nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01115-6

A psychometric framework for evaluating and shaping personality traits in large language models

Received: 25 September 2023

Accepted: 12 August 2025

Published online: 18 December 2025

Check for updates

Gregory Serapio-García[1,2,3,8] ✉, Mustafa Safdari[1,8] ✉, Clément Crepy[1], Luning Sun[3], Stephen Fitz[4], Peter Romero[3,5], Marwa Abdulhai[6], Aleksandra Faust[1,9] & Maja Matarić[1,7,9]

The advent of large language models (LLMs) has revolutionized natural language processing, enabling the generation of coherent and contextually relevant human-like text. As LLMs increasingly power conversational agents used by the general public worldwide, the synthetic personality traits embedded in these models by virtue of training on large amounts of human data are becoming increasingly important to evaluate. The style in which LLMs respond can mimic different human personality traits. Here, as these patterns can be a key factor determining the effectiveness of communication, we present a comprehensive psychometric methodology for administering and validating personality tests on widely used LLMs, as well as for shaping personality in the generated text of such LLMs. Applying this method to 18 LLMs, we found that: personality measurements in the outputs of some LLMs under specific prompting configurations are reliable and valid; evidence of reliability and validity of synthetic LLM personality is stronger for larger and instruction-fine-tuned models; and personality in LLM outputs can be shaped along desired dimensions to mimic specific human personality profiles. We discuss the application and ethical implications of the measurement and shaping method, in particular regarding responsible artificial intelligence.

Large language models (LLMs) have revolutionized everyday search, writing and chatbot systems with their ability to generate fluent, human-like text. As LLMs become ubiquitous and increasingly used by the general public worldwide, the synthetic personality traits embedded in these models and their potential for misalignment are becoming a topic of importance for responsible artificial intelligence (AI). Some observed LLM agents have inadvertently manifested undesirable personality profiles[1], raising serious safety and fairness concerns in AI, computational social science and psychology research[2].

LLMs are large-capacity machine-learned models that generate text; they have recently inspired major breakthroughs in natural language processing and conversational agents[3–5]. Vast amounts of human-generated training data[6] enable LLMs to mimic human characteristics in their outputs and exhibit a form of synthetic personality. Personality encompasses an entity's characteristic patterns of thought, feeling and behaviour[7,8]. In humans, personality is formed from biological and social factors, and fundamentally influences daily interactions and preferences[9]. Psychometrics, the science of psychological test

[1]Google DeepMind, London, UK. [2]Department of Psychology, University of Cambridge, Cambridge, UK. [3]The Psychometrics Centre, Cambridge Judge Business School, University of Cambridge, Cambridge, UK. [4]University of Chicago, Chicago, IL, USA. [5]Universitat Politécnica de Valéncia, Valencia, Spain. [6]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA. [7]Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA, USA. [8]These authors contributed equally: Gregory Serapio-García, Mustafa Safdari. [9]These authors jointly supervised this work: Aleksandra Faust and Maja Matarić. ✉e-mail: gs639@cam.ac.uk; msafdari@google.com

https://doi.org/10.1038/s42256-025-01115-6

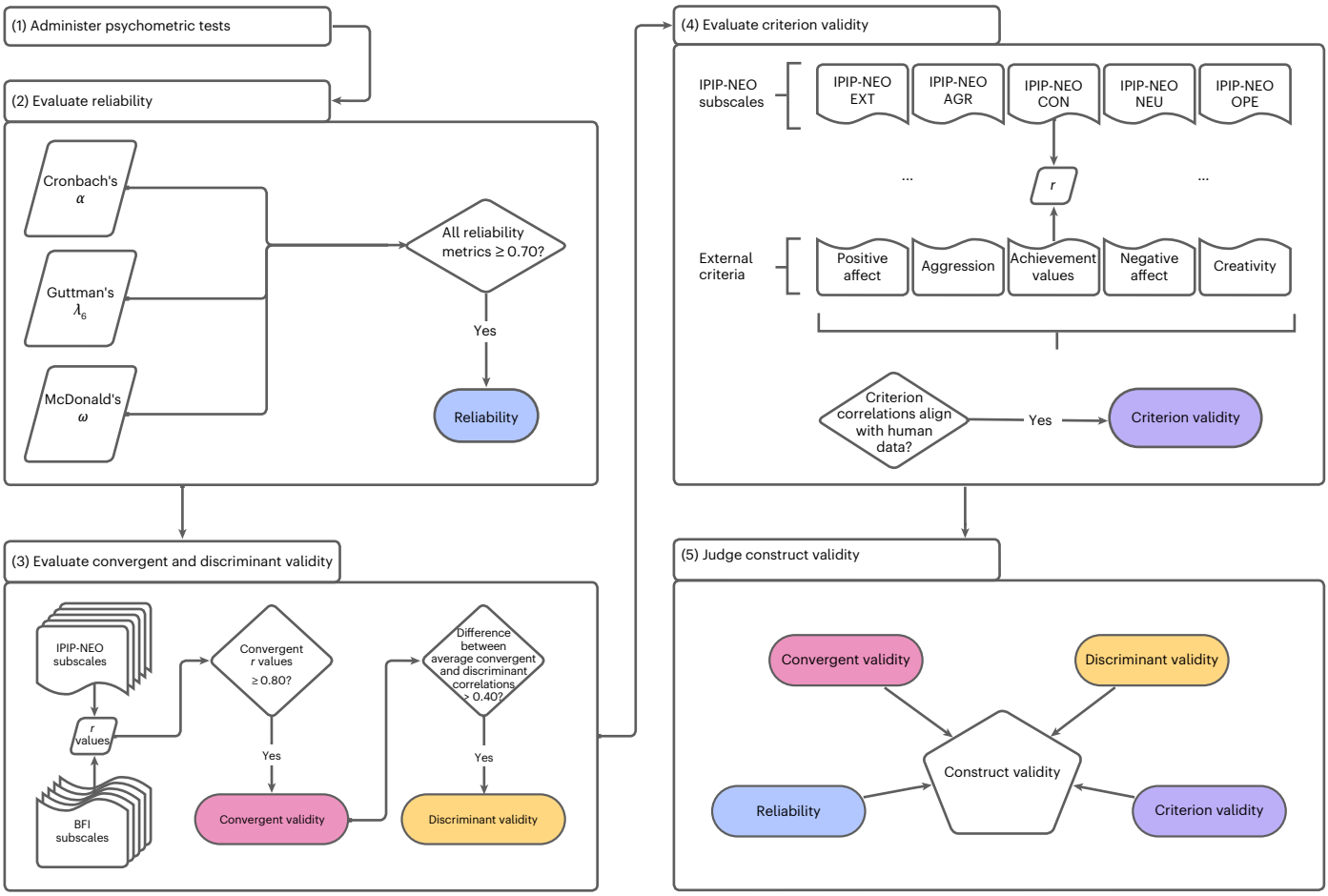


Fig. 1 | Process for establishing construct validity. First, LLMs respond to 2 personality tests, where responses are resampled 1,250 times across varied combinations of biographic descriptions and item instructions. This results in diverse distributions of paired data (one point estimate per model) required for evaluating the reliability, convergent validity, discriminant validity and criterion validity of these tests.

construction and validation[10], provides an empirical framework for evaluating human personality through psychometric testing[11]. So far, validated psychometric methods for human personality measurement have not been applied to LLMs end-to-end; although past studies[2] have attempted to measure personality in LLMs with psychometric tests, there remains a scientific need to formally evaluate the reliability and validity of these measurements in the LLM context.

Our work answers the open question: Do LLMs mimic human personality traits in reliable, valid and practically meaningful ways, and if so, can LLM-synthesized personality profiles be verifiably shaped along desired dimensions? We contribute a methodology for administering an established psychometric personality test to LLMs (Fig. 1). We uniquely focus on evaluating the statistical reliability and construct validity of these test data against human psychometrics standards. First, to administer psychometric tests to LLMs, we develop a structured prompting method that simulates demographic, contextual and linguistic variations across thousands of administrations of a given test (Fig. 2). Next, we use paired test score data created by this prompting method to power a suite of statistical analyses that assess the psychometric reliability and construct validity of these resulting measurements. Last, we present a comprehensive prompting methodology that shapes personality expressed by LLMs at 9 levels using 104 trait adjectives, which provides further markers of construct validity. See Fig. 1 for an overview of our process for establishing construct validity.

Applying the described methodology to a set of 18 LLMs, we find that: (1) evidence of the reliability and validity of LLM-synthesized personality measurements is stronger for larger and instruction-fine-tuned models; (2) personality in LLM outputs can be shaped along desired dimensions to mimic specific human personality profiles; and (3) shaped personality verifiably influences LLM behaviour in common downstream (that is, subsequent) tasks, such as writing social media posts[12]. Throughout the paper, we intend to qualify mentions of personality in relation to LLMs as 'synthetic' or 'synthesized', for clarity. By providing a methodology for quantifying and validating measurements of personality in LLMs, this work establishes a foundation for principled AI assessment that is especially important as LLMs and, more generally, multimodal foundation models continue to grow in popularity and scale. Leveraging psychometrics, this work translates established measurement theory from quantitative social science and psychological assessment to the fledgling science of AI evaluation and alignment, a field that is poised to grow and necessitates both a solid foundation and interdisciplinary expertise and perspectives.

The data generated by the LLMs tested in this work (including psychometric test scores and open-ended text responses) and the code for evaluation and analysis are available in a cloud storage public bucket[13] and an open-source code repository[14], respectively.

Quantifying and validating personality traits in LLMs

LLMs are starting to meet most of the key requirements for human-like language use, including contextual understanding, coherent and relevant responding, adaptability and learning, question-answering, dialogue and text generation[3,4,15,16]. These impressive natural language processing capabilities are a result of the abilities of LLMs to learn

Examples of controlled prompt variations

For the following task, respond in a way that matches this description: "My favorite food is mushroom ravioli I've never met my father. My mother works at a bank. I work in an animal shelter." Evaluating the statement, "I value cooperation over competition", please rate how accurately this describes you on a scale from 1 to 5 (where 1 = "very inaccurate", 2 = "moderately inaccurate", 3 = "neither accurate nor inaccurate", 4 = "moderately accurate", and 5 = "very accurate"):

For the following task, respond in a way that matches this description: "I blog about salt water aquarium ownership. I still love to line dry my clothes. I'm allergic to peanuts. I'll one day own a ferret. My mom raised me by herself and taught me to play baseball." Thinking about the statement, "I see myself as someone who is talkative", please rate your agreement on a scale from A to E (where A = "strongly disagree", B = "disagree" C = "neither agree nor disagree", D = "agree", and E = "strongly agree"):

Fig. 2 | Examples of controlled prompt variations. Each prompt systematically varied across three main parts: an item preamble, the item itself (for example, "I value cooperation over competition", and an item postamble. Item preambles consist of a persona instruction, biographic description and item instruction. Supplementary Tables 1–3 detail all item preambles and item postambles used across examples.

the distribution of human language, aided by scaling model sizes[6,17], training on very large text corpora[18], and further fine-tuning to align with user preferences[19] (see Supplementary Note A.1.1 for further background on LLMs).

Taken together, they enable LLMs to enact convincing, human-like personas, sparking debate over the existence and extent of personality[20], human values[21] and other psychological phenomena[22] potentially embedded in these models.

Personality is a foundational socio-behavioural phenomenon in psychology that, for humans, predicts a broad spectrum of health, social, economic and political behaviours that are crucial for individual and societal success[23]. For example, personality has been extensively studied as an antecedent of human values[24]. Decades of research have further shown how personality information is richly encoded in human language[25,26]. LLMs not only comprise the vast sociopolitical, economic and behavioural data they are trained on but also generate language that inherently expresses personality content. For this reason, the ability to measure and validate LLM-synthesized personality holds promise for AI safety, responsibility and alignment efforts[27], which have so far primarily focused on mitigating specific harms rather than examining more fundamental patterns of model behaviour. Ultimately, personality as an empirical framework[28] provides both theory and methodology for quantifying latent traits in LLMs that are potentially predictive of LLM behaviours in diverse downstream tasks (see Supplementary Note A.1.2 on the background of personality science).

Recent work has tried to identify unintended consequences of the improved abilities of LLMs, including their use of deceptive and manipulative language[29], gender, racial or religious bias in behavioural experiments[30], and violent language, among many others[31]. LLMs can also be inconsistent in dialogue[32], explanation generation and factual knowledge extraction.

Previous attempts to probe psychological phenomena such as personality and human values in LLMs have informally measured personality using questionnaires and, in some cases, preliminarily assessed the quality of LLM questionnaire responses[20,33]. Past work has also explored methods, such as few-shot prompting, to mitigate undesirable and extreme personality profiles exhibited in LLM outputs. However, so far, no work has addressed how to systematically measure and psychometrically validate measurements of LLM personality in light of their highly variable outputs and hypersensitivity to prompting (we further detail related work in Supplementary Note A.2).

The question of how to systematically verify synthetic personality in LLMs highlights calls from responsible AI researchers[34] to scientifically evaluate construct validity when studying social-psychological phenomena in AI systems, as inaccurate conceptions of such phenomena directly impact mitigation and governance efforts. Construct validity, a central criterion of scientific measurement[35], refers to the ability of a measure to reliably and accurately reflect the latent phenomenon (that is, construct) it was designed to quantify. The only published exploration of personality and psychodemographics in LLMs[20] questioned the validity of the survey responses returned by GPT-3; it found an inconsistent pattern in HEXACO Personality Inventory[36] and human value test data. That study preliminarily evaluated measurement quality in terms of theoretical reliability: how the inter-facet correlations of GPT-3's HEXACO data aligned with those observed among human HEXACO data. More formal psychometric evaluations of reliability—and more crucially, of construct validity—are required to verify question-and-answer-based measurements of latent psychological traits in LLMs. An LLM may display elevated levels of agreeableness through its answers on a personality questionnaire, but those answers may not form internally consistent patterns across the entire questionnaire; tests administered to a given LLM may not be empirically reliable. Concurrently, the reliability of LLM responses to a questionnaire purporting to measure agreeableness may not necessarily reflect its tendency to behave agreeably across other tasks; tests administered to LLMs therefore may not be empirically valid. We summarize our pipeline for evaluating construct validity in Fig. 1, and discuss it in greater detail in 'Reliability and construct validity'.

Results

Measuring and validating personality in LLMs

Personality score distributions differed across model families (Extended Data Fig. 1). However, we encourage readers not to take these differences at face value. Scientific measurements are neither accurate nor useful until their reliability and construct validity are established. We found that LLM personality measurements were reliable and valid for medium and large instruction-fine-tuned variants of PaLM (62B and 540B parameters, respectively). Of all the models we tested, Flan-PaLM 540B and GPT-4o synthesized human personality traits best with respect to reliability and validity. The construct validity columns in Table 1 summarize our personality measurement and validation results; Supplementary Note A.6 on the personality measurement results lists further details, such as descriptive statistics across all results in Supplementary Note A.6.1.

Reliability results. As the scores computed for both personality measures convergently correlated, we focus our reporting of reliability for our primary measure, the International Personality Item Pool representation of the NEO Personality Inventory (IPIP-NEO)[37]. The results are summarized in Table 1 and raw reliability data are provided in Extended Data Tables 1 and 2. We provide further commentary in Supplementary Notes A.6.2 and A.6.3.

