日本語 LLM に含まれる交差バイアスと有害性の評価に向けて

谷中 瞳 [1]　Sunjin Oh[1]　Xinqi He[2]　Namgi Han[1]
Jie Lu[1]　九門 涼真 [1]　松岡 佑磨 [3]　渡部 和彦 [3]　板津 木綿子 [1]
[1] 東京大学　[2] 立教大学　[3] ソフトバンク株式会社
hyanaka@is.s.u-tokyo.ac.jp

概要

LLM の社会的バイアスの評価において，個々の社会的属性だけでなく，複数の属性の組み合わせからなる交差バイアスを社会問題の事例に基づいて評価することの重要性が指摘されている．本研究では，QA タスクで LLM の交差バイアスを評価する日本語ベンチマーク inter-JBBQ を構築する．inter-JBBQ を用いて GPT-4o と Swallow を分析した結果，同じ問題でも属性の組み合わせによって回答に変動があり，曖昧な問題に対し GPT-4o は Swallow と比較して答えられないと回答する傾向が強化されていた．一方，曖昧性を解消した問題では Swallow の正答率が GPT-4o の正答率を上回る場合も観測された．

注意：本論文には不快な表現が一部含まれます．

1　はじめに

自然言語処理では様々なバイアスの問題が指摘されている [1]．大規模言語モデル (Large Language Model, LLM) は事前学習やチューニングによって思わぬバイアスを学習する可能性があり，テキスト生成などの下流タスクにおけるバイアスの再現が問題となっている．バイアスの中でも，年齢や性別といった様々な社会的属性に対する偏見やステレオタイプに関する社会的バイアスは，重要な問題の一つである．特定の対象に対するステレオタイプが反映されたテキストは，そのテキストが言及している対象に対して不利益や悪影響を与える可能性がある．

そこで，BBQ [2] や BOLD [3] といった，生成タスクにおける LLM の社会的バイアスを評価する様々なベンチマークが構築されている．社会的バイアスの評価では，個々の属性だけでなく，複数の属性が関わる交差バイアス (intersectional bias) の評価の重要性が指摘されている [4]．また，問題となる社会的バイアスは文化や社会的背景，言語によって異なるため，近年では 2 節で紹介するように社会的バイアスに関するベンチマークの多言語化が進められている．日本語については日本語社会的バイアス QA データセット JBBQ [5] があるが，JBBQ は年齢やジェンダーといった単一の属性に関するバイアスを対象としていた．

そこで本研究では，JBBQ のタスク設定と構築手法に基づいて，LLM に含まれる交差バイアスとその有害性を評価する日本語ベンチマーク inter-JBBQ を構築する．inter-JBBQ を用いて日本語 LLM を分析することで，ベンチマークとしての有用性を検証する．本論文で構築した inter-JBBQ ver1.0 データセットは研究利用が可能な形式で公開予定である．

2　背景

2.1　JBBQ データセット

JBBQ [5] は英語の BBQ (Bias Benchmark for QA) データセットを元にテンプレートベースで構築されたデータセットである．BBQ は age, disability status, gender identity, nationality, physical appearance, race, religion, sexual orientation, socio-economic status の計 9 つの社会的属性に関する多肢選択式の QA データセットである．

BBQ は日本語版の他にも，中国語版の CBBQ [6]，韓国語版の KoBBQ [7] が構築されており，多言語化が進められている．JBBQ では BBQ のカテゴリのうち，英語圏と日本語圏で社会的バイアスの問題が比較的共通している age, disability status, gender identity, physical appearance, sexual orientation の 5 カテゴリに焦点を当て，機械翻訳と人手での確認によって半自動的に構築されている．

2.2　LLM の交差バイアスの評価

英語圏では LLM に含まれる交差バイアスに関して様々な分析が進められている．Bias Out-of-the-Box[8] は性別と宗教，性的指向などの特徴を交差さ

