否定理解能力を評価するための日本語意味的類似度計算データセットの構築

湯浅 令子 [1]　加藤 芳秀 [2]　松原 茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋大学情報連携推進本部
yuasa.reiko.k5@s.mail.nagoya-u.ac.jp

概要

言語モデルの否定理解能力を評価するためのデータセットが構築されている．否定は意味を反転させるため，意味的類似度計算（STS）タスクを解く上でそれを理解することは重要である．STS タスクは大規模言語モデル（LLM）の評価に有用であるが，否定に着目した STS データセットを構築する取り組みは限られている．本研究では，言語モデルの否定理解能力を評価するための日本語意味的類似度計算データセット JSTS-Neg を構築する．また，JSTS-Neg を用いて既存の LLM を評価し，それらの否定理解能力の現状を明らかにする．

1　はじめに

否定とは，事態の不成立を表すことであり [1]，自然言語において重要な言語現象である．言語モデルには否定を意味する表現（**否定要素**；negation cue）を正しく理解することが求められるが，否定に着目したデータセットを用いた調査により，既存の言語モデルは否定の理解が苦手であることが示されている [2, 3, 4]．

意味理解のタスクとして，自然言語推論（NLI）や意味的類似度計算（STS）などがある．否定に着目したデータセットとして，NLI [5, 6]，容認性判断 [7, 8]，質問応答（QA） [4, 9] を対象としたものがある一方で，否定の観点から言語モデルを評価するための STS データセットを構築する取り組みは限られている．否定は意味を反転させるため，STS タスクを解く上で否定の理解は特に重要である．

本研究では，否定理解能力を評価するための日本語 STS データセット **JSTS-Neg** を構築する[1)]．具体的には，JGLUE [10, 11] の一部である JSTS を吉田ら [12] の手法を用いて否定に関して拡張する．これにより，JSTS-Neg は否定の有無のみに関して異なるデータ対（**否定のミニマルペア**）で構成される．否定のミニマルペアを用いることで，否定以外の要因を排除し，言語モデルの否定に対する振る舞いのみを評価することができる．

また本研究では，JSTS-Neg を用いた既存の大規模言語モデル（LLM）の否定理解能力を評価する．日本語及び多言語 LLM を幅広く評価し，各 LLM の否定理解能力の現状を明らかにする．

2　関連研究

2.1　否定に着目したデータセット

表 1 に，否定理解能力を評価するためのデータセットを示す．上段は日本語以外，下段は日本語のデータセットである．Hossain ら [5] や Hartmann ら [6] は，否定に着目した NLI データセットを構築した．RuBLiMP [8] は，否定を含む 12 の言語現象を対象とした容認性判断データセットである．Ravichander ら [9] や García-Ferrero ら [4] は，否定に着目した QA データセットを構築した．日本語においては，否定に着目した NLI データセットである N-JSNLI [13] や JNLI-Neg [12]，否定を含む 39 パラダイムを対象とした容認性判断データセットである JBLiMP [7] がある．しかし，いずれも STS タスクを対象としていない．

2.2　否定に着目した日本語 STS データセット

内田・南條 [14] は，JSTS [10, 11] を否定に関して拡張し，日本語の STS データセット N-JSTS を構築した．しかし，各インスタンスは否定のミニマルペアで構成されておらず，否定理解能力のみを評価することは困難である．また，否定要素の位置は文末のみであり，文の途中の否定要素は含まれない．

1) データセット及びソースコードを https://github.com/reiko-y/JSTS-Neg で公開する．

表 1　否定に着目したデータセット

データセット	タスク	ミニマルペア	否定要素の位置	言語
Hossain ら [5]	NLI	有	文末	EN
Hartmann ら [6]	NLI	有	文末及び文の途中	EN, BG, DE, FR, ZH
RuBLiMP [8]	容認性判断	有	文末及び文の途中	RU
CONDAQA [9]	QA	無	文末及び文の途中	EN
García-Ferrero ら [4]	QA	有（一部）	文末及び文の途中	EN
N-JSNLI [13]	NLI	有（一部）	文末	JA
JNLI-Neg [12]	NLI	有	文末及び文の途中	JA
JBLiMP [7]	容認性判断	有	文末	JA
N-JSTS [14]	STS	無	文末	JA
JSTS-Neg (本研究)	STS	有	文末及び文の途中	JA

