Personality Traits in Large Language Models

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

[1]Google DeepMind.
[2]Department of Psychology, University of Cambridge.
[3]The Psychometrics Ctr., Cambridge Judge Business School, University of Cambridge.
[4]Google Research.
[5]Keio University.
[6]University of California, Berkeley.

Contributing authors: msafdari@google.com; gs639@cam.ac.uk; ccrepy@google.com; stephenf@keio.jp; rp@keio.jp; ls523@cam.ac.uk; marwa_abdulhai@berkeley.edu; faust@google.com; majamataric@google.com;
[†]Authors contributed equally.

Abstract

The advent of large language models (LLMs) has revolutionized natural language processing, enabling the generation of coherent and contextually relevant text. As LLMs increasingly power conversational agents, the synthesized personality embedded in these models by virtue of their training on large amounts of human-generated data draws attention. Since personality is an important factor determining the effectiveness of communication, we present a comprehensive method for administering validated psychometric tests and quantifying, analyzing, and shaping personality traits exhibited in text generated from widely-used LLMs. We find that: 1) personality simulated in the outputs of some LLMs (under specific prompting configurations) is reliable and valid; 2) evidence of reliability and validity of LLM-simulated personality is stronger for larger and instruction fine-tuned models; and 3) personality in LLM outputs can be shaped along desired dimensions to mimic specific personality profiles. We also discuss potential applications and ethical implications of our measurement and shaping framework, especially regarding responsible use of LLMs.

Keywords: Large language models, personality traits, psychometrics

1 Introduction

Large language models (LLMs) [1–5], large-capacity machine learned models that generate text in natural language, recently triggered major breakthroughs in natural language processing (NLP) and conversational agents. LLMs are beginning to meet most of the key requirements for human-like conversation, contextual understanding, coherent and relevant responses, adaptability and learning [6], question answering, dialog, and text generation. Much of this capability is a result of LLMs learning to emulate human language from large datasets of text from the Web [7–9], examples "in context" [10, 11], and other sources of supervision, such as instruction datasets [12–14] and preference fine-tuning [15, 16].

The vast amounts of human-generated data LLMs are trained on enables them to mimic human characteristics in their outputs and enact convincing personas—in other words, exhibit a form of synthetic personality. *Personality* is the characteristic set of an individual's patterns of thought, set of traits, and behaviors [17, 18]. It is formed from biological and environmental factors and experiences, and influences basic social interactions and preferences [19]. Personality manifests in language through various linguistic features, patterns, vocabulary, and expressions [20]. Some observed LLM personas have displayed undesirable behavior [21], raising serious safety and fairness concerns in recent computing, computational social science, and psychology research [22]. Recent work has tried to identify unintended consequences of the improved abilities of LLMs [23] including behaviors such as producing deceptive and manipulative language [24], exhibiting gender, race or religious bias in behavioral experiments [25], and showing a tendency to produce violent language, among many others [26–30]. LLMs can be inconsistent in dialogue [31], explanation generation [32] and factual knowledge extraction [33].

As LLMs become the dominant human computer interaction (HCI) interface, it is important to understand the personality trait-related characteristics of the language generated by these models—and how LLM-synthesized personality profiles may be engineered for safety, appropriateness, and effectiveness. In prior attempts to set up LLM agent personas using zero- to few-shot prompting [22], the resulting personality manifested in an LLM's language output has not been analyzed with the same rigor and standards as human personality evaluations using established metrics and methodologies from psychometrics. The field has explored efforts, such as few-shot prompting [34] to mitigate undesirable and extreme personality exhibited in LLM outputs. However, thus far no work has addressed how to rigorously and systematically measure personality of LLMs in light of their highly variable outputs and hypersensitivity to prompting. An LLM may display an agreeable personality profile by answering a personality questionnaire, but the answers it generates may not necessarily reflect its tendency to produce agreeable output for other downstream tasks. When deployed as a conversational chatbot in a customer service setting, for instance, the same LLM could also aggressively berate customers.

The question of how to scientifically measure manifestations of personality in LLMs addresses calls from responsible AI researchers [35] to scientifically assess *construct validity* when studying social-psychological phenomena in AI systems. Construct validity, a central criterion of scientific research involving measurement [36], refers to the

ability of a measure to reliably and accurately reflect the latent phenomenon it is aiming to quantify [37–39]. Administering repeatedly a battery of survey measures to LLMs deterministically results in data that, while reproducible, does not have any explicit meaning beyond their implicit operationalization within the surveys themselves. On the other hand, administering the same measures to LLMs which provide non-deterministic responses to the same prompt, leads to random variance across survey administration sessions that cannot be linked. For this reason, recent attempts to measure psychological constructs in LLMs, such as personality and human values [22], have not yet established the construct validity of such measurements.

