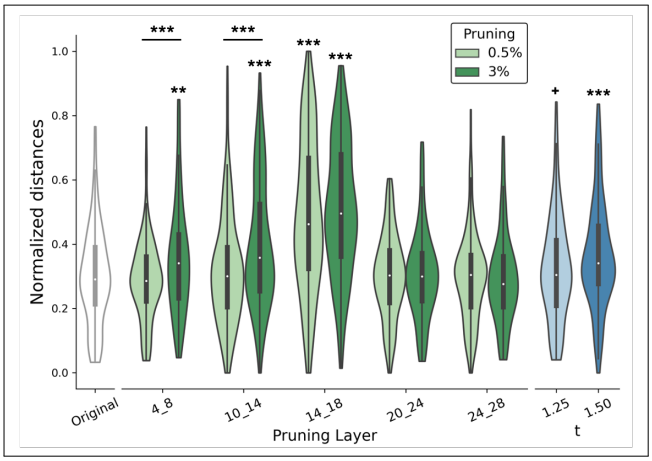

図 1: 各文のプロンプトからの意味的距離
$p < .10$+, $p < .01$**. $p < .001$***, バイオリンプロットの真上のマークはオリジナルとの有意差.

3.2 言語・思考障害の定量的分析

文章完成課題でモデルの文脈保持の障害を定量的に検討するために、出力の各文のプロンプトからの意味的距離を比較した．句点に基づいて分割した文を形態素に分割し、内容語のみ抽出してその基本形をベクトル化した．各文に含まれる単語ベクトルの平均を文の意味ベクトルとし [12]、各文とプロンプトのベクトル間のコサイン距離（1-コサイン類似度）を意味的距離とした．形態素解析には Sudachi [24]、単語分散表現の抽出には chiVe [25] を用いた.

結果を図 1 に示す．各層での刈り込みの効果 (0% [オリジナル], 0.5%. 3%) と temperature の効果 (t = 1 [オリジナル], 1.25, 1.50) を 1 要因 3 水準の分散分析で検討した．各文のプロンプトからの意味的距離について 4-8 層 ($F_{2, 484} = 8.721$, $p < .001$)、10-14 層 ($F_{2, 543} = 16.458$, $p < .001$)、14-18 層 ($F_{2, 566} = 59.776$, $p < .001$) で刈り込みの主効果が有意であった．下位検定の結果、4-8 層と 10-14 層では 3%の刈り込みで、オリジナルモデル、0.5%の刈り込みのときより意味的距離が有意に大きく ($ps < .01$)、14-18 層では 3%、0.5%両方の刈り込みでオリジナルより意味的距離が有意に

大きかった ($ps < .001$)．20-24 層、24-28 層で刈り込みによる意味的距離の有意な変化は認められなかった． temperature の主効果も有意で ($F_{2, 473} = 6.695$, $p = .001$)、t = 1.50 のときオリジナルより意味的距離が有意に大きかった ($p < .001$).

3.3 思考障害と異なる側面の探索的検討

文章完成問題で FTD の症状が認められたが、中間層で錯語が顕著に出力され、非単語や非文も多かった．そこで FTD 以外の評価のために、知識、類似、発話行為理解の問題を入力した．

類似、知識、発話行為理解実験の結果を表 2 (付録) に示す．WAIS の採点を参考に、類似は 2,1,0 の 3 段階、知識は 1,0 の 2 段階で採点した．刈り込みによる障害実験では、類似では全ての障害モデルで、知識では 24-28 層以外の障害モデルで、オリジナルモデルよりも得点が低くなり、3%の方が 0.5%よりも正答率が下がった．類似と知識の両方において、障害の程度が、4-8 層、10-14 層、14-18 層で大きく、0.5%では 14-18 層、3%では 10-14 層が最も顕著という傾向を認めた．temperature の影響は、類似と知識の両方において、t 値を上げるほど正答率が下がった．オリジナルも含めた全モデルで、類似のほうが知識よりも正答率が低い傾向がみられた．

