文脈の違いに着目した言外の意図理解 日本語ベンチマークの構築

小方雅子 [1]　菊池英明 [1]
[1] 早稲田大学 人間科学研究科
m.ogata_2491@asagi.waseda.jp　kikuchi@waseda.jp

概要

対話における言外の意図は文脈に応じて変化しうるため，大規模言語モデル（LLM）が人間と円滑にコミュニケーションを行うためには，文脈を踏まえた言外の意図理解能力が重要である。本研究では，同一の発話文に対して異なる文脈を与え言外の意図を回答させることで文脈依存的な言外の意図理解を評価可能な日本語ベンチマークを構築した。構築したベンチマークを用いて複数の LLM を評価した結果，いずれのモデルも文脈依存事例において高い正解率を示し，文脈情報を利用した言外の意図理解が可能であることが示された。さらに，ベンチマーク規模の影響を分析した結果，構築段階で無作為に並べ替えられたベンチマークの先頭 400 件程度のみを用いた評価であっても，全体評価を安定して近似できることが確認された。

1　はじめに

大規模言語モデル（Large Language Model; LLM）の著しい発展が続く中で，LLM の性能を様々な観点から評価することが求められている。LLM の性能を評価するにあたっては，主にベンチマークと呼ばれる評価用データセット群が用いられる [1] 。英語におけるベンチマークは GLUE[2] を先駆けとして数多くのデータセットが構築されている。一方で日本語においては GLUE の日本語版である JGLUE[1] などのベンチマークの構築が急速に進められているものの，依然として英語と比べ日本語のベンチマークは少なく，日本語におけるベンチマークのさらなる構築が期待される。

LLM 評価の観点のひとつとして挙げられるのが対話における曖昧な意図や要求を読み取る能力である。人間の対話では，意図や要求を直接的にではなく間接的に表現する発話が見られる（間接発話行為）[3]。人間は対話の文脈をもとに，このような間接的な発話によって表現された意図を推測する。同じ文であっても文脈に応じて発話の意図は異なり得る [3]。LLM が人間と円滑にコミュニケーションをとるためには，文脈に応じて変化する曖昧な意図を理解することが重要である。しかしながら，対話における曖昧な意図に関連するデータセットはいくつか構築されている [4] ものの，文脈が異なることによる意図解釈の変化に着目しているものは少ない。そこで本研究では，LLM が対話において文脈に応じて変化する言外の意図を理解できているのかを評価するための日本語のベンチマークを構築し，LLM の理解能力を評価することを目的とする。

元来，ベンチマーク構築研究ではデータサイズが大きいほど信頼性があるとされ，多くのデータを収集することが目指されてきた。しかしながら巨大なベンチマークでの LLM 評価は評価そのものに多大なコストがかかることが指摘されている。そこで，無作為抽出や項目反応理論を用いた抽出により信頼性を損なわずにベンチマークの問題数を削減し，LLM 評価のコストを抑える研究がなされている [5][6]。そこで本研究では異なる規模のベンチマークを用いてベンチマークの大規模化の必要性についても検証する。

2　ベンチマーク構築

本研究で構築するベンチマークは，対話文およびその対話文内で生じる言外の意図の選択肢・正解ラベルの組から構成した。対話文は言外の意図を生じさせる発話（コア文と呼称する）と，その発話の意図を解釈するために必要となる対話文脈（文脈文と呼称する）を収集した。一つのコア文に対し，そのコア文からそれぞれ異なる意図が読み取られるような複数の文脈文を用意した。データセット構築の手順は以下の通りである。次節以降で手順の詳細を説

明する。

1. コア文・文脈文の抽出
2. 追加の文脈文の作成
3. フィルタリング・正解ラベル付与

2.1 コア文・文脈文の抽出

日本語日常対話コーパス [7] をもとに，LLM によってコア文・文脈文となりうる対話文を抽出した。同時に，その中で生じる言外の意図を記述させた。その結果，抽出するデータとして不適切なものが観察されたため，著者らの観察によって以下の条件を満たすものを不適切なデータとして除外した。

- 対話文から読み取られる言外の意図としては直接的過ぎるもの
- 日本語対話文として不自然であるもの

抽出したデータの例を以下に示す。

- 文脈文：入院中の病院の環境はどうでしたか？
- コア文：ベッドの上での生活は大変でした。
- 言外の意図：入院中の環境が快適ではなかった

2.2 追加の文脈文の作成

2.1 節で抽出したデータをもとに，LLM によって 2.1 節に示した方法で記述した言外の意図と異なる意図を生じさせるような文脈文を複数件ずつ作成させた。同時に，各文脈文において生じる言外の意図を記述させた。その結果，2.1 節と同様に不適切なデータが作成されたため，著者らの観察によって同様の条件で不適切なデータを除外した。

