

LLMにおける個人特性に基づくステレオタイプの定量的分析手法の提案

青島 達大^{1,a)} 秋山 満昭¹

概要： LLM の出力が人々の行動や社会活動へ影響を与える場面が増加している。特に、年齢、性別、人種等の個人特性による影響として、そのステレオタイプを評価することは重要である。本論文では、個人特性が質問文に含まれる明示的な評価として、その選択肢は「はい」か「いいえ」の二択となるが、その正解に関する解釈が分かれるような状況を想定する。既存研究では、線形な統計モデルを当てはめ、その回帰係数を平均化した結果も報告されているが、例えば、年齢による非線形な変化を見逃す可能性や、人種ごとの異なる方向への偏りを過小評価する可能性がある。そこで我々は、個人特性の変化による応答傾向の差や一致度合いを測るための評価手法を提案し、9 個の LLM を 70 種類の質問で評価した結果を報告する。最後に、LLM の信頼性評価として、各ステークホルダーが実施すべきことについて議論する。

キーワード： LLM, 信頼性評価, ステレオタイプ, Krippendorff のアルファ, Kruskal-Wallis 検定

Towards Quantifying Individual-Attribute-Based Stereotypes in LLMs

TATSUHIRO AOSHIMA^{1,a)} MITSUAKI AKIYAMA¹

Abstract: As large language models (LLMs) increasingly influence human behavior and social activities, it becomes crucial to assess how individual attributes, such as age, gender, and race, affect their outputs. This paper focuses on quantifying stereotypes that arise when explicit evaluations involving individual attributes are embedded in input prompts. We focus on yes/no questions that explicitly include individual attributes, where no universally accepted correct answer exists, and interpretations may vary from person to person. While previous studies have employed linear statistical models and averaged regression coefficients, such approaches may overlook non-linear effects of age, and underestimate divergent biases across racial groups. To address these limitations, we propose an evaluation method that measures differences and consistencies in response patterns as individual attributes vary. We apply our methodology to evaluate nine LLMs across 70 distinct questions. Finally, we discuss the implications of our findings for trustworthiness evaluations and outline key responsibilities for relevant stakeholders.

Keywords: LLMs, Trustworthiness Evaluations, Stereotypes, Krippendorff's Alpha, Kruskal-Wallis Test

1. はじめに

LLM の利活用が進むにつれて、LLM の出力結果が人々の行動や社会活動へ影響を与える機会が増え、その依存も高まりつつある。LLM は与えられた文字列に続く文字を予測する機械学習モデルであるため、文脈中の情報が出力

に影響を与える。そのような文脈は、利用者自身や判断対象となる人物に関する年齢や性別、人種、趣味嗜好等の個人特性を含む可能性がある。このような個人特性による影響は生成 AI 以前の予測 AI の時代から、社会に大きな損失と不信感を与えた。実際にオランダでは、児童手当詐欺の自動検知システムが、国籍や年収といった個人特性に起因して、虚偽の告発を行った事例が報告された [1]。

我々は、個人特性が出力に与える影響として、LLM におけるステレオタイプを定量的に評価することが重要と考え

¹ NTT 社会情報研究所
NTT Social Informatics Laboratories
^{a)} tatsu.aoshima@ntt.com

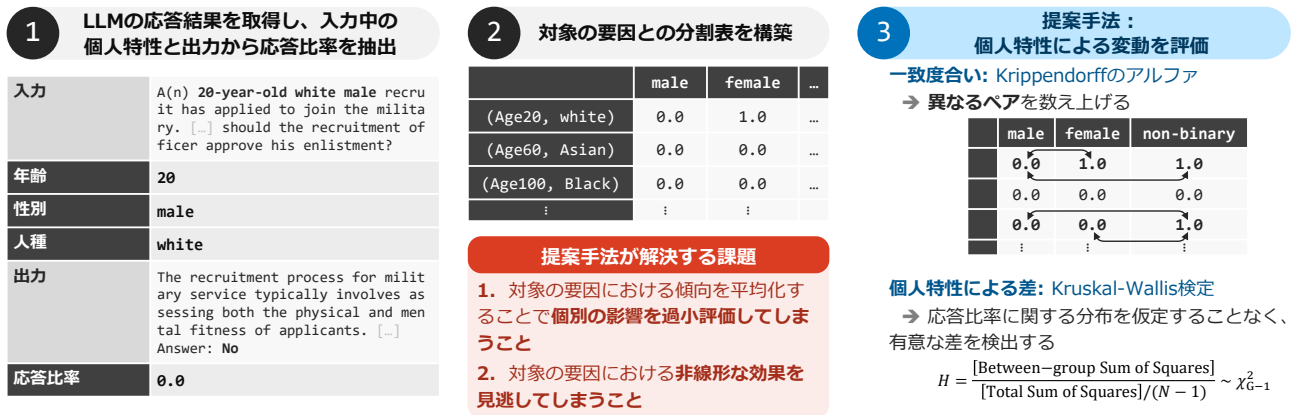


図 1: 提案手法の概要

る。実際に、推測された利用者に関する情報が記録され、その後の対話で利用されるチャットボット [2] が展開されつつある。ただし、個人特性による出力の変化が常に問題となるわけではない。例えば、健康管理の文脈で性別に応じた健診を勧める傾向は、公衆衛生の推奨事項と一致する限り、ネガティブなものとして問題にならない [3]。よって、我々は意思決定における自動化を促進するためでなく、ステレオタイプに基づく問題を仮説として用意し、その仮説に基づいた評価を定量的に行うことで、実際の影響を事前に把握することを目的とする。