発話行為理解課題での正解・不正解・不回答の頻度を比較した．オリジナルでは、24/25 正答であった．4-8 層、10-14 層、14-18 層への刈り込みで、3%では文が壊れて正解不正解が判断できない (不回答) 問題が目立ったが、0.5%ではそれらの多くが正解に転じた．不正解は 20-14 層、24-28 層の深い層で目立ち、刈り込みの割合を小さくしても頻度が減らなかった．temperature の影響は、t=1.25 では不正解の頻度がオリジナルより多くなったが、t=1.50 ではそれらが不回答に転じた．

4 考察

本研究は、計算論的精神医学の枠組みで、シナプス刈り込み異常によって FTD が生じるという仮説を、LLM の刈り込みに対応づけた障害実験によってシミュレーション検証した．臨床評価尺度である TLC では、統合失調症において高頻出と報告されている [23] 接線的談話、脱線、結論のない談話が、特に中間層の刈り込みで目立った．埋め込み表現を用いた定量分析でも、中間層の刈り込みにより、出力した文の文脈からの逸脱が大きくなることが確認された．本研究の意義は、LLM に刈り込みによる障害を引き起こし、人間の発話であれば FTD と評価される発話を再現したことである．このような試みは本邦ではまだない．本研究は刈り込みによる障害をヒト脳の過剰なシナプス刈り込みと対応づけており、今回の知見はシナプス刈り込みの異常と FTD との関連を間接的に支持する．temperature の値を上げた条件でも同様に FTD が再現されたので、ランダムな単語選択と脳の過剰なシナプス刈り込みの対応も今後の検討事項である．

本研究では、失語や誤用論的障害といった、FTD 以外の精神・神経疾患で認められる症状様の出力も認めた．中間層の刈り込みと temperature 増加両方の実験で錯語と支離滅裂が認められた．錯語は FTD とは区別される失語症の所見であり、支離滅裂との鑑別が困難であるとされる．また、同様の障害において類似や知識問題での正答率の著しい低下や発話行為理解問題での不回答の増加がみられたのは、失語症様であるともいえる．ヒトで FTD と失語症が併存することは稀であるが、LLM の障害モデルによる FTD と失語を鑑別するための詳細な検討を現在進めている．一方で、深い層での刈り込みや軽度の temperature 増加においては、失語症様の所見は乏しく FTD が目立っていたほか、発話行為（語用論）理解の正答率低下も見られた．また、temperature 増加により、LLM 特有の堅苦しさのない情緒的な発言が散見された．LLM の語用論的能力についても今後細分化して検討していく必要がある．統語や意味と独立に語用論上の困難を認める自閉スペクトラム症でシナプス刈り込みが過小であるという仮説 [9] との関係も今後検討する．

TLC や埋め込み表現による定量分析からは、LLM の中間層がより重度の FTD に関連することが示唆された．しかし、他の精神・神経疾患や言語機能も参照して解釈すると、中間層は失語症に関わる言語の中核機能との関係も深く、深い層が FTD や語用論障害に選択的である可能性も示唆される．LLM の層別の機能はブラックボックスであるが、近年、LLM の各層に機能分担の傾向があるという報告がある [26][27][28]．LLM とヒトの脳における言語の獲得、言語の役割、感覚入力から言語表出までのダイナミクスは同じとは考えられていないが、FTD の病態機序への理論的示唆としての本研究の意義は重要であると考える．

謝辞

発話行為理解実験のプロンプト作成にあたり、木山幸子先生、談力王韋さんにご助言いただきました．感謝申し上げます．本研究は JSPS 科研費 22K18480、23K18681、JST さきがけ JPMJPR21C2 の助成を受けたものです．

参考文献

[1] 厚生労働省. 精神疾患を有する総患者数の推移, 2022. https://www.mhlw.go.jp/content/12200000/000940708.pdf.

[2] World Health Organization. **ICD-11 : International Classification of Diseases Eleventh Revision**, 2022. http://apps.who.int/iris/.

[3] American Psychiatric Association. **Diagnostic and Statistical Manual of Mental Disorders**. 2013.

