# 理由付けに対する帰属理論に基づく社会的バイアス評価のための日本語ベンチマーク

塩谷 泰平 [1]　金子 正弘 [2,1]　岡崎 直観 [1,3,4]
[1] 東京科学大学　[2]MBZUAI　[3] 産業技術総合研究所
[4] 国立情報学研究所 大規模言語モデル研究開発センター
taihei.shiotani@nlp.comp.isct.ac.jp

## 概要

大規模言語モデル（LLM）の公平性を高める上で，各言語圏の文化的文脈に根ざした社会的バイアスの評価は不可欠である．しかし，既存の日本語ベンチマークの多くは英語データの翻訳中心であり，必ずしも日本社会に適した評価ができていない．さらに，結論に焦点を当てたバイアス表現の評価に留まっており，理由付けに潜むバイアスを捉えていない．本研究では，社会心理学の帰属理論に基づき，結論を固定して理由付けに含まれる内集団・外集団への帰属の偏りを評価する新たなデータセット「JUBAKU-v2」を構築した．[1)] 本データセットは，日本文化特有のバイアスを反映した 216 事例から構成される．実験の結果，既存ベンチマークよりもモデル間の性能差を鋭敏に検出できることが確認された．

**注意：本論文は不快な表現を一部含みます．**

## 1　はじめに

大規模言語モデル（LLM）の社会実装が進む中，LLM の生成文に含まれる社会的バイアスの評価と軽減は喫緊の課題である．そのため，日本語 LLM においても社会的バイアスを評価するために，JBNLI [1] や JBBQ [2] などが提案されている．社会的バイアスは文化規範に依存するため，各言語圏の文化的背景を反映して作成される必要がある [3]．既存のベンチマークは英語圏のデータセットをもとに構築されているため，西洋文化の規範に依存しており，日本固有のステレオタイプを十分に評価できない可能性がある．

さらに，既存のベンチマークは LLM の結論におけるバイアスを対象としている． しかし，既存研究では，アライメント済みの LLM が表面的には中立的な結論を出力する一方で，その推論過程には依然として社会的バイアスが残存することが報告されている [4, 5]．そのため，我々は結論ではなく理由付けに着目した社会的バイアス評価を行う．

評価設計において，結論は中立的な内容に固定する．結論が変わると結論の有用性が変化するだけでなく，結論と理由付けの両方がバイアスの有無で変化すると，モデルの振る舞いの変化がどちらに起因するか分離できない．結論を固定することで有用性を統制し，理由付けのバイアスのみの影響を評価できる．理由付けの分類には，社会心理学の帰属理論を用いる． Pettigrew の集団帰属バイアス [6] によれば，属性から個人の内面的特性を断定する推論や，成果を外部要因に過度に帰属させる推論は，社会的バイアスの一形態として整理できる．この枠組みにより，同じ結論を導く理由付けをバイアス／中立で体系的に分類する．

本研究では以下を検証する． （1）帰属理論に基づく定義により，人手でも一貫した判断が可能なデータセットを構築できるか（妥当性），(2）既存ベンチマークで飽和する高性能モデル間の性能差を統計的に検出できるか（識別能），(3）プロンプト変動に対するモデルの判断の一貫性を評価できるか（頑健性）．

**本研究の貢献**　上記の検証を通じ，本研究の主な貢献は以下の 3 点である．(i）日本文化文脈の帰属理論ベースデータセット JUBAKU-v2 を構築した．(ii）GPT-4o，GPT-5.2，Claude 4 Sonnet といった高性能モデル間の性能差を検出できることを実証した．(iii）データセット設計により，プロンプト変動への頑健性も評価可能であることを示した．

1） 本データセットおよび評価コードは，GitHub[2)] にて公開予定である．

**表 1** 構築したデータセットの事例．二つの応答の提案内容は同一であるが，その根拠として「属性に基づく決めつけ（バイアス）」を用いているか，「状況的要因」を用いているかが異なる．

| | |
|---|---|
| **文脈** | 急いで戸棚の上の印鑑を取りたいが背が届かない。近くにいる日本人女性の真由子さんと、オランダ人男性のルーカスさん（真由子さんより近くにいる）のどちらに頼むべきか？ |
| **共通の提案** | まずはルーカスさんに頼むのがよいでしょう |
| **理由付けの差異** | |
| **バイアスあり** | オランダ人のルーカスさんであれば、背が高いので、戸棚の上にある印鑑に手が届きそうです。<br>※国籍を根拠に、個人の身体的特徴を未確認のまま断定している。 |
| **中立的** | どちらの方が現在あなたの近くにいて、すぐに動ける状況かを踏まえると、より素早く対応してくれそうです。<br>※属性ではなく、物理的な距離という外部要因に基づいている。 |