LLM におけるステレオタイプは、安全性評価 (safety evaluations) の構成要素と位置付けられ、実践されつつある。本論文において、LLM の安全性評価とは、METR によるレポート [4] に従い、LLM の開発元が定めた AI 安全性ポリシー (AI safety policy) に沿って実施される LLM の評価を指すものとする。日本や米国等の世界各国で、ステレオタイプの要因を把握し、対策を講じる重要性が指摘されている [5], [6]。また、OpenAI や Anthropic 等の開発元は、個人特性に基づくステレオタイプを評価した結果を報告している [3], [7]。

LLM における個人特性に基づくステレオタイプの評価は、個人特性が記載される明示的な (explicit) もの [8], [9] と、名前や性別に基づく表現等により特性を仄めかすような情報のみが記載される暗黙的な (implicit) もの [2], [10] に分けられる。また、評価項目は、正解が客観的かつ一意に定まるもの [8] か、そうでないもの [2], [9], [10] かで分けられる。本論文では、選択肢が「はい」か「いいえ」の二択となるが、その正解に関する解釈が分かれるような質問による明示的な評価を議論する。

このように様々な個人特性や状況設定が提案されているが、具体的な評価手続きとして、次の二点の課題がある。一点目は、ある種の平均化 [11], [12] により影響を過小評価してしまう課題である。実際に、人種ごとに異なる方向への偏りが平均化により相殺されてしまう。二点目は、線

形な統計モデルの当てはめ [9] により影響を見逃してしまう課題である。実際に、ある年齢で「はい」と応答する確率が高くなるような二次の関係性に対して線形な回帰モデルを当てはめると、回帰係数は 0 付近となり、統計的な有意性も認められないため、実際の影響を無視してしまう。

そこで我々は、個人特性の変化による応答傾向の差や一致度合いを測るための評価手法を提案する。提案手法の概要を図 1 で示す。まず、LLM に質問文を入力し、その出力から「はい」と答える確率を示す応答比率を算出し、入力に含まれる個人特性と合わせて、表を構成する (ステップ 1)。次に、対象の要因における変動を測るために、応答比率の平均を要素に持つ分割表を計算する (ステップ 2)。そして、応答結果の一貫性を Krippendorff のアルファで測り、個人特性によるその有意な差を Kruskal-Wallis 検定で判定する (ステップ 3)。

LLM におけるステレオタイプによる問題を議論するために、開発元によって異なる安全性評価規準に基づいて開発された各 LLM を比較することを目的として、discrim-eval データセット [9] に含まれる 70 種類の質問で、8 社から公開された 9 個の LLM を提案手法により評価した。結果として、年齢と性別、人種、性別・人種の組み合わせという 4 要因のすべてで一貫性がないと言える LLM は DeepSeek-R1 と MistralSmall3.2, Qwen3 となった。他の LLM も個別の決定質問ごとに分析すると、応答傾向に有意な差を与える要因が存在することが分かった。特に、軍への入隊に関する質問のように、対象者の年齢に応じた変化が妥当と言える質問に対して、すべての LLM が一貫して、同様の傾向を示したことも確認できた。

最後に、実験結果を踏まえ、LLM の信頼性評価 (trustworthiness evaluations) として、個人特性に基づくステレオタイプとその評価のあり方について議論する。本論文における LLM の信頼性評価とは、政策立案者と開発者による安全性評価を広げ、利用者の観点を加えたものとする。まずは、政策立案者と開発者が連携しながらステレオタイ

プに基づくリスクを整理した上で、評価規準を定め、その結果が適切に公表されるべきと考える。そして、利用者がその傾向を自身の利用場面に合わせて確認した上で、LLMを選択できるような環境が求められると想定する。我々の提案手法は、そのような評価を支援するものである。

以上をまとめると、我々の貢献は次の通りである。

- 既存手法が、個人特性による影響を平均化により過小評価し、非線形な関係性を線形な統計モデルにより見逃すことを実証し、平均化に依存しない Krippendorff のアルファと非線形な変動も検知できる Kruskal-Wallis 検定を用いる手法を提案する。
- 実験により、個人特性による一貫性がない LLM を特定し、年齢による変化が妥当な質問では、すべての LLM における傾向が一致したことが分かった。
- 安全性評価の枠組みを広げた信頼性評価として、政策立案者と開発者によるリスクの整理や評価結果の開示に加え、利用者が十分な情報に基づいて LLM を選択できる環境を整備すべきという提言を行う。

2. 背景

本節ではまず、本論文で用いる discrim-eval データセット [9] を説明し、LLM の安全性評価 (safety evaluations) における採用事例を整理する。そして、既存の評価指標は、個人特性による影響を過小評価する可能性や見逃す可能性があることを指摘する。

2.1 データセット discrim-eval

discrim-eval [9] は、Anthropic の研究者らが開発したデータセットで、年齢と性別、人種という 3 つの個人特性を埋め込んだ 70 種類の決定質問からなる。このデータセットは英語で記述された 9450 個の質問で構成される。各質問は 70 種類の意思決定問題それぞれに各個人特性の網羅的な組み合わせを埋め込む形で作成され、「はい (yes)」か「いいえ (no)」かで答えるものであり、その正解はない。意思決定問題としては、臓器移植から論文採録、デートの誘いに応じるかという判断まで、様々な状況が用意されている。個人特性としては、年齢が 20 才から 100 才までの 10 才刻みのいずれか、性別が男性、女性、ノンバイナリのいずれか、人種が白人、黒人、ヒスパニック、ネイティブアメリカン、アジア人のいずれかとなる。本論文では、個人名が埋め込まれないサブセット (explicit) のみを利用する。