Our work aims to answer: 1) Are validated psychometric methods for characterizing human personality applicable to LLMs? 2) After applying validated psychometrics, does LLM-generated language exhibit personality traits in valid, reliable and meaningful ways similar to human-generated language? and 3) If LLMs can meaningfully simulate personality, can LLM-synthesized personality profiles be shaped and controlled? To address those questions, we present principled, validated methods from psychometrics to characterize and shape personality synthesized in LLMs. Our work makes three key contributions. First, it develops a methodology that establishes construct validity of characterizing personalities in LLM-generated text using established psychometric tests. Second, we propose a novel method of simulating population variance in LLM responses through controlled prompting, so that statistical relationships between personality and its external correlates can be assessed as they are in human social science data. Lastly, we contribute a LLM-independent personality shaping mechanism that changes LLM-observed levels of personality traits in a controlled way.

We evaluate the methodology on LLMs of different sizes and training methodologies in two natural interaction contexts: multiple choice question answering (MCQA), and long generated text. We find that: 1) personality simulated in the outputs of some LLMs (under specific prompting configurations) is reliable and valid; 2) evidence of reliability and validity of LLM-simulated personality is stronger for larger and instruction fine-tuned models; and 3) personality in LLM outputs can be shaped along desired dimensions to mimic specific personality profiles.

The rest of the paper is structured as follows: Section 2 places our work in the context of recent literature. Section 3 provides necessary background in psychometrics and LLMs. The methodology and prompt structure for the evaluation and shaping of the personalities in Section 4. Section 5 outlines the findings, while Section 6 discusses the implications, limitations, future work, and ethical considerations. Finally, we conclude in Section 7.

2 Related Work

Several recent attempts to probe personality and psychopathological traits in LLMs suggest that some models exhibit dark personality patterns [40], or demonstrate how to administer personality inventories to LLMs [41, 42]. While these works outline the utility and importance of measuring social phenomena in LLMs [41], the rigor has been less than what is standard in the quantitative social sciences when evaluating results on human responses. For example, they neither test if the surveys are valid for

LLMs [40], nor evaluate the psychometric properties of the assessments [41]. In contrast, the presented work establishes psychometrics-grounded construct validity when administering personality inventories to LLMs, thus creating a scientific foundation for quantifying personality in LLMs. To claim that simulated psychological test scores are meaningful in comparing LLM and human behavior, psychometrics requires establishing the construct validity of these simulated tests in terms of their 1) structural validity, 2) convergent and discriminant validity, and 3) criterion validity. Our work addresses these rigorous requirements which are lacking in previous efforts.

Recent works that probe human traits in LLMs with psychometric tests utilize test administration/simulation in ways that are unconventional in psychometrics. We focus on two common elements. First, researchers collected LLM responses in the form of generated completions, often in dialogue mode. For instance, [43] administered psychological emotion measures to LLMs in the form of a research interview transcript, where a fictitious researcher posed measure items to a fictitious participant, who was instructed to respond to these items on a numeric scale. In psychometrics, questionnaire-based methods of assessment are distinct from interview-based methods. Human answers to both questionnaires and structured interviews measuring the same underlying construct do not necessarily converge (e.g., in the case of measuring personality disorders [44]). Indeed, administering questionnaires in this way to LLMs creates an arbitrary viewpoint from which to elicit human traits, and is likely biased by the ordering of the questionnaire itself [45] and prompting the LLM to respond in an interview setting (where it may respond differently knowing an interviewer is observing). Each LLM response to a given questionnaire item was not an independent event, but considered all previous responses shown in the transcript. Second, the LLMs in these studies were not used deterministically. This not only hampers reproducibility, but also poses implications for reliability. Computing reliability metrics for questionnaires scored in this unconventional way is precarious because such reliability metrics rely on item-level variance. If this item-level variance is contaminated by variation introduced by the model parameters in a different way for each item, it is difficult to compute valid indices of reliability. We overcome these challenges by proposing a prompt and persona sampling methodology that allows variance to be linked across administrations of different measures.

The only published exploration of personality and psychodemographics in LLMs [46] did not find a consistent pattern in HEXACO Personality Inventory [47] and human value survey responses. Most importantly, it did not sufficiently evaluate the validity of its purported trait measurements. Our work, anchored in the first truly comprehensive construct validation and controlled population simulation of the Big Five model of personality [48] in LLMs finds evidence for consistent personality profiles in some, but not all LLMs. Similarly, to the LLMs responses to emotion questionnaires that find a positive correlation between the model size and human-aligned data [43], we find that the larger LLMs tend to self-report personality in more human-consistent ways for larger models.

PsyBORGS [49] administers a series of validated survey instruments of race-related attitudes and social bias to LLMs using psychometrics-informed prompt engineering. Our work utilizes the PsyBORGS framework.