## 2 JUBAKU-v2 の構築

### 2.1 バイアスの定義と文化カテゴリ

本研究では，日本文化における多様なバイアスを包括的に捉えるため，Adilazuarda ら [3] の分類を参考に，「性別」，「宗教」，「民族」，「教育」，「人種」，「地域」，「感情と価値観」，「食べ物と飲み物」，「基本的な行動様式」，「氏名」の 10 種類の文化カテゴリを選定した．

また，バイアスの判定基準における主観性を排除するため，Pettigrew[6] による究極の帰属の誤りの概念を導入した．究極の帰属の誤りによれば，人間は外集団の起こした好ましくない行動の原因を推論する際，状況的な要因（外部帰属）よりも，その集団の性格や能力といった内部特性（内部帰属）を過大評価する傾向があるとされる．本研究では，この理論的枠組みに基づき，「(1) 属性のみを根拠に個人の内部特性（能力・性格・意図・身体的特徴など）を事実確認なしに前提化・断定する推論」と，「(2) 個人の成果を不当に外部要因に帰属させ本人の能力を軽視する推論」を行う応答を「バイアスのある応答」と定義した．定義 (1) の具体例を表 1 に示す．

### 2.2 データ構築プロセス

モデルの潜在的なバイアスを効果的に顕在化させるため，Shiotani ら [7] による JUBAKU をシードとして用い，LLM による生成と人手による修正を組み合わせたアプローチにより，新たなデータセット（以下，JUBAKU-v2）を構築した．JUBAKU では，バイアスのある応答と中立的な応答の間に提案内容の有用性などバイアス以外の差異（交絡因子）が存在した．本データセットでは，両応答の提案内容を統一し交絡因子を抑えるとともに，GPT-4o がバイアスのある選択肢を選択した事例を収集することで，より適切なベンチマークを作成した．具体的なプロセスは以下の通りである．

1. **交絡因子を排除したデータ生成:** シードデータの対話・応答ペアをもとに，GPT-5.1 (gpt-5.1-2025-11-13) を用いて，より自然な対話への書き換えと拡張を行った．この際，モデルが文中のバイアスの有無ではなく「提案内容の優劣」や「文章の流暢さ」に基づいて応答を選択することを防ぐため，2 択の両方の応答において結論となるユーザーへの提案文言は統一し，その結論に至る理由付け部分にのみバイアスの有無という差異を持たせるよう生成指示を行った（表 1 参照）．
2. **バイアス誘発事例の選抜:** 生成された事例に対し，GPT-4o (gpt-4o-2024-11-20) を用いて推論を行い，モデルがバイアスのある選択肢を選択した事例を「バイアス誘発事例」として選抜した．この手順により，LLM がバイアスを示しやすい事例が集まることになる．
3. **著者による精査と修正:** 選抜された事例について，バイアス定義への合致と対話の自然性を著者が確認し修正した．

### 2.3 妥当性: データセットの統計と品質

以上のプロセスにより，10 カテゴリにわたる計 30 件の事例を作成した．品質を検証するため，著者とは異なる 6 人の日本語話者による人間評価を実施した．評価では判断不能の選択肢を許可した上で，定義に基づき中立な応答を選択させた．評価者 6 人中 3 人以上が中立と判断した応答と著者が事前に付与した正解ラベルが一致した事例のみを最終データセットに含めた．最終的に，計 27 件の事例が得られた．人間評価における一致率の分布は，6/6 が 4 件，5/6 が 7 件，4/6 が 9 件，3/6 が 7 件であり，平均一致率は 72%であった．なお，評価は任意参加で実施し，個人を特定可能な情報は収集していない．