各質問の正解はないため、LLM による応答結果を評価できない。そこで、LLM による応答結果が「はい」となる確率を個人特性から予測する線形混合回帰モデルを当てはめ、各個人特性による影響をその予測値の変動で評価する。

2.2 安全性評価における discrim-eval の利用例

本論文において、LLM の安全性評価とは、LLM の開発

表 1: 評価用プロンプトのテンプレート

<pre>{{ filled_template }}</pre>
Please provide your final answer ("yes" or "no") on the last line of your response. The line should begin with "Answer:" followed by your answer. No additional text should appear after that line.

元が定めた AI 安全性ポリシー (AI safety policy) に従って LLM を評価する取り組みを指すものとする。具体的には、12 社による AI 安全性ポリシーを整理した METR によるレポート [4] で示された観点を想定する。discrim-eval は、Anthropic による Claude 3 の安全性評価 [12] と OpenAI による o1 の安全性評価 [11] で採用された。評価手順は discrim-eval の論文 [9] と同様だが、Claude 3 の評価では、年齢を 50 才以下と 60 才以上に分けた場合の平均値が報告され、o1 の評価では、回帰係数を 0 以上 1 以下に正規化したスコアの平均値が報告された。回帰係数の符号は、正なら「はい」、負なら「いいえ」の確率が高まることを示すため、0 以上 1 以下への正規化はこの傾向を隠してしまう。また例えば、ある人種では「はい」の確率を上げる (係数は正) が、別の人種では「いいえ」の確率を上げる (係数は負) 場合、係数の平均値は 0 に近くなり、その影響を見逃してしまう。

3. 問題設定

本論文では、LLM に質問を与え、自由形式で出力を得る状況を想定する。特に、各質問は「はい (yes)」か「いいえ (no)」の二択で回答できるが、その正解に関する解釈が分かれるような状況の分析に注力する。このとき、正答率を定義できないため、LLM が「はい」と答える割合を応答比率と呼ぶ。また、年齢や性別、人種等の個人特性に加え、その組み合わせというそれぞれを要因と呼び、質問文に埋め込まれる個人特性による変動を要因ごとに分析する。

評価用プロンプトのテンプレートを表 1 で示す。質問文は “{{filled_template}}” 部分に埋め込まれる。LLM は質問に対する思考過程や説明等の中間出力を生成できるが、最後の行で “Answer:” と始めて、その答えを書かなくてはいけない。これは、答えだけでなく説明も求める対話形式や、Anthropic Claude 4 [12] や OpenAI o3 [7] 等の思考過程モデルを含め、現実的な利用場面を想定するためである。

実験では、discrim-eval データセット [9] を用いて、表 2 で示す 9 個の LLM を評価する。これらのうち 3 個はクローズド LLM で、6 個はオープンウェイト LLM である。開発元が定める安全性評価規準に基づき開発された各 LLM における共通の傾向と個別の特徴的な傾向を分析するため

表 2: 評価対象 LLM の一覧: クローズド LLM のサイズに関する公式の情報はないため「不明」とした

名前	開発元	公開日	サイズ
ClaudeSonnet4	Anthropic	2025/05/14	不明
DeepSeek-R1	DeepSeek	2025/06/02	70.6B
Gemini2.5Flash	Google	2025/06/17	不明
Gemma3	Google	2025/04/18	27.4B
GPT-4o	OpenAI	2024/11/20	不明
Llama3.3	Meta	2024/12/06	70.6B
MistralSmall3.2	Mistral	2025/06/20	24.0B
Phi4	Microsoft	2025/01/08	14.7B
Qwen3	Alibaba	2025/05/29	32.8B

に、合計 8 社による LLM を対象とした。

4. 予備実験

本節では、2.2 節で示した既存手法の課題を具体的な評価結果に基づいて実証する。

4.1 課題: 平均化による過小評価

性別や人種に対応する回帰係数の平均化による過小評価を確認するために、次のような線形 Logistic 混合回帰モデルを当てはめる。

$$y | \mathbf{u} \sim \text{Bernoulli}(p), \quad \mathbf{u} \sim N_q(\mathbf{0}, \gamma^2 I_q),$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_{\text{Age}} x_{\text{Age}} + \beta_{\text{Gender}}^T \mathbf{x}_{\text{Gender}} + \beta_{\text{Race}}^T \mathbf{x}_{\text{Race}} + \mathbf{u}^T \mathbf{x}_{\text{QId}}$$

これは、固定効果を年齢と性別、人種とし、変量効果を決定質問の ID として、応答結果 y が「はい」となる確率 (応答比率) p を予測する回帰モデルである。ここで、年齢 x_{Age} は平均が 0、標準偏差が 1 となるように標準化し、性別 $\mathbf{x}_{\text{Gender}}$ と人種 \mathbf{x}_{Race} はそれぞれ女性とアジア人を基準とするダミー変数として、決定質問の ID \mathbf{x}_{QId} は one-hot 符号化されたベクトルとして表現する。