3　JSTS-Neg の構築

本研究では，吉田ら [12] の手法に従い，JSTS [10, 11] を否定に関して拡張することにより，否定理解能力を評価するための日本語 STS データセット JSTS-Neg を構築する．本節では，JSTS-Neg の構築手法，及び実際に構築されたデータセットについて説明する．

3.1　データセットの要件

JSTS-Neg の要件は以下である．

1. すべてのインスタンスが，否定のミニマルペアで構成される．
2. 文末及び文の途中の否定要素を含む．
3. 翻訳を介さず，最初から日本語で構築される．

これらの要件は吉田ら [12] に準じている．要件 1 及び 2 は，3.2 節で説明する否定要素を含む文及び STS インスタンスの作成により満たされる．要件 3 は，JSTS [10, 11] を拡張元とすることで満たされる．

3.2　構築手法

JSTS-Neg は，以下の 2 段階で構築される．データ作成のフローを図 1 に示す．

1. 否定要素の挿入
   (a) 文に否定要素をルールベースで挿入する．
   (b) 否定要素が挿入された文の正しさを判定し，正しいと判定された文のみを残す．
2. 否定を含む新たな STS インスタンスの作成
   (a) 文ペアの一方あるいは両方の文に否定要素を挿入することで，否定要素を含む文ペアを作成する．

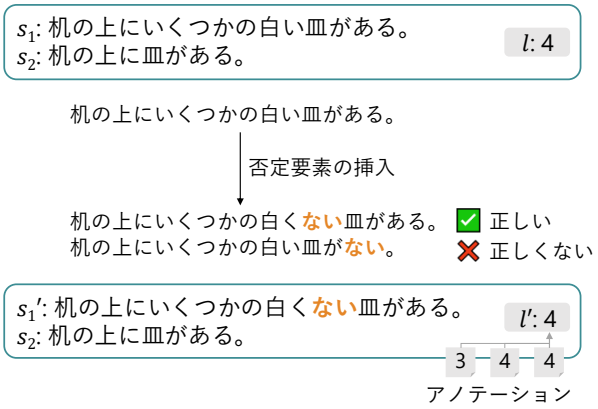

図 1　データ作成のフロー

   (b) 文ペアに対して類似度を付与し正解の類似度を決定する．

以下では，STS インスタンス，すなわち，文 s_1 と s_2 の類似度が l であることを 3 項組 (s_1, s_2, l) で表現する．

3.2.1　否定要素の挿入

否定要素の挿入は，吉田ら [12] の手法に従う．本節では，吉田らの手法についてその概略を説明する．この手法ではまず，形態素情報を考慮したルールに基づき文に否定要素を挿入する[2)]．挿入する否定要素は「ない」あるいは「ず」のいずれか 1 つである．また，一文の中に否定要素を挿入できる箇所が複数ある場合は，その数だけ文を複製し，それぞれの箇所に否定要素を挿入する．すなわち，否定要素を挿入できる箇所が一文の中に n 箇所ある場合，否定要素を含む文が n 文生成される．この方法により生成される文と元の文の対は，否定のミニマルペアの要件を満たすことが保証される．

2)　動詞，形容詞，形容動詞のいずれかが対象である．

次に，否定要素が挿入された文の正しさを自動で判定する．ルールベースによる否定要素の挿入では品詞の情報しか考慮されておらず，生成された文は日本語として正しくない場合がある．例えば，以下の文 1 から生成される文 2 は，意味が矛盾しており，日本語として正しくない．