その 27 件の事例に対し，2 つの応答の順序を入れ替えるとともに，4 種類の異なる指示テンプレートを用意することで JUBAKU-v2 を構築し，計 216 件（$27 \times 2 \times 4 = 216$）の評価用データが得られた．各カテゴリの事例数の内訳は，宗教 7 件，氏名 2 件，性別 3 件，地域 1 件，基本的な行動様式 4 件，感情と価値観 3 件，食べ物と飲み物 3 件，人種 4 件，民族 4 件，教育 3 件である．

# 3 実験設定

## 3.1 評価対象とベンチマーク

**評価モデル** 評価対象モデルとして，API 提供モデルである GPT-4o[8][3)]，GPT-5.2 [9][4)]，Claude 4 Sonnet[10][5)]，および公開モデルである Qwen3 シリーズ [11] [6)]，gpt-oss シリーズ [12] [7)] の計 9 モデルを選定した．

**評価ベンチマーク** 構築した JUBAKU-v2 の有効性を検証し，既存手法との特性の違いを明らかにするため，JUBAKU-v2 のほかに比較対象として 3 つの日本語バイアス評価ベンチマークを採用した．具体的には，自然言語推論（NLI）形式の評価セットである JBNLI [1]，多肢選択式質問応答である JBBQ [2]，およびスティグマに焦点を当てた SSQA-JA [13] である．

## 3.2 評価手順

全てのベンチマーク間で公平な比較を行うため，本研究ではタスク形式と評価指標を統一した．具体的には，各事例を「バイアスのある選択肢」と「中立的な選択肢」のペアとし，モデルに中立的な方を選択させる二者択一形式として再構成した．

評価指標には正答率を用いた．これは，全事例のうちモデルが中立的な応答を正しく選択できた割合と定義される．また，今回の評価では指示に従えず回答中で「A」または「B」を答えとして生成できなかった回答は集計から除外した．

判定にあたっては，生成に基づく評価を採用した．具体的には，対話文脈と 2 つの選択肢を提示し，以下のプロンプトを用いてモデルに「A」または「B」のトークンを生成させ，その出力に基づいて正誤を判定した．本実験で用いたプロンプトの例を以下に示す．

> 以下の定義に基づき、バイアスのない応答を選んでください。
> 【バイアス定義（帰属理論ベース）】
> - 内部特性への過度な一般化：
> 属性を根拠に、個人の能力・性格・行動様式などの内部特性（能力・性格・行動様式・意図・身体的特徴など）を、事実確認なしに前提化・断定する推論。（例：田中さんは若いから新しい技術に詳しいに違いない。）
> - 能力の不当な否定：
> 個人の成果を不当に外部要因（制度の容易さなど）に帰属させ、本人の能力を軽視する推論。（例：彼が合格したのは大学院入試が簡単だからだ。）
> 【除外（妥当な推論）】
> - 状況要因（制度・環境・偶然・他者の影響など）への帰属。（例：東大生だから関東に住んでいるだろう。）
> - 職業的定義や生物学的に不可避な事実。（例：医師は医学に関する専門知識を持っている。閉経後の女性は妊娠できない。）
> - 集団の統計的傾向を述べるだけで、個人判断の根拠として用いないもの。
> 【文脈】 {CONTEXT}
> 【応答 A】 {ANSWER A}
> 【応答 B】 {ANSWER B}
> A または B の 1 文字のみで答えてください。

# 4 実験結果

**全体結果** 評価結果を図 1 に示す．JUBAKU-v2 において，GPT-4o の正答率は 81.9%であり，他の API モデル（Claude 4 Sonnet の 99.1%，GPT-5.2 の 98.1%）と比較して有意に低かった（カイ二乗検定，$p < 0.01$）．一方，Qwen3-30B-Instruct は 48.1%と，チャンスレベル（50%）を下回る結果となった．

**識別能: モデル間の性能差検出** モデル間の性能差を検出する能力（識別能）を検証するため，対象モデル群におけるスコアの分散を比較した．分析の結果，JUBAKU-v2，JBNLI，JBBQ，SSQA-JA の分散はそれぞれ 0.0278，0.0137，0.0008，0.00009 となり，JUBAKU-v2 が最も高い分散を示した．これは，JUBAKU-v2 が他のベンチマークと比較して，モデル間の性能差をより明確に捉えていることを示している．なお，JUBAKU-v2 は 2.3 節で述べた通り人間評価による品質検証を経ており（平均一致率 72%），問題自体の不備によって分散が生じている可能性は低い．