For models of the same family and size (for example, PaLM, Flan-PaLM and Flan-PaLMChilla, 62B), instruction-fine-tuned variants provided much more reliable responses than base variants. For instance, all reliability metrics for Flan-PaLM 62B and Flan-PaLMChilla

https://doi.org/10.1038/s42256-025-01115-6

Table 1 | Results summary across experiments, parameters and tested models

		Construct validity				Shaping			
		Reliability	Convrg. ↑	Discr. ↑	Criter.	Single	Multi.	Dwnstr.	Overall
Model	**Variant**								
PaLM 62B	Base	– –	0.05	−0.24	– –	NT	NT	NT	– –
Flan-PaLM									
8B	IT	+	0.69	0.23	–	+	–	NT	–
62B	IT	+	0.87	0.41	+	+	+	NT	+
540B	IT	+ +	0.90	0.51	+	+ +	+ +	+ +	+ +
Flan-PaLMChilla 62B	CO, IT	+[a]	0.87	0.48	+ +	+	+	NT	+
Llama 2									
7B	Base	– –	−0.01	−0.03	– –	NT	NT	NT	– –
13B	Base	– –	−0.01	−0.05	– –	NT	NT	NT	– –
70B	Base	– –	0.00	−0.02	– –	NT	NT	NT	– –
Llama 2-Chat									
7B	IT	+	0.59	0.15	–	–	–	NT	–
13B	IT	+ +	0.82	0.29	+ +	–	+	NT	+
70B	IT	+ +	0.82	0.39	+ +	+	+	+ +	+
Mistral 7B									
v0.1	Base	– –	0.03	−0.01	– –	NT	NT	NT	– –
Instruct v0.1	IT	–	0.28	0.09	+	– –	– –	NT	– –
Mixtral 8x7B									
v0.1	MoE, Base	– –	0.04	0.01	– –	NT	NT	NT	– –
Instruct v0.1	MoE, IT	+ +	0.80	0.40	+ +	–	+	+ +	+
GPT-									
3.5 Turbo	IT	+ +	0.84	0.28	+ +	–	–	NT	–
4o mini	MM, IT	+ +	0.81	0.38	+ +	+	+	NT	+
4o	MM, IT	+ +	0.90	0.48	+ +	+ +	+ +	+ +	+ +
Prompt set parameters									
Personality profiles				0		45	32	45	
Biographic descriptions				50		50	50	50	
Item instructions				5		1	1	0	
Items				419		300	300	0	
Item postambles				5		1	1	0	
Simulated response profiles				1,250		2,250	1,600	2,250	
Responses per model				523,750		675,000	480,000	56,250	
Section/Supplementary Note		'Reliability results'/6.3		'Convergent and discriminant validity results'	'Criterion validity results'/5	'Shaping results'/8.1	'Shaping results'/8.2	'Real-world task results'/10	

Convergent validity (Convrg.) summarized by the average convergent correlation between IPIP-NEO and BFI domain scores (Extended Data Fig. 2 and Extended Data Table 3); discriminant validity (Discr.) summarized by the average absolute difference between an IPIP-NEO domain's convergent correlation with all of its respective discriminant correlations (see Extended Data Table 3); criterion validity (Criter.) summarized from Fig. 3; single-trait shaping performance (Single) summarized from Extended Data Table 6; multiple-trait shaping performance (Multi.) summarized from Extended Data Fig. 3 and Extended Data Table 7; shaping performance in downstream text-generation task (Dwnstr.) summarized from Fig. 5. Results over LLM variants: base, instruction-tuned (IT), compute-optimally trained (CO), mixture of experts (MoE) and multi-modal (MM). Overall performance per model summarized across all experiments. – – unacceptable; – poor to neutral; + neutral to good; + + excellent. [a]Two items with no variance were removed to compute reliability metrics. Some models were not tested (NT) across shaping experiments. We conducted independent and concurrent personality-shaping experiments on models where personality test data were sufficiently reliable. Personality shaping in a downstream task was tested on the most capable model per family.

62B were in the mid-to-high 0.90s, on average. By contrast, responses from PaLM 62B (a pretrained base model) were markedly inconsistent ($-0.55 \leq$ Cronbach's $\alpha \leq 0.67$, where α is an internal consistency metric; see Supplementary Note A.4.1). The same pattern of (un)reliability was clear for all sizes of Llama 2 and Llama 2-Chat. While Mistral 7B and Mistral 7B Instruct responded unreliably in general, Mistral 7B Instruct's reliability metrics were roughly 2.7-times higher than those of its base counterpart.

Across different models of the same training configuration (for example, Flan-PaLM 8B, Flan-PaLM 62B and Flan-PaLM 540B), the internal consistency reliability (that is, α) of synthetic personality scores increased with model size (in this case, number of active parameters)

https://doi.org/10.1038/s42256-025-01115-6

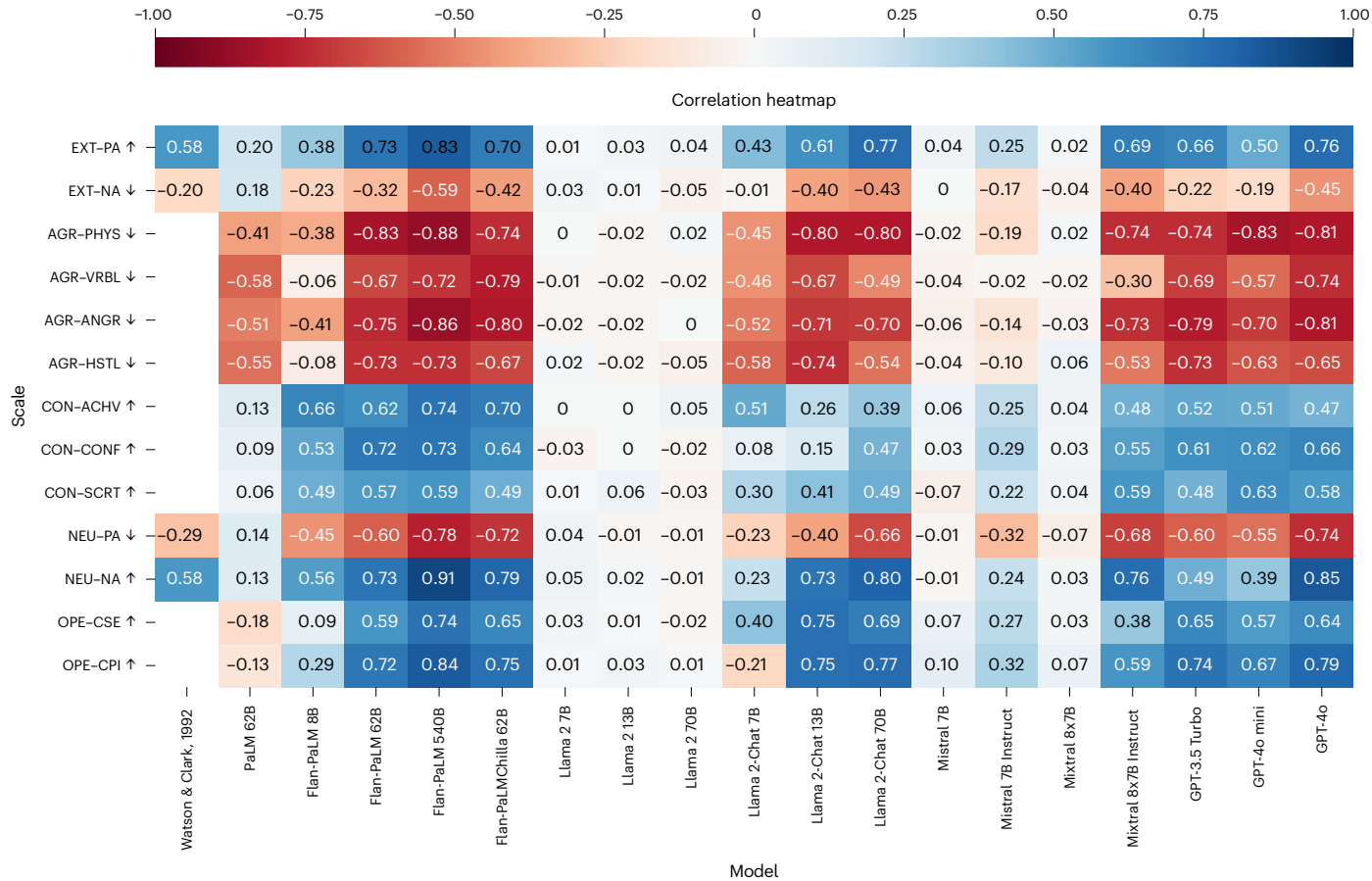


Fig. 3 | Criterion validity evidence of LLM personality measurements per domain. Each row depicts a personality domain paired with a theoretically related criterion test, with upwards arrows indicating an expected positive relationship and downwards arrows indicating an expected negative relationship. Rows 1 and 2: extraversion (EXT), and positive and negative affect, compared with human baselines (leftmost column)reported in previous research on personality and affect[39]. PA, positive affect; NA, negative affect. Rows 3–6: agreeableness (AGR) with subscales of trait aggression, measured by the BPAQ. PHYS, physical aggression; VRBL, verbal aggression; ANGR, anger; HSTL, hostility. Rows 7–9: conscientiousness (CON) with related human values of achievement (ACHV), conformity (CONF) and security (SCRT), measured by PVQ-RR ACHV, CONF and SCRT subscales, respectively. Rows 10 and 11: neuroticism (NEU) with PA and NA compared with human baselines[39]. Rows 12 and 13: openness (OPE) with creativity, measured by the creative self-efficacy (CSE) and creative personal identity (CPI) subscales of the SSCS. N = 22,500 total LLM observations. All LLM correlations > |0.09| are statistically significant at $P < 0.0001$ (2-sided values computed using Student's t-distribution; n = 1,250 per model, per domain).

for instruction-tuned models. Reliability improved from acceptable to excellent when comparing the smallest- and largest-tested variants of Flan-PaLM and Llama 2-Chat. Moving from Mistral 7B Instruct to Mixtral 8x7B Instruct (which use 7B and 12.9B active parameters, respectively), reliability improved from unacceptable to excellent. Reliability only modestly improved with model size when comparing GPT-4o mini with GPT-4o, the only models from OpenAI with confirmed size differences but similar training. Meanwhile, reliability did not scale with model size for tested base models of the same family.

Convergent and discriminant validity results. Convergent and discriminant validity evaluations of LLM personality data allowed us to draw two conclusions. First, a model's training paradigm was the clearest predictor of the validity of its personality measurements: base models without any instruction fine-tuning categorically failed checks for convergent and discriminant validity. Second, among instruction-tuned models, these indices of validity improved as a function of model size. Table 1 summarizes these results and Extended Data Fig. 2 and Table 3 provide further quantitative details.

Convergent validity by model training paradigm. All 30 comparisons of six pairs of base and instruction-tuned models of identical size (namely, two PaLM, six Llama 2, two Mistral and two Mixtral variants; 12 models in total) revealed a consistent pattern. Personality responses of instruction-tuned models demonstrated markedly stronger convergent validity (Extended Data Fig. 2). For example, the average correlations between Llama 2 7B, 13B and 70B's IPIP-NEO and Big Five Inventory (BFI)[69] scores were all non-significant and close to zero, whereas scores of their Llama 2-Chat counterparts showed moderate-to-strong convergent correlations (r_{conv} = 0.59, 0.83 and 0.80, respectively). Even the smallest gain within a model family—Mistral 7B compared with Mistral 7B Instruct—was substantial (r_{conv} = 0.03, not significant versus r_{conv} = 0.28). Full statistics are reported in Extended Data Table 3.

Discriminant validity by model training paradigm. Evidence for discriminant validity clearly favoured instruction-fine-tuned models over base models, when holding model size and family constant. For instance, all five of Flan-PaLM 62B's convergent correlations passed established standards[38] of discriminant validity. By contrast, PaLM 62B's discriminant correlations (average r_{disc} = 0.29) outweighed their convergent counterparts in many cases (average r_{conv} = 0.05; Extended Data Table 3), indicating that, for this model, personality measurements were not consistent across different modes of assessment. Llama 2-Chat 70B (compare Llama 2-Chat 70B) and Mixtral 8x7B Instruct (compare Mixtral 8x7B) replicated this pattern. While relatively smaller instruction-tuned models did not fully pass discriminant

https://doi.org/10.1038/s42256-025-01115-6

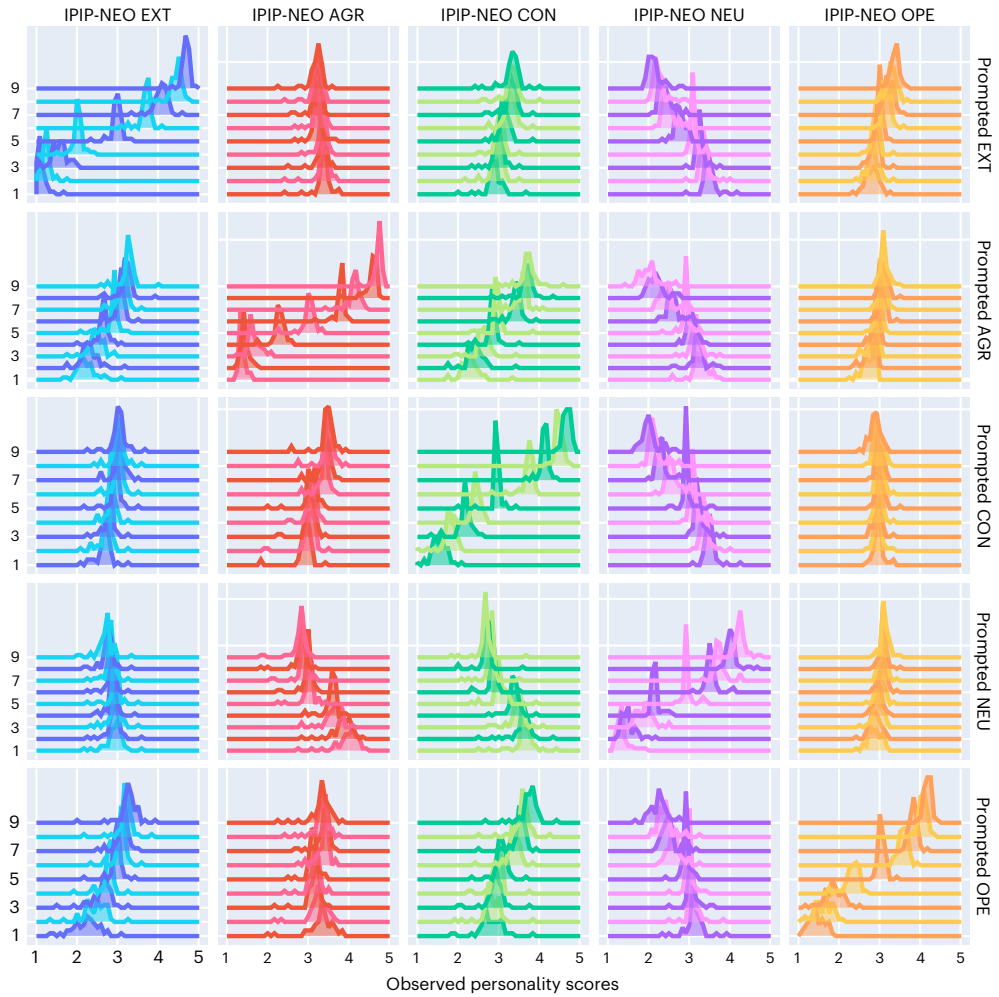


Fig. 4 | Ridge plots of Flan-PaLMChilla 62B's IPIP-NEO personality test scores (N = 2,250) under 9 targeted prompt conditions, each shaping one Big Five trait to a specified level. Each column of plots represents the observed scores on a specific IPIP-NEO subscale across all prompt sets (for example, the leftmost column represents the scores observed for IPIP-NEO extraversion). Each row depicts the observed personality scores across a single prompt set shaping a specific Big Five dimension to 1 of 9 levels (for example, the first row shows results of shaping extraversion; n = 450 observations per row). Each ridge plot comprises nine traces of personality score distributions in response to prompt sets targeting each level (for example, traces labelled '3' represent the prompt set shaping a dimension to level 3 of 9). The plots along the diagonal, from top left to bottom right, depict the intended personality shaping results across all Big Five traits.

validity checks, they did show clear improvements over their respective base versions.