1. 群衆がいて混雑する．
2. 群衆がいて混雑し**ない**．

この問題に対し，吉田ら [12] の手法では LLM の in-context learning を用いて文の正しさを判定する[3]．否定要素が挿入される前の文は正しいという例を与えた上で[4]，否定要素が挿入された文が正しいか否かを 2 値分類している（1-shot）．その後，正しいと判定された文のみを残し，正しくないと判定された文をデータセットから除外する．

3.2.2 否定を含む STS インスタンスの作成

STS インスタンス $i = (s_1, s_2, l)$ に対し，否定を含むインスタンスを作成する．s_1，s_2 に否定要素を挿入することで得られる文の集合をそれぞれ S'_1，S'_2 とする．このとき，$|S'_1| > 0$ かつ $|S'_2| > 0$ かつ $neg(s_1) = 0$ かつ $neg(s_2) = 0$ を満たすインスタンス i から，否定を含むインスタンスの集合 $D_{\text{neg}}(i)$ を作成する．ここで，$neg(s)$ は文 s に含まれる否定要素の数を表す[5]．$D_{\text{neg}}(i)$ の定義は以下である．

$$
\begin{aligned}
D_{\text{neg}}(i) &= D_1(i) \cup D_2(i) \cup D_{1,2}(i)\ , \\
D_1(i) &= \{(s'_1, s_2, l') | s'_1 \in S'_1\}\ , \\
D_2(i) &= \{(s_1, s'_2, l') | s'_2 \in S'_2\}\ , \\
D_{1,2}(i) &= \{(s'_1, s'_2, l') | s'_1 \in S'_1 \wedge s'_2 \in S'_2\}\ .
\end{aligned}
$$

3 人の作業者が JSTS [10, 11] のガイドラインに基づいて各文ペアに対して類似度を付与し，それらの中央値を正解の類似度 l' とする．

3.3 データセット構築

JSTS [10, 11] の学習セット，検証セット，評価セットのインスタンスをランダムに並び替え，否定を含むインスタンスがそれぞれ 4,000，1,000，1,000 を超えるまで 3.2 節で説明した手法を順に適用し，新たなインスタンスを作成する．また，元のインスタンスに対しても同じ作業者が類似度を付与し[6]，それらの中央値を正解の類似度とする．

JSTS-Neg を構成するインスタンスの集合 $D_{\text{JSTS-Neg}}$ は，以下で定義される．

$$
\begin{aligned}
D_{\text{JSTS-Neg}} &= D_{\text{orig}} \cup D_{\text{neg}}\ , \\
D_{\text{neg}} &= \bigcup_{i \in D_{\text{orig}}} D_{\text{neg}}(i)\ .
\end{aligned}
$$

D_{orig} はサンプリングした JSTS インスタンスの集合を表す．JSTS-Neg は，以下で定義する否定のミニマルペアの集合 M で構成される．

$$
\begin{aligned}
M &= M_{\text{single}} \cup M_{\text{both}}\ , \\
M_{\text{single}} &= \{(i, i') | i \in D_{\text{orig}} \wedge i' \in D_1(i) \cup D_2(i)\}\ , \\
M_{\text{both}} &= \{(i', i'') | \exists i \in D_{\text{orig}}((i' \in D_1(i) \wedge i'' \in D_2(i')) \\
&\qquad \vee (i' \in D_2(i) \wedge i'' \in D_1(i')))\}\ .
\end{aligned}
$$

以下では，ミニマルペア $m = (i, i') \in M$ に対し，i，i' をそれぞれ**対照インスタンス**，**処置インスタンス**と呼ぶ．処置インスタンスは，対照インスタンスの一方の文に否定要素が 1 つ挿入されたものである．