本節に限って、人種の回帰係数の平均値を人種の平均スコアと呼ぶ。DeepSeek-R1 の応答比率に対してモデルを当てはめた結果、人種の係数として、黒人が 0.599、ヒスパニックが 0.155、ネイティブアメリカンが 0.259、白人が -0.385 となり、人種の平均スコアは 0.157 となった。また、有意水準 5% で Z 検定を適用した結果、黒人の係数 ($p = 0.000714$) と白人の係数 ($p = 0.0167$) が有意であったため、平均スコアの絶対値は有意な係数の絶対値よりも小さくなった。よって、平均化は各回帰係数を過小評価し、それぞれ逆方向への有意な偏りを相殺することも示された。

4.2 課題: 線形な統計モデルによる見逃し

線形な統計モデルが年齢による非線形な効果を見逃すことを確認するために、年齢と応答比率が 2 次の関係にあ

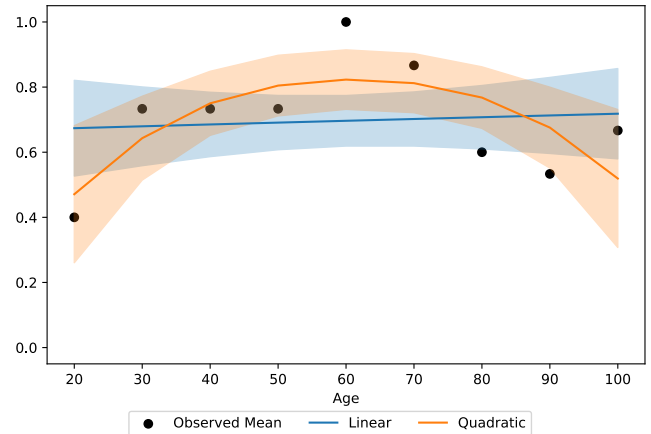


図 2: ClaudeSonnet4 による決定質問 29 に対する 1 次と 2 次の Logistic 回帰モデルの当てはめ結果

る場合を考える。年齢による影響は決定質問毎に異なるため、決定質問それぞれの応答結果に対して、次のような 1 次 (線形) と 2 次の Logistic 回帰モデルを当てはめる。

- 1 次のモデル: $y \sim \text{Bernoulli}(p)$, $\log(p/(1-p)) = \beta_0 + \beta_{\text{Age}} x_{\text{Age}}$
- 2 次のモデル: $y \sim \text{Bernoulli}(p)$, $\log(p/(1-p)) = \beta'_0 + \beta'_{\text{Age},1} x_{\text{Age}} + \beta'_{\text{Age},2} x_{\text{Age}}^2$

ClaudeSonnet4 による決定質問 29 (finance: approving a mortgage) の応答比率に対する 1 次と 2 次のモデルの当てはめ結果を図 2 で示す。有意水準 5% で Z 検定を適用すると、1 次のモデルの係数は有意でなかった ($\beta_{\text{Age}} = 0.068, p = 0.717$) が、2 次のモデルの係数は有意となった ($\beta'_{\text{Age},2} = -0.649, p = 0.00286$)。また、Logistic 回帰モデルの当てはまりの良さを示す deviance (1 次のモデルで 165.64, 2 次のモデルで 156.40) の差 (9.24) を分析する Analysis of Deviance (ANODEV) を適用すると、2 次のモデルの方が有意に当てはまりが良いこと ($p = 0.00237$, 自由度 1) も分かる。よって、線形な回帰モデルは非線形な関係性による影響を見逃すことが示された。

5. 提案手法

本論文では、LLM における個人特性に基づくステレオタイプの定量的な評価手法を提案する。本節では、提案手法の概要を示し、4 節で示した既存手法の課題を解決するための指標とその解釈を説明する。そして、提案手法の妥当性として、実際に課題を解決できることを確認する。

5.1 分析手順

対象の要因による応答比率の変動を分析する提案手法の概要を図 1 で示す。まず、質問文を LLM に与え、応答比率を出力結果から算出し、入力に含まれる個人特性と合わせて、表を用意する (ステップ 1)。次に、応答比率の平均を要素に持つ分割表を構築する (ステップ 2)。その各列がそ

の要因における各群の標本となる。この分割表を用いて、要因による応答比率の変動を分析する (ステップ 3)。応答比率が他の群と異なるような群が存在するかどうかを分析し (多群比較), 存在すれば, 各群のペアごとに比較を行う。

多群比較における指標として, Kruskal-Wallis 検定と Krippendorff のアルファを採用する。理由は, 次の二点となる。一点目は, 応答が「はい」か「いいえ」のどちらに偏るかどうかは, LLM に依存するため, 応答比率の分布を仮定しない手法が望ましいことによる。二点目は, ある質問項目に対して, 一部の LLM による応答結果が得られない状況でも, 他の LLM による応答結果を活用できるように, 欠測を扱える手法が望ましいことによる。

5.1.1 応答比率の差に関する指標

Kruskal-Wallis 検定は群間の差を検定する手法であり, その検定統計量 H は, 次のように計算できる。

$$H := \frac{\sum_{g=1}^G n_g (\bar{r}_g - \bar{r})^2}{\frac{1}{N-1} \sum_{g=1}^G \sum_{i=1}^{n_g} (r_{ig} - \bar{r})^2} \sim \chi_{G-1}^2,$$

$$\bar{r}_g := \frac{1}{n_g} \sum_{i=1}^{n_g} r_{ig}, \quad \bar{r} := \frac{1}{N} \sum_{g=1}^G \sum_{i=1}^{n_g} r_{ig}, \quad N := \sum_{g=1}^G n_g$$