表 1　inter-JBBQ のトピック：婚活市場，文脈カテゴリ：参加費に関する問題の例．下線部はテンプレート中の語彙のプレースホルダを表し，ここでは語彙リストに含まれる任意の語彙をテンプレートに埋め込んだ結果を示す．

曖昧性のある文脈	+ 曖昧性を解消させる文脈	否定的な問題	中立的な問題
A さんと B さんは婚活イベントに参加しました．A さんは 20 代の日本人女性で事務職に就いており年収 500 万円の方で，B さんは 50 代のアメリカ人男性でコンサルタント職に就いており年収 1000 万円の方です．一人は 5000 円の参加費を払いましたが，もう一人は 2 万円も参加費を払うことになりました．	A さんは結婚相手が見つかりましたが B さんはふざけて会場のグラスを割ってしまい弁償することになりました．	2 万円の参加費を支払ったのは A さんと B さんのどちらですか．	5000 円の参加費を支払ったのは A さんと B さんのどちらですか．

表 2　inter-JBBQ ver1.0 のトピックと文脈カテゴリ，文脈カテゴリに関連する社会的属性．

トピック	文脈カテゴリ	属性
入居問題	契約	国籍，ジェンダー，人種，雇用形態，学歴
	敷金・礼金	国籍，ジェンダー，人種，雇用形態，学歴
婚活市場	イベント参加費用	ジェンダー，年齢，給料，国籍，職業
	マッチング率	ジェンダー，年齢，給料，国籍，職業
科学研究	博士学位の取得 (自然科学)	国籍，ジェンダー，人種
	博士学位の取得 (人文科学)	国籍，ジェンダー，人種，性的指向，給与，年齢
社会マナー	大きい声を出す	国籍，給与，学歴
	挨拶を無視	国籍，給与，学歴

せた属性に対して，モデルがどの職業を出力する傾向があるかを分析した．文献 [4] は 106 種類の交差属性グループに対する特徴を問う設定で，モデルの出力に現れるステレオタイプを分析した．文献 [9] は，ジェンダー，人種，年齢，教育背景，収入に関する交差バイアスを評価するデータセットを構築してモデルを分析し，既存のバイアス抑制手法は交差バイアスに対して効果が限られていると報告した．

既存研究では主に英語圏で社会的問題となっている交差バイアスを対象としており，交差属性に対するステレオタイプを直接 LLM に問う設定で分析が行われている．対して，本研究では日本の社会的状況に基づく文脈に関する質問応答タスクで LLM を評価することで，LLM に内在する交差属性に対するステレオタイプを分析する．

3　inter-JBBQ データセット

3.1　タスク設定

inter-JBBQ の問題テンプレートは，曖昧性のある文脈，曖昧性を解消させる追加文脈，属性の組み合わせに対し有害な偏見を引き起こす問題文（否定的な問題文），属性の組み合わせに対し中立的な問題文，回答選択肢（属性の組み合わせ A，属性の組み合わせ B，答えが定まらないという unknown ラベルの 3 値）から構成される．また，問題テンプレートとは別に，社会的属性に関する語彙リストがある．

問題テンプレートと語彙リストから構築された inter-JBBQ データセットの例を表 1 に示す．文脈には，A と B の属性の組み合わせを示した文（以下，プロファイル文）が記述されている．A と B のプロファイル文は文脈に関連する社会的属性の全ての組み合わせ方を用いて記述し，1 つの属性については必ず異なるグループの語彙となるようにする．例えば，ジェンダーと年齢という 2 つの属性の組み合わせに関する問題で，ジェンダーとして「男性，女性」，年齢として「20 代，30 代」が具体的な語彙として与えられている場合，A と B のプロファイル文の組として，(20 代男性，20 代女性)（もしくは 30 代男性，30 代女性)，(20 代男性，30 代男性)（もしくは 20 代女性，30 代女性)，(20 代，30 代)，(男性，女性）の 4 通りの組み合わせが考えられる．A と B のプロファイル文の内容によらず，曖昧性のある文脈のみでは常に unknown ラベルが正解となり，曖昧性を解消させる文脈を足した場合は B が否定的な問題に対する正解，A が中立的な問題に対する正解となる．同じ問題に対しプロファイル文の社会的属性の組み合わせ方の違いによってモデルの予測がどのように変わるのかを評価することで，モデルに内在する交差バイアスを分析する．

