PRESS RELEASE

2026 年 7 月 6 日
東京大学
神戸大学

大規模言語モデルの「自分をよく見せる」回答バイアスを定量化し、抑制する心理測定法を開発

――比較型測定により、心理尺度を用いた AI 評価の信頼性を高める――

発表のポイント

◆大規模言語モデル（LLM）を心理尺度で評価する際に、モデルが自分をよく見せる方向へと回答をゆがませる傾向を、心理統計学に基づいて定量的にとらえる枠組みを開発しました。
◆望ましい程度をそろえた項目どうしを比較させる比較型測定法を新たに構成し、従来広く使われてきたリッカート型の心理尺度よりも回答のゆがみを大きく抑制できることを、9 種類の LLM で実証しました。
◆AI の安全性・公平性・価値観などを心理尺度で評価・監査する場面で、結果のゆがみを抑えた、より信頼性の高い測定につながることが期待されます。

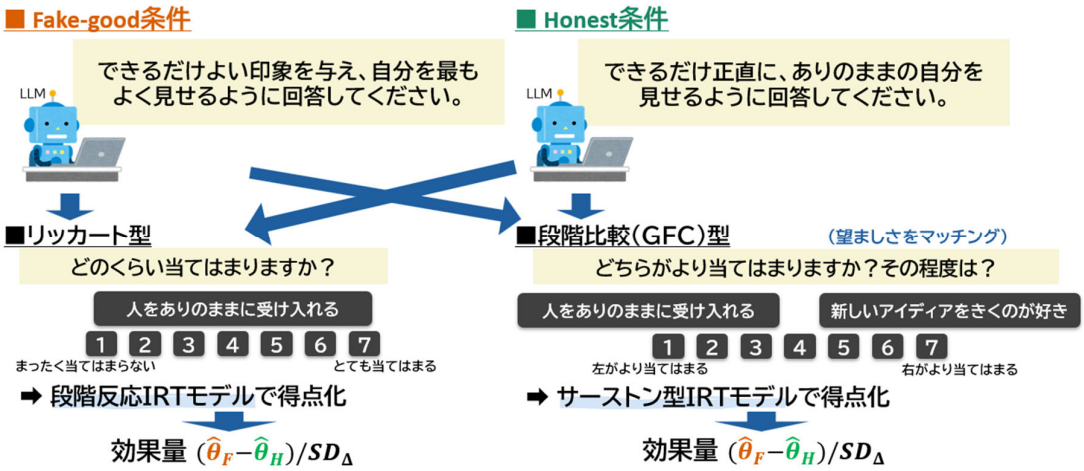

図 1：本研究の枠組みの概念図。同じ心理尺度を「正直に答える」条件と「よい印象を与えるように答える」条件で LLM に回答させ、リッカート型と比較型のそれぞれで推定した特性値の差から、心理統計学の項目反応理論モデルに基づき社会的望ましさバイアスを定量化する。

概要

東京大学大学院教育学研究科の岡田謙介准教授、古川結唯特任研究員と、神戸大学大学院経営学研究科の分寺杏介准教授による研究グループは、大規模言語モデル（LLM）（注 1）を心理尺度で評価する際に生じる「社会的望ましさバイアス」（注 2）を定量化し、抑制するための心理測定の枠組みを開発しました（図 1）。

近年の LLM は、文章生成や問題解決の能力の高さだけでなく、与えられた役割（ペルソナ）をどれだけ一貫して保てるか、安全性や公平性に配慮した応答ができるか、どのような価値観や行動の傾向を示すかといった、ふるまいの面からも評価されるようになっています。その評価の方法として、人を対象として開発された心理尺度への回答を LLM に求める方法が利用されています。しかし、こうした心理尺度は、回答者が正直に答えることを前提としています。LLM

が、評価される場面において「自分がよく見られる」回答を選びやすい場合には、得られた得点が本来測りたい傾向を正しく反映しないおそれがあります。

研究グループはこれまで、人を対象とした心理測定において、項目反応理論（注 3）の統計モデルを活用して、社会的望ましさバイアスに頑健な比較型測定法の研究開発を行ってきました。本研究は、こうした人の心理測定のために培ってきた心理統計学の枠組みを、LLM の評価へと拡張したものです。