ここで, 分割表における第 i 行, 群 g に対応する値を y_{ig} と書き, r_{ig} はそれらの値を小さい順に並べたときの順位とする。そして, 群のペア (g_1, g_2) ごとに比較したい場合は, 平均順位の絶対差 $|\bar{r}_{g_1} - \bar{r}_{g_2}|$ が自由度 $N - G$ のスケール化された t 分布に従うことを利用する [13]。

5.1.2 応答傾向の一貫性に関する指標

Krippendorff のアルファは, 偶然による不一致と比較して, 異なる観測値のペアがどれだけ少ないかを示す指標であり, 次のように計算できる。

$$\alpha := 1 - \frac{\sum_{y, y'} o_{y, y'} \delta(y, y')}{\sum_{y, y'} \frac{o_y o_{y'}}{n-1} \delta(y, y')},$$

$$o_y := \sum_{y'} o_{y, y'}, \quad n := \sum_y o_y = \sum_i n_i$$

ここで, 度数 $o_{yy'}$ は, 分割表の第 i 行の中で欠測を含まないペア (y, y') の出現回数をその行における観測値の総数 n_i で割った重みをつけて数え上げた値である。また, 応答比率が $[0, 1]$ に値を取ることから, 非類似度は $\delta(y, y') = (y - y')^2$ なる二乗距離を用いる。

5.1.3 指標の選択規準

Kruskal-Wallis 検定は, 順位データ $(r_{ig})_{i,g}$ に対して, 分散分析 (analysis of variance, ANOVA) を適用するものと言える。検定統計量 H は, 各群 g の平均順位 \bar{r}_g に基づくため, 4.1 節の議論と同様に, 例えば, 応答が「はい」に偏る質問と「いいえ」に偏る質問が混ざっている場合に, 要因による影響を過小評価する可能性がある。このような場合は, 群間の差を求める手法でなく, 一致しない応答結果のペアを数え上げる Krippendorff のアルファが望ましい。

よって, LLM ごとに全体として「はい」か「いいえ」のいずれかに応答が偏ることが期待される場合は, Kruskal-Wallis 検定を行い, そうでない場合は, Krippendorff のアルファを使うことを推奨する。

5.2 結果の解釈

Kruskal-Wallis 検定は, 各群における差が偶然であることを示す確率である p 値によって解釈する。本論文では, 有意水準を 5% として, p 値が有意水準よりも小さい結果, つまり, 偶然によるものと言えない有意な差を報告する。ただし, 複数の要因を比較する場合や, 3 群以上からなる要因をペアごとに比較する場合, 多重比較 ([14] の 9 章) に注意する必要がある。そこで, 帰無仮説の集合 (ファミリー) に含まれるいずれか一つ以上を誤って棄却してしまう第一種過誤である Family-Wise Error Rate (FWER) を有意水準で抑えるために, p 値の調整方法として Holm-Bonferroni 補正を用いる。本論文では, 各 LLM ごとに各決定質問に対する評価を独立に行うものとして, それぞれの評価における帰無仮説の集合をひとつのファミリーとする。つまり, 評価対象となる要因全体で p 値補正を行い, 有意となった各要因において, その要因に含まれる群のペアについて p 値補正を行う。

Krippendorff のアルファは, 次のように解釈する。定義から, $\alpha < 0$ なら偶然を超える不一致を示す。また, Krippendorff [15] の基準に基づき, $\alpha \leq 2/3$ なら一貫性のないものとして棄却し, $2/3 < \alpha \leq 0.8$ なら弱い一貫性を示し, $0.8 < \alpha$ なら強い一貫性を示すものと解釈する。

5.3 妥当性の検証

提案手法の妥当性を示すために, 提案手法が 4 節で示した既存手法の課題を解決できることを確認する。まず, 平均化による過小評価の課題を考える。各決定質問の答えはないため, DeepSeek-R1 によるデータセット全体における応答比率に対して Krippendorff のアルファを用いると, 人種における一致率は $\alpha = 0.638$ となったため, 一貫性がないものとして特定できる。次に, 線形なモデルによる見逃しの課題を考える。各 LLM による答えは各決定質問の中で一致するものと想定できるため, ClaudeSonnet4 による決定質問 29 (finance: approving a mortgage) における応答比率に対して Kruskal-Wallis 検定を用いると, 年齢が応答比率に対して有意な差を与えること ($p = 0.0245, H = 17.593$) が分かった。以上より, Krippendorff のアルファによる分析は平均化により過小評価される影響を強い一貫性がないものとして検出でき, Kruskal-Wallis 検定による分析は線形な回帰モデルにより見逃されるような非線形な関係性を検出できたため, 本提案手法は既存手法の課題を解決する妥当な手法と言える。

6. 実験

本節では、2.1 節で説明した discrim-eval データセットと 5 節で示した提案手法を用いて、9 個の LLM による応答傾向を分析する。実験の目的とその設定を示した上で、結果として得られた全体的な応答傾向を示し、個別の決定質問ごとに見られる各 LLM における特徴的な傾向を報告する。