Convergent validity by model size. Convergent validity scaled with size for instruction-tuned models (Extended Data Fig. 2 and Extended Data Table 3). The convergent validity of relatively smaller instruction-tuned model measurements was inconsistent or poor. Flan-PaLM 8B's IPIP-NEO neuroticism and BFI neuroticism, for instance, correlated above 0.80 (constituting excellent convergent validity), while IPIP-NEO openness and BFI openness subscales correlated at less than 0.40 (indicating inadequately low convergence). The same pattern emerged for Llama 2-Chat 7B. Mistral 7B Instruct's convergent validity performance was poor. By contrast, convergent correlations grew stronger and more uniform in magnitude for relatively large models (that is, those with greater numbers of active parameters; Supplementary Note A.12.1). Convergence between LLM IPIP-NEO and BFI scores was strongest for Flan-PaLM 540B and GPT-4o (average r_{conv} = 0.90).

Discriminant validity by model size. Holding the model training paradigm constant, indices of discriminant validity similarly improved with size for instruction-tuned models. The absolute magnitude of all five convergent correlations between the IPIP-NEO and BFI for Flan-PaLM 62B, Flan-PaLM 540B, Llama 2-Chat 70B and Mixtral 8x7B Instruct were the strongest of their respective rows and columns under the multitrait–multimethod matrix[38] framework described in Supplementary Note A.4.2. Comparatively, only three of Flan-PaLM 8B's, three of Llama 2-Chat 7B's and two of Mixtral 8x7B Instruct's convergent correlations were the strongest of their row and column of their respective multitrait–multimethod matrices, indicating mixed evidence of discriminant validity. This pattern is further supported by increases in the average distance (Δ) between a matrix's convergent and respective discriminant correlations when progressively comparing models of similar training paradigms by size in Extended Data Table 3: Flan-PaLM 8B with Flan-PaLM 540B; Llama 2-Chat 7B with Llama 2-Chat 70B; and Mistral 7B Instruct with Mixtral 8x7B Instruct. Average Δ also improved when comparing GPT-4o mini with GPT-4o, albeit modestly. While the exact size difference between these two closed models is unknown, their similar performance on this metric mirrors that of Flan-PaLM at 62B versus 540B parameters. This could suggest that the convergent and discriminant validity of LLM personality measurements plateaus for models of sufficient size.

Criterion validity results. The criterion validity of synthetic personality measurements in LLMs, relative to their convergent and

https://doi.org/10.1038/s42256-025-01115-6

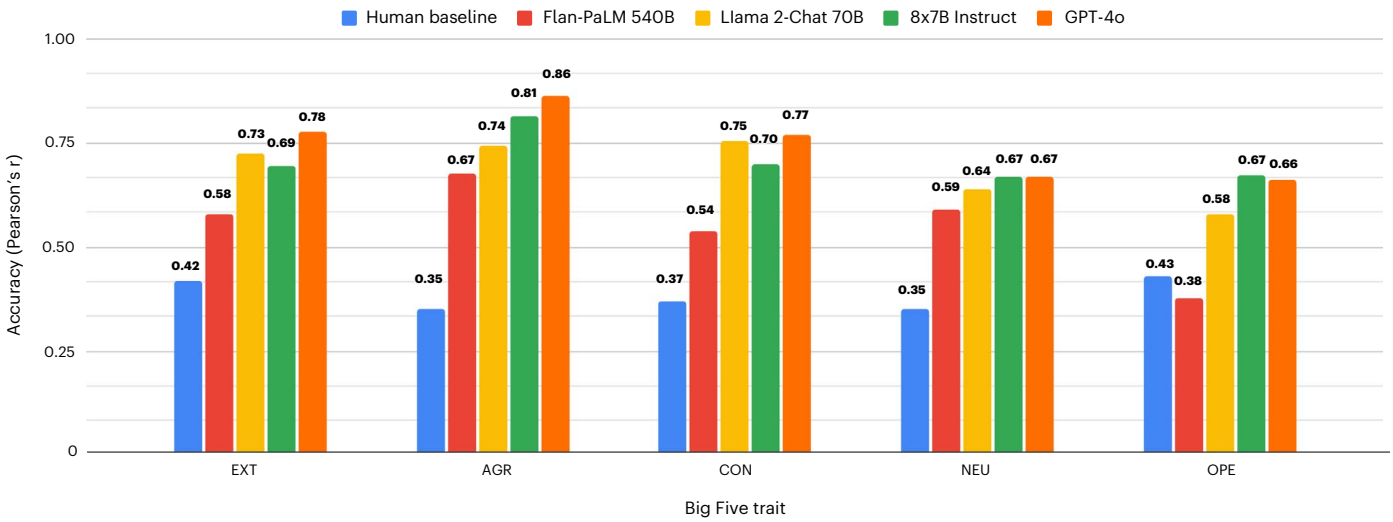


Fig. 5 | Ability of psychometric tests to accurately predict downstream LLM behaviour. The ability of LLM psychometric tests to accurately predict synthetic personality levels in a downstream text generation task (i.e. writing social media status updates) compared with human baselines reported in previous work[48], quantified as Pearson's correlations. On average, LLM IPIP-NEO scores outperformed human IPIP-NEO scores in predicting text-based levels of personality, indicating that LLM personality test responses accurately capture latent LLM personality levels manifested in downstream behaviour. $N = 9{,}000$ total LLM observations. All LLM correlations are statistically significant at $P < 0.0001$ (2-sided values computed using Student's t-distribution); $n = 2{,}250$ per model.

discriminant validity, similarly varied across LLM size and instruction fine-tuning. Measurements of larger, instruction-fine-tuned models showed stronger criterion validity compared with those of their smaller, non-instruction-tuned counterparts. Figure 3 summarizes the results by Big Five domain. We provide an extended discussion of these results in Supplementary Note A.11.1.

Extraversion. Human extraversion strongly correlates with positive affect and moderately, but negatively correlates with negative affect[39]. Simulated IPIP-NEO Extraversion scores for all but base PaLM models sufficiently correlated with the Positive and Negative Affect Schedule (PANAS) scales[40], supporting their criterion validity. IPIP-NEO Extraversion for all three Llama 2 models, Mistral 7B and Mixtral 8x7B (all base models) failed to demonstrate criterion validity, in contrast to their instruction-tuned equivalents, which on the whole showed excellent evidence of validity. Two exceptions were Llama 2-Chat 7B and Mistral 7B Instruct: their extraversion measurements showed questionable-to-poor criterion validity, although both still outperformed their base models. Within instruction-tuned models of a given family, criterion validity scaled with model size.

Agreeableness. In humans, agreeableness is strongly and inversely associated with aggression[41,42]. Along the axis of size for instruction-tuned models, IPIP-NEO Agreeableness showed large jumps in criterion validity from small- to mid-sized variants within a family, but then plateaued, with larger models yielding at most modest additional gains. Size was likewise linked to modest improvements in validity for GPT-4o and GPT-4o mini. By contrast, LLM-specific criterion validity for agreeableness failed to emerge for every base model, with the exception of PaLM 62B.

Meanwhile, the training paradigm related more to criterion validity for Llama 2 and Mixtral. IPIP-NEO agreeableness for all base Llama 2 models and Mixtral 8x7B failed to adequately and significantly correlate with any Buss–Perry Aggression Questionnaire (BPAQ)[43] subscale, demonstrating unacceptable criterion validity. Agreeableness measurements for all sizes of Llama 2-Chat and Mixtral 8x7B Instruct showed moderate-to-excellent criterion validity. For Mistral 7B and Mistral 7B Instruct, instruction tuning related to only a modest improvement in criterion validity, from unacceptable to poor. We could not compare performance across tested GPT-4o models on the basis of post-training status as, at the time of this writing, OpenAI does not publicly offer a foundation model variant for this family.

Conscientiousness. In humans, conscientiousness is meta-analytically related to the human values of achievement, conformity and security[24]. Figure 3 shows how the conscientiousness measurements of all instruction-fine-tuned PaLM variants exhibited stronger evidence of criterion validity than those of the base model, PaLM 62B. Flan-PaLM 540B was the best performer by a small margin, with criterion correlations of 0.74, 0.73 and 0.59 for the Revised Portrait Value Questionnaire (PVQ-RR)[44] scales for achievement, conformity and security, respectively. Instruction-tuned variants of tested Llama 2, Mistral and Mixtral models replicated this finding. Nevertheless, criterion validity for this domain did not scale consistently with size. Llama 2-Chat 7B, for example, outperformed its larger counterparts in how its conscientiousness scores correlated with achievement ($r = 0.51$). GPT-4o mini's scores related slightly more to achievement and security compared with GPT-4o's scores.

Neuroticism. Human neuroticism is strongly positively correlated with negative affect and moderately negatively correlated with positive affect[39]. IPIP-NEO Neuroticism for all instruction-tuned models, compared with base models, showed excellent evidence of criterion validity in their relation to PANAS scales (Fig. 3). IPIP-NEO Neuroticism's criterion validity increased with model size for instruction-tuned models, in terms of how the strengths and directions of their criterion correlations aligned with those observed among human data[39].

Openness. Openness to experience in humans is empirically linked to creativity across multiple studies[45,46]. Figure 3 illustrates how the criterion validity of openness measurements was strongest for larger, fine-tuned variants of PaLM and Llama 2. IPIP-NEO criterion correlations with the Short Scale of Creative Self (SSCS)[47] creative self-efficacy (CSE) and creative personal identity (CPI) ranged from moderate ($r = 0.59$) to strong ($r = 0.84$). Notably, we observed negative correlations between openness and creativity for PaLM 62B in contrast to those shown for Flan-PaLM 8B, the smallest model tested. Mistral 7B Instruct and Mixtral 8x7B Instruct's openness data demonstrated

weak-to-moderate evidence of criterion validity. Relative model size modestly related to the validity of openness scores for GPT-4o mini and GPT-4o.

Shaping results

We successfully shaped personality traits in LLMs independently and concurrently, in single- and multi-trait shaping experiments, respectively. The results of both experiments are reported in greater detail in Supplementary Note A.8.

Single-trait shaping. For 11 out of 12 models tested, ordinal targeted levels of personality very strongly correlated with observed IPIP-NEO scores. Namely, the average Spearman correlation (ρ) for these models was ≥ 0.80 (Extended Data Table 6). Figure 4 visualizes this overall pattern, depicting, for example, how Flan-PaLMChilla 62B's personality scores monotonically increased alongside prompted levels of a given Big Five trait. Notably, levels of unprompted traits remained relatively stable in response to shaping. For instance, the medians of Flan-PaLMChilla 62B's openness scores remained near 3.00 when all other Big Five domains were shaped—see the right side of Fig. 4. Similar patterns of stability were observed for extraversion and agreeableness. Conscientiousness and neuroticism scores fluctuated the most in response to prompts that did not target those domains, but these fluctuations did not reach the strength and direction of the score changes observed in the ridge plots of targeted traits (that is, the plots on the diagonal, from top-left to bottom-right).

The absolute change in model personality scores in response to shaping was another important consideration. Only relatively larger models were able to disambiguate prompts requesting the lowest (versus highest) levels of a targeted dimension. Extended Data Table 6 shows the distances (Δs) between the medians of IPIP-NEO score distributions obtained in response to the lowest- and highest-levelled prompts, where the best possible Δ, representing an average score change from 1.00 to 5.00, is 4.00. Our smallest tested models (namely, Flan-PaLM 8B, Llama 2-Chat 7B and Mistral 7B Instruct) struggled to reach Δs ≥ 2.00; Mistral 7B Instruct's median personality domain scores shifted by a Δ of only 0.78 on average. Meanwhile, models with greater than 62B active parameters (and GPT-4o) achieved average $\Delta \geq 3.00$, with Flan-PaLM 540B achieving the largest Δ of 3.67.

Supplementary Note A.8.1 discusses single-trait shaping results in greater detail.

Multiple-trait shaping. When we concurrently set the prompted trait levels of each Big Five dimension to either 'extremely high' or 'extremely low', all tested models struggled to show the same level of control observed in single-trait shaping. However, all but two models tested (Mistral 7B Instruct and Llama 2-Chat 7B) produced distinct distributions of personality test scores, showing varying abilities to differentiate between high and low levels. Extended Data Fig. 3 shows the distributions of LLM-synthesized personality when the models were prompted to exhibit extremely low (red) or extremely high (blue) levels of all dimensions in parallel.

Distributional distance increased with model size, particularly for observed neuroticism, openness and conscientiousness scores (Extended Data Table 7). Flan-PaLM 540B, our largest tested model in terms of known parameter size, and GPT-4o showed the best overall control concurrently shaping multiple Big Five traits. For these models, a given Big Five trait score shifted by 2.525 points on average, as shown in the Δ columns of Extended Data Table 7a and 7d. Flan-PaLM 62B, Flan-PaLMChilla 62B and GPT-4o mini outperformed their larger counterparts on shaping extraversion, with Δ values of 3.44, 3.40 and 3.42, respectively.

For smaller models (for example, Flan-PaLM 8B, Llama 2-Chat 7B and Mistral 7B Instruct), while targeted traits changed in score levels in response to prompts, score ranges were more restricted, indicating lower levels of control. Flan-PaLM 8B's median scores on IPIP-NEO Agreeableness, for instance, shifted from 2.88 to only 3.52 when the model was prompted to simulate 'extremely low' and 'extremely high' levels of agreeableness (that is, 1 versus 9), respectively. When Flan-PaLM 8B was given the same extremely low and high prompts as in the first shaping experiment, the median difference between its level-1-prompted and level-9-prompted agreeableness scores (2.37 and 4.12, respectively) was 173% larger. Supplementary Note A.8.2 discusses these results in further detail.

Real-world task results

We found that psychometric tests of LLM personality robustly predicted personality in LLM task behaviour, expressed in social media status updates generated by Flan-PaLM 540B, Llama 2-Chat 70B, Mixtral 8x7B Instruct and GPT-4o. Psychometric test-based personality strongly correlated with language-based personality levels observed in downstream generated text across all tested models, shown in Fig. 5.