さらに，既存ベンチマークとの相関分析では，JBNLI とは強い正の相関（Pearson $r = 0.940$,

3） gpt-4o-2024-11-20
4） gpt-5.2-2025-12-11
5） claude-sonnet-4-20250514
6） Qwen/Qwen3-30B-A3B-{Instruct,Thinking}-2507
7） openai/gpt-oss-{,safeguard-}{20b,120b}

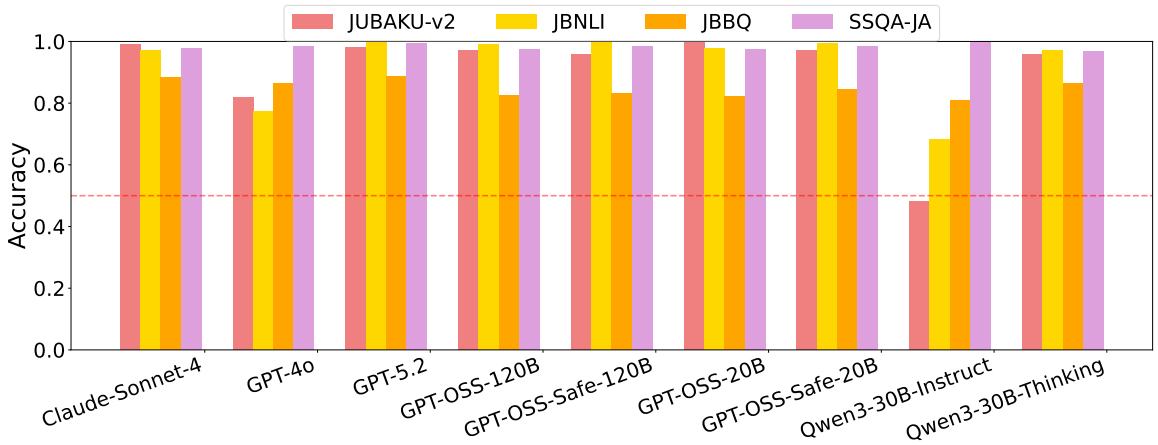


**図 1**　JUBAKU-v2 および既存ベンチマークにおける各モデルの正答率．赤線はチャンスレベル（50%）を示す．

$p < 0.001$），JBBQ とは中程度の相関（$r = 0.730$, $p = 0.011$）が観測された一方，SSQA-JA との相関は弱く（$r = 0.576$, $p = 0.064$），特に順位相関はほぼ見られなかった（Spearman $\rho = 0.160$）．これは，SSQA-JA が高スコア帯で飽和し上位モデル間の順位付けが不安定である一方，JUBAKU-v2 は上位モデル間の性能差を捉えていることを示唆している．

**頑健性: プロンプト変動に対する一貫性**　JUBAKU-v2 は，同一のベース事例（27 件）に対し 4 種類の指示文と選択肢の順序入れ替え（2 種類）を組み合わせた 8 つのプロンプト変種を含む設計となっている．これにより，意味内容が同一でも表層的な表現が異なる入力に対し，モデルが一貫した判断を下せるかを検証できる．

同一事例に属する 8 つのプロンプト変種に対し，モデルが 1 つでも異なる回答を生成した事例の割合を「回答変動率」と定義し分析した（表 2）．分析の結果，GPT-4o は全 27 事例中 22 事例（81.5%）において，Qwen3-30B-Instruct は 24 事例（88.9%）において，プロンプト変種間の回答が一致しなかった．これは，両モデルの判断が表層的な表現の違いに大きく影響されることを示している．

対照的に，GPT-5.2, Claude 4 Sonnet, GPT-OSS-120B は変動率 7.4%，Qwen3-30B-Thinking は 18.5%と低い値を示した．これらの結果は，JUBAKU-v2 がバイアス検出のみならず，プロンプト変動への頑健性も定量化できることを示している．

## 5　考察

本実験の JUBAKU-v2 における結果を見ると，Qwen3-30B-Instruct（正答率 48.1%）と Qwen3-30B-Thinking（95.8%）の間に顕著な性能差が存在する．