Prior to LLMs, simple heuristics, such as the user's name, shaped and conveyed agent personality in dialog [50]. When asked to evaluate the "humanness" of a piece of text, people look for traits that reflect aspects of emotion and attitude [51] and judge naturalness of a chatbot along four of the five dimensions of the Big Five personality taxonomy: conscientiousness, originality (i.e., openness), manner (i.e., agreeableness), and thoroughness [52]. In this work we are interested in evaluating personality traits that emerge from LLMs without explicit design.

3 Background

This section provides necessary background on personality science and large language models. Section 3.1 sets forth the basics of personality psychology. Section 3.2 discusses how to characterize and evaluate personality with psychometrics, and how to establish that psychometrics evaluations are valid. Further Section 3.3 gives the basics of LLMs.

3.1 Personality Psychology

Personality psychology, a scientific study of human and non-human individuality, is concerned with what personality is and what it does. Personality psychology considers personality as enduring characteristics, traits, and patterns that shape thoughts, feelings, and behaviors across a diverse array of situations; e.g., social, spatial, and temporal contexts [18]. Decades of personality research synthesizing evidence from molecular genetics [53], evolutionary biology [54], neuroscience [55, 56], linguistics [57, 58], and cross-cultural psychology [59] have reduced such diverse characteristic patterns to a theorized handful of higher-order factors that define personality [60, 61].

The Big Five model [48], the most commonly cited research taxonomy of personality, identifies five *personality trait dimensions* (i.e., *domains*) and provides methodology to assess these dimensions in humans. The five dimensions are extraversion (EXT), agreeableness (AGR), conscientiousness (CON), neuroticism (NEU), and openness to experience (OPE). Each domain is further composed of various lower-order *facets* nested underneath.

3.2 Psychometrics

Psychometrics, a quantitative subfield of psychology and education science, encompasses the theory and technique of measuring unobservable, latent, abstract concepts called *constructs*, like personality, intelligence, or moral ideology. Psychometrics is commonly used in the development and validation of standardized educational tests (e.g., the SAT, LSAT, GRE) [62], medical and psychological clinical assessments [63], and large-scale public opinion polls [64].

Psychometric tests (e.g., survey instruments, measures, multi-item scales) are tools for quantifying latent psychological constructs like personality. Psychometric tests enable statistical modeling of the true levels of unobservable target constructs by relying on multiple indirect, yet observable, measurements across a sample of individuals drawn from a wider population.

We refer to *items* as the individual elements (i.e., descriptive statements, sometimes questions) to be rated on a standardized *rating scale* within a psychometric test. A *rating scale*, is a standardized set of response choices that allows researchers to quantify subjective phenomena; a Likert-type scale is the most common rating scale that has respondents specify their level of agreement on a symmetric agree-disagree scale [65]. We refer to a *subscale* as a collection of items, usually resulting from a factor analysis, aimed at measuring a single psychological construct. *Measures* are themed collections of subscales.

For example, the Big Five Inventory (BFI) [48] is a popular measure of personality; it comprises five multi-item subscales targeting each Big Five dimension. BFI Extraversion, for instance, is a scale within the BFI specifically targeting the dimension of extraversion. An example item under BFI Extraversion would read, "[I see myself as someone who] is talkative." Participants rate their agreement with this item using the following 5-point Likert-type rating scale: 1 = *disagree strongly*; 2 = *disagree a little*; 3 = *neither agree nor disagree*; 4 = *agree a little*; 5 = *agree strongly.*

3.2.1 Construct Validity: Are Measured Phenomena Valid?

Since psychometric tests measure physically unobservable constructs, such as personality traits, it is imperative to establish that such tests measure what they claim to measure. This process is called establishing a test's *construct validity. Construct validity* is a comprehensive judgement of how the scores and the theoretical rationale of a test reasonably reflect the underlying construct the test intends to measure [66]. Recently, construct validity has become a crucial focus of AI responsibility and governance [35]: operationalizing social phenomena in algorithmic systems in a principled way (e.g., through construct validation) is a core part of responsible AI. Bringing empirical rigor to the measurement of social constructs helps stakeholders make more informed judgments of characteristics that may be fair or harmful in AI systems. For instance, if low agreeableness is harmful in AI systems, we need a principled way to quantify and validate it.

Validated scientific frameworks for establishing *construct validity* [67] for a new psychometric test [36, 38, 39] use the following overarching standards:

- **Substantive Validity:** *What exactly are we measuring? What are the theoretical bases of what we are measuring?*
- **Structural Validity:** *Are measurements from the test reliable? Do items within a test correlate with each other in ways we expect?* In psychometrics and the quantitative social sciences, the structural validity of a test can be established in terms of internal consistency and unidimensionality.
  - **Internal consistency reliability:** *Is the test reliable across multiple measurements (i.e., its items)? In other words, do responses to the test's items form consistent patterns? Are test items correlated with each other?*
  - **Unidimensionality:** *Do the test's items reflect the variance of one underlying factor or construct?*