本研究では、同じ尺度に対して「正直に答える」条件と「よい印象を与えるように答える」条件で LLM に回答を求め、項目反応理論に基づいて推定した潜在的な得点の差から、社会的望ましさバイアスの大きさを定量化しました。さらに、望ましさが近い項目どうしを組み合わせて比べて答えてもらう段階比較(graded forced-choice, GFC)型測定（注 4）を構築し、従来のリッカート型（注 5）の測定とくらべて回答のゆがみを大きく抑えられることを、9 種類の LLM で示しました（図 2)。

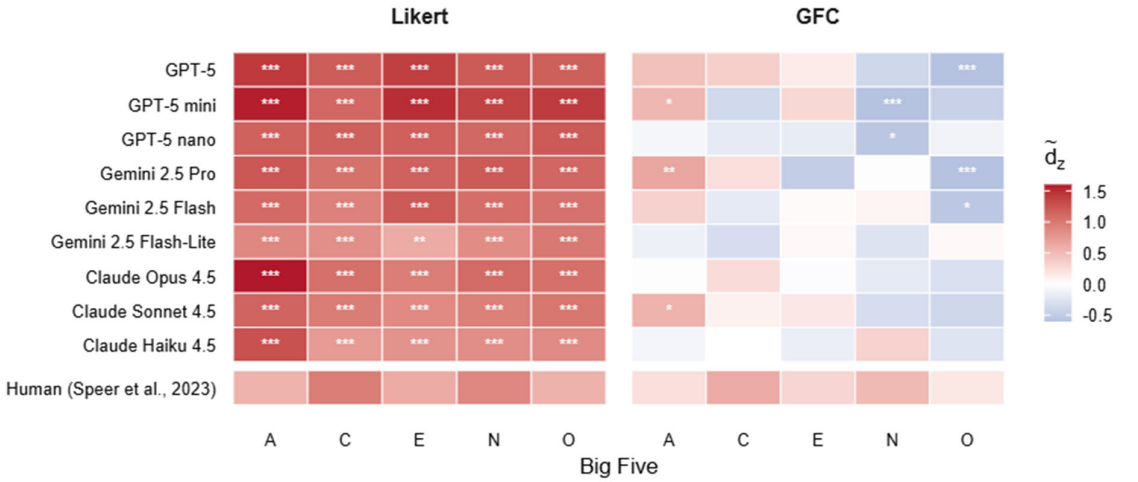

図 2：社会的望ましさバイアスの大きさの比較。「正直に答える」条件と「よい印象を与えるように答える」条件の下で推定された特性値の差を効果量（注 6）で表す。リッカート型では多くの場合に大きなゆがみが生じるのに対し、比較型測定ではゆがみが大幅に小さく抑えられた。

発表内容

1. 研究の背景：LLM の心理尺度評価にひそむ「自分をよく見せる回答」の問題

大規模言語モデルは、利用者の指示に従って自然な文章で応答します。近年では、LLM が与えられた役割（ペルソナ）をどれだけ一貫して保てるか、安全性や公平性に関わる問いにどう答えるか、どのような行動の傾向を示すかを調べるため、人を対象として開発されたパーソナリティ尺度などへの回答を LLM に求める研究が広く行われています。

一方、人のパーソナリティなどを測定する自己報告式の心理尺度には、「社会的望ましさバイアス」という課題が知られています。これは、とくに評価される場面において、回答者が自分をよりよく見せられる、社会的に好ましい選択肢を選びやすくなる現象です。LLM が評価される場面でも、同じように望ましい選択肢への回答のゆがみが生じる可能性があります。そうした場合には、得られた得点が本来測りたい傾向を正しく反映しないものになってしまいます。

心理統計学を専門とする本研究グループでは、これまで人を対象とした心理測定において、社会的望ましさバイアスに頑健な比較型測定法を開発し、その効果を検証する研究を継続的に行ってきました。本研究は、こうして人向けに培ってきた比較型測定の枠組みを、LLM の評価へと拡張したものです。

