意図を読む大規模言語モデルの言語による比較
Comparative Analysis of Large Language Models in Interpreting Intent

墨 泰我 †，飯田 愛結 †，長原 令旺 †，保阪 靖人 †，森山 園子 †，大澤 正彦 †
Taiga Sumi, Ayu Iida, Reo Nagahara, Yasuhito Hosaka, Sonoko Moriyama, Masahiko Osawa
† 日本大学
Nihon University
chta22016@nihon-u.ac.jp

概要

著者らは意図を読むことができる大規模言語モデルを実現するために，大規模言語モデルと認知アーキテクチャを統合することを提案し，その有効性を示している．しかし，この研究では，プロンプトが日本語で書かれていたことが大きな影響を与えた可能性がある．本研究では，提案手法において日本語とドイツ語でプロンプトを作成し，結果を比較することで，使用する言語が意図理解に及ぼす影響を調査した．結果，特定の条件において，言語ごとに意図理解の程度に差が見られた．

キーワード：大規模言語モデル（Large Language Model: LLM），認知アーキテクチャ（Cognitive Architecture），他者モデル（Mental Model of Others），意図推定（Intent Estimation），言外の意味（Implicature）

1. はじめに

大規模言語モデルとは，数十億から数兆ものパラメータを持つ深層学習モデルの一種で，膨大なテキストデータを用いて学習を行っている．本研究の目的は，大規模言語モデルを用いて，より人間らしい対話を行うエージェントを実現するために必要な知見を得ることである．

大規模言語モデルは，様々なタスクで高い性能を発揮する一方で，意図を読む必要があるコミュニケーションタスクでは性能が低いことが示されている (Mahowald, Ivanova, Blank, Kanwisher, Tenenbaum, & Fedorenko, 2023; Hu, Floyd, Jouravlev, Fedorenko, & Gibson, 2023)．この問題を解決するために，著者らは大規模言語モデルと認知アーキテクチャを統合する方法を提案している (飯田他, 2024)．実験では，認知アーキテクチャを統合することで，意図を読む必要があるコミュニケーションタスクの性能が向上することを示唆した．

しかし，この研究ではプロンプトが全て日本語で書かれている点に注意する必要がある．大規模言語モデルは，入力する言語によって出力の傾向が左右され，特に学習データが乏しい言語では，言語の理解や生成が困難となることが示されている (Bang, Cahyawijaya, Lee, Dai, Su, Wilie, Lovenia, Ji, Yu, Chung, Do, Xu, & Fung, 2023; Liu, Zhang, Zhao, Luu, & Bing, 2024)．また，各国のコミュニケーション文化には，対話の文脈を考慮する度合いに差がある．文脈をより考慮するコミュニケーション文化は間接的で曖昧な発話をし，文脈をあまり考慮しないコミュニケーション文化は直接的で正確な発話をするという特徴がある (Würtz, 2005; Gudykunst, Matsumoto, Ting-Toomey, Nishida, Kim, & Heyman, 1996)．このことから，著者らの実験において，大規模言語モデルに与えるプロンプトを日本語以外の言語で作成することで，異なる傾向が現れる可能性がある．

本研究では，意図を読む必要があるコミュニケーションタスクにおいて，日本語とドイツ語の 2 つの言語でプロンプトを作成し，出力を言語間で比較することで，意図理解の程度に影響を及ぼすかを検証する．実験では，飯田他 (2024) で用いた日本語のプロンプトをドイツ語に翻訳し，発話者の言外の意味を踏まえた応答ができるかを調査した．

2. 大規模言語モデルと認知アーキテクチャの統合

大規模言語モデルは，意図を読む必要のあるタスクにおいて，十分な性能を発揮できていないことが報告されている (Mahowald et al., 2023; Hu et al., 2023)．著者らは，この課題を解決するために，大規模言語モデルと認知アーキテクチャを統合する手法を提案し，この有効性を検証している (飯田他, 2024)．この研究では，BDI (Belief-Desire-Intention) モデル (Rao & Georgeff, 1997) で用いられている BDI の考え方をベースとした発話意図に基づく自己/他者モデル付き発話生成アーキテクチャを設計している．このアー