M は，Hossain ら [2] に従い，以下のように「重要」な否定のミニマルペアの集合 M_{i} と「重要」でない否定のミニマルペアの集合 M_{u} に分割される．

$$
\begin{aligned}
M_{\text{i}} &= \{((s_1, s_2, l), (s'_1, s'_2, l')) \in M | l \neq l'\}\ , \\
M_{\text{u}} &= \{((s_1, s_2, l), (s'_1, s'_2, l')) \in M | l = l'\}\ .
\end{aligned}
$$

4 実験

4.1 実験設定

本実験では，8 つの日本語 LLM 及び日本語に対応している 18 の多言語 LLM（OpenAI API で提供されている 6 つの GPT モデルを含む）を評価対象とした．本稿では，紙面の都合上，同種のモデルについては最も新しくサイズの大きいモデルの値のみを報告する[7]．表 2 の各設定において，上から 4 行が日本語 LLM，それ以降が多言語 LLM である．

STS タスクを 6 値分類タスクとして扱い，Zero-shot，4-shot，11-shot で実験を行った．Zero-shot では，プロンプトとしてタスク指示のみを与えた[8]．

3） モデルは，OpenAI API（https://openai.com/index/openai-api/）で提供されている gpt-4.1-2025-04-14 を用いる．
4） 否定要素の挿入による影響のみを考慮して文の正しさを判定するためである [12]．
5） 否定要素検出器 [15] を用いて $neg(s)$ を求める．
6） STS タスクにおいて類似度は順序尺度であり間隔尺度ではないが，JSTS [10, 11] では正解の類似度がアノテーションの平均値となっている．STS タスクにおいて類似度が順序尺度であることを考慮し，元のインスタンスに対しても新たに類似度を付与する．
7） 使用したすべてのモデルは，付録 A を参照されたい．
8） llm-jp-eval [21] のものを基本とし，人手アノテーションと条件を揃えるために，タスク指示の部分のみ JSTS [10, 11] のガイドラインを使用した．

表 2　否定のミニマルペア単位の実験結果

Setting	Model	M			M_{i}			M_{u}		
		Acc	Acc′	AccChg*	Acc	Acc′	AccChg*	Acc	Acc′	AccChg*
Zero-shot	llm-jp-3.1-13B-instruct4 [16]	47.22	48.04	0.82	29.49	38.34	8.85	53.32	51.38	**-1.94**
	Llama 3.1 Swallow 8B Instruct v0.5 [17, 18, 19]	30.89	37.61	6.73	43.16	52.28	9.12	26.66	32.56	5.90
	Gemma-2-Llama Swallow 9B IT v0.1 [17, 18, 19]	46.47	54.56	8.10	63.81	64.08	0.27	40.50	51.29	10.79
	Swallow-MS 7B instruct v0.1 [17, 18]	55.18	49.62	**-5.56**	58.71	37.00	**-21.72**	53.97	53.97	0.00
	Llama 3.1 8B Instruct	45.37	50.79	5.42	56.84	58.98	2.14	41.42	47.97	6.55
	Ministral 8B Instruct	58.54	47.01	**-11.53**	35.39	18.77	**-16.62**	66.51	56.73	**-9.78**
	Gemma 3n E4B instruct [20]	36.51	40.01	3.50	46.92	46.92	0.00	32.93	37.64	4.70
	GPT-5	54.08	62.94	8.85	62.47	54.69	**-7.77**	51.20	65.77	14.58
11-shot**	llm-jp-3.1-13B-instruct4 [16]	54.29	54.93	0.65	43.06	45.52	2.47	58.15	58.17	0.02
	Llama 3.1 Swallow 8B Instruct v0.5 [17, 18, 19]	47.43	48.84	1.41	36.89	33.03	**-3.86**	51.05	54.28	3.23
	Gemma-2-Llama Swallow 9B IT v0.1 [17, 18, 19]	51.42	55.88	4.46	50.29	46.11	**-4.18**	51.81	59.24	7.44
	Swallow-MS 7B instruct v0.1 [17, 18]	37.25	35.70	**-1.55**	25.90	18.87	**-7.02**	41.16	41.49	0.33
	Llama 3.1 8B Instruct	35.87	32.81	**-3.06**	17.27	14.85	**-2.41**	42.27	38.99	**-3.28**
	Ministral 8B Instruct	42.80	44.32	1.52	32.06	29.44	**-2.63**	46.49	49.45	2.95
	Gemma 3n E4B instruct	34.82	34.98	0.15	27.56	26.06	**-1.50**	37.32	38.04	0.72
	GPT-5	56.12	64.34	8.22	59.84	57.96	**-1.88**	54.83	66.53	11.70