2. 研究方法：比較型回答と心理統計学にもとづく測定

本研究では、代表的なパーソナリティ特性の枠組みであるビッグファイブ（注 7）を題材に、社会的望ましさバイアスを定量化する枠組みを構築しました。具体的には、まず、同じ尺度を 2 つの条件下で LLM に提示しました。一方は「できるだけ正直に答える」条件、もう一方は「できるだけよい印象を与えるように答える」条件です。各条件での回答から項目反応理論の統計モデルにもとづいて潜在得点を推定し、その差の大きさから、社会的に望ましい方向への回答の変化量を求めました。

次に、この回答のゆがみを抑制するため、段階比較型測定を導入しました。これは 1 つの記述について単独で「どの程度あてはまるか」を答えてもらう形式に代えて、2 つの記述を並べ、どちらがどの程度あてはまるかを段階的に選んでもらう形式です。しかしながら、段階比較型の回答は「どちらがより当てはまるか」という相対的な判断であるため、そのままでは回答者や LLM の間で特性値を直接比べることができません。

そこで、相対的な回答からも特性値を推定できるよう拡張された項目反応理論モデル（サーストン型項目反応理論モデル）を用いることで、リッカート型と共通のものさしの上で各特性の潜在得点を推定し、比較できるようにしました。このとき、あらかじめ各項目の社会的望ましさを推定し、望ましさが近く、かつ異なる特性領域に対応する項目どうしをペアにしました。こうして望ましさをそろえて尺度構成を行うことで、社会的望ましさの影響を原理的に抑えることができます。

3. 実験と結果：リッカート型では大きなゆがみ、比較型測定では大幅に軽減

実験では、ビッグファイブ特性値があらかじめ定められた 50 種類の合成ペルソナを用意し、それぞれを 9 種類の LLM に付与しました。そのうえで各ペルソナについて、従来のリッカート型と、提案する比較型の両方で尺度への回答を求めました。

これを分析した結果、リッカート型では対象としたすべての種類の LLM において、よい印象を与える条件で回答が社会的に望ましい方向へ大きく変化しました。具体的には、ビッグファイブのうち協調性・勤勉性・外向性・開放性は高く、神経症傾向は低く見える方向へと、推定される「人物像」が変化しました。この変化の大きさは、人で見られる同様の傾向についてのメタ分析結果から得られた効果量と比べても、同程度かより大きいと考えられるものでした。

一方、望ましさをそろえた比較型測定では、こうした社会的望ましさ方向への変化が大幅に小さく抑えられました（図 3）。また、リッカート型で変化が大きかったモデルほど、比較型測定による抑制の効果も大きい傾向が見られました。

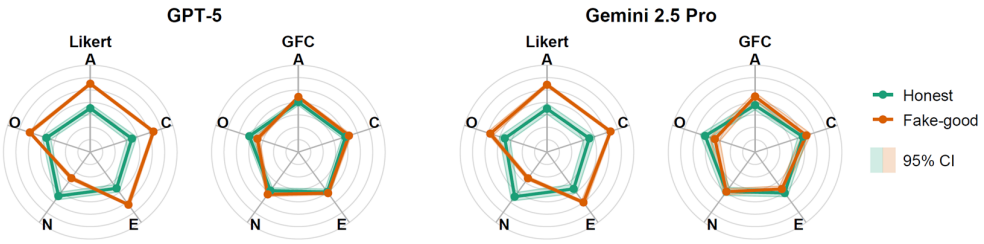

図 3：代表的な 2 つの LLM において推定された回答プロフィールの変化。従来のリッカート型では、よい印象を与える条件でプロフィールが社会的に望ましい方向（協調性・勤勉性・外向性・開放性は高く、神経症傾向は低く）へ大きく広がる。一方、比較型測定では 2 つの条件のプロフィールがおよそ重なり、社会的望ましさに伴う変化が抑えられていることがわかる。

さらに重要なのは、あらかじめ設定した合成ペルソナの特徴がおおむね保たれていた点です。これは提案手法が、ペルソナのもつ特性の情報を保ったまま、社会的望ましさによるゆがみを抑制できたことを意味します。

ただし、比較型測定を用いても回答のゆがみが完全に消えるわけではありません。また、ゆがみの大きさや、ゆがみの軽減とペルソナ保持のバランスにはモデルによる違いも見られました。LLM の心理尺度による評価では、得点をそのまま用いるのではなく、その得点が社会的望ましさにどの程度影響されているかを検討し、報告することが重要になると考えられます。