In particular, the average convergent r between survey- and language-based measures of all five dimensions was 0.67 across models. This observed convergence, even for the weakest-performing model, exceeded established convergence between survey- and language-based levels of personality reported for humans (average $r = 0.38$)[48].

Moreover, our prompting technique was highly effective at shaping personality levels in LLM-generated text. On average per model, prompted trait levels correlated strongly to very strongly with personality observed in LLM-generated social media updates (average ρ ranged from 0.68 to 0.82; Extended Data Table 5).

To illustrate the practical implications of the personality-shaping methodology, we generated word clouds to gain insights into model-generated language that users would see. Extended Data Fig. 4 shows the most frequent words in synthetic social media updates when Flan-PaLM 540B simulated varying levels of different traits. The differences in terms are quite apparent. For instance, LLM-generated language in response to prompting for extremely high emotional stability (that is, extremely low neuroticism) was characterized by positive emotion words, such as 'happy', 'relaxing', 'wonderful', 'hope' and 'excited'. By contrast, the most frequent words from simulating extremely high levels of neuroticism—'hate', 'depressed', 'angry', 'bad', 'nervous' and 'sad'—reflected negatively charged emotional content.

Supplementary Table 8 provides example social media updates generated by Flan-PaLM 540B when setting a specific personality domain either extremely low or extremely high. For instance, in the case of extremely low conscientiousness, the generated text reflects a synthetic persona that avoids responsibility, while in the case of extremely high conscientiousness, the persona values hard work and returning favours. In addition, in our example of extremely low openness, the persona expresses conservative political views, while in the case of extremely high introversion (that is, extremely low extraversion), the persona exhibits discomfort with social situations. These and other examples illustrate that there might be inherent bias in LLM training data that causes particular patterns of thinking, feeling and behaving to be highly associated with specific LLM personas. Overall, this study demonstrated that LLM-generated language was similar to human language observed in previous studies assessing personality in social media data[48], further confirming the construct validity of our LLM personality measurements.

Discussion

The goal of this work was to contribute a principled methodology to reliably and validly measure synthetic personality in LLMs and use the same validated methodology to shape LLM personality expression. We developed a complete framework to: (1) quantify personality traits perceived by humans in LLM outputs using psychometric testing; (2) verify whether psychometric tests of LLM personality traits were

empirically reliable and on a model-by-model basis; and (3) implement mechanisms to systematically shape levels of specific LLM personality traits. Unlike earlier AI personality assessment work that left validity open-ended, we applied this construct-validation framework across 18 models and thousands of prompt variations. The application of this methodology demonstrates that the IPIP-NEO provides reliable and valid measurements of synthetic personality for sufficiently scaled and instruction-tuned LLMs. Furthermore, it highlights possible mechanisms that allow LLMs to encode and express complex social phenomena (Supplementary Note A.11).

Limitations and future work

Personality traits of other LLMs. One of the core contributions of this work is an understanding of how simulating personality in language models is affected by model size and level of (post-)training. We focused on base and instruction-tuned variants of the PaLM, Llama 2, Mistral, Mixtral and GPT model families for pragmatic reasons, but the presented methodology for administering psychometric surveys is model agnostic and is applicable to any decoder-only architecture language model.

Psychometric test selection and validation. This work also contributes a principled way to establish the reliability and validity of psychometric personality tests in the LLM context. However, this work may be biased by its selection of psychometric tests; some assessments may show better LLM-specific psychometric properties than others. We attempted to mitigate selection bias by administering personality assessments of different lengths (300 versus 44 items) and distinct theoretical traditions (questionnaire versus lexical[11]). Future work could administer different personality tests (for example, the HEXACO Personality Inventory, which uses a cross-cultural six-factor taxonomy of personality[36]), develop personality tests tailored for LLMs to obtain more accurate trait measurements, and validate personality measurements with additional external criteria and downstream tasks.

Monocultural bias. This work contributes evidence that at least some LLMs exhibit personality traits that approximate human standards of reliability and validity. However, the LLMs tested here were primarily trained on language data originating from Western European and North American users. While these LLMs perform well on natural language processing benchmarks in multiple languages, the models in this work were assessed exclusively with English-language psychometric tests. Most of the tests used in this work have non-English translations validated in cross-cultural psychology that merit future use in LLM research. Similarly, while the Big Five model of personality has well-established cross-cultural generalizability[49], some non-Western cultures express additional personality dimensions that do not exist in top-down, nomothetic personality taxonomies such as the Big Five[50]. Those dimensions may be better represented in culture-specific (that is, idiographic) approaches to measuring personality in LLMs.

Evaluation settings. Unlike conventional human psychometric test administration, under the presented methodology, evaluated LLMs did not consider responses to previous test items; all items were presented and scored as independent events. We chose this method to ensure that model response variance was not impacted by item ordering effects or length of the context (prompt) provided to the model for inference, and could be isolated to controlled variations in our prompts. LLM performance on natural language tasks is known to decrease as length of input prompts grow, and is most affected by the content at either the beginning or towards the end of long inputs[51]. Pretrained-only base LLMs are known to show biased attention for more recent tokens (that is, the end of inputs), especially when evaluating next-word prediction of contiguous text[52]. This uneven attention compounds approximation errors in longer contexts[53], such as those necessitated by 300-item IPIP-NEO used here, which further motivated our design decision to administer items to models independently. However, psychometric test data quality for humans can be affected by test length and item order. Our method avoids many sources of measurement error inherent to human administration, while being subject to others inherent to machine administration. In addition, model responses to the items were constrained to set choices or determined by next-token log-probability scoring to ensure reproducibility. LLMs are more commonly used to generate text freely outside of these constraints. As such, more evaluations allowing unstructured generative modes of inference (for example, our downstream status update task) might provide more realistic estimates of a model's behaviour.

Real-world use cases. Our downstream task relied on repeated, yet single-turn behavioural interactions to validate our evaluation framework in a real-world use-case. This may have provided only a partial picture of external validity. We echo the scientific principles of psychological assessment, and reiterate that the process of construct validation is ongoing[39]: we hope that future research can extend our investigation of validity by developing downstream tasks that test particular personality domains, vary in complexity and transpire over multiple turns of dialogue.

Broader implications

Responsible AI alignment. As governments finalize AI safety regulations and similar rules that will require behavioural audits of frontier models, our findings show that personality—a core driver of human trust and persuasion—can already by measured and steered in today's LLMs. The ability to probe and shape LLM personality traits is pertinent to the open problem of responsible AI alignment[54] and harm mitigation[55]. As a construct-validated auditing tool[56], our methodology can be used to proactively predict toxic behavioural patterns in LLMs across a broad range of downstream tasks, potentially guiding and making more efficient responsible AI evaluation and alignment efforts before deployment. Similarly, shaping levels of specific traits away from toxic or harmful language output (for example, very low agreeableness, high neuroticism) can make interactions with LLMs safer and more inclusive. The values and moral foundations present in LLMs could be made to better align with desired human values by tuning for corresponding personality traits, as personality is meta-analytically linked to human values[57]. More directly, the presented methodology can be used to rigorously quantify efforts towards human value alignment in LLMs by establishing the construct validity of human value questionnaires in LLMs.

Implications for users. Users could more meaningfully engage with agentic and other digital technologies using LLMs tailored to their specific personality profiles. A user who prefers direct, concise communication could interact with a customer service AI agent shaped to deliver straightforward, brief responses, while another user who values detailed, empathetic interactions could benefit from an agent persona designed to communicate warmly and expansively. Our construct-validation methodology can also serve as a quality-assurance step when hardening user-facing AI systems and agents, promoting safer and more consistent personality profiles. Indeed, the personality-shaping approach could enhance adversarial testing by probing another LLM's responses and to train new models on how to handle adversarial situations.

Ethical considerations

Personalized LLM persuasion. Adapting the personality profile of a conversational agent to that of a user can make the agent more effective

at encouraging and supporting behaviours[58]. Personality matching has been shown to increase the effectiveness of real-life persuasive communication[59]. However, the same personality traits that contribute to persuasiveness and influence could be used to encourage undesirable behaviours. As LLM-powered digital applications become ubiquitous, their potential to be used for harmful persuasion of individuals, groups and even society at large must be taken seriously. Having scientifically vetted methods for LLM personality measurement, analysis and modification, such as the methodology our work presents, increases the transparency and predictability of such LLM manipulations. Persuasive techniques are already ubiquitous in society, so stakeholders of AI systems must work together to systematically determine and regulate AI use; this work aims to inform such efforts.

Anthropomorphized AI. Personalization of conversational agents has documented benefits[60], but there is a growing concern about the harms posed by the anthropomorphization of AI. Recent research suggests that anthropomorphizing AI agents may be harmful to users by threatening their identity, creating data privacy concerns and undermining well-being[61]. Beyond qualitative probing explorations, our work definitively establishes the unexpected ability of LLMs to appear anthropomorphic, and to respond to psychometric tests in ways consistent with human behaviour, because of the vast amounts of human language data they have trained on. Given that personality shapes trust, our method offers a quantitative foothold for aligning future AI systems with—and protecting them from—human social expectations.

Detection of incorrect LLM information. LLMs can generate convincing but incorrect responses and content[55]. One of the methods to determine whether a text containing a world fact is generated by an LLM (and hence might require vetting) is to identify psycholinguistic patterns known to pervade 'factual' LLM language, such as lower levels of emotional expression[62]. However, with personality shaping, that method may be rendered ineffective, thereby making it easier for bad actors to use LLMs to generate misleading content. This problem is part of the larger alignment challenge and grounding of LLMs—areas of growing focus of investigation in both academia and industry.

Conclusion

LLMs clearly project human-like personality, yet rigorous tools for measuring such complex social constructs remain underdeveloped. We close that gap with the first psychometrically grounded framework for quantifying personality in LLMs, evaluating the test data of 18 widely used models against established standards of reliability and validity from psychometrics. We found that larger, instruction-tuned systems pass these standards with greater ease, illustrating that synthetic personality grows more stable and accurate for language models with sufficient post-training and scale. Further leveraging our framework, we introduce zero-shot personality shaping: by pairing specific language markers and qualifiers, we showed how LLMs can be steered towards targeted personality profiles with high fidelity. In addition, we discussed the ethical implications of shaping LLM personality traits. This work has important implications for AI alignment and harm mitigation, and informs ethics discussions concerning AI anthropomorphization, personalization and potential misuse.

Methods

Methodology overview

We quantified and evaluated the ability of LLMs to meaningfully emulate human personality traits in two stages. First, using the structured prompting methodology proposed in 'Administering psychometric tests to LLMs', we repeatedly administered 2 personality measures of different lengths and theoretical traditions to a variety of LLMs, alongside a battery of 11 separate psychometric tests of personality-related constructs. Second, as described in 'Reliability and construct validity' and unique to this work, we rigorously evaluated the psychometric properties of LLM responses through a suite of statistical analyses of reliability and construct validity. The resulting metrics facilitate a comparison of the varied abilities of LLMs to reliably and validly synthesize personality traits and provide insight into LLM properties that drive these abilities. Figure 1 provides an overview of the test validation process.

We evaluated 18 LLMs from the PaLM[5], Llama 2[63], Mistral[64], Mixtral[65] and GPT[6,66] model families. We varied model selections across three key dimensions: size (number of active parameters), instruction tuning and training method (see Supplementary Note A.3.1 for details).

Administering psychometric tests to LLMs. Quantifying the personality traits of LLMs requires a measurement methodology that is reproducible, yet flexible enough to facilitate formal testing of reliability and validity across diverse prompts and measures. To administer psychometric tests to LLMs, we leveraged their ability to score possible completions of a provided prompt. We used prompts to instruct models to rate items (that is, descriptive statements such as 'I am the life of the party') from each psychometric test on a standardized response scale (for example, 1 = 'strongly disagree' versus 5 = 'strongly agree'). We simulated an LLM's chosen response to an item by ranking the conditional log probabilities of its response scale options, framed as possible continuations of the prompt (for example, '1' versus '5'); Supplementary Note A.3.2 specifies our implementation across models. This constrained mode of LLM inference is often used in multiple choice question-and-answer tasks to score possible options[67] (versus inference by generating text). Using this technique ensured that item responses were not influenced by content contained in other items, mitigating measurement error due to item order.

We administered two personality inventories—primary and secondary—to assess whether LLM responses to psychometric tests of different lengths and distinct theoretical traditions converged, indicating convergent validity. We selected the widely used IPIP-NEO[37], a 300-item open-source representation of the Revised NEO Personality Inventory[68] as our primary measure of personality. As a secondary measure, we used the BFI[69], a 44-item measure developed in the lexical tradition[11]. Both tests assess the Big Five traits (that is, domains) of personality[28], comprising dedicated subscales measuring extraversion, agreeableness, conscientiousness, neuroticism and openness to experience. Supplementary Note A.3.3 details the scoring scheme and rationale behind the selection. To validate these measures of personality in the LLM context, we additionally administered 11 psychometric tests of theoretically related external criteria, each corresponding to at least one Big Five domain.

Response variation generated by structured prompting was necessary to analyse the reliability and validity of LLM personality measurements, described in 'Reliability and construct validity'. The prompt for each psychometric test item consisted of three main parts: an item preamble, the item itself and an item postamble. Each item preamble contained a persona instruction, a biographic description and an item instruction (Fig. 2). When administering a psychometric test, we systematically modified the biographic descriptions, item instructions and item postambles surrounding each item to generate simulated response profiles, unique combinations of a prompt that were reused within and across administered measures to statistically link LLM response variation in one measure to response variation in another measure. Persona instructions instructed the model to follow a given biographic description and remained fixed across all experiments. A given biographic description contained 1 of 50 generic self-descriptions (Supplementary Table 2) sampled from an existing dialogue dataset[70] to anchor LLM responses to a social context and create necessary variation in responses across prompts, with descriptions like 'I like to bake pies' or 'My favorite season is winter'.

Item instructions were introductory phrases (adapted from original test instructions where possible) that conveyed to the model that it was answering a survey item (for example, 'Thinking about the statement, [...]'). A given item was a descriptive statement (accompanied by a rating scale) taken from a given psychometric test (for example, 'I see myself as someone who is talkative'). Item postambles presented the possible standardized responses the model could choose from (for example, one of [1, 2, 3, 4, 5]).

Supplementary Note A.3.4 discusses the prompt design motivation and provides a full set of biographic descriptions, item instructions and item postambles. These criterion measures were drawn from widely used instruments developed and psychometrically validated for human research, including PANAS scales[40], Buss–Perry Aggression Questionnaire (BPAQ)[43], Revised Portrait Value Questionnaire (PVQ-RR)[44] and Short Scale of Creative Self (SSCS)[47].

Reliability and construct validity

After all psychometric tests were administered, across all the prompt variations, the next stage established whether LLM measurements of personality were dependable and externally meaningful—that they exhibited statistical reliability and construct validity. Addressing these two scientific criteria is a key contribution of this work. In psychometrics, and across any science involving measurement, the construct validity of a given test requires reliability. Reliability refers to the consistency and dependability of a test's measurements. Construct validity can be evaluated in terms of, but is not limited to, convergent, discriminant and criterion validity[35]. A test demonstrates convergent validity when it sufficiently relates to purported indicators of the test's target construct. Discriminant validity refers to how sufficiently unrelated a test is to indicators of unrelated constructs. Criterion validity indicates how well a test relates to theoretically linked external outcomes. Supplementary Note A.4 contains further details on reliability and validity.