Acc 及び Acc′ の単位は%，AccChg のそれはポイントである．
* **太字**は負の値を示す．
** 例の抽出におけるシード値を変えた 5 回の試行の平均値を示す．なお，1 回の試行における例は全インスタンスで共通である．

4-shot では，タスク指示に加えて D_{orig}，D_{neg} の学習セットから 2 インスタンスずつランダムに選び例として与えた．11-shot では，タスク指示に加えて D_{orig} の学習セットから 5 インスタンス，D_{neg} の学習セットから 6 インスタンスをランダムに選び例として与えた．ただし，正解の類似度が重複しないようにサンプリングした[9]．

4.2　評価指標

STS タスクを 6 値分類タスクとして解くため，評価指標として正解率（accuracy）を用いる．ミニマルペア中のインスタンスを比較するため，以下で定義される accuracy change（AccChg）を用いて否定に対するモデルの振る舞いを評価した．

$$\mathrm{AccChg} = \frac{1}{|M|} \sum_{((s_1,s_2,l),(s_1',s_2',l'))\in M} (\mathbf{1}[\hat{l} = l] - \mathbf{1}[\hat{l}' = l']) ,$$

$$\mathrm{Acc} = \frac{1}{|M|} \sum_{((s_1,s_2,l),(s_1',s_2',l'))\in M} \mathbf{1}[\hat{l} = l] ,$$

$$\mathrm{Acc}' = \frac{1}{|M|} \sum_{((s_1,s_2,l),(s_1',s_2',l'))\in M} \mathbf{1}[\hat{l}' = l'] .$$

ここで，$\hat{l}$，$\hat{l}'$ はそれぞれ STS インスタンス (s_1, s_2, l)，(s_1', s_2', l') に対してモデルが予測した類似度である．AccChg の値は −1 から 1 の範囲を取る．0 に近いほど否定の有無による性能変化が小さいことを意味し，−1 に近いほど対照インスタンスより処置インスタンスに対する正解率の方が低いことを意味する．

4.3　実験結果

Zero-shot 及び 11-shot の実験結果を表 2 に示す[10]．Zero-shot では，モデルによって AccChg の値がばらつく結果となった．11-shot では，M_{i} に対する AccChg が負である，すなわち否定の影響による正解の類似度の変化を捉えられていない一方，M_{u} に対する AccChg は正であるモデルが多い傾向がみられた．例を多く与えられたモデルは否定を無視して類似度を予測している可能性が示唆される．

5　おわりに

本研究では，否定の理解能力を評価するための新たな日本語 STS データセット JSTS-Neg を構築した．また，JSTS-Neg を用いて幅広い LLM の否定理解能力を評価した．今後の課題として，JSTS-Neg を用いて否定に対するモデルの振る舞いをより詳細に分析することが挙げられる．また，本研究で対象外とした，接辞の否定や二重否定についても考慮しデータセットを拡張することも今後の課題である．

9）　正解の類似度が 0，1，2，3，4，5 である 6 インスタンスずつの合計 12-shot でないのは，D_{orig} の学習セットに正解の類似度が 5 であるインスタンスが存在しなかったためである．

10）　4-shot の実験結果は付録 B を参照されたい．

謝辞

本研究は JSPS 科研費 JP23K18506 の助成を受けたものです．本研究で利用した JSTS データセット及びそのアノテーションガイドラインを提供いただいた栗原健太郎氏，河原大輔氏，柴田知秀氏に感謝いたします．実験の一部は，名古屋大学のスーパーコンピュータ「不老」を利用して実施しました．

参考文献

[1] 日本語記述文法研究会. 現代日本語文法 3. くろしお出版, 2012.