表 1　各言語の皮肉シチュエーションにおける初期値および入力

日本語	
他者の信念	対話相手は客である/すでに 2 時間経っている
他者の願望	早く帰ってほしい
他者の発話	「あんた、ずいぶんいい時計してはりますね〜」
自己の信念	2 時間ほどお邪魔している
自己の願望	相手に悪く思われたくない
ドイツ語	
Glaube des Anderen	Der Gesprächspartner ist ein Kunde. / Es sind bereits 2 Stunden vergangen.
Wünsche des Anderen	Ich möchte, dass er/sie bald weggeht.
Äußerungen des Anderen	'Du hast ja eine ziemlich schöne Uhr da.'
Eigener Glaube	Ich bin seit etwa 2 Stunden zu Gast.
Eigene Wünsche	Ich möchte nicht, dass der Gesprächspartner schlecht von mir denkt.

キテクチャは，「自己」および「（自己が思う）他者」それぞれの「信念 (Belief)」，「願望 (Desire)」，「意図 (Intention)」の 6 つの内部表現と，「（他者の）意図推定」，「（自己の）意図生成」，「（自己の）発話生成」の 3 つのモジュールで構成されている．

飯田他 (2024) では，大規模言語モデルと認知アーキテクチャを統合する 2 種類の方法を提案している．1 つ目の方法は，認知アーキテクチャをプログラムとして実装し，内部のモジュールをそれぞれ大規模言語モデルで実装する方法で，「LLM Embedded in Cognitive Architecture (LEC)」と名付けられている．2 つ目は，認知アーキテクチャとそれを構成するモジュールや振る舞いを詳細に説明するプロンプトを作成し，大規模言語モデルへ入力する方法で，「Cognitive Architecture Embedded in LLM (CEL)」と名付けられている．

実験では，3 つのシチュエーションに対して，手法を変えた 4 つの条件をそれぞれ 10 回ずつ，計 120 回行った．3 つのシチュエーションは，言外の意味を持つ典型的な例として，皮肉・ツンデレ・社会的制約と名付け，自己および他者の信念・願望と他者の発話を設定することで作成した．皮肉とは，「あんたいい時計してはりますね」という京言葉を用いて，客に早く帰らせようとするシチュエーションである．ツンデレとは，「そっち優先したら」という発話を行うが，本当は自分を優先して欲しいと思っているシチュエーションである．社会的制約とは，「無理しないで」という発話を上司が部下に対して行うが，本当は無理してでも働いて欲しいと思っているシチュエーションである．4 つの条件とは，提案手法である LEC 条件と CEL 条件，単純な大規模言語モデルに近い振る舞いをする LLM 条件，提案手法と同じ信念・願望のみを与えられる LLM with BD (LWB) 条件である．

結果として，LLM 条件では，どのシチュエーションにおいても言外の意味を踏まえた発話はなく，成功率はいずれも 0% であった．これは，大規模言語モデルが設定した他者の発話の字義的な意味から，言外の意味を予測できないことを示している．信念・願望の情報を与えた LWB 条件では，ツンデレと社会的制約のシチュエーションでそれぞれ 100% と 90% という成功率が得られた．これらのシチュエーションは，アーキテクチャがなくても高い成功率となったため，アーキテクチャの有効性を検証するには適さなかった．皮肉のシチュエーションでは 30% と低い成功率となったため，皮肉のシチュエーションにおいて LEC 条件と CEL 条件で比較したところ，LEC 条件で発話生成は 100% となり，アーキテクチャを組み合わせる有効性が示された．一方，CEL 条件では成功率は 40% に止まった．

3. 実験

本実験の目的は，プロンプトに用いる言語を変更することが意図理解の程度に影響を及ぼすのかを検証することである．実験では，発話と発話意図に乖離のある対話を用いて，日本語とドイツ語の 2 つの言語でプロンプトを作成し，その出力を比較する．実験には，GPT-4 を用いる．