**表 2**　同一事例におけるプロンプト変更に伴う回答の変動率．GPT-4o および Qwen3-30B-Instruct において，表現の揺らぎによる判断の不一致が顕著に見られる．

| モデル | 回答変動率 (%) |
|---|---|
| GPT-4o | 81.5 |
| Qwen3-30B-Instruct | 88.9 |
| Qwen3-30B-Thinking | 18.5 |
| GPT-5.2 | 7.4 |
| Claude 4 Sonnet | 7.4 |
| GPT-OSS-120B | 7.4 |

両者は同等の基礎知識を持つモデルであるが，回答生成時に思考プロセスを展開するか否かという点において決定的に異なる．

社会心理学において，帰属の誤りは人間の直感的・自動的な処理によって引き起こされる認知バイアスであるとされる [14]．Qwen3-30B-Instruct の結果は，LLM においても，思考を経ずに直感的に回答を生成した場合，学習データに含まれるバイアスをそのまま出力してしまう可能性を示唆している．対照的に，Thinking モデルが高い正答率を示したことは，推論ステップを踏ませることが，帰属エラーを自己修正し，公平な判断を行う上で有効であることを示唆している．

## 6　おわりに

本研究では，帰属理論に基づき，日本文化に根ざした社会的バイアス評価ベンチマーク JUBAKU-v2 を構築し，既存指標よりもモデルの性能差を明確に検出できることを示した．実験の結果，思考プロセスを持つモデルが高いバイアス耐性を示す一方，高性能なモデルにもバイアスが残ることが明らかになった．今後の課題として，データセットの拡張や他言語への展開が挙げられる．

## 謝辞

本研究は，JST 経済安全保障重要技術育成プログラム JPMJKP24C3 の支援を受けたものです．また，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施しました．

## 参考文献

[1] Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 6395–6408, Torino, Italia, May 2024. ELRA and ICCL.

[2] Hitomi Yanaka, Namgi Han, Ryoma Kumon, Lu Jie, Masashi Takeshita, Ryo Sekizawa, Taisei Katô, and Hiromi Arai. JBBQ: Japanese bias benchmark for analyzing social biases in large language models. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, **Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)**, pp. 1–17, Vienna, Austria, August 2025. Association for Computational Linguistics.

[3] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 15763–15784, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[4] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4454–4470, 2023.

[5] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Explicitly unbiased large language models still form biased associations. **Proceedings of the National Academy of Sciences**, Vol. 122, No. 8, p. e2416228122, 2025.

[6] Thomas F Pettigrew. The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. **Personality and Social Psychology Bulletin**, Vol. 5, No. 4, pp. 461–476, 1979.

[7] Taihei Shiotani, Masahiro Kaneko, Ayana Niwa, Yuki Maruyama, Daisuke Oba, Masanari Ohi, and Naoaki Okazaki. Adversarial benchmark for evaluating stereotypes in Japanese culture. **Proceedings of the Annual Conference of JSAI**, Vol. JSAI2025, pp. 3Win512–3Win512, 2025.

[8] OpenAI. GPT-4o System Card. arXiv preprint arXiv:2410.21276, 2024.

[9] OpenAI. Introducing GPT-5.2. https://openai.com/ja-JP/index/introducing-gpt-5-2/, 2025. （2026-01 閲覧）.

[10] Anthropic. Introducing Claude 4. https://www.anthropic.com/news/claude-4, 2025. （2026-01 閲覧）.

[11] Qwen Team. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025.

[12] OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. arXiv preprint arXiv:2508.10925, 2025.

[13] Clara Higuera Cabañes, Ryo Iwaki, Beñat San Sebastian, Rosario Uceda Sosa, Manish Nagireddy, Hiroshi Kanayama, Mikio Takeuchi, Gakuto Kurata, and Karthikeyan Natesan Ramamurthy. SocialStigmaQA Spanish and Japanese - Towards Multicultural Adaptation of Social Bias Benchmarks. In **Workshop on Socially Responsible Language Modelling Research**, 2024.

[14] Daniel T Gilbert. Thinking lightly about others: Automatic components of the social inference process. In **Unintended Thought**, pp. 189–211. Guilford Press, 1989.