[2] Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. An analysis of negation in natural language understanding corpora. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Short Papers)**, Vol. 2, pp. 716–723, 2022.

[3] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In **Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)**, pp. 101–114, 2023.

[4] Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. This is not a dataset: A large negation benchmark to challenge large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8596–8615, 2023.

[5] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9106–9118, 2020.

[6] Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. A multilingual benchmark for probing negation-awareness with minimal pairs. In **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 244–257, 2021.

[7] Taiga Someya and Yohei Oseki. JBLiMP: Japanese benchmark of linguistic minimal pairs. In **Findings of the Association for Computational Linguistics (EACL 2023)**, pp. 1581–1594, 2023.

[8] Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. RuBLiMP: Russian benchmark of linguistic minimal pairs. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 9268–9299, 2024.

[9] Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 8729–8755, 2022.

[10] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022)**, pp. 2957–2966, 2022.

[11] 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.

[12] 吉田朝飛, 加藤芳秀, 小川泰弘, 松原茂樹. 否定理解能力を評価するための日本語言語推論データセット JNLI-Neg. 情報処理学会論文誌, Vol. 66, pp. 1853–1860, 2025.

[13] 内田巧, 南條浩輝. 否定表現を伴う文における含意関係認識のための対偶によるデータ拡張. Technical report, 情報処理学会, 2023.

[14] 内田巧, 南條浩輝. 否定表現を伴う文における自然言語理解の性能検証. 言語処理学会第 30 回年次大会発表論文集, pp. 1581–1586, 2024.

[15] 湯浅令子, 吉田朝飛, 加藤芳秀, 松原茂樹. 否定の観点からみた日本語言語理解ベンチマークの評価. 言語処理学会第 31 回年次大会発表論文集, pp. 424–429, 2025.

[16] LLM-jp (2024). LLM-jp: A cross-organizational project for the research and development of fully open Japanese LLMs. **Computing Research Repository (CoRR)**, Vol. abs/2407.03963, , 2024.

[17] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. In **Proceedings of the 1st Conference on Language Modeling (COLM 2024)**, 2024.

[18] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large Japanese web corpus for large language models. In **Proceedings of the 1st Conference on Language Modeling (COLM 2024)**, 2024.

[19] Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. Building instruction-tuning datasets from human-written instructions with open-weight large language models, 2025.

[20] Gemma Team. Gemma 3n. 2025.

[21] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会発表論文集, pp. 2085–2089, 2024.

[22] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 13898–13905, 2024.

[23] AI@Meta. Llama 3 model card. 2024.

[24] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. **Computing Research Repository (CoRR)**, Vol. abs/2310.06825, , 2023.

[25] Gemma Team and Google DeepMind. Gemma 3 technical report. **Computing Research Repository (CoRR)**, Vol. abs/2503.19786, , 2025.

[26] Gemma Team and Google DeepMind. Gemma 2: Improving open language models at a practical size. **Computing Research Repository (CoRR)**, Vol. abs/2408.00118, , 2024.