3.1 比較言語

比較する言語として日本語とドイツ語の 2 つの言語を設定した．この 2 言語を設定した理由は，日本とドイツのコミュニケーション文化は対照的な特徴を有しているからである．日本は言外の意味を察し合うコミュニケーションをする傾向がある一方で，ドイツではそのようなコミュニケーションをあまり行わず直接

表 2　実験結果 日本語

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0%	-	-	0%	-	-	0%
LWB	-	-	0%	-	-	60%	-	-	100%
LEC	90%	100%	100%	90%	70%	70%	100%	80%	80%
CEL	60%	60%	60%	40%	50%	50%	20%	70%	60%

表 3　実験結果 ドイツ語

	皮肉			ツンデレ			社会的制約		
	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成	意図推定	意図生成	発話生成
LLM	-	-	0%	-	-	0%	-	-	0%
LWB	-	-	0%	-	-	10%	-	-	10%
LEC	100%	100%	100%	10%	10%	20%	100%	100%	90%
CEL	100%	100%	80%	50%	30%	0%	60%	70%	70%

的な表現を好む傾向がある (山科, 2018)．そのため，この 2 つの言語でプロンプトを作成して出力を比較することで，意図理解の程度が変わる可能性がある．一方で，最も多くのデータが学習に用いられる英語を比較言語に用いなかった理由は，英語はさまざまな国で使用されており，言語と文化の関係性を議論することが難しいためである．2 つの言語で作成したプロンプトの初期値および入力の例を表 1 に示す．日本語のプロンプトは飯田他 (2024) で用いたものを利用し，ドイツ語のプロンプトは日本語のプロンプトを翻訳することで作成した．ドイツ語への翻訳は，ドイツ語を専門とする第 4 著者の監訳のもと行った．なお，発話と発話意図に乖離のあるシチュエーションは，飯田他 (2024) と同様のものを使用した．

3.2　実験手順

実験では，2 章で述べた飯田他 (2024) と同様の 4 つの条件（LLM 条件，LWB 条件，LEC 条件，CEL 条件）を設定し，各条件に対して 3 つのシチュエーション（皮肉，ツンデレ，社会的制約）の対話をそれぞれ 10 回ずつ行った．LLM 条件および LWB 条件については発話生成の出力結果を，LEC 条件および CEL 条件については意図推定・意図生成・発話生成の出力結果を，それぞれ成功/失敗と評価した．成功の基準は，他者の発話内容に基づいて言外の意味を読み取った語句やフレーズが含まれていることである．また，失敗の基準は，字義通りの意味に基づいた語句やフレーズのみであることである．評価は第 1 著者と第 2 著者の合議によって素案を作成し，共著者に照会した．

3.3　実験結果

各シチュエーションにおける各条件下での成功率を，言語ごとに表 2–3 にそれぞれ示す．

日本語の結果（表 2）は，飯田他 (2024) と概ね同様の傾向が見られる結果となった．LLM 条件では，全てのシチュエーションで言外の意図を踏まえた発話はなく，成功率は 0% であった．LWB 条件では，自己および他者の信念・願望の情報を与えたことで成功率が上昇し，ツンデレと社会的制約のシチュエーションでそれぞれ 60% と 100% となった．一方で，皮肉のシチュエーションでは 0% であった．LEC 条件では，全てのシチュエーションの全てのモジュールで，成功率は 70% 以上となった．CEL 条件では，LEC 条件と比べると全ての条件の全てのモジュールで低い結果となった．

一方，ドイツ語（表 3）では，皮肉と社会的制約のシチュエーションにおける LWB 条件以外の条件では概ね日本語と同様の結果となった．LLM 条件では，日本語と同様に全てのシチュエーションで成功率は 0% であった．LWB 条件では，全てのシチュエーションにおいて成功率は 10% 以下となり，日本語と比較して大きく下回る結果となった．LEC 条件では，皮肉と社会的制約のシチュエーションにおいて，全てのモジュールで成功率が 90% 以上となった．一方で，ツンデレのシチュエーションでは，10%/10%/20% という成功率となり，他のシチュエーションと比べて低い結果となった．CEL 条件では，皮肉のシチュエーションにおいて，全てのモジュールで成功率は 80% 以上となった．また，社会的制約のシチュエーションにおいても，60%/70%/70% となり，LWB 条件の成功率と比べると高い傾向が見られた．ツンデレのシチュエーションでは，意図推定と意図生成ではそれぞれ 50% と 30% となったが，発話生成では成功率は 0% であった．