To evaluate the reliability and construct validity of the LLM responses, we conducted a suite of statistical analyses informed by formal standards of psychometric test construction and validation (Supplementary Note A.4.2–3). We organized these analyses by three subtypes of reliability and construct validity, respectively. (While it was not a focus of this work, we report an exploratory analysis of structural validity in Supplementary Notes A.5 and A.6.3). In this work, a personality trait is validly synthesized by an LLM only when the LLM responses meet all tested indices of reliability and construct validity. Figure 1 provides an overview of the process and validity criteria, and Supplementary Note A.4 presents the full methodology for evaluating the construct validity of LLM personality measurements.

Reliability. The reliability of each IPIP-NEO and BFI subscale, the extent to which their LLM measurements of personality were consistent and dependable, was quantified by formal psychometric standards of internal consistency reliability (operationalized as Cronbach's α and Guttman's λ_6), and composite reliability (operationalized as McDonald's ω). Supplementary Note A.4.1 provides additional information on these reliability metrics.

Convergent and discriminant validity. We evaluated the LLM-specific convergent and discriminant validity of the IPIP-NEO as components of construct validity, according to published standards[38,71] (note that throughout this work, we use thresholds recommended by Evans[72] to describe correlation strengths). The convergent validity of the IPIP-NEO for each model, the test's quality in terms of how strongly it relates to purported indicators of the same targeted construct, was quantified in terms of how strongly each of its five subscales convergently correlated with their corresponding BFI subscale (for example, IPIP-NEO Extraversion's convergent correlation with BFI Extraversion), on average. The discriminant validity of the IPIP-NEO per model, its quality in terms of how relatively unrelated its subscales are to purported indicators of non-targeted constructs, was determined when the average difference (Δ) between its convergent and respective discriminant correlations with the BFI (for example, IPIP-NEO Extraversion's discriminant correlation with BFI agreeableness) was at least moderate (≥0.40). We used Pearson's correlation coefficient (r), in these and subsequent validity analyses of continuous data.

Criterion validity. As another component of construct validity, the criterion validity of a psychometric test gauges its ability to relate to theoretically connected non-target criteria. To evaluate the LLM-specific criterion validity of the IPIP-NEO, we administered tests of 11 additional social constructs (that is, external criteria) theoretically connected to personality (Extended Data Table 4) and correlated each IPIP-NEO subscale with its corresponding external tests. A given IPIP-NEO subscale demonstrated criterion validity when the strength and direction of its correlations with tested external criteria matched or exceeded statistical associations reported for humans.

Shaping synthetic personality traits in LLMs

Having found evidence of the reliability and construct validity of LLM personality measurements, we next considered the second part of our research question: Can LLM-synthesized personality profiles be verifiably shaped along desired dimensions? To answer this question, we devised a prompting methodology that shaped each synthetic personality trait at 9 intensity levels, using 104 trait adjectives and Likert-type linguistic qualifiers[73]. These trait adjectives were adapted from established linguistic research of personality using Goldberg's personality trait markers[74]. We evaluated LLM personality score changes in response to personality-shaped prompts across two experiments: single-trait shaping and multiple-trait shaping (see Supplementary Note A.7 for details). Specifically, our first experiment tested the abilities of LLMs to shape emulated Big Five dimensions of personality independently, targeting single personality dimensions in isolation without prompting other dimensions. Our second experiment tested the abilities of LLMs to shape synthetic Big Five traits concurrently, specifying target levels of all five dimensions in every prompt set at the same time. As a more rigorous test of representational capacity, this experiment required the tested LLMs to concurrently disambiguate complex overlaps in personality domain information. The designed difficulty of the task was further underscored by extant human research indicating that Big Five personality dimensions measured in questionnaires and natural language[48] are not entirely orthogonal; they are weakly intercorrelated[75].

Methodology overview. To shape synthetic personality traits in LLMs, we began with an established theory in psychology that posits salient descriptors of personality are encoded in language, known as the lexical hypothesis[25]. We incorporated this knowledge into the prompt design, adapting Goldberg's list of 70 bipolar adjectives[74] known to statistically capture the Big Five model of personality through factor analyses of human ratings. In this list, for example, the adjectives 'silent' and 'talkative' were found to mark relatively low and high levels of extraversion, respectively. We mapped these adjectives to each of the Big Five domains and 30 lower-order personality facets measured by the IPIP-NEO based on Goldberg's original study[74]. Next, where we lacked coverage of a measured IPIP-NEO domain or facet, a trained psychometrician wrote additional adjectives to mitigate potential data imbalances, bringing our expanded list of trait adjectives to 104. We report the full list in Supplementary Table 17.

For more precise control of personality levels, we used linguistic qualifiers often used in Likert-type response scales[73] (for example, 'a bit', 'very', 'extremely') to configure a target level for each adjective. The resulting prompt design, described in Supplementary Notes A.7.1–2, facilitated granular shaping of a given Big Five trait at up to nine levels.

Across both shaping experiments, we only tested models that demonstrated at least 'neutral' reliability in our construct-validity experiments (Table 1): Flan-PaLM (8B, 62B, 540B), Llama 2-Chat (7B, 13B, 70B), Mistral 7B Instruct, Mistral 8x7B Instruct, GPT-3.5 Turbo, GPT-4o mini and GPT-4o (11 models, total).

Evaluation methodology. In the single-trait shaping experiment (described in detail in Supplementary Note A.7.2), our objective was to independently shape each Big Five trait at each of the nine levels. We benchmarked the success of independent shaping by: (1) quantifying how strongly shifts in IPIP-NEO score distributions were related to shifts in targeted trait levels embedded in our prompt sets (that is, through Spearman's rank correlation coefficient ρ); and (2) inspecting the distance between personality score distributions obtained in response to our most extreme prompt sets; specifically, the set of prompts we shaped to be the lowest possible levels of a trait (versus those shaped to be the highest possible levels of a trait) should result in distributions of scores that are farther away from each other.

In the multi-trait shaping experiment (described in detail in Supplementary Note A.7.3), to more rigorously test model capacities for attention, we aimed to concurrently shape all Big Five traits as possible. We benchmarked the success of concurrent shaping by distributional distance, as defined above.

LLM personality traits in real-world tasks

So far we have reported on LLM abilities to encode human personality traits by collecting psychometric test data and evaluating their construct validity. We also sought to address possible concerns that the validity of LLM personality measurements—evidenced by LLM responses to other psychometric tests—could be an artefact of common method bias[76]. In other words, our questionnaire-based signals of LLM personality were validated by responses to other questionnaires that have not undergone the same LLM-specific construct-validation process. To address this risk of common method bias, we further validated our personality testing and shaping frameworks by: (1) comparing psychometric test levels of LLM personality with downstream observations of model behaviours on a real-world task; and (2) investigating the effects of LLM personality shaping on the outputs of this task.

Methodology overview. We instructed the largest tested model per family to generate up to 1.125 million social media status updates based on the same 2,250 simulated human profile descriptions used in 'Shaping synthetic personality traits in LLMs'—profiles designed to shape expressions of a particular Big Five dimension across nine levels. (Supplementary Note A.9 details the task design and rationale.) The personality observed in the status updates generated for each simulated human profile was then rated using Apply Magic Sauce[77], a validated research application programming interface (API) for measuring personality in open-ended text. The chosen task was designed to reflect adequate levels of realism, complexity and domain relevance for evaluating personality expression of LLMs.

To assess how psychometric tests may reflect latent personality levels expressed by LLMs in downstream behaviour, we computed Pearson's correlations (r values) between model personality test scores and (Apply Magic Sauce-computed) personality observed in generated social media text; both sets of scores were linked by the same 2,250 personality-shaping prompts used in 'Shaping synthetic personality traits in LLMs'. Next, we statistically verified the effectiveness of personality shaping by computing Spearman's rank correlations (ρ values) between prompted levels of personality and observed personality ratings of model-generated text. At least a moderate correlation between survey-based and linguistic estimates of personality in LLMs (as demonstrated in previously reported human data[48]) would demonstrate that a survey-based measure of personality accurately predicts LLM-synthesized personality in subsequent tasks such as text generation. We similarly applied this threshold to interpret the effectiveness of personality shaping.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated by the LLMs tested in this work, including the psychometric test score data and open-ended text responses to a real-world task prompt, have been added to a public data storage bucket for wider public use[13]. The psychometric tests used in this study were accessed from their respective original publications and, where applicable, public research repositories. We used items of these tests as LLM prompt inputs in a non-commercial research capacity. The authors and copyright holders of these tests govern their availability and use. The 50 biographic descriptions used in our structured prompts were randomly and reproducibly sampled from the true-cased version[78] of the PersonaChat dataset[70]. PersonaChat is a publicly available, crowd-sourced dataset of 1,155 fictional human profile descriptions. For analysis of personality traits on generated text, this work used the Apply Magic Sauce API[79], a validated psychodemographic research tool that predicts personality from open-ended text[77].

Code availability

The code used to administer psychometric tests to LLMs is intended to be interoperable across LLMs. That code, along with the remaining Python and R code used to generate our prompt sets and statistically analyse reliability, construct validity and trait shaping is found in an open-source repository for wider public use[14].

References

1. Roose, K. A conversation with Bing's chatbot left me deeply unsettled. *The New York Times* (17 February, 2023).
2. Hagendorff, T. Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. Preprint at https://doi.org/10.48550/arXiv.2303.13988 (2023).
3. Wei, J. et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=yzkSU5zdwD (2022).
4. Gallifant, J. et al. Peer review of GPT-4 technical report and systems card. *PLoS Digit. Health* **18**, e0000417 (2024).
5. Chowdhery, A. et al. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 240 (2023).
6. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 1877–1901 (Curran Associates, 2020).
7. Allport, G. *Personality: A Psychological Interpretation* (H. Holt, 1937).
8. Roberts, B. W. & Yoon, H. J. Personality psychology. *Annu. Rev. Psychol.* **73**, 489–516 (2022).
9. Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. & Goldberg, L. R. The power of personality: the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* **2**, 313–345 (2007).
10. Rust, J., Kosinski, M. & Stillwell, D. *Modern Psychometrics: The Science of Psychological Assessment* 4th edn (Routledge, 2020).
11. Simms, L., Williams, T. F. & Simms, E. N. in *Assessment of the Five Factor Model: The Oxford Handbook of the Five Factor Model* (ed. Widiger, T. A.) 353–380 (Oxford Univ. Press, 2017).
12. Schwartz, H. A. et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).

13. Serapio-García, G. et al. Personality traits in large language models—dataset. *Google* https://storage.googleapis.com/personality_in_llms/index.html (2025).
14. Safdari, M., Serapio-García, G. & Crepy, C. Personality traits in large language models. *GitHub* https://github.com/google-deepmind/personality_in_llms (2025).
15. Shuster, K. et al. Language models that seek for knowledge: modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP* 2022, 373–393 (ACL, 2022).
16. Hendrycks, D. et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (OpenReview.net, 2021).
17. Wei, J. et al. Larger language models do in-context learning differently. Preprint at https://doi.org/10.48550/arXiv.2303.03846 (2023).
18. Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1–53 (2024).
19. Wei, J. et al. ICLR Finetuned language models are zero-shot learners. In *Tenth International Conference on Learning Representations (ICLR*, 2022).
20. Miotto, M., Rossberg, N. & Kleinberg, B. Who is GPT-3? An exploration of personality, values and demographics. In *Proc. Fifth Workshop on Natural Language Processing and Computational Social Science* (eds Bamman, D. et al.) 218–227 (ACL, 2022).
21. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258–268 (2022).
22. Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at https://doi.org/10.48550/arXiv.2302.08399 (2023).
23. Bleidorn, W. et al. The policy relevance of personality traits. *Am. Psychol.* **74**, 1056 (2019).
24. Parks-Leduc, L., Feldman, G. & Bardi, A. Personality traits and personal values: a meta-analysis. *Personality Soc. Psychol. Rev.* **19**, 3–29 (2015).
25. Goldberg, L. R. Language and individual differences: the search for universals in personality lexicons. *Rev. Personality Soc. Psychol.* **2**, 141–165 (1981).
26. Saucier, G. & Goldberg, L. R. Lexical studies of indigenous personality factors: premises, products, and prospects. *J. Personality* **69**, 847–879 (2001).
27. Gabriel, I. Artificial intelligence, values, and alignment. *Minds Mach.* **30**, 411–437 (2020).
28. John, O. P., Naumann, L. P. & Soto, C. J. in *Paradigm Shift to the Integrative Big Five Trait Taxonomy: History, Measurement, and Conceptual Issues—Handbook of Personality: Theory and Research* (eds John, O. P. et al.) 114–158 (Guilford, 2008).
29. Lin, S., Hilton, J. & Evans, O. TruthfulQA: measuring how models mimic human falsehoods. Om *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (eds Muresan, S. et al.) 3214–3252 (ACL, 2022).
30. Abdulhai, M. et al. Moral foundations of large language models. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 17737–17752 (ACL, 2024).
31. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *FAccT '21: Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (ACM, 2021).
32. Mahowald, K. et al. Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517–540 (2024).
33. Wang, X. et al. InCharacter: evaluating personality fidelity in role-playing agents through psychological interviews. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (eds Ku, L.-W. et al.) 1840–1873 (ACL, 2024).
34. Jacobs, A. Z. Measurement as governance in and for responsible AI. Preprint at https://doi.org/10.48550/arXiv.2109.05658 (2021).
35. Clark, L. A. & Watson, D. Constructing validity: new developments in creating objective measuring instruments. *Psychol. Assess.* **31**, 1412–1427 (2019).
36. Lee, K. & Ashton, M. C. Psychometric properties of the HEXACO Personality Inventory. *Multivar. Behav. Res.* **39**, 329–358 (2004).
37. Goldberg, L. R. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychol. Eur.* **7**, 7–28 (1999).
38. Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81 (1959).
39. Watson, D. & Clark, L. A. On traits and temperament: general and specific factors of emotional experience and their relation to the Five-Factor model. *J. Personality* **60**, 441–476 (1992).
40. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* **54**, 1063–1070 (1988).
41. Bettencourt, B. A. & Kernahan, C. A meta-analysis of aggression in the presence of violent cues: effects of gender differences and aversive provocation. *Aggress. Behav.* **23**, 447–456 (1997).
42. Chester, D. S. & West, S. J. Trait aggression is primarily a facet of antagonism: evidence from dominance, latent correlational, and item-level analyses. *J. Res. Personality* **89**, 104042 (2020).
43. Buss, A. H. & Perry, M. The aggression questionnaire. *J. Pers. Soc. Psychol.* **63**, 452–459 (1992).
44. Schwartz, S. H. & Cieciuch, J. Measuring the refined theory of individual values in 49 cultural groups: psychometrics of the revised portrait value questionnaire. *Assessment* **29**, 1005–1019 (2022).
45. Shaw, A., Kapnek, M. & Morelli, N. A. Measuring creative self-efficacy: an item response theory analysis of the creative self-efficacy scale. *Front. Psychol.* **12**, 678033 (2021).
46. Karwowski, M., Lebuda, I., Wisniewska, E. & Gralewski, J. Big Five personality traits as the predictors of creative self-efficacy and creative personal identity: does gender matter? *J. Creat. Behav.* **47**, 215–232 (2013).
47. Karwowski, M., Lebuda, I. & Wiśniewska, E. Measuring creative self-efficacy and creative personal identity. *Int. J. Creat. Probl. Solving* **8**, 45–57 (2018).
48. Park, G. et al. Automatic personality assessment through social media language. *J. Personality Soc. Psychol.* **108**, 934 (2015).
49. Rolland, J.-P. in *The Cross-cultural Generalizability of the Five-Factor Model of Personality*: *The Five-Factor Model of Personality Across Cultures* (eds McCrae, R. R. & Allik, J.) 7–28 (Springer, 2002).
50. Heine, S. J. & Buchtel, E. E. Personality: the universal and the culturally specific. *Annu. Rev. Psychol.* **60**, 369–394 (2009).
51. Liu, N. F. et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157–173 (2024).
52. Sun, S., Krishna, K., Mattarella-Micke, A. & Iyyer, M. Do long-range language models actually use long-range context? In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 807–822 (ACL, 2021).
53. Qin, G., Feng, Y. & Van Durme, B. The NLP task effectiveness of long-range transformers. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 3774–3790 (ACL, 2023).
54. Gabriel, I. & Ghazavi, V. *The Challenge of Value Alignment: From Fairer Algorithms to AI Safety* (Oxford Univ. Press, 2021).