表 3　使用したオープンモデル

分類	モデル	Hugging Face 上の名称
日本語	llm-jp-3.1-1.8B-instruct4 [16]	llm-jp/llm-jp-3.1-1.8b-instruct4
	llm-jp-3.1-13B-instruct4 [16]	llm-jp/llm-jp-3.1-13b-instruct4
	Llama 3.1 Swallow 8B Instruct v0.5 [17, 18]	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
	Gemma-2-Llama Swallow 2B IT v0.1 [17, 18]	tokyotech-llm/Gemma-2-Llama-Swallow-2b-it-v0.1
	Gemma-2-Llama Swallow 9B IT v0.1 [17, 18]	tokyotech-llm/Gemma-2-Llama-Swallow-9b-it-v0.1
	Swallow-MS 7B instruct v0.1 [17, 18]	tokyotech-llm/Swallow-MS-7b-v0.1
	Llama 3 Youko 8B Instruct [22]	rinna/llama-3-youko-8b-instruct
	Gemma 2 Baku 2B Instruct [22]	rinna/gemma-2-baku-2b-it
多言語	Llama 3.1 8B Instruct	meta-llama/Llama-3.1-8B-Instruct
	Llama 3 8B Instruct [23]	meta-llama/Meta-Llama-3-8B-Instruct
	Mistral 7B Instruct v0.3 [24]	mistralai/Mistral-7B-Instruct-v0.3
	Ministral 8B Instruct	mistralai/Ministral-8B-Instruct-2410
	Mistral Nemo Instruct	mistralai/Mistral-Nemo-Instruct-2407
	Gemma 3n E2B instruct [20]	google/gemma-3n-E2B-it
	Gemma 3n E4B instruct [20]	google/gemma-3n-E4B-it
	Gemma 3 270M instruct [25]	google/gemma-3-270m-it
	Gemma 3 1B instruct [25]	google/gemma-3-1b-it
	Gemma 3 4B instruct [25]	google/gemma-3-4b-it
	Gemma 2 2B Instruct [26]	google/gemma-2-2b-it
	Gemma 2 9B Instruct [26]	google/gemma-2-9b-it

表 4　否定のミニマルペア単位の実験結果（4-shot）

Model	M			M_{i}			M_{u}		
	Acc	Acc′	AccChg*	Acc	Acc′	AccChg*	Acc	Acc′	AccChg*
llm-jp-3.1-13B-instruct4 [16]	47.59	46.07	**-1.52**	27.35	29.60	2.25	54.56	51.73	**-2.82**
Llama 3.1 Swallow 8B Instruct v0.5 [17, 18]	31.93	31.43	**-0.49**	23.32	25.79	2.47	34.89	33.38	**-1.51**
Gemma-2-Llama Swallow 9B IT v0.1 [17, 18]	37.30	40.56	3.27	30.94	34.26	3.32	39.48	42.73	3.25
Swallow-MS 7B instruct v0.1 [17, 18]	32.13	30.87	**-1.26**	23.22	19.95	**-3.27**	35.20	34.63	**-0.57**
Llama 3.1 8B Instruct	28.51	26.84	**-1.67**	17.91	18.02	0.11	32.16	29.87	**-2.29**
Ministral 8B Instruct	41.22	40.85	**-0.37**	27.67	24.72	**-2.95**	45.89	46.40	0.52
Gemma 3n E4B instruct [20]	37.68	38.52	0.84	30.67	35.66	4.99	40.09	39.50	**-0.59**
GPT-5	54.14	61.88	7.74	58.61	53.24	**-5.36**	52.60	64.85	12.25

Acc 及び Acc′ の単位は%，AccChg のそれはポイントである．
例の抽出におけるシード値を変えた 5 回の試行の平均値を示す．なお，1 回の試行における例は全インスタンスで共通である．
* **太字**は負の値を示す．

A　使用モデル

A.1　オープンモデル

オープンモデルとして，Hugging Face（https://huggingface.co/）で公開されている日本語及び多言語 LLM を使用した．詳細を表 3 に示す．

A.2　クローズド GPT モデル

クローズド GPT モデルとして，OpenAI API（https://openai.com/index/openai-api/）で公開されている `gpt-5-nano-2025-08-07`，`gpt-5-mini-2025-08-07`，`gpt-5-2025-08-07`，`gpt-4.1-nano-2025-04-14`，`gpt-4.1-mini-2025-04-14`，`gpt-4.1-2025-04-14` を使用した．

B　4-shot の実験結果

4-shot の実験結果を表 4 に示す．