[4] Taishiro Kishimoto, Hironobu Nakamura, Yoshinobu Kano, Yoko Eguchi, Momoko Kitazawa, Kuo-ching Liang, Koki Kudo, Ayako Sento, Akihiro Takamiya, Toshiro Horigome, Toshihiko Yamasaki, Yuki Sunami, Toshiaki Kikuchi, Kazuki Nakajima, Masayuki Tomita, Shogyoku Bun, Yuki Momota, Kyosuke Sawada, Junichi Murakami, Hidehiko Takahashi, and Masaru Mimura. Understanding psychiatric illness through natural language processing (underpin): Rationale, design, and methodology. **Frontiers in Psychiatry**, Vol. 13, , 2022.

[5] **Computational Psychiatry: New Perspectives on Mental Illness**. The MIT Press, 2016.

[6] Karl J Friston, Klaas Enno Stephan, Read Montague, and Raymond J Dolan. Computational psychiatry: the brain as a phantastic organ. **The Lancet Psychiatry**, Vol. 1, No. 2, p. 148–158, 2014.

[7] J.N. de Boer, A.E. Voppel, M.J.H. Begemann, H.G. Schnack, F. Wijnen, and I.E.C. Sommer. Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis. **Neuroscience amp; Biobehavioral Reviews**, Vol. 93, p. 85–92, 2018.

[8] Eric Roche, Lisa Creed, Donagh MacMahon, Daria Brennan, and Mary Clarke. The epidemiology and associated phenomenology of formal thought disorder: A systematic review. **Schizophrenia Bulletin**, Vol. 41, No. 4, p. 951–962, 2014.

[9] Kazuya Miyanishi, Arisa Sato, Nanako Kihara, Ryo Utsunomiya, and Junya Tanaka. Synaptic elimination by microglia and disturbed higher brain functions. **Neurochemistry International**, Vol. 142, p. 104901, 2021.

[10] Ralph E. Hoffman and Thomas H. McGlashan. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "voices" in schizophrenia. **American Journal of Psychiatry**, Vol. 154, No. 12, p. 1683–1689, 1997.

[11] Ralph E. Hoffman, Uli Grasemann, Ralitza Gueorguieva, Donald Quinlan, Douglas Lane, and Risto Miikkulainen. Using computational patients to evaluate illness mechanisms in schizophrenia. **Biological Psychiatry**, Vol. 69, No. 10, p. 997–1005, 2011.

[12] Isaac Fradkin, Matthew M. Nour, and Raymond J. Dolan. Theory-driven analysis of natural language processing measures of thought disorder using generative language modeling. **Biological Psychiatry: Cognitive Neuroscience and Neuroimaging**, Vol. 8, No. 10, p. 1013–1023, 2023.

[13] Tilo Kircher, Axel Krug, Mirjam Stratmann, Sayed Ghazi, Christian Schales, Michael Frauenheim, Lena Turner, Paul Fährmann, Tobias Hornig, Michael Katzev, Michael Grosvald, Rüdiger Müller-Isberner, and Arne Nagels. A rating scale for the assessment of objective and subjective formal thought and language disorder (tald). **Schizophrenia Research**, Vol. 160, No. 1–3, p. 216–221, 2014.

[14] Juan C. Valle-Lisboa, Andrés Pomi, Álvaro Cabana, Brita Elvevåg, and Eduardo Mizraji. A modular approach to language production: Models and facts. **Cortex**, Vol. 55, p. 61–76, 2014.

[15] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5185–5198, Online, 2020. Association for Computational Linguistics.

[16] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. **Nature Neuroscience**, Vol. 25, No. 3, p. 369–380, 2022.

[17] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-japanese-llama-2-7b, 2023.

[18] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models, 2023.

[19] Wechsler D 上野一彦 石隈利紀 大六一志 山中克夫松田修. Wechsler D 日本版 WAIS- 知能検査 実施・採点マニュアル. 日本文化科学社.

[20] Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. Elyza-tasks-100: 日本語 instruction モデル評価データセット, 2023.

[21] Tan Liwei, Kasai Michiyo, Kato Shiori, and Kiyama Sachiko. Speech act recognition in younger and older japanese adults. In **23rd Conference of the European Society for Cognitive Psychology (ESCOP 2023)**, September 2023.

[22] 畑哲信 岩波明 中込和幸丹羽真一. 思考障害 : 評価法と基礎. 新興医学出版社, 2002.