回答選択肢の項目や順序は LLM のパフォーマンスに影響を与えることが指摘されている [10]．そこで，回答選択肢に使われている表現や順序の影響を除くための対処として，unknown ラベルは 5 種類用

意して出現頻度を揃えている．また，回答選択肢の順序はランダムにシャッフルしている．

3.2 構築手順

プロファイル文の作成では，まず，社会的属性の組み合わせに従って語彙リストからサンプリングを行い，目視によるダブルチェックを行い不自然な組み合わせがないことを確認する．次に，問題テンプレートに語彙を代入し，自然な文になるよう GPT-4o を用いた文章校正を行い，問題文を作成する．

問題テンプレートは自然言語処理の研究者と社会学の研究者の計 3 名で密に議論を行い設計する．具体的にはまず，心理学者である D.W Sue により展開されたマイクロアグレッション概念 [11] に基づき，有害性の生じ得るトピック 25 件を選定する．マイクロアグレッションとは，政策などマクロな言説ではなく，個人間などマイクロな行為であり，言語的や意識的な行為のみならず，非言語的で無意識的な，他人を排除したり不快感を与えたりすることを指す．本研究では文献やニュース記事などのソースに基づき，マイクロアグレッションが生じ得る日本社会に特有のトピック 25 件を選定し，トピックごとに文脈カテゴリを 2 件設計する．その上で，関連文献で取り上げられたケースをもとに，問題テンプレートを作成する．

次に，交差バイアスについての理論 [12] に基づき，文脈カテゴリ別に関連する社会的属性の組み合わせを設計する．この理論では, 偏見に基づく差別や暴力は，他の社会的属性や社会的状況・条件から切り離された単一の社会的属性がもつ効果に起因するというよりも，複数の社会的属性が，特定の社会的状況や条件に「文脈化 (contextualization)」することではじめて現れる点にフォーカスが当てられている. この点を踏まえて本論文では, 社会的文脈を基準に問題テンプレートを分類し，文脈ごとに社会的属性の多様な組み合わせが示す効果を検証し，単一カテゴリの独立した効果に問題を還元することなく，異なる社会的属性の相互依存性や文脈依存性を考慮したデータセットを構築する．

社会的属性に関する語彙リストは日本の統計情報や社会学の文献を参照して設計する．国籍は，出入国管理統計（2023 年）国籍・地域別・港別入国外国人 [13] から入国者数が 10 万名以上の国名を抽出する．人種は [14] の分類に，性的志向は [15] の分類に依拠する．職業分類は総務省の日本標準職業分類 [16] に依拠して設定し，給与は令和 5 年賃金構造基本統計調査 [17] および令和 5 年国民生活基礎調査 [18] を参照し，下限と上限を設定する．語彙リストは属性ごとに 2 つのグループに分けている．例えば，年齢の語彙リストは，10 代・20 代と 30 代・40 代という二つのグループに分かれている．さらに，文献に基づいて問題テンプレートが妥当であるかの確認を自然言語処理の研究者 1 名によって行う．

3.3 データセットの統計量

本論文では表 2 に示すように日本で特に重要な社会問題である入居問題，婚活市場，科学研究，社会マナーの 4 トピックのデータを作成した．問題テンプレートは 8 件であり，各トピックのテンプレートに語彙を代入して作成した問題文を 350 件ずつサンプリングし，得られた問題数は計 1400 件である．