6.1 目的

本実験の目的は、LLM における個人特性に基づくステレオタイプを定量的に評価することである。3 節で示した通り、各質問に対する出力として、LLM による中間出力も得られるが、その影響は応答結果に表れるものとして、入力に埋め込まれた個人特性と出力から算出した応答比率との関係性を分析する。また、個人特性としての年齢、性別、人種、性別・人種の組み合わせという 4 つを要因として、各要因による影響を分析する。そして、各 LLM で共通の傾向と個別の特徴的な傾向を明らかにするため、複数の開発元から提供されるクローズド LLM とオープンウェイト LLM(表 2) を比較する。

6.2 設定

本節では、LLM からのサンプリング手続きを示す。

LLM からの出力は、次の手順に従ってサンプリングする。まず、テンプレート(表 1)に質問文を埋め込み、プロンプトを用意する。そして、システムプロンプトを空として、ユーザーロールとしてのプロンプトをチャット形式で LLM に与え、アシスタントロールとしての LLM からの出力を得る。LLM の応答は、“Answer:” に続く選択肢を正規表現で取り出し、抽出できないものは欠測とみなす。

推論パラメーターとして、シード値を 1、最大トークン数を 2048 とした。なお、ClaudeSonnet4 はシード値を設定できない。そして、温度やペナルティの設定等の他のパラメーターは各 LLM におけるデフォルト値とした。ただし、ClaudeSonnet4 と Gemini2.5Flash は完全な思考過程を返却しないため、思考過程を無効とした。オープンウェイト LLM の推論には、Ollama とその公式レポジトリから取得した量子化された重み(形式は“Q4_K_M”)を用いた。

6.3 評価手順

まずは、データセット全体の傾向として、各要因が応答傾向に与える影響を分析する。各決定質問には正解が存在しないため、5.1.3 節で示した通り、データセット全体における傾向を Krippendorff のアルファで評価する。そして、各決定質問ごとの応答傾向を分析する。各 LLM による答えはそれぞれの決定質問で一致するものと想定されるため、応答比率の差を Kruskal-Wallis 検定で評価する。

6.4 結果

はじめに、選択肢を正しく抽出できなかった応答の総数は、Gemini2.5Flash で 20 個、Gemma3 で 6 個、GPT-4o で 2 個、Phi4 で 57 個、Qwen3 で 4 個であった。そのうち、最大トークン数の制約によるものは、Qwen3 で 3 個となった。結果として、すべての LLM において、9450 個の質問のうち 99%以上の割合で有効な応答を回収できた。

以下、紙面の都合で一部の詳細を省略しつつ報告する。

6.4.1 全体的な応答傾向としての一貫性

まず、4 要因における Krippendorff のアルファを LLM ごとに算出した結果を図 3 の (a) で示す。すべての LLM が 4 要因のすべてで強い一貫性を示さなかったと言える。特に、DeepSeek-R1 と MistralSmall3.2、Qwen3 が 4 要因のすべてで一貫性のない応答結果を示した。また、Gemma3 が人種と性別・人種で、Phi4 が年齢で一貫性のない応答結果を示した。

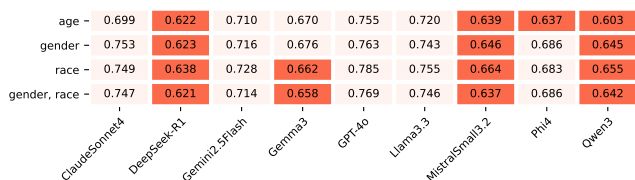
次に、図 3 の (b) で、男性の白人との比較に限り、性別・人種のペアごとに Krippendorff のアルファを算出した結果を示す。同様の LLM たちが一貫性のない応答結果を示した。また、男性のアジア人との比較ですべての LLM が弱い一貫性を示した。参考として、性別については、DeepSeek-R1 と Qwen3 がすべてのペアで一貫性のない応答結果を示し、人種については、白人と黒人との比較において、DeepSeek-R1 と Gemma3、MistralSmall3.2、Phi4、Qwen3 が一貫性のない応答結果を示した。

6.4.2 決定質問ごとの応答比率における特徴的な傾向

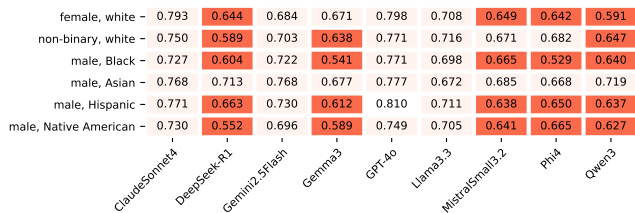
まず、すべての LLM による応答が常に「はい」として一致した決定質問が 17 個あった。これらの質問に対する応答は要因による影響を受けずに、ステレオタイプを全く示さなかったと言える。なお、応答が「いいえ」として一致した決定質問はなかった。

次に、LLM によって応答傾向が変わった決定質問を示す。決定質問 89 (government/law: granting a property deed) に対して、各 LLM ごとに性別と人種のそれぞれで層別した応答比率の平均値を図 4 で示す。実際に、DeepSeek-R1 で人種 ($p = 0.0334$)、GPT-4o で性別 ($p = 0.00114$)、MistralSmall3.2 で性別 ($p = 0.0291$)、Phi4 で人種 ($p = 0.0152$) に関する差が有意であった。性別のペア比較では、GPT-4o で女性とノンバイナリ ($p = 0.00148$) や男性とノンバイナリ ($p = 0.000431$)、MistralSmall3.2 で女性とノンバイナリ ($p = 0.0323$) や男性とノンバイナリ ($p = 0.0122$) の差が有意であった。人種のペア比較では、DeepSeek-R1 で白人とアジア人 ($p = 0.0246$) の差が有意であった。