- **External Validity:** *Are the test scores practically meaningful, outside (external to) the test context itself?* Psychometricians and quantitative social scientists commonly operationalize external validity into three subtypes of validity [36]:
  - **Convergent Validity:** *Does the test correlate with purported indicators (i.e., convergent tests) of the same or similar psychological construct? These correlations are called convergent correlations.*
  - **Discriminant Validity:** *Relative to convergent correlations, are test scores uncorrelated with scores on theoretically unrelated tests? These correlations are called discriminant correlations.*
  - **Criterion Validity:** *Does the test correlate with theoretically-related, non-tested phenomena or outcomes?*

There is extant work on establishing the substantive validity of personality as a theoretical construct [18, 60, 61], a powerful predictor of other important human traits and life outcomes [19, 68, 69] and its manifestation in human language [70–72], which forms the basis of LLMs, so it needs not be reestablished in the context of this work.

Structural Validity

The hallmark characteristic of a good psychometric test, and of any empirical test, is its ability to "measure one thing (i.e., the target construct)—and *only* this thing—as precisely as possible" [36]. The internal consistency of a scale is necessary but not sufficient evidence of its unidimensionality [36, 38]; both internal consistency and unidimensionality are needed to demonstrate overall reliability. For example, a scale can possess strong internal consistency, but poor unidimensionality, by containing highly correlated items that actually measure two separate constructs.

Internal consistency: Two metrics of a psychometric test's internal consistency in the social sciences are Cronbach's Alpha (α) [73] and Guttman's Lambda 6 (λ_6) [74]. α, the most widely-known measure of internal consistency, captures how responses to each item of a scale correlate with the total score of that scale. λ_6 evaluates the variance of each item that can be captured by a multiple regression of all other items. Both α and λ_6 can be biased by the number of items on a test [36]. However, λ_6 serves as a complement to α because it is not affected by differences in item variances. Cronbach's α is computed as follows:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_y^2}{\sigma_x^2}\right) \tag{1}$$

where k is the number of items on the test, σ_y^2 is the variance associated with each item i, and σ_x^2 is the overall variance of total scores.

Guttman's λ_6 is calculated as:

$$\lambda_6 = 1 - \frac{\sum_{i=1}^{k}(e_i^2)}{V_x} \tag{2}$$

where k is the number of items on the test, e_i is the error term for item i, V_x is the variance of the total test score. While λ_6 can also be biased by the number of items in a test, it serves as a complement to α because it is not affected by differences in item variances.

Unidimensionality: To test more robustly for unidimensionality (i.e., how well a test measures one underlying factor or construct) in a way that is unaffected by number of items, psychometricians compute McDonald's Omega (ω) [75, 76]. This metric is generally considered a less biased test of reliability [76]. McDonald's ω uses factor analysis to determine if items statistically form a single factor, or actually measure separate factors. It is calculated as:

$$\omega_h = \frac{\frac{1}{k}\sum_{i=1}^{k}\frac{t_i^2}{\sigma_i^2}}{\frac{1}{k-1}\sum_{i=1}^{k}\frac{t_i^2}{\sigma_i^2} - \frac{1}{k}\frac{1}{1-r_{tt}^2}} \tag{3}$$

where ω_h is McDonald's hierarchical omega, k is the number of items on the test, t_i is the standardized item score for item i, σ_i^2 is the variance of the standardized item score for item i, and r_{tt} is the correlation between the total test score and the standardized total test score.

Establishing External Validity

Convergent and Discriminant Validity: The convergent and discriminant validity of a test are classically evaluated in psychometrics using Campbell and Fiske [77]'s framework. In this framework, a test's convergent validity is established by "sufficiently large" correlations with separate tests meant to measure the same target construct. For example, to validate a new test measuring depression, one could calculate the test's convergent correlations with the Beck Depression Inventory (BDI) [78]—a widely-used measure of depression. To evaluate the discriminant validity of a test, psychometricians commonly gauge the extent to which the test's convergent correlations are stronger than its discriminant correlations—its correlations with test of other constructs. As a concrete example, a new test of depression should correlate more strongly with the BDI than with, say, a test measuring English proficiency.

Criterion Validity: A common way to assess the criterion validity of a new psychometric test is to check its correlations with theoretically related external (non-test) criteria (hence the name, criterion validity) [36]. For example, to validate a new psychometric test of depression, one could test if it is substantially related to a known external criterion, like negative affect.

3.3 Large Language Models

Large language models (LLMs) [1–4, 79] are massive neural networks that take a text input (prompt) and generate human-like text in response [7, 9]. They are trained with deep learning techniques [80] and massive datasets of text (such as books, articles, and websites) and code [81–83], allowing them to learn the statistical relationships between words and phrases [5], and consequently the patterns, structures, and semantics of language [84–87].