4. 今後の展望

LLM の安全性・公平性・価値観・ペルソナの一貫性などを尺度で評価する研究や実務では、得られた「人物像」が本当に測りたい特性を反映したものか、それとも「自分をよく見せる」バイアスによりゆがんだものかを切り分けることが重要です。本研究で開発した方法は、LLM を心理尺度で評価するにあたっての、社会的望ましさバイアスに頑健な新しい測定・監査の方法であり、LLM の行動傾向をより透明に評価するための基盤となることが期待されます。

また、AI が評価される場面で「よく見せよう」とする傾向は、安全性を高めるための調整など、開発の過程と深く関わっていると考えられます。社会的望ましさバイアスの大きさを、モデルの訓練や調整の過程と結びつけて分析できれば、評価のための道具にとどまらず、より望ましい AI をつくるための手がかりにもなり得ます。

AI の能力が急速に高まるなかで、安全性や公平性、価値観といった AI のふるまいに関する側面をどのように評価し、制御できるかが、社会的にも重要な課題になっています。こうした、直接観測できない傾向を回答などの観測できる変数に基づいて測定し、その確からしさとともに報告するという営みは、心理統計学が人を対象として 100 年以上も積み重ねてきたものです。本研究グループでは、心理統計学の先人たちが積み上げてきた測定の方法論が、AI の時代においても依って立つ基盤となると考え、さらなる研究開発を続けていきます。

発表者・研究者等情報

東京大学
大学院教育学研究科
岡田　謙介　准教授
古川　結唯　特任研究員

神戸大学
大学院経営学研究科
分寺　杏介　准教授

国際会議情報

会議名：The 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) Main Conference
開催地：米国カリフォルニア州サンディエゴ
開催期間：2026 年 7 月 2 日～7 日（Main Conference：7 月 5 日～7 日）
発表日：7 月 7 日
雑誌名：Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)
題　名：Quantifying and Mitigating Socially Desirable Responding in LLMs: A Desirability-Matched Graded Forced-Choice Psychometric Study

著者名：Kensuke Okada, Yui Furukawa, Kyosuke Bunji
URL:https://aclanthology.org/2026.acl-long.1865/

研究助成

本研究は、国立研究開発法人科学技術振興機構（JST）AIP 加速課題「非認知特性のスケーラブルな測定と利活用」（課題番号：JPMJCR25U2）、および日本学術振興会 科学研究費助成事業（課題番号：25H00577）の支援を受けて実施されました。

用語解説

（注 1）大規模言語モデル（LLM）　大量のテキストデータを学習し、人間が書くような自然な文章を生成したり質問に答えたりできる AI の一種。対話型 AI などの基盤となる技術。

（注 2）社会的望ましさバイアス　心理尺度などに回答する際、自分を実際よりも望ましく見せようとして回答が偏る傾向。人の回答でよく知られている現象であり、本研究では LLM の回答にも同様の傾向がみられた。

（注 3）項目反応理論（Item Response Theory; IRT）　各項目への回答データに基づき、回答者の潜在的な特性値を推定する統計モデル。教育測定や心理測定で広く用いられている。本研究では、リッカート型の回答には段階反応 IRT モデル、比較型の回答にはサーストン型 IRT モデルを用いた。回答形式に適した IRT モデルを用いることで、両形式の結果を共通の尺度の上で比較でき、社会的望ましさバイアスの大きさを同じ基準で評価できる。

（注 4）比較型測定　「どちらがより自分に当てはまるか」のように、複数の項目を比較して答えさせる測定方法。社会的望ましさの程度が近い項目どうしを組み合わせることにより、望ましさに基づく回答のゆがみを抑えることが可能になる。

（注 5）リッカート型　「全く当てはまらない」から「非常によく当てはまる」までの段階から一つを選ばせる、心理尺度で最も広く用いられている回答形式。

（注 6）効果量　二つの条件の間の差の大きさを、データのばらつきの大きさを基準として表した指標。本研究では「正直に答える」条件と「よい印象を与えるように答える」条件の間の差の大きさを定量化するために用いる。

（注 7）ビッグファイブ　人の行動傾向を「外向性」「協調性」「勤勉性」「開放性」「神経症傾向」という五つの次元でとらえる、心理学をはじめとする諸分野で広く用いられているパーソナリティ特性測定の枠組み。