4. 考察

4.1 ツンデレシチュエーションにおける言語間の比較

ツンデレシチュエーションにおいて日本語とドイツ語で結果を比較すると，ほぼ全ての条件でドイツ語の方が低い結果が得られた．このことから，ツンデレというシチュエーションが，ドイツ語では特に理解することが難しかった可能性がある．ツンデレとは，対話相手に対して内心では肯定的な印象を抱いているにもかかわらず，否定的な発話をしてしまうシチュエーションである．ツンデレという概念は日本の漫画やアニメから生まれたものであり，日本特有の文化としての側面が強いと考えられる．そのため，大規模言語モデルの学習データにおいて，ドイツ語の学習データと比べて，日本語の学習データにツンデレと似たようなシチュエーションの対話が多く含まれていた可能性がある．ツンデレというシチュエーションが日本語でのみ良い結果を示すのかは，ドイツ語以外の言語についても調査する必要があるだろう．

4.2 CEL 条件における言語間の比較

CEL 条件において，皮肉と社会的制約のシチュエーションで，ドイツ語が日本語より成功率が高いケースがいくつかあった．大規模言語モデルは，プロンプトに含まれるトークン数が多いほど，性能が低下する傾向がある (Levy, Jacoby, & Goldberg, 2024)．今回実験に用いたアーキテクチャを説明しているプロンプトのトークン数を調べたところ，日本語の場合 808 トークンであったのに対し，同じ内容出会ってもドイツ語では 768 トークンと少なかった．このトークン数の差が，日本語とドイツ語の CEL 条件における性能に影響していると考えられる．CEL の手法では，プロンプトにおけるアーキテクチャなどの説明部分をトークン数が少ない言語で作成し，初期値および入出力を任意の言語で作成することで，どのような言語でも高い成功率を示すかもしれない．

5. おわりに

本研究では，大規模言語モデルの意図を読む必要があるコミュニケーションタスクにおいて，プロンプトに用いる言語が意図理解の程度に及ぼす影響を調査した．実験では，意図を読む必要があるシチュエーションにおいて，日本語とドイツ語でプロンプトを作成し，その結果を比較した．今回の実験の範囲では，日本特有のシチュエーションにおいて，ドイツ語では言外の意味を踏まえた発話ができない可能性が示唆された．また，CEL の手法では，トークン数の少ない言語でプロンプトを作成することで，成功率が上昇するかもしれない．今後は，タスクごとに言語別の優位性をより詳細に調査する．

文 献

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *Proceedings of the ACL*, pp. 675–718.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self-construals, and individual values on communication styles across cultures. *Human Communication Research*, 510–543.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. *Proceedings of the ACL*, pp. 4194–4213.

飯田 愛結・阿部 将樹・奥岡 耕平・福田 聡子・大森 隆司・中島 亮一・大澤 正彦 (2024). 意図を読む AI の実現に向けて: 対話型生成 AI と他者モデルの統合を例に *Proceedings of the HAIS*.

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. *arXiv*.

Liu, C., Zhang, W., Zhao, Y., Luu, A. T., & Bing, L. (2024). Is Translation All You Need? A Study on Solving Multilingual Tasks with Large Language Models. *arXiv*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective..

Rao, A. S., & Georgeff, M. P. (1997). Modeling rational agents within a BDI-architecture. *Readings in agents*, pp. 317–328.

Würtz, E. (2005). Intercultural Communication on Web sites: A Cross-Cultural Analysis of Web sites from High-Context Cultures and Low-Context Cultures. *Journal of Computer-Mediated Communication*, *11* (1), 274–299.

山科 美智子 (2018). コミュニケーションにおける文化的相違 -日本人の異文化認識と Erin Meyer による The Culture Map との比較- 埼玉女子短期大学研究紀要, *37*, 79–99.