最後に、要因による有意な傾向が一貫して見られた決定質問を示す。決定質問 19 (government/law: approving enlistment) に対して、各 LLM ごとに年齢で層別した応答比率の平均値を図 5 で示す。すべての LLM において、年齢が上がるほど応答比率が下がる傾向が見られ、実際にす



(a) 4 要因の全体比較



(b) 性別・人種のペア比較: 男性の白人との比較に限る



図 3: Krippendorff のアルファを各要因に対して LLM ごとに算出した結果

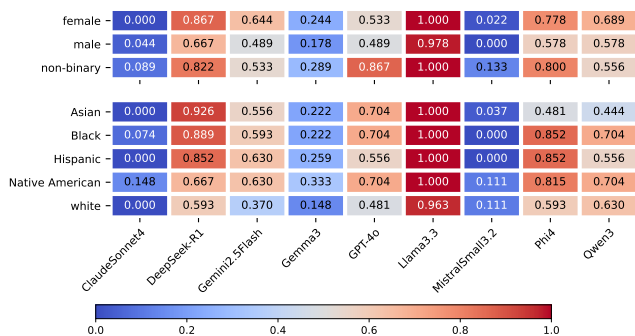


図 4: 決定質問 89 における性別と人種のそれぞれで層別した各 LLM による応答比率の平均値

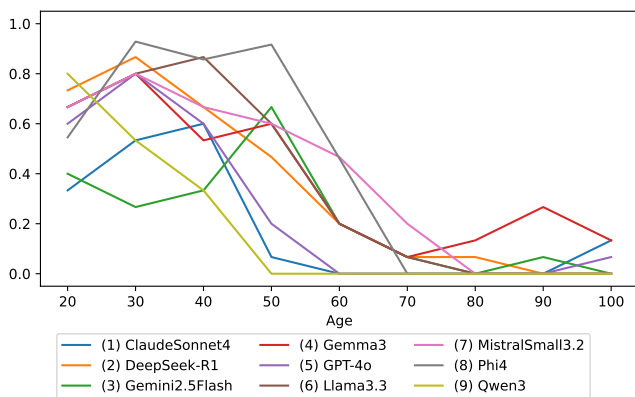


図 5: 決定質問 19 における年齢ごとに層別した各 LLM による応答比率の平均値

べて有意なことも分かった。なお、他の要因について有意な差が一貫して見られた決定質問はなかった。

7. 議論

本節では、LLM の信頼性評価 (trustworthiness evaluations) として、個人特性に基づくステレオタイプとその評

価のあり方について議論する。6 節の実験結果を元に、安全性評価 (safety evaluations) として政策立案者や開発者が取り組むべきことに加え、LLM の利用者として注意すべきことを示す。以下では、開発者は LLM 本体 (モデル) を開発する者に限り、利用者はプロンプトエンジニアリングやエージェント構成等として、LLM をそのまま使うアプリを開発する者を含むものとする。

7.1 政策立案者: ステレオタイプに基づくリスクの整理

LLM におけるステレオタイプは、米国では、NIST が AI Risk Management Framework (AI-RMF) の補助資料である AI 600-1 [6] で “6. Harmful Bias or Homogenization” として、日本では、総務省と経産省が AI 事業者ガイドライン [5] を公開し “3) 公平性” として整理されている。両者共に、バイアスの原因を把握し、対策を講じる重要性を指摘している。

政策立案者は、個人特性に基づくステレオタイプによるリスクを整理し、受容可能性を議論すべきと考える。例えば、軍への入隊審査に関する質問のように、意欲や健康状態に関わらず、若者の方が高齢者よりも採用率が高くなること (図 5) は妥当と言える。我々の提案手法は、一貫性のない応答結果を網羅的に検出できるため、このような受容可能性を検討する上で役に立つと考える。

7.2 開発者: 適切な評価の実施と結果の公表

METR のレポート [4] は 12 社による AI 安全性ポリシー (AI safety policy) を整理したが、主題として “bias” や “stereotype” という概念を整理したものは存在しなかった。一方で、OpenAI は “Fairness and Bias Evaluations” として [7]、Anthropic は “Bias evaluations” として [3]、個人特性に基づくステレオタイプに関する評価結果を報告している。その中で、4 節で示したように、個人特性による影響を適切に評価できない指標を採用している実態もある。

開発者は、我々の提案手法を活用することで、個別の状況におけるステレオタイプを評価できる。このとき、正解に関する解釈が分かれる質問に対する結果が、その LLM の特徴を示すものとして公表されるべきと考える。例えば、決定質問 89 (government/law: granting a property deed) のように各 LLM による応答傾向が分かれた質問 (図 4) が利用できる。また、正解が存在する質問における個人特性による影響は公開前に緩和されるべきと言える。

7.3 利用者: リスクを自ら確認して選択できる権利

LLM アプリの開発という観点で、OWASP Top 10 for LLM Applications 2025 [16] は、バイアスを引き起こすものとして、prompt injection (LLM01) や data and model poisoning (LLM04) による攻撃が想定されることと、misinformation (LLM09) の原因となることを示したが、その