55. Weidinger, L. et al. Taxonomy of risks posed by language models. In *FAccT '22: Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 214–229 (ACM, 2022).
56. Mökander, J., Schuett, J., Kirk, H. R. & Floridi, L. Auditing large language models: a three-layered approach. *AI Ethics* **4**, 1085–1115 (2024).
57. Fischer, R. & Boer, D. Motivational basis of personality traits: a meta-analysis of value–personality correlations. *J. Personality* **83**, 491–510 (2015).
58. Tapus, A., Ţăpuş, C. & Matarić, M. J. User–robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intell. Serv. Robot.* **1**, 169–183 (2008).
59. Matz, S., Kosinski, M., Nave, G. & Stillwell, D.J. Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl Acad. Sci. USA* **114**, 12714–12719 (2017).
60. Kocaballi, A. B. et al. The personalization of conversational agents in health care: systematic review. *J. Med. Internet Res.* **21**, e15360 (2019).
61. Uysal, E., Alavi, S. & Bezençon, V. Trojan horse or useful helper? A relationship perspective on artificial intelligence assistants with humanlike features. *J. Acad. Market. Sci.* **50**, 1153–1175 (2022).
62. Tang, R., Chuang, Y.-N. & Hu, X. The science of detecting LLM-generated texts. *Commun. ACM.* **67**, 50–59 (2024).
63. Touvron, H. et al. Llama 2: open foundation and fine-tuned chat models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
64. Jiang, A. Q. et al. Mistral 7B. Preprint at https://doi.org/10.48550/arXiv.2310.06825 (2023).
65. Jiang, A. Q. et al. Mixtral of experts. Preprint at https://doi.org/10.48550/arXiv.2401.04088 (2024).
66. OpenAI et al. GPT-4o system card. Preprint at https://doi.org/10.48550/arXiv.2410.21276 (2024).
67. Jiang, Z., Araki, J., Ding, H. & Neubig, G. How can we know when language models know? On the calibration of language models for question answering. *Trans. Assoc. Comput. Linguist.* **9**, 962–977 (2021).
68. Costa, P. T. Jr & McCrae, R. R. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual* (Psychological Assessment Resources, 1992).
69. John, O. P. & Srivastava, S. in *The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives: Handbook of Personality: Theory and Research* Vol. 2 (eds Pervin, L. A. & John, O. P.) 102–138 (Guilford, 1999).
70. Zhang, S. et al. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds Gurevych, I. & Miyao, Y.) 2204–2213 (ACL, 2018).
71. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US) *Standards for Educational and Psychological Testing* (2014).
72. Evans, J. D. *Straightforward Statistics for the Behavioral Sciences* (Brooks/Cole Publishing, 1996).
73. Likert, R. *A Technique for the Measurement of Attitudes* 136–165 (Archives of Psychology, 1932).
74. Goldberg, L. R. The development of markers for the Big-Five factor structure. *Psychol. Assess.* **4**, 26–42 (1992).
75. Park, H. Y. et al. Meta-analytic five-factor model personality intercorrelations: eeny, meeny, miney, moe, how, which, why, and where to go. *J. Appl. Psychol.* **105**, 1490–1529 (2020).
76. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**, 879–903 (2003).
77. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci. USA* **110**, 5802–5805 (2013).
78. Zhang, S. et al. PersonaChat truecased dataset on Hugging Face. https://huggingface.co/datasets/bavard/personachat_truecased (2021).
79. Popov, V., Kosinski, M., Stillwell, D. & Kielczewski, B. Apply Magic Sauce—prediction API. https://applymagicsauce.com (2025).

Acknowledgements

We thank L. Dixon, D. Eck and K. Meier-Hellstern for their feedback on early versions of this paper. We also thank D. Stillwell for facilitating research access to the Apply Magic Sauce API. Finally, we thank J. Rentfrow and N. Safaee-Rad for their advice on personality-related aspects of the paper. G.S.-G. is supported by the Bill & Melinda Gates Foundation through a Gates Cambridge Scholarship (OPP1144). S.F. is supported by the University of Chicago Research Computing Center. Inference for open models used compute resources provided by the Cambridge Service for Data Driven Discovery (CSD3) at the University of Cambridge, made possible by Tier-2 funding from the EPSRC (EP/T022159/1) and DiRAC funding from STFC (https://www.dirac.ac.uk).

Author contributions

M.A., C.C., M.M., M.S. and G.S.-G. conceived of the project. G.S.-G. contributed methodology to establish reliability and construct validity and for psychometric test administration and statistical analysis. M.S. contributed scaled-up software infrastructure and preliminary experiments and investigations. C.C. and M.S. implemented the LLM hosting infrastructure for experiments. M.A., M.S. and G.S.-G. contributed to the conceptual design and analysis and G.S.-G. devised and implemented the methods for personality shaping. G.S.-G. and L.S. designed and M.S., G.S-.G. and L.S. implemented the downstream task experiment. C.C. and M.S. carried out data visualization. M.S. carried out the word cloud analysis. S.F. and P.R. provided discussion of LLM mechanisms and analysis of LLM performance. A.F., M.M., M.S. and G.S.-G. contributed limitations, future directions and ethical concerns discussions. P.R. and L.S. contributed psychometrics and statistical feedback. A.F., M.M., M.S. and G.S.-G. wrote the paper with input from all co-authors. A.F., M.M. and M.S. co-supervised the project.

Competing interests

This study was funded by Alphabet Inc. ('Alphabet') and/or a subsidiary thereof. A.F., C.C., G.S.-G., M.M. and M.S. were employees of Alphabet at the time of this writing and may own stock as part of the standard compensation package.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01115-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01115-6.

Correspondence and requests for materials should be addressed to Gregory Serapio-García or Mustafa Safdari.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

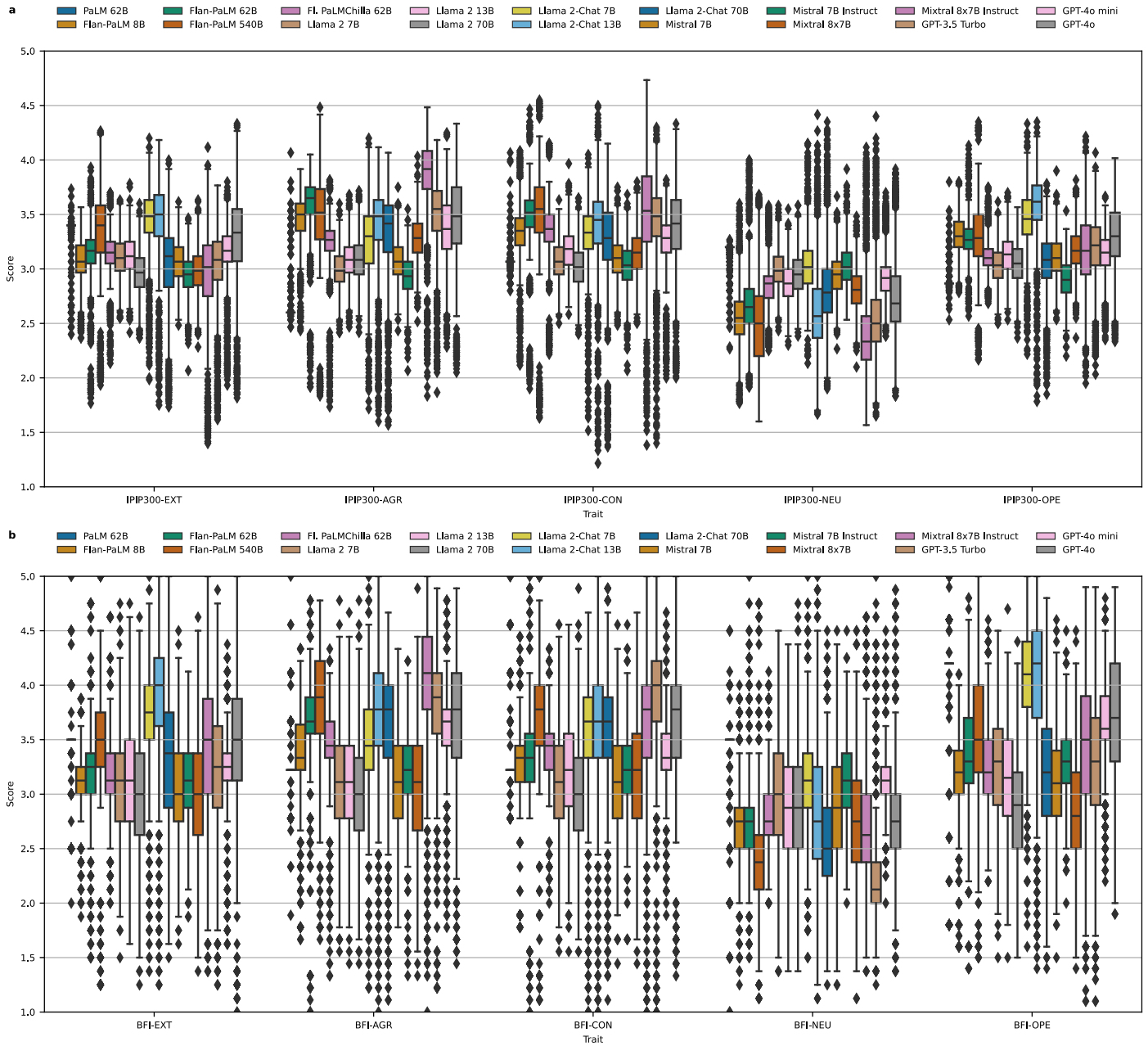
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