作成したデータの例を以下に示す。

- 文脈文：骨折してギプスをしていた時期はどうでしたか？
- コア文：ベッドの上での生活は大変でした。
- 言外の意図：ギプス期間中は安静にしていなければならず辛かった

2.1 節および本節の方法で記述した言外の意図を 2.3 節および 3 章で用いる言外の意図選択肢とした。

2.3 フィルタリング・正解ラベル付与

前節までの手順によって得られた 779 件のデータに対し，人手によるフィルタリングおよび正解ラベル付与を行った。まず，得られたデータを 8 分割し，各分割に対して 3 名ずつ，合計 24 名の作業者にアノテーションを依頼した。各作業者は互いに独立して作業を行った。

作業者には，各対話文について質問文を提示し，文脈文およびコア文を読んだ上で，コア文から読み取られる言外の意図として最も適切であると感じた選択肢を 1 つ選択するよう求めた。質問文のフォーマットを以下に示す。

以下の対話文を読んで，その後の質問に答えてください。
【対話文】
B: {文脈文}
A: {コア文}
【質問】
発話者 A は，「{コア文}」と言うことによって，以下のどのような意図を伝えていると感じましたか？もっともよく当てはまると感じたものを選んでください。
【選択肢】
- {言外の意図選択肢 1}
- {言外の意図選択肢 2}
- {言外の意図選択肢 3}
- {言外の意図選択肢 4}

質問文の例を以下に示す。

以下の対話文を読んで，その後の質問に答えてください。
【対話文】
B: 入院中の病院の環境はどうでしたか？
A: ベッドの上での生活は大変でした。
【質問】
発話者 A は，「ベッドの上での生活は大変でした。」と言うことによって，以下のどのような意図を伝えていると感じましたか？もっともよく当てはまると感じたものを選んでください。
【選択肢】
- ギプス期間中は安静にしていなければならず辛かった
- リハビリテーションに長期間かかった
- 入院中の環境が快適ではなかった
- 該当なし

各対話文に対して 3 名の作業者の回答を集計し，同一の選択肢が 2 名以上によって選択された場合に，当該選択肢を正解ラベルとして採用した。一方で，3 名の回答がすべて異なる事例については，分析対象から除外した。

この手順により，言外の意図について一定の合意が得られた事例（735 件）のみをベンチマークに

表 1 文脈依存性別の正解率

モデル	文脈非依存（n=45）	文脈依存（n=690）
GPT-5	0.733	0.842
GPT-4	0.800	0.826
Claude-4.5	0.689	0.883

含めることで，評価データとしての信頼性を担保した。

3 LLM 評価実験

構築したベンチマークを用いて LLM の評価実験を行う。LLM に対話文（文脈文およびコア文）を提示し，コア文から解釈される言外の意図を選択肢から選択させる。その正解率によって LLM を評価する。本研究では LLM 評価実験結果を以下の観点から分析する。

1. 文脈依存的な言外の意図理解能力の検証
2. ベンチマーク規模の比較

3.1 文脈依存的な言外の意図理解能力の検証

各コア文に対し，文脈文ごとに正解選択肢を選択できているかを観察することで，LLM が文脈に応じて変化する言外の意図を理解できているのかを検証した。同一のコア文を持つデータの中で正解ラベルが複数存在する事例を文脈依存事例，単一の正解ラベルのみが存在する事例を文脈非依存事例とした。

表 1 に，各モデルにおける文脈依存事例および文脈非依存事例での正解率を示す。GPT-5 および GPT-4 はいずれも文脈依存事例において高い正解率を示し，それぞれ 0.842，0.826 であった。Claude-4.5 は文脈依存事例において最も高い正解率（0.883）を示した。

一方，文脈非依存事例は 45 件と少数であり，正解率の解釈には注意が必要であるが，いずれのモデルにおいても文脈依存事例の方が高い正解率を示す傾向が確認された。

3.2 ベンチマーク規模の比較

ベンチマーク規模の影響を検証するため，構築段階で無作為に並べ替えられたベンチマークの先頭 n 件のみを用いた評価結果を，全体ベンチマークでの評価結果と比較した。

表 2 および図 1 に示すように，n=50 や n=100 といった小規模な設定では，全体正解率との差が比較的大きく，評価結果が不安定になる傾向が確認された。一方，n を増やすにつれて差は減少し，n=300 以降ではいずれのモデルにおいても全体ベンチマークとの差は 0.02 未満となった。特に n=400 では，すべてのモデルで全体評価をほぼ再現できている。

これらの結果から，ベンチマーク構築段階で無作為な並べ替えを行っておけば，大規模なベンチマークを構築した後に無作為抽出を行う必要はなく，構築段階で得られたベンチマークの先頭部分のみを用いることで，十分に安定した評価が可能であることが示唆された。

4 おわりに

本研究では，対話における文脈依存的な言外の意図理解能力を評価するための日本語ベンチマークを構築し，複数の大規模言語モデルを用いて評価実験を行った。本ベンチマークは，同一のコア発話に対して異なる文脈文を与えることで，文脈に応じて変化する言外の意図を明示的に扱える点に特徴がある。