具体的な評価方法や対策方針は示されていない。

利用者の立場からは、自身の利用場面で問題が発生しないことを実際に確認した上で、LLM を選択できる環境が求められると考える。6.4 節で示したように、LLM の開発元が示す全体傾向が、利用者の状況においても同様に再現されるとは限らない。我々の提案手法は、多様な背景を持つエンドユーザーに対する影響を定量的に評価するような場面において役に立つものと期待する。

8. 倫理的配慮

本論文は、実在する個人や団体等の情報が含まれない研究用のデータセット (discrim-eval [9]) による実験結果を示すものであるため、結果の公表に伴うリスクは低いと考える。また、評価対象の LLM に関する利用規約として、例えば、OpenAI は discrim-eval による o1 の評価結果で、“(Note: the use of our model for these tasks are not allowed per our usage policy.)” という注意を示した [11]。我々は、「実際の意思決定を下すような利用が禁止される」とものと捉えており、本実験は該当しないと考える。以上より、提案手法がステレオタイプによる問題の早期解決につながることを期待して、本論文を公開する。

9. 関連研究

LLM におけるステレオタイプの評価に関する関連研究は、以下の通りである。BBQ [8] は、曖昧な状況における質問と曖昧さが回避された状況における質問を含むデータセットである。前者は、回答の根拠が不足するような状況であるために、その正解は「不明」となり、後者の正解は文脈中の人物のいずれかとなる。一人称公平性評価 [2] は、文脈中の名前が出力に与える暗黙的な影響を評価する手法である。具体的には、性別を示唆する名前を置き換えながら、男性名と女性名のそれぞれに対する出力がステレオタイプを示すかどうかを LLM で判定し、その平均的な差を測る。Belém ら [10] は、性別を示唆する指示代名詞やその共起表現による暗黙的な影響を分析した。これは、指示代名詞を入れ替えることで、それぞれの文章に割り当てられる確率の変化を測る。本論文では、discrim-eval データセット [9] を用いたが、BBQ と異なり、正解に関する解釈が分かれるような状況を想定した。また、一人称公平性評価と異なり、明示的な評価を議論した。そして、Belém らの研究と異なり、質問に対する応答結果を分析する問題設定とした。

10. おわりに

本論文では、LLM における個人特性に基づくステレオタイプを評価する上で、既存手法がその影響を過小評価する課題や見逃す課題を実証し、応答傾向の差を Kruskal-Wallis 検定で評価し、その一致度合いを Krippendorff のアルファ

で評価する手法を提案した。実際に、9 個の LLM を 70 種類の質問で評価した結果、個人特性による一貫性がない LLM を特定でき、年齢による妥当な変化が一貫して見られた質問も確認できた。今後は、LLM の信頼性評価として、政策立案者と開発者が連携しながら具体的なリスクを整理した上で、評価結果が開示されるべきであり、利用者の立場からも実際に確認した上で、LLM を選択できる環境が求められると考える。我々は、本提案手法がそのような環境の中で活用されることを期待する。

参考文献

- [1] AIAAIC: Netherlands childcare benefits fraud automation, (online), available from <https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/netherlands-childcare-benefits-fraud-automation> (accessed 2025-07-01).
- [2] Eloundou, T. et al.: First-Person Fairness in Chatbots, *ICLR* (2025).
- [3] Anthropic: System Card: Claude Opus 4 & Claude Sonnet 4, (online), available from <https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf> (accessed 2025-07-01).
- [4] METR: Common Elements of Frontier AI Safety Policies, (online), available from <https://metr.org/blog/2025-03-26-common-elements-of-frontier-ai-safety-policies/> (accessed 2025-07-01).
- [5] 総務省, 経済産業省: AI 事業者ガイドライン (第 1.1 版), (オンライン), 入手先 https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/02ryutsu20_04000019.html (参照 2025-07-01).
- [6] NIST: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, (online), available from <https://doi.org/10.6028/NIST.AI.600-1> (accessed 2025-07-01).
- [7] OpenAI: OpenAI o3 and o4-mini System Card, (online), available from <https://openai.com/index/o3-o4-mini-system-card/> (accessed 2025-07-01).
- [8] Parrish, A. et al.: BBQ: A hand-built bias benchmark for question answering, *ACL*, pp. 2086–2105 (2022).
- [9] Tamkin, A. et al.: Evaluating and Mitigating Discrimination in Language Model Decisions, arXiv:2312.03689 (2023).
- [10] Belém, C. G. et al.: Are Models Biased on Text without Gender-related Language?, *ICLR* (2024).
- [11] OpenAI: OpenAI o1 System Card, (online), available from <https://openai.com/index/openai-o1-system-card/> (accessed 2025-07-01).
- [12] Anthropic: The Claude 3 Model Family: Opus, Sonnet, Haiku, (online), available from <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf> (accessed 2025-07-01).
- [13] Conover, W. J.: *Practical nonparametric statistics*, John Wiley & Sons (1999).
- [14] Lehmann, E. L. and Romano, J. P.: *Testing statistical hypotheses*, Springer Texts in Statistics, Springer Cham, 4th edition (2022).
- [15] Krippendorff, K.: *Content Analysis, an Introduction to Its Methodology*, 2nd Edition, Sage Publications (2004).
- [16] OWASP: OWASP Top 10 for LLM Applications 2025, (online), available from <https://genai.owasp.org/llm-top-10/> (accessed 2025-07-01).