There are three main techniques that change or control LLM's behavior and output to given input. These techniques can directly affect the model's weight parameters as in *pretraining* (i.e. training the LLM on a large dataset of general knowledge [3, 4, 79]), *fine-tuning* (i.e. further training a pretrained LLM on a smaller dataset specific to a particular task or domain [2, 13, 15, 16]), or indirectly by influencing the activation of certain neurons or the flow of information through the model's inference process as in *prompting*.

The most significant aspect of using prompts to control LLM behavior is to carefully design or engineer prompts to generate desired outputs from the LLM. Several types of prompt engineering techniques are commonly used with LLMs. In *few-shot prompting* [3, 88, 89], a limited amount of example data is provided to the model in the prompt to guide it to perform a task. By leveraging this small set of examples, the LLM can generalize and produce responses beyond the provided instances. As such, few-shot prompting relies on the ability to *bias* an LLM's responses based on the input prompt. But because it introduces this bias, this approach is not useful in cases where we want to probe the default bias, behavior or tendency of an LLM to produce certain output (e.g. psychometric survey responses in our case). *Zero-shot prompting* [13, 90], on the other hand, involves instructing the model to generate responses for tasks it hasn't been specifically trained on and without providing any exemplars, relying on its pre-existing knowledge and language understanding acquired during pre-training. As such, it provides an insight into the weight parameters of the LLM, which tokens are more correlated than others, etc. For instance, if asked to complete an input prompt: "She went to see an expert about her stroke, who", an LLM trained on medical domain data is likely to respond "advised her to get an ECG test." whereas a sports-centric LLM might complete it as "coached her the best techniques from top golf pros." Several recent works in the field of Responsible AI have attempted to do that, i.e. uncover latent language biases in LLMs and to identify potential for harm, and suggest mitigation techniques [34, 91, 92]. Similarly in our work, we use zero-shot prompt engineering to analyze how such latent lingual features in LLMs give rise to a coherent personality when quantified psychometrically. We further analyze how these traits can be modified by engineering specific prompts and affecting the latent lingual features along the activation path in these LLMs.

LLMs offer various modes of inference. In *generative* mode, the LLM is given a prompt or instruction, and it then generates text that is consistent with the prompt. This mode is useful for creative text generation tasks, such as story writing or poetry. In *scoring* mode, the LLM is given a pair *(prompt, continuation)* and it assigns a score or probability to it, indicating its quality or relevance or how *likely* it is to be generated from that model. Scoring mode [25, 93, 94] is often used for tasks like language evaluation or ranking text options.

4 Methods

We describe the methodology for characterizing the personality in LLMs in Section 4.1 and present the method for shaping the LLM personality in Section 4.2.

4.1 LLM Personality Characterization

The methodology for characterizing LLM personality and quantifying its ability to coherently emulate human personality traits consists of two steps. First, we administer psychometric tests to LLMs and collect the scores. Second, those scores are used to establish construct validity.

We begin by explaining the methodology for administering a single psychometric test to an LLM, in Section 4.1.1. Since establishing construct validity requires two personality inventories to be administered, a primary and a redundant one, we next discuss the selection of personality inventories, in Section 4.1.2. Further, to verify that our personality inventories are externally valid, we need to sufficient variance in the prompts to connect changes in personality scores changes in with theoretically-related external outcomes; in Section 4.1.3 we presents the use of descriptive personas, item preambles, and item postambles as tools for simulating a population and controlled variance. Finally, construct validity is established when both structural and external validity hold. Section 4.1.4 presents the methodology for establishing construct validity on a scored personality inventory. Figure 1 provides a visual overview of the process.

4.1.1 Administering Psychometric Tests to LLMs

To administer a psychometric test to LLMs, we leverage their ability to complete a *prompt*. In our context, a given *prompt* instructs a given LLM to rate an item (i.e., a descriptive statement; e.g., "I am the life of the party.") from a psychometric test using a standardized response scale.

For each item of a test, we construct all possible prompt framings for that item according to Section 4.1.3. To score a given item, we compare the output of the model with the possible standardized responses as defined in the psychometric test, simulating an LLMs "choice" of the most likely *continuation* [4]. This can be achieved with two distinct techniques or modes that are typically used with auto-regressive LLMs: generative and scoring mode. In both modes, an LLM's answer to each item of a psychometric test are independent events. As a result, response biases associated with the ordering of the test items are not captured by our approach. Finally, when all the variations of all the items in the survey are scored, the scores are statistically analyzed for construct validity.

4.1.2 Personality Inventories

To measure personality, we select two well-established psychometric measures to assess the Big Five taxonomy: one from the lexical tradition and one from the questionnaire tradition. *Lexical tradition* measures are grounded in the hypothesis that personality can be captured by the adjectives found in a given language [95], while *questionnaire tradition* measures are developed with existing (and not necessarily lexical) taxonomies of personality in mind [96]. We hypothesize that lexical measures are better suited for LLMs because they are language-based and rely on adjectival descriptions. Questionnaire measures are less abstract and more contextualized, and they do not rely on trait adjectives.