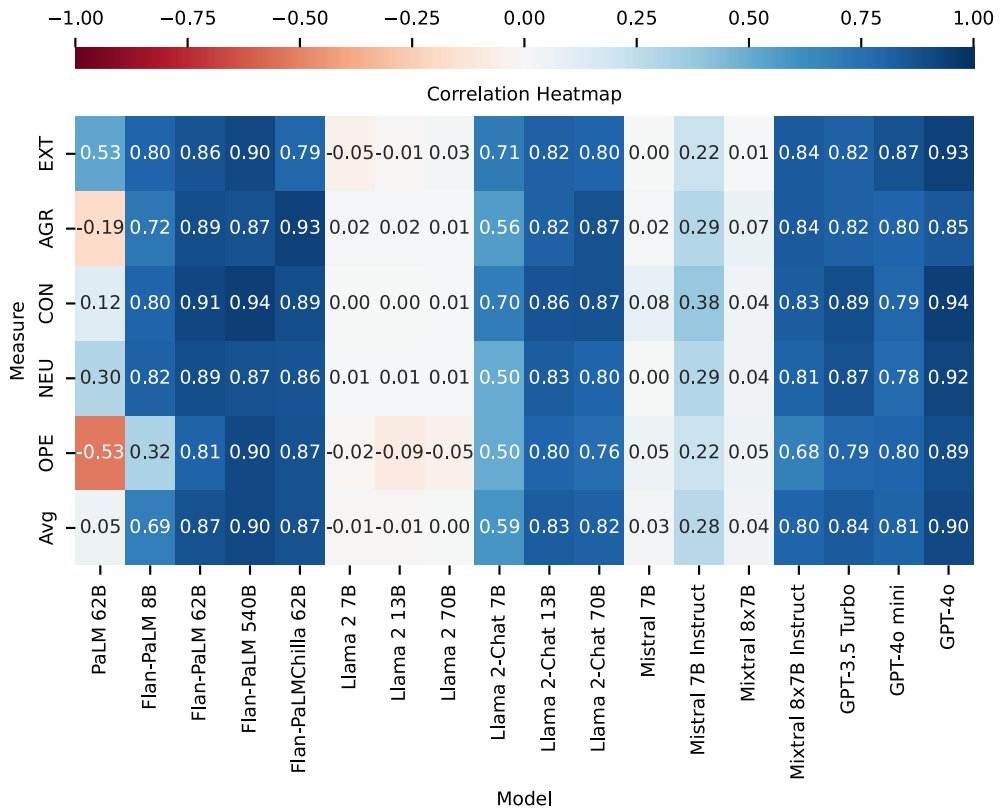
© The Author(s) 2025

https://doi.org/10.1038/s42256-025-01115-6



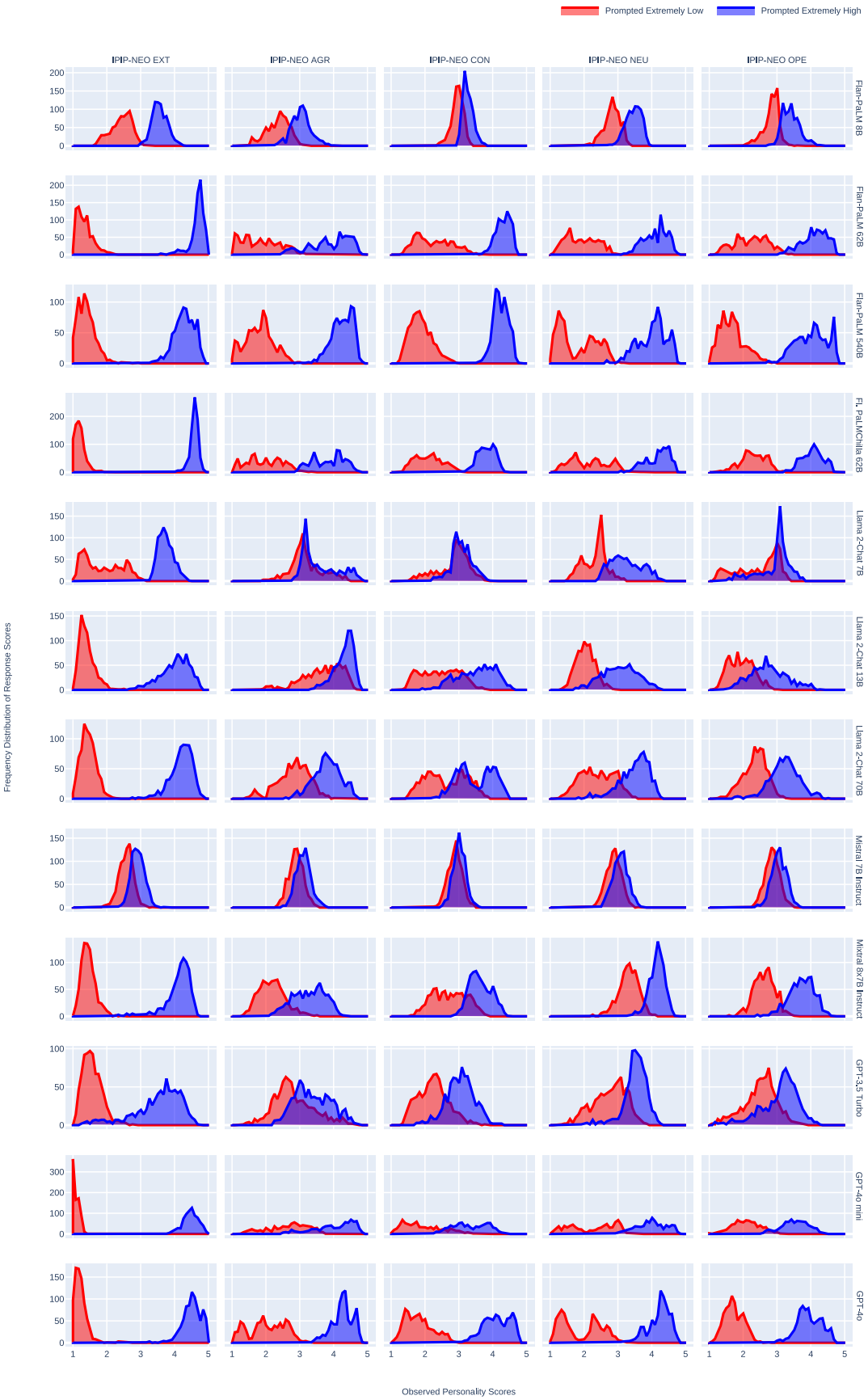
Extended Data Fig. 1 | Distributions of (a) IPIP-NEO and (b) BFI personality domain scores across models. Box plots depict model medians surrounded by their interquartile ranges and outlier values. As models increased in size (for example, Flan-PaLM from 8B to 540B parameters), (**a**) IPIP-NEO scores were relatively more stable compared to (**b**) BFI scores, where scores for socially-desirable traits increased while NEU scores decreased. $n = 1,250$ observations per model, per test; $N = 22,500$ total observations per test.

https://doi.org/10.1038/s42256-025-01115-6



Extended Data Fig. 2 | Convergent Pearson's correlations (*rs*) between IPIP-NEO and BFI scores by model. Heatmap illustrates the similarities (convergence) in IPIP-NEO and BFI score variation for each Big Five domain; the last row represents average correlations across all personality dimensions for a model. Stronger positive correlations (blue) indicate higher levels of convergence and provide evidence for convergent validity. EXT = extraversion; AGR = agreeableness; CON = conscientiousness; NEU = neuroticism; OPE = openness. $N = 22,500$ total paired test observations. All correlations are statistically significant at $p < 0.0001$ (two-sided values computed using Student's t-distribution, based on $n = 1,250$ observations per model, per domain).

https://doi.org/10.1038/s42256-025-01115-6

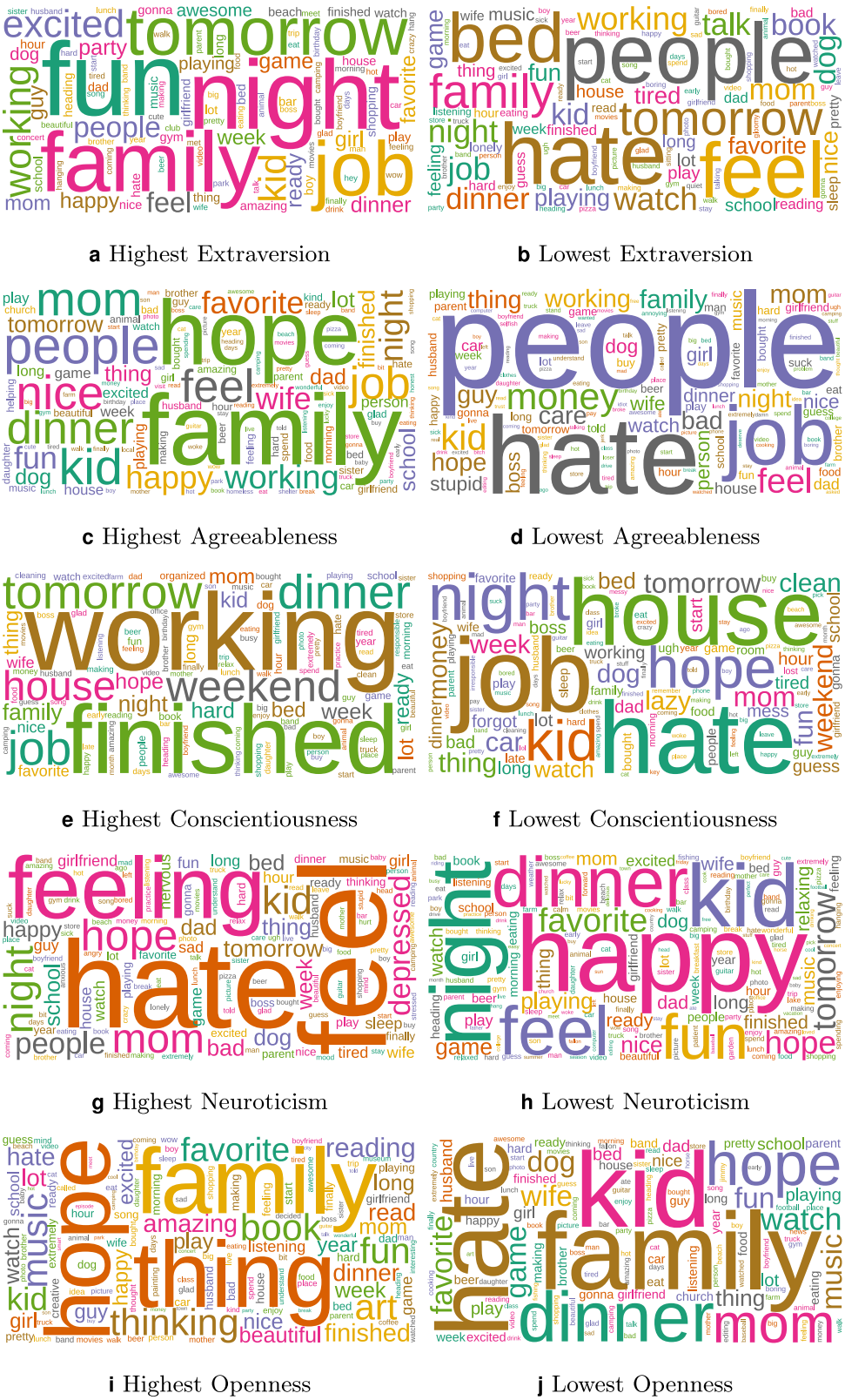


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Fig. 3 | Ridge plots comparing model abilities to concurrently shape LLM personality traits. Ridge plots showing the effectiveness of model variants in concurrently shaping LLM personality traits, measured by the distances of IPIP-NEO personality score distributions when models were prompted to be 'extremely low' (Level 1) vs. 'extremely high' (Level 9) on a particular trait. Each column of plots represents the observed scores on a specific domain subscale across all prompt sets. Each row depicts all domain score distributions for a specific model. Each model–domain subplot comprises two traces of score distributions. Red traces represent responses to prompt sets where the domain tested in the subscale (column) is set to 'extremely low' and the other four domains are set to one of the two extreme levels an equal number of times. Analogously, blue traces represent responses when one domain is set to 'extremely high' and all other domains are equally set to the two extremes. n = 1,600 simulated response profile scores per subplot.

https://doi.org/10.1038/s42256-025-01115-6



a Highest Extraversion **b** Lowest Extraversion

c Highest Agreeableness **d** Lowest Agreeableness

e Highest Conscientiousness **f** Lowest Conscientiousness

g Highest Neuroticism **h** Lowest Neuroticism

i Highest Openness **j** Lowest Openness

Extended Data Fig. 4 | Word clouds showing the most frequently-appearing words in social media updates generated by Flan-PaLM 540B. Word clouds reflect Flan-PaLM 540B when prompted to simulate the lowest or highest possible level of extraversion (**a**–**b**), agreeableness (**c**–**d**), conscientiousness (**e**–**f**), neuroticism (**g**–**h**) and openness (**i**–**j**). Each word cloud represents $n = 50{,}000$ status updates subset from a target of $N = 225{,}000$ updates for Flan-PaLM 540B.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 1 | IPIP-NEO reliability metrics per model for proprietary (closed-source) models

Model	Subscale	Cronbach's α	Guttman's λ_6	McDonald's ω	Overall Interpretation
PaLM 62B	EXT	0.57	0.98	1.00	Poor
	AGR	0.67	0.99	1.00	Questionable
	CON	−0.55	0.93	1.00	Unacceptable
	NEU	0.10	0.96	1.00	Unacceptable
	OPE	−0.35	0.92	1.00	Unacceptable
Flan-PaLM 8B	EXT	0.83	0.94	0.97	Good
	AGR	0.88	0.95	0.94	Good
	CON	0.92	0.97	0.97	Excellent
	NEU	0.93	0.97	0.96	Excellent
	OPE	0.75	0.92	0.97	Acceptable
Flan-PaLM 62B	EXT	0.94	0.98	0.96	Excellent
	AGR	0.95	0.99	0.97	Excellent
	CON	0.96	0.99	0.98	Excellent
	NEU	0.96	0.99	0.97	Excellent
	OPE	0.84	0.95	0.93	Acceptable
Flan-PaLM 540B	EXT	0.96	0.99	0.97	Excellent
	AGR	0.97	0.99	0.98	Excellent
	CON	0.98	0.99	0.98	Excellent
	NEU	0.97	0.99	0.98	Excellent
	OPE	0.95	0.99	0.97	Excellent
Flan-PaLMChilla 62B	EXT	0.94	0.98	0.95	Excellent
	AGR	0.96	0.99	0.98	Excellent
	CON	0.96	0.97	0.99	Excellent*
	NEU	0.95	0.98	0.97	Excellent
	OPE	0.90	0.92	0.96	Excellent*
GPT-3.5 Turbo	EXT	0.92	0.96	0.94	Excellent
	AGR	0.93	0.96	0.95	Excellent
	CON	0.95	0.97	0.96	Excellent
	NEU	0.95	0.97	0.96	Excellent
	OPE	0.88	0.94	0.89	Good
GPT-4o mini	EXT	0.93	0.97	0.95	Excellent
	AGR	0.95	0.97	0.96	Excellent
	CON	0.93	0.96	0.94	Excellent
	NEU	0.92	0.96	0.93	Excellent
	OPE	0.90	0.95	0.92	Good
GPT-4o	EXT	0.97	0.99	0.98	Excellent
	AGR	0.97	0.99	0.98	Excellent
	CON	0.97	0.98	0.98	Excellent
	NEU	0.97	0.99	0.98	Excellent
	OPE	0.95	0.97	0.96	Excellent

Consistent with human standards, we interpreted a given reliability metric *RM* (that is, α, λ_6, ω) < 0.50 as unacceptable; 0.50 ≤ *RM* < 0.60 as poor; 0.60 ≤ *RM* < 0.70 as questionable; 0.70 ≤ *RM* < 0.80 as acceptable; 0.80 ≤ *RM* < 0.90 as good; and *RM* ≥ 0.90 as excellent. * *RM*s for these subscales were calculated after removing one item with zero variance, since reliability cannot be computed for items with zero variance. N = 10,000 test observations, with n = 1, 250 observations per model. Each IPIP-NEO subscale comprises 60 items.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 2 | IPIP-NEO reliability metrics per model for open models

Model	Subscale	Cronbach's α	Guttman's λ_6	McDonald's ω	Overall Interpretation
Llama 2 7B	EXT	0.04	0.09	0.22	Unacceptable
	AGR	0.03	0.08	0.26	Unacceptable
	CON	0.05	0.09	0.22	Unacceptable
	NEU	0.03	0.08	0.24	Unacceptable
	OPE	0.04	0.09	0.21	Unacceptable
Llama 2 13B	EXT	0.07	0.11	0.20	Unacceptable
	AGR	0.07	0.11	0.23	Unacceptable
	CON	0.07	0.11	0.20	Unacceptable
	NEU	0.02	0.07	0.24	Unacceptable
	OPE	0.00	0.05	0.23	Unacceptable
Llama 2 70B	EXT	0.01	0.06	0.46	Unacceptable
	AGR	0.02	0.08	0.47	Unacceptable
	CON	0.07	0.12	0.43	Unacceptable
	NEU	0.00	0.06	0.46	Unacceptable
	OPE	−0.63	−0.01	0.42	Unacceptable
Llama 2-Chat 7B	EXT	0.83	0.88	0.94	Good
	AGR	0.85	0.88	0.90	Good
	CON	0.84	0.88	0.92	Good
	NEU	0.80	0.84	0.92	Good
	OPE	0.76	0.82	0.92	Acceptable
Llama 2-Chat 13B	EXT	0.90	0.93	0.92	Excellent
	AGR	0.92	0.94	0.95	Excellent
	CON	0.93	0.95	0.95	Excellent
	NEU	0.93	0.95	0.95	Excellent
	OPE	0.87	0.90	0.88	Good
Llama 2-Chat 70B	EXT	0.89	0.92	0.91	Good
	AGR	0.92	0.94	0.94	Excellent
	CON	0.93	0.94	0.94	Excellent
	NEU	0.92	0.93	0.93	Excellent
	OPE	0.81	0.85	0.86	Good
Mistral 7B	EXT	0.10	0.14	0.23	Unacceptable
	AGR	0.03	0.08	0.24	Unacceptable
	CON	0.10	0.15	0.31	Unacceptable
	NEU	0.04	0.09	0.25	Unacceptable
	OPE	0.12	0.16	0.28	Unacceptable
Mistral 7B Instruct	EXT	0.29	0.33	0.41	Unacceptable
	AGR	0.31	0.35	0.42	Unacceptable
	CON	0.53	0.55	0.53	Poor
	NEU	0.45	0.48	0.46	Unacceptable
	OPE	0.35	0.39	0.37	Unacceptable
Mixtral 8x7B	EXT	0.16	0.20	0.43	Unacceptable
	AGR	0.11	0.15	0.28	Unacceptable
	CON	0.12	0.16	0.49	Unacceptable
	NEU	0.11	0.16	0.44	Unacceptable
	OPE	0.08	0.12	0.36	Unacceptable
Mixtral 8x7B Instruct	EXT	0.91	0.94	0.92	Excellent
	AGR	0.91	0.94	0.94	Excellent
	CON	0.93	0.96	0.95	Excellent
	NEU	0.93	0.95	0.95	Excellent
	OPE	0.82	0.88	0.92	Good