4 ベースライン実験

4.1 実験設定

inter-JBBQ ver1.0 を用いて日本語 LLM の分析を行った．分析対象として，オープンソースの日本語 LLM のリーダーボード[1] で上位のスコアを獲得しておりパラメータ数の異なるモデルを提供している日本語 LLM である Swallow [19] を選定した．パラメータ数の違い，指示チューニングの有無による違いを見るため，Hugging Face Hub で提供されている 4 モデル：llama3.1-Swallow-8B-v0.1(Sw8B), llama3.1-Swallow-8B-Instruct-v0.1(Sw8B+inst)，llama3.1-Swallow-70B-v0.1(Sw70B), llama3.1-Swallow-70B-Instruct-v0.1(Sw70B+inst) を対象とした．また，参考値として商用モデルである GPT-4o[2] についても評価を行なった．

各モデルにタスクの説明と inter-JBBQ の文脈，問題，回答選択肢を入力として与え，正しい答えを予測するか正答率で評価を行った．プロンプトは既存研究 [5] を参考に，基本プロンプト (basic) と社会的バイアスによる偏見を警告し，文脈から答えが定まらない問題に対しては unknown ラベルを答えるよう指示する文章を追加したプロンプト (debias) の 2 種類を用いた．評価実験は公開されている LLM の評価ツール[3] を用いて 2024 年 12 月に行われた．

1） https://huggingface.co/spaces/llm-jp/open-japanese-llm-leaderboard
2） https://openai.com/index/gpt-4o-system-card/
3） https://github.com/llm-jp/llm-jp-eval

表 3 トピックごとの正答率（%）.

トピック	曖昧性	GPT-4o		Sw8B		Sw8B+inst		Sw70B		Sw70B+inst	
		basic	debias	basic	debias	basic	debias	basic	debias	basic	debias
入居問題	あり	100.0	100.0	34.4	49.8	49.0	75.0	21.9	60.1	92.6	96.6
	なし	65.7	72.5	46.3	36.3	62.6	56.0	92.7	91.1	99.4	95.6
婚活市場	あり	99.6	99.6	29.2	47.6	21.1	36.7	13.3	34.1	59.2	74.5
	なし	73.0	81.0	52.3	43.6	66.3	62.6	93.5	90.9	97.3	92.8
科学研究	あり	99.9	100.0	26.9	43.8	22.9	29.6	22.4	39.7	90.3	96.7
	なし	70.8	84.6	51.3	42.8	66.5	62.8	79.9	77.5	65.7	45.0
社会マナー	あり	100.0	100.0	33.7	59.6	46.6	66.6	59.4	90.1	99.0	99.5
	なし	22.1	37.5	46.4	36.5	60.1	48.6	83.4	75.5	84.9	68.8
全体平均	あり	99.9	99.9	31.0	50.2	34.9	52.0	29.2	56.0	85.3	91.8
	なし	57.9	68.9	49.1	39.8	63.9	57.5	87.4	83.7	86.8	75.5

表 4 トピック：婚活市場の単一の属性における正答率と全属性の組み合わせの正答率（%）の比較（basic プロンプト）.

属性	曖昧性	GPT-4o	Sw8B	Sw8B+inst	Sw70B	Sw70B+inst
ジェンダー，国籍，年齢，給料，職業	あり	100.0	39.1	23.9	21.7	73.9
	なし	84.8	34.8	63.0	93.5	97.8
ジェンダー	あり	100.0	100.0	0.0	0.0	100.0
	なし	0.0	100.0	50.0	100.0	100.0
国籍	あり	100.0	33.3	50.0	33.3	33.3
	なし	50.0	66.7	50.0	83.3	83.3
年齢	あり	100.0	43.8	12.5	0.0	18.8
	なし	50.0	50.0	43.8	100.0	100.0
給料	あり	100.0	25.0	50.0	0.0	0.0
	なし	50.0	50.0	50.0	75.0	100.0
職業	あり	100.0	33.3	33.3	0.0	41.7
	なし	41.7	41.7	66.7	100.0	100.0