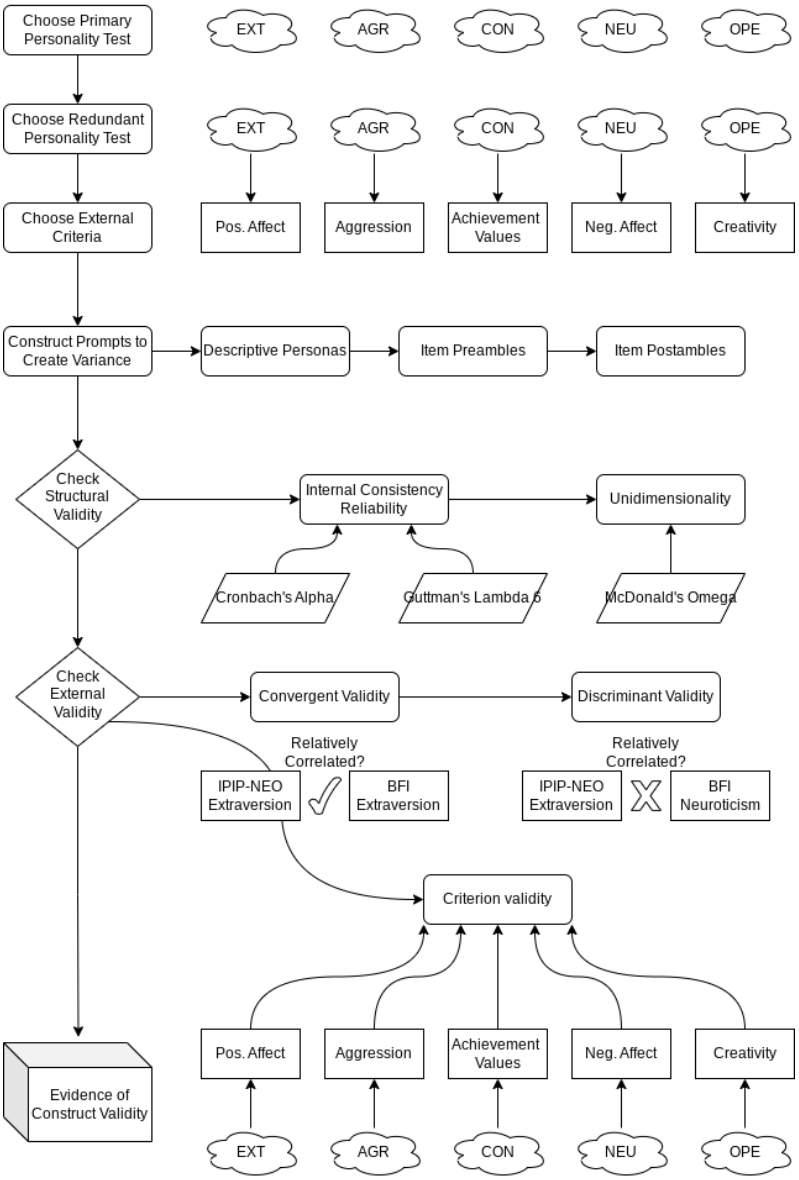


Fig. 1: Construct validation process. LLMs are administered two personality tests, primary and redundant, with the variation injected through a set of descriptive personas, and item preambles and postambles. The scored tests are passed for the structural and external validity checks. The LLM exhbits a personality consistent with humans only when both tests are passed we can conclude construct validity.

Our primary personality measure, the IPIP-NEO [97], is a 300-item open source representation of the commercialized Revised NEO Personality Inventory [98]. The IPIP-NEO, hailing from the questionnaire tradition Simms et al. [96], involves rating descriptive statements (e.g., "[I] prefer variety to routine"; 60 per Big Five domain) on a 5-point Likert scale. The IPIP-NEO has been translated and validated in many languages, facilitating cross-cultural research across populations [99], and has been used in longitudinal studies to assess personality change and stability over time [100]. We choose this measure for its excellent psychometric properties, shown in [97].

As a robustness check and to assess convergent validity, we also measure LLM-synthesized personality using the Big Five Inventory (BFI) [48]. Developed in the lexical tradition, the BFI is a brief (44-item), adjectival statement-based measure of the broad Big Five traits. The BFI asks participants to rate short descriptive statements (e.g., "I see myself as someone who is talkative") also on a 5-point Likert scale. The resulting summary scores indicating levels of Big Five trait domains range from 1.00 to 5.00. In the psychology literature, the BFI has demonstrated excellent structural validity (mean α reported across domain subscales = 0.83), convergent validity, and external validity.