評価実験の結果，いずれのモデルにおいても，文脈依存事例において高い正解率が得られ，LLM が文脈情報を手がかりとして言外の意図をある程度適切に理解できていることが示された。一方で，文脈非依存事例は件数が少なく，今後さらなる検証が必要である。

また，ベンチマーク規模の影響に関する分析から，構築段階で無作為化を行っておけば，ベンチマーク全体を用いなくとも，先頭部分のみを用いた評価で十分に安定した結果が得られることが確認された。この結果は，大規模なベンチマークを構築した後に無作為抽出を行う必要性を低減し，評価コストを抑えた実験設計の可能性を示唆している。

今後の課題としては，より多様な対話場面を含むデータの拡充や，言外の意図の粒度や種類に着目した詳細な分析が挙げられる。また，本研究で構築したベンチマークを用いて，さらなるモデル間比較や学習手法の検証を行うことも重要である。

参考文献

[1] 栗原健太郎, 河原大輔, 柴田知秀. Jglue: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[2] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. **arXiv preprint arXiv:1804.07461**, 2018.

[3] John R. Searle. Indirect speech acts. In **Speech Acts**, pp.

表 2　ベンチマークの各規模での正解率と全体正解率との差

モデル	n=50	n=100	n=200	n=300	n=400	n=500	n=600	全体
GPT-5	0.005	0.015	0.035	0.039	0.020	0.019	0.010	0.000
GPT-4	0.024	0.054	0.009	0.021	0.007	0.006	0.002	0.000
Claude-4.5	0.031	0.031	0.016	0.007	0.001	0.001	0.001	0.000

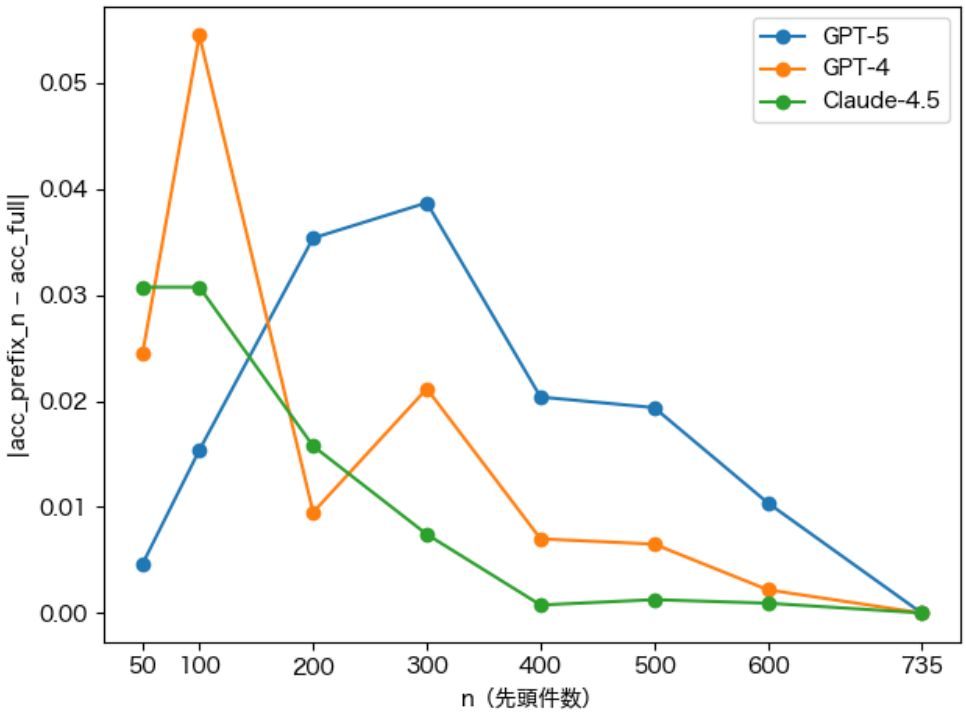

図 1　ベンチマークの各規模での正解率と全体正解率との差

59–82. Brill, 1975.

[4] Louisa Pragst and Stefan Ultes. Changing the level of directness in dialogue using dialogue vector models and recurrent neural networks. In **Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue**, pp. 11–19, 2018.

[5] Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. **arXiv preprint arXiv:2503.13335**, 2025.

[6] 山崎友大, 谷口仁慈, 山際愛実, 原田憲旺, 小島武, 岩澤有祐, 松尾豊. Jamse：日本語 llm 評価用の高品質な少サンプル日本語ベンチマークの作成および評価 — geniac llm 開発コンペティションからの知見—. 言語処理学会第 31 回年次大会 発表論文集, pp. 2481–2485, 2025.

[7] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会第 29 回年次大会 発表論文集, pp. 108–113, 2023.