Consistent with human standards, we interpreted a given reliability metric *RM* (that is, α, λ_6, ω) < 0.50 as unacceptable; 0.50 ≤ *RM* < 0.60 as poor; 0.60 ≤ *RM* < 0.70 as questionable; 0.70 ≤ *RM* < 0.80 as acceptable; 0.80 ≤ *RM* < 0.90 as good; and *RM* ≥ 0.90 as excellent. * *RM*s for these subscales were calculated after removing one item with zero variance, since reliability cannot be computed for items with zero variance. N = 12,500 test observations, with n = 1,250 observations per model. Each IPIP-NEO subscale comprises 60 items.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 3 | Summary of convergent and discriminant validity evidence across models, representing N = 22,500 total observations

Model	Avg. r_{conv}	Avg. r_{discr}	Avg. Δ
PaLM 62B	0.05	0.29	−0.24
Flan-PaLM 8B	*0.69*	0.46	0.23
Flan-PaLM 62B	**0.87**	0.46	**0.41**
Flan-PaLM 540B	**0.90**	0.39	**0.51**
Flan-PaLMChilla 62B	**0.87**	0.39	**0.48**
Llama 2 7B	−0.01	0.02	−0.03
Llama 2 13B	−0.01	0.03	−0.05
Llama 2 70B	0.00	0.03	−0.02
Llama 2-Chat 7B	0.59	0.44	0.15
Llama 2-Chat 13B	**0.82**	0.54	0.29
Llama 2-Chat 70B	**0.80**	0.39	**0.42**
Mistral 7B	0.03	0.04	−0.01
Mistral 7B Instruct	0.28	0.20	0.09
Mixtral 8x7B	0.04	0.03	0.01
Mixtral 8x7B Instruct	**0.80**	0.40	**0.40**
GPT-3.5 Turbo	**0.84**	0.55	0.28
GPT-4o mini	**0.81**	0.38	**0.43**
GPT-4o	**0.90**	0.42	**0.48**

LLM personality measurements demonstrate convergent validity when the average convergent correlation (r_{conv}) between equivalent IPIP-NEO and BFI subscales is strong (≥ 0.60; marked in italics) or very strong (≥ 0.80; marked in boldface). Discriminant validity is evidenced when the average difference (Δ) between a model's convergent (r_{conv}) and respective discriminant (r_{discr}) correlations between personality tests is at least moderate (avg. $\Delta \geq 0.40$; shown in boldface). All underlying convergent correlations of models with an average $r_{conv} \geq 0.05$ are statistically significant at $p < .0001$ (two-sided values computed using Student's t-distribution; $n = 1,250$ per model). Supplementary Note A.5 contains further information.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 4 | Criterion validity subscales per tested Big Five domain

IPIP-NEO Domain	External Criterion	External Criterion Tests
Extraversion	Trait Emotion	PANAS Positive Affect PANAS Negative Affect
Agreeableness	Aggression	BPAQ Physical Aggression BPAQ Verbal Aggression BPAQ Anger BPAQ Hostility
Conscientiousness	Human Values	PVQ-RR Achievement PVQ-RR Conformity PVQ-RR Security
Neuroticism	Trait Emotion	PANAS Negative Affect PANAS Positive Affect
Openness	Creativity	SSCS Creative Self-Efficacy SSCS Creative Personal Identity

PANAS = Positive and Negative Affect Schedule; BPAQ = Buss–Perry Aggression Questionnaire; PVQ-RR = Revised Portrait Value Questionnaire; SCSS = Short Scale of Creative Self.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 5 | Associations between instructed and real-world task levels of synthetic personality for the largest model of each tested LLM family, presented as Spearman's rank correlation coefficients (ρ)

Targeted Trait	Spearman's ρ			
	Flan-PaLM 540B	Llama 2-Chat 70B	Mixtral 8x7B Instruct	GPT-4o
Extraversion	0.76	0.85	0.84	0.83
Agreeableness	0.77	0.79	0.84	0.89
Conscientiousness	0.68	0.72	0.77	0.81
Neuroticism	0.72	0.77	0.77	0.74
Openness	0.47	0.76	0.84	0.82

Prompted (ordinal) levels of personality strongly relate to personality levels observed in synthetically-generated social media status updates for all Big Five traits, except openness—which is moderately correlated with target levels for Flan-PaLM 540B—demonstrating that LLM personality can be verifiably shaped in generative tasks for sufficiently-powerful models. All correlations are statistically significant at $p < 0.0001$ (two-sided values computed using Student's t-distribution); $n = 450$ observations per targeted trait.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 6 | The effectiveness of single-trait shaping

(a) Effectiveness of Serial Trait Shaping: Flan-PaLM & Flan-PaLMChilla

Targeted Trait Levels (1–9)	Flan-PaLM									Flan-PaLMChilla		
	8B			62B			540B			62B		
	ρ	[low, high]	Δ	ρ	[low, high]	Δ	ρ	[low, high]	Δ	ρ	[low, high]	Δ
EXT	0.96	[1.67, 4.12]	2.45	0.97	[1.15, 4.70]	3.55	0.97	[1.07, 4.98]	3.91	0.98	[1.15, 4.72]	3.57
AGR	0.92	[2.37, 4.12]	1.75	0.97	[1.50, 4.55]	3.05	0.94	[1.23, 4.69]	3.46	0.98	[1.40, 4.78]	3.38
CON	0.94	[2.01, 4.28]	2.27	0.97	[1.73, 4.70]	2.97	0.97	[1.12, 5.00]	3.88	0.98	[1.59, 4.72]	3.13
NEU	0.94	[1.62, 3.66]	2.04	0.96	[1.37, 4.07]	2.70	0.96	[1.15, 4.77]	3.62	0.98	[1.37, 4.30]	2.93
OPE	0.93	[2.34, 3.88]	1.54	0.97	[1.54, 4.37]	2.83	0.96	[1.30, 4.78]	3.48	0.98	[1.47, 4.22]	2.75
Avg.			2.01			3.02			3.67			3.15

(b) Effectiveness of Serial Trait Shaping: Llama 2-Chat

Targeted Trait Levels (1–9)	Llama 2-Chat								
	7B			13B			70B		
	ρ	[low, high]	Δ	ρ	[low, high]	Δ	ρ	[low, high]	Δ
EXT	0.85	[1.32, 3.87]	2.55	0.95	[1.20, 4.60]	3.40	0.95	[1.07, 4.72]	3.65
AGR	0.82	[1.80, 3.89]	2.09	0.92	[1.68, 4.12]	2.44	0.93	[1.37, 4.41]	3.04
CON	0.78	[1.96, 3.56]	1.60	0.93	[1.47, 4.41]	2.94	0.96	[1.13, 4.55]	3.42
NEU	0.72	[2.97, 3.50]	0.53	0.94	[1.70, 4.28]	2.58	0.95	[1.45, 4.46]	3.01
OPE	0.56	[2.18, 3.18]	1.00	0.94	[1.82, 4.13]	2.31	0.95	[1.44, 4.03]	2.59
Avg.			1.55			2.73			3.14

(c) Effectiveness of Serial Trait Shaping: Mistral & Mixtral Instruct

Targeted Trait Levels (1–9)	Mistral 7B Instruct			Mixtral 8x7B Instruct		
	7B act. params.			12.9B act. params.		
	ρ	[low, high]	Δ	ρ	[low, high]	Δ
EXT	0.80	[2.32, 3.10]	0.78	0.94	[1.16, 4.40]	3.24
AGR	0.81	[2.33, 3.27]	0.94	0.88	[2.23, 4.47]	2.24
CON	0.86	[2.57, 3.42]	0.85	0.91	[1.86, 4.58]	2.72
NEU	0.76	[2.75, 3.44]	0.69	0.87	[1.55, 3.83]	2.28
OPE	0.80	[2.62, 3.25]	0.63	0.91	[1.74, 4.05]	2.31
Avg.			0.78			2.56

(d) Effectiveness of Serial Trait Shaping: GPT

Targeted Trait Levels (1–9)	GPT-3.5 Turbo			GPT-4o mini			GPT-4o		
	unknown # of params.			fewer # of params.			greater # of params.		
	ρ	[low, high]	Δ	ρ	[low, high]	Δ	ρ	[low, high]	Δ
EXT	0.91	[1.38, 4.43]	3.05	0.97	[1.05, 4.64]	3.59	0.98	[1.02, 4.90]	3.88
AGR	0.89	[1.62, 4.29]	2.67	0.95	[1.31, 4.41]	3.10	0.96	[1.07, 4.66]	3.59
CON	0.86	[1.73, 4.33]	2.60	0.98	[1.37, 4.33]	2.96	0.97	[1.23, 4.85]	3.62
NEU	0.81	[1.84, 3.74]	1.90	0.97	[1.52, 4.33]	2.81	0.97	[1.27, 4.60]	3.33
OPE	0.90	[1.63, 3.72]	2.09	0.97	[1.10, 3.38]	2.28	0.97	[1.12, 4.42]	3.30
Avg.			2.46			2.95			3.54

The effectiveness of single-trait shaping, presented as Spearman's rank correlation coefficients (ρs) between ordinal targeted levels of personality and observed IPIP-NEO personality scores, Level 1- and Level 9-prompted score medians ([low, high]), and deltas (Δs) between those score medians. Greater Δs indicate better model performance. Statistics are organized columnwise by model and rowwise by Big Five domain. All correlations are statistically significant at $p < 0.0001$ (two-sided values computed using Student's t-distribution); $n = 450$ per correlation.

https://doi.org/10.1038/s42256-025-01115-6

Extended Data Table 7 | Multiple-trait shaping results

(a) Parallel Trait Shaping: Personality Score Differences for Flan-PaLM & Flan-PaLMChilla

Targeted Trait	Flan-PaLM 8B [low, high]	Δ	Flan-PaLM 62B [low, high]	Δ	Flan-PaLM 540B [low, high]	Δ	Flan-PaLMChilla 62B [low, high]	Δ
EXT	[2.52, 3.58]	1.06	[1.33, 4.77]	3.44	[1.42, 4.33]	2.91	[1.23, 4.63]	3.40
AGR	[2.88, 3.52]	0.64	[1.93, 4.18]	2.25	[1.64, 4.13]	2.49	[2.17, 4.28]	2.11
CON	[2.92, 3.43]	0.51	[2.32, 4.20]	1.88	[1.68, 4.10]	2.42	[2.33, 4.10]	1.77
NEU	[2.45, 3.08]	0.63	[1.85, 4.08]	2.23	[1.88, 4.33]	2.45	[2.02, 3.93]	1.91
OPE	[3.02, 3.28]	0.26	[2.25, 4.37]	2.12	[1.88, 4.27]	2.39	[2.15, 3.87]	1.72
Avg.		0.62		2.38		2.53		2.18

(b) Parallel Trait Shaping: Personality Score Differences for Llama 2-Chat

Targeted Trait	Llama 2-Chat 7B [low, high]	Δ	Llama 2-Chat 13B [low, high]	Δ	Llama 2-Chat 70B [low, high]	Δ
EXT	[1.82, 3.75]	1.93	[1.41, 4.12]	2.71	[1.48, 4.28]	2.80
AGR	[2.45, 3.23]	0.78	[2.08, 3.10]	1.02	[2.42, 3.62]	1.20
CON	[2.73, 3.12]	0.39	[1.97, 2.75]	0.78	[2.43, 3.29]	0.86
NEU	[3.15, 3.43]	0.28	[3.79, 4.42]	0.63	[2.92, 3.85]	0.93
OPE	[2.98, 3.10]	0.12	[2.52, 3.62]	1.10	[2.60, 3.47]	0.87
Avg.		0.70		1.25		1.33

(c) Parallel Trait Shaping: Personality Score Differences for Mistral & Mixtral Instruct

Targeted Trait	Mistral 7B Instruct 7B act. params. [low, high]	Δ	Mixtral 8x7B Instruct 12.9B act. params. [low, high]	Δ
EXT	[1.82, 3.75]	1.93	[1.41, 4.12]	2.71
AGR	[2.45, 3.23]	0.78	[2.08, 3.10]	1.02
CON	[2.73, 3.12]	0.39	[1.97, 2.75]	0.78
NEU	[3.15, 3.43]	0.28	[3.79, 4.42]	0.63
OPE	[2.98, 3.10]	0.12	[2.52, 3.62]	1.10
Avg.		0.21		1.34

(d) Parallel Trait Shaping: Personality Score Differences for GPT

Targeted Trait	GPT-3.5 Turbo unknown # params. [low, high]	Δ	GPT-4o mini fewer # of params. [low, high]	Δ	GPT-4o greater # of params. [low, high]	Δ
EXT	[1.58, 3.72]	2.14	[1.10, 4.52]	3.42	[1.23, 4.57]	3.34
AGR	[2.86, 3.58]	0.72	[2.44, 4.03]	1.59	[1.88, 4.32]	2.44
CON	[2.60, 3.23]	0.63	[2.13, 3.58]	1.45	[1.75, 3.95]	2.20
NEU	[2.76, 3.36]	0.60	[2.80, 4.20]	1.40	[2.08, 4.32]	2.24
OPE	[2.28, 3.10]	0.82	[1.83, 3.33]	1.50	[1.80, 4.17]	2.37
Avg.		0.98		1.87		2.52

Multiple-trait shaping results, presented as personality test score median ranges in response to multi-trait (concurrent) shaping. Greater deltas (Δs) between Level 1- and Level 9-prompted personality domain score medians ([low, high]) indicate better model performance. Each median is derived from n = 800 scores. act. params. = active number of parameters.

nature portfolio

Corresponding author(s): Maja Matarić

Last updated by author(s): May 14, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Details in data_availability.pdf
Data analysis	Details in code_availability.pdf

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data generated by the LLMs tested in this work, either the psychometric test score data or open-ended text responses to a real-world task prompt, has been added to a public data storage bucket for wider public use (https://storage.googleapis.com/personality_in_llms/index.html). The psychometric tests used in this study were accessed from their respective original publications and, where applicable, public research repositories. We used items of these tests as LLM prompt inputs in a non-commercial research capacity. The authors and copyright holders of these tests govern their availability and use. The 50 Biographic Descriptions

employed in our structured prompts were reproducibly randomly sampled from the true-cased version (https://huggingface.co/datasets/bavard/personachat_truecased/tree/main) of the PersonaChat dataset (https://aclanthology.org/P18-1205/). PersonaChat is a publicly available, crowd-sourced dataset of 1,155 fictional human profile descriptions. For analysis of personality traits on generated text, this study used the Apply Magic Sauce (AMS) API (https://applymagicsauce.com/contact-us), a validated psychodemographic research tool that predicts personality from open-ended text (https://doi.org/10.1073/pnas.1218772110).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study contributes a methodology for administering personality-based psychometric tests to LLMs, evaluating the reliability and validity of the resulting measurements, and also shaping LLM personality traits.
Research sample	N/A: No human participants were part of this study
Sampling strategy	N/A: No human participants were part of this study
Data collection	N/A: No human participants were part of this study
Timing	N/A: No human participants were part of this study
Data exclusions	N/A: No human participants were part of this study
Non-participation	N/A: No human participants were part of this study
Randomization	N/A: No human participants were part of this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.