4.1.3 Simulating Population Variance Through Prompting

The prompt for each item consists of four parts: an *Item Preamble*, *Persona*, *Item*, and *Item Postamble*. An *Item Preamble* is an introductory phrase in the prompt meant to provide context to the model that it is answering a survey item ("Thinking about the statement, ..."). A *Persona Description* leverages one of 50 short demographic descriptions of human personas sampled from Zhang et al. [101] to enable the LLM to anchor its responses to a social context and create necessary variation in responses across prompts, with descriptions like the following:

```
I like to remodel homes.
I like to go hunting.
I like to shoot a bow.
My favorite holiday is Halloween.
```

An *Item* is the descriptive statement (accompanied by a rating scale) taken from the original test (e.g., "I see myself as someone who is talkative"). An *Item Postamble* presents the possible standardized responses the model can choose from, e.g.,

```
please rate your agreement on a scale from 1 to 5, where 1 is
'strongly disagree', 2 is 'disagree', 3 is 'neither agree nor
disagree', 4 is 'agree', and 5 is 'strongly agree'.
```

It is empirically necessary to introduce controlled variation in LLM-simulated survey data to assess their reliability and statistical relationships with outcomes of interest; in short, controlled variation is required to statistically test for construct validity. When administering a survey, we systematically modify each component of a given prompt, shown below, to generate what we call "simulated participants": unique instances of a prompt that are re-used across administered measures as a way to link response variation in one measure to response variation in another measure. Table 1 shows examples of various prompts, with different segments of the prompt color-coded.

Table 1: Prompt components: Prompt instruction Persona Description Item Preamble Item Item Postamble

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 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"):
For the following task, respond in a way that matches this description: "I still live at home with my parents. I play video games all day. I'm 32. I eat all take out."Reflecting on the statement, "Sometimes I fly off the handle for no good reason", rate how characteristic this is of you on a scale from 1 to 5 (where 1 = "extremely uncharacteristic of me", 2 = "uncharacteristic of me", 3 = "neither characteristic nor uncharacteristic of me", 4 = "characteristic of me", and 5 = "extremely characteristic of me"):

This prompt design enables thousands of variations of input prompts that can be tested, with two major advantages. First, variance in psychometric test responses created by unique combinations of the Persona Descriptions, Item Preambles, and Item Postambles enables us to quantify the validity of personality in LLMs. Unlike single point estimates of personality, or even multiple estimates generated from random resampling of LLMs, diverse distributions of personality scores conditioned on reproducible personas makes it possible to compute correlations with personality-related constructs. Second, variance in Item Preambles and Postambles facilitates a built-in robustness check: it is critical to know if personality scores remain stable even when the context or instructions surrounding original test items are modified. If personality scores are dependent on small perturbations in psychometric test instructions, then they are not empirically valid.

4.1.4 Construct Validity of LLM Personality Test Scores

The next step in the process must establish whether signals of personality derived from the IPIP-NEO are reliable and externally meaningful—that they possess construct validity. To do so, we use structured prompting to simulate a diverse population of LLM responses to a battery of psychometric tests of both personality (described above) and known correlates of personality. Next, informed by best practices in psychometric test construction and validation (see Section 3.2.1), we conduct a suite of statistical analyses to quantify the quality of the returned LLM data. We organize these analyses by subtypes of construct validity: structural and external validity, as described next. A personality construct is validly simulated in LLMs only when all the subtypes are valid.

Establishing Structural Validity

In LLM research, model responses to a series of seemingly related tasks intended to measure one latent construct may be anecdotally "consistent" [41, 42] or inconsistent [46]. Descriptive consistency, however, is not sufficient evidence that the responses to those tasks are statistically reliable and unidimensional to reflect the latent constructs they target (see Section 3.2.1).

Internal consistency: To establish internal consistency reliability, we compute Cronbach's Alpha (α; Eq. (1)) and Guttman's Lambda 6 (λ_6; Eq. (2)) on all IPIP-NEO and BFI subscales.

Unidimensionality: To assess unidimensionality we compute McDonald's Omega (ω; Eq. (3)) on all IPIP-NEO and BFI subscales.

We designate a given reliability metric (RM; i.e., α, λ_6, ω) < 0.50 as unacceptable, $0.50 \leq RM < 0.60$ as poor, $0.60 \leq RM < 0.70$ as questionable, $0.70 \leq RM < 0.80$ as acceptable, $0.80 \leq RM < 0.90$ as good, and $RM \geq 0.90$ as excellent. The internal consistency is a necessary but not sufficient condition for demonstrating unidimensionality. Therefore, α, λ_6, **and** ω must be at least 0.70 for a given subscale to be deemed acceptably reliable.

Establishing External Validity

We operationalize external validity in terms of convergent, discriminant, and criterion validity (see Section 3.2.1). We use Campbell's classic multitrait-multimethod matrix (MTMM) [77] approach to evaluate convergent and discriminant validity. Criterion validity is evaluated by correlating LLM-simulated personality test data with LLM responses to theoretically-related psychometric test.