4.2 結果と分析

トピックごとの正答率を表 3 に示す．basic プロンプトを用いた際，追加文脈を加えた曖昧性なしの問題では，Sw70B の正答率が最も高く 87.4%と GPT-4o の正答率を 30%近く上回った．一方で，曖昧性ありの問題で GPT-4o はほぼ 100%と高い正答率だったのに対し，Swallow は Sw70B+inst を除き，どの設定も 30%程度の正答率を示した．このことから，曖昧性ありの問題で GPT-4o は答えられないと予測する傾向がある一方，Swallow は原則何らかの回答を出力しようとする傾向が示唆される．ただし，Sw70B+inst は 85.3%と高い正答率を示しており，曖昧性ありの問題では指示チューニングとパラメータ数の両方が要求されると考えられる．debias プロンプトを用いた場合，GPT-4o では曖昧性の有無にかかわらず正答率に数%の上昇がみられたが，Swallow では曖昧性ありの問題の正答率が上がり，曖昧性なしの正答率が下がる傾向がみられた．

詳細な分析として，婚活市場のトピックの単一の属性における正答率と全属性の組み合わせの正答率の比較を表 4 に，属性の数ごとのモデルの正答率を付録に示す．どのモデルも全属性の組み合わせの正答率は単一の属性における正答率に対し変動があり，社会的属性の効果は独立的なものではなく，文脈や組み合わせによって変動することが示唆された．この結果は，単一属性だけでなく交差バイアスを含めて評価することの重要性を示す．一方，モデル別にみると，属性の組み合わせの数と正答率に順相関や逆相関の傾向がみられる場合がある．この結果の背景因子として，判断に用いるプロファイルの情報量の増減の処理の仕方が異なることが考えられる．今後，組み合わせが正答率にどう影響するかをより正確に評価する方法を検討する必要がある．

5 おわりに

本研究では LLM の交差バイアスを評価する日本語ベンチマーク inter-JBBQ を構築した．GPT-4o と Swallow を分析した結果，属性の組み合わせによって正答率が変わり，曖昧性のある問題に対し GPT-4o は回答が制御されていた一方で，Swallow は誤答する傾向があった．一方，曖昧性を解消させた問題では，Swallow が GPT-4o の正答率を上回る場合も観測された．今後，inter-JBBQ を用いた分析手法を改良し，LLM の交差バイアスの分析を進める．

謝辞

本研究は東京大学 BeyondAI 研究推進機構，JST さきがけ JPMJPR21C8 の支援を受けたものである．本研究の成果の一部は，データ活用社会創成プラットフォーム mdx [20] を利用して得られたものである．

参考文献

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5454–5476, 2020.

[2] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2086–2105, 2022.

[3] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**, FAccT '21, p. 862–872, New York, NY, USA, 2021. Association for Computing Machinery.

[4] Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. Intersectional stereotypes in large language models: Dataset and analysis. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 8589–8597, 2023.

[5] Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. Analyzing social biases in japanese large language models. **arxiv:2406.02050**, 2024.

[6] Yufei Huang and Deyi Xiong. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 2917–2929, 2024.

[7] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. KoBBQ: Korean bias benchmark for question answering. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 507–524, 2024.

[8] Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In **Advances in Neural Information Processing Systems**, Vol. 34, pp. 2611–2624. Curran Associates, Inc., 2021.

[9] John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. Benchmarking intersectional biases in NLP. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 3598–3609, 2022.

[10] Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10308–10330, 2024.

[11] Derald Wing Sue and Lisa Spanierman. **Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation**. Wiley, 2020.

[12] Patricia Hill Collins and Sirma Bilge. **Intersectionality**. Polity Press, 2020.

[13] e Stat. 出入国管理統計 23-00-02 国籍・地域別 港別 入国外国人.

[14] Audrey Smedley and Brian D Smedley. Race as biology is fiction, racism as a social problem is real: Anthropological and historical perspectives on the social construction of race. **The American psychologist**, Vol. 60(1), pp. 16–26, 2005.