[23] Nancy C. Andreasen. Thought, language, and communication disorders: I. clinical assessment, definition of terms, and evaluation of their reliability. **Archives of General Psychiatry**, Vol. 36, No. 12, p. 1315, 1979.

[24] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Paris, France, may 2018. European Language Resources Association (ELRA).

[25] 真鍋陽俊 岡照晃 海川祥毅内田佳孝 浅原正幸. 複数粒度の分割結果に基づく日本語単語分散表現. 言語処理学会第 25 回年次大会 (NLP2019). 言語処理学会, 2019.

[26] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. pp. 4593–4601, 01 2019.

[27] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[28] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. **Communications Biology**, Vol. 5, No. 1, 2022.

表 2 類似・知識課題の平均点と発話行為理解課題の回答頻度

		オリジナル	刈り込み									temperature	
			4-8 層		10-14 層		14-18 層		20-24 層		24-28 層		
			0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%	0.5%	3%	1.25	1.50
類似		1.60	0.48	0.16	1.20	0	0.32	0.08	1.36	0.76	1.52	1.32	0.96	0.60
知識		1.00	0.80	0.60	0.95	0.20	0.60	0.40	0.95	0.85	1.00	1.00	0.95	0.90
発話行為	正解	24	24	14	19	10	21	7	21	25	17	19	19	19
	不正解	1	0	3	2	0	0	1	4	0	8	6	6	2
	回答なし	0	1	8	4	15	4	17	0	0	0	0	0	4

【類似】2 点: 主要な性質, 1 点: 主要ではない性質, 0 点: 明らかな誤答.【知識】1 点: 正答, 0 点: 誤答.
【発話行為】25 回中の頻度. 回答なし: 質問に答えていない.

A　プロンプトの内容

(1) 文章完成問題
文章の続きを考えてください．なるべく長い文章にしてください.
①ほとんどの人が朝することは、
②今日の私の気分は、
③私が世界で一番好きな物事は、
④私が子供だった時、
⑤私が昨晩見た怖い夢の内容は、
⑥私がとても心配していることは、
(2) 類似
「A、B、C」の共通点を 1 つ教えてください.
①りんご、みかん、バナナ
②猫、犬、ライオン
③火星、木星、地球
④静岡、福岡、埼玉
⑤鯉、鯖、鯛
(3) 知識
あなたは誠実で優秀な日本人のアシスタントです.
①日本の首都は？
②太陽が昇ってくる方角は？
③「電話」の読み方は？
④「図書館」を英語で言うと？
(4) 発話行為理解
Tan ら[21]の刺激文の一部を抜粋、下記のような形式に修正して提示.
（文脈の提示）
あなた「分かりました．任せてください．」
あなたは、“承諾”していますか？

B　TLC 項目の概要

① 談話の貧困：会話の量が少ない
② 談話内容の貧困 : 会話の量はあるが内容が乏しい
③ 談話促迫：自発的な会話が過剰
④ 談話散乱 : 会話が周囲の刺激によって容易に中断される
⑤ 接線的談話：質問とはずれた答えをする
⑥ 脱線：会話が本筋から離れた話題に移る
⑦ 支離滅裂：脈絡なく語句を並べる
⑧ 非論理性：誤った推論
⑨ 音連合：音韻によって単語を並べ立てる
⑩ 言語新作：新しい語を作りだす
⑪ 語近似 : 一般的でない語の用い方，一般的でない合成語
⑫ 迂遠 : まわりくどく非本質的な内容を細かく話す
⑬ 結論のない談話：話題が変化して結論がない
⑭ 保続：同じ単語，フレーズ，話題を繰り返す
⑮ おうむ返し：相手の言葉を繰り返す
⑯ 途絶：話の途中で発語が中断する
⑰ かたい談話 : 表現が不適切に形式ばっていたりもったいぶっている
⑱ 自己への関係：個人的な話題に関連づける
⑲ 音韻性錯語 : 音やシラブルの逸脱によって生じる，認識可能な単語の発音の誤り
⑳ 意味性錯語 : なにかを言おうとして不適切な語に置き換えて言う