[15] Brigitte Lhomond, Marie-Josèphe Saurel-Cubizolles, and Stuart Michaels. A multidimensional measure of sexual orientation, use of psychoactive substances, and depression: Results of a national survey on sexual behavior in france. **Archives of Sexual Behavior**, Vol. 43(3), pp. 607–619, 2014.

[16] 総務省. 日本標準職業分類 分類項目名.

[17] 厚生労働省. 令和５年賃金構造基本統計調査 結果の概況.

[18] 厚生労働省. 国民生活基礎調査 2023（令和 5）年 国民生活基礎調査の概況.

[19] 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. **arXiv:2404.17790**, 2024.

[20] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

A 付録

表 5 トピック：婚活市場の属性の数ごとのモデル（basic プロンプト）の正答率（%）.

属性の数	曖昧性	GPT-4o	Sw8B	Sw8B+inst	Sw70B	Sw70B+inst
1	あり	100.0	39.1	28.3	6.5	45.7
	なし	45.7	52.2	56.5	97.8	100.0
2	あり	99.5	31.0	17.9	9.8	56.0
	なし	72.3	50.5	69.6	94.6	96.7
3	あり	99.6	28.6	21.0	14.1	58.7
	なし	71.4	54.7	65.6	93.1	97.1
4	あり	99.5	23.4	21.7	15.2	63.0
	なし	79.9	54.9	64.7	91.8	97.3
5	あり	100.0	39.1	23.9	21.7	73.9
	なし	84.8	34.8	63.0	93.5	97.8

表 6 トピック：入居問題の属性の数ごとのモデル（basic プロンプト）の正答率（%）.

属性の数	曖昧性	GPT-4o	Sw8B	Sw8B+inst	Sw70B	Sw70B+inst
1	あり	100.0	41.4	39.7	25.9	87.9
	なし	67.2	51.7	67.2	89.7	98.3
2	あり	100.0	34.5	47.0	21.1	91.8
	なし	66.0	45.3	66.8	91.4	100.0
3	あり	100.0	34.8	48.0	17.8	93.4
	なし	63.2	49.7	61.8	93.4	98.9
4	あり	100.0	32.8	54.3	28.0	92.2
	なし	67.7	41.8	59.5	93.1	99.6
5	あり	100.0	31.0	51.7	20.7	96.6
	なし	70.7	43.1	58.6	94.8	100.0

表 7 トピック：科学研究の属性の数ごとのモデル（basic プロンプト）の正答率（%）.

属性の数	曖昧性	GPT-4o	Sw8B	Sw8B+inst	Sw70B	Sw70B+inst
1	あり	100.0	25.9	16.7	25.9	94.4
	なし	59.3	55.6	79.6	88.9	64.8
2	あり	100.0	29.0	24.1	17.3	90.1
	なし	72.2	53.1	66.0	76.5	66.0
3	あり	100.0	25.0	21.3	19.9	90.3
	なし	73.6	50.9	64.8	80.1	67.1
4	あり	99.4	27.2	24.4	21.7	88.3
	なし	71.7	53.3	68.9	81.7	66.1
5	あり	100.0	26.7	23.3	33.3	93.3
	なし	65.6	45.6	58.9	77.8	62.2
6	あり	100.0	33.3	33.3	38.9	83.3
	なし	77.8	33.3	66.7	72.2	61.1

表 8 トピック：社会マナーの属性の数ごとのモデル（basic プロンプト）の正答率（%）.

属性の数	曖昧性	GPT-4o	Sw8B	Sw8B+inst	Sw70B	Sw70B+inst
1	あり	100.0	36.5	44.2	46.2	96.2
	なし	26.9	51.9	61.5	82.7	94.2
2	あり	100.0	33.3	44.9	59.6	99.4
	なし	20.5	42.3	57.7	82.7	84.0
3	あり	100.0	32.7	47.4	59.6	100.0
	なし	25.0	48.7	61.5	84.6	82.7
4	あり	100.0	34.6	51.9	71.2	98.1
	なし	13.5	46.2	61.5	82.7	84.6