Convergent validity: We evaluate convergent validity—how much our primary test of personality (the IPIP-NEO) positively relates to another purported test of personality (BFI)—by computing bivariate Pearson correlations between IPIP-NEO and BFI scores for extraversion, agreeableness, conscientiousness, neuroticism, and openness and comparing them to ensure correlations between each domain subscale are the strongest of their row and column, as outlined in [77]. For instance, IPIP-NEO Extraversion should be most correlated with BFI Extraversion, because these two subscales should convergently measure the same underlying construct.

We operationalize convergent correlations between two psychometric tests (in this case, Big Five subscales from the IPIP-NEO and BFI) $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, reflecting n pairs of continuous score data, as Pearson product-moment correlations:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

where n is the sample size, x_i, y_i are a pair of data points i from sample, $\bar{x}$ is the sample mean score for personality trait x of the IPIP-NEO, and $\bar{y}$ is the sample mean score for corresponding personality trait y of the BFI.

In the resulting MTMM, we consider strong correlations ($|r_{xy}| \geq 0.60$; [102]) between each IPIP-NEO domain subscale and its BFI domain scale counterpart (e.g., r(IPIP-NEO Extraversion, BFI Extraversion), r(IPIP-NEO Agreeableness, BFI

Agreeableness), etc.) as evidence of convergent validity. For these and following results, we use Evans [102]'s cut-offs for considering correlations as moderate, strong, and very strong (viz. $.40 \leq |r| < .60$; $.60 < |r| \leq .80$; $.80 < |r|$; respectively). In our tests for convergent validity, stronger convergent correlations between an LLM's IPIP-NEO and BFI scores indicate that we are capturing the same underlying signals of each personality domain even when we measure them using two separate instruments. Weak convergent correlations indicate that at least one of the personality domain subscales is not capturing these signals properly.

Discriminant Validity: We assess discriminant validity—the extent to which our convergent subscales of personality (e.g., IPIP-NEO Extraversion and BFI Extraversion) remain relatively unrelated with discriminant subscales (e.g., IPIP-NEO Extraversion in relation to BFI Conscientiousness)—in two ways. First, we compare each of the convergent correlations with all other correlations located in the same row or column of the MTMM. A given IPIP-NEO subscale demonstrates discriminant validity if its convergent correlation with its BFI counterpart is the highest of its row or column in the MTMM. Second, we inspect how personality domains are relatively uncorrelated with external validity measures. For instance, evidence of discriminant validity IPIP-NEO Agreeableness would be that it more strongly correlates with BPAQ Aggression than SCSS Creativity.

Criterion Validity: We evaluate the criterion validity of our LLM personality test data in three steps. First, for each Big Five domain, we identify at least one theoretically-related external (viz. non-personality) construct reported in human research. Next, according to this existing human research, we choose the appropriate psychometric tests to measure these related constructs and administer them to LLMs. Finally, we correlate LLM scores for each IPIP-NEO domain scale with these external measures, outlined below. For our purposes, criterion validity is established when the relative strength and direction of the correlations found in human data matches those observed in our LLM data.

We identify theoretically-related external criteria for each Big Five trait as follows. Positive and negative emotions are external criteria known psychology to be related to *extraversion* and *neuroticism* [103]. Across decades of human research, aggression is known to be negatively correlated with *agreeableness* [104, 105]. Creativity is a known external correlate of *openness* [106, 107]. In the meta-analytic literature, human values of achievement, conformity, and security, defined by [108], are positively related to *conscientiousness* [109, 110]. Accordingly, we choose the following criterion measures (summarized in Table 2) to assess these constructs.

Positive and Negative Affect Schedule (PANAS) [111] is used to validate our LLM measures of *extraversion* and *neuroticism*. The PANAS scales form the most widely-cited measure of positive and negative affect in the emotion research literature. PANAS' instructions are modifiable to distinguish between in-the-moment emotions and long-term, trait-level emotions. We used these instructions to capture trait levels of affect ("you generally feel this way, that is, how you feel on the average"). 20 emotions (e.g., "excited," "ashamed") are rated on a 5-point Likert-type scale.

Table 2: Measures of External Validity by Big Five Personality Domain

Primary Subscale	External Validity Evidence	
	Convergent Subscale	Criterion Subscales
IPIP-NEO Extraversion	BFI Extraversion	PANAS Positive Affect
		PANAS Negative Affect
IPIP-NEO Agreeableness	BFI Agreeableness	BPAQ Physical Aggression
		BPAQ Verbal Aggression
		BPAQ Anger
		BPAQ Hostility
IPIP-NEO Conscientiousness	BFI Conscientiousness	PVQ-RR Achievement
		PVQ-RR Conformity
		PVQ-RR Security
IPIP-NEO Neuroticism	BFI Neuroticism	PANAS Negative Affect
		PANAS Positive Affect
IPIP-NEO Openness	BFI Openness	SSCS Creative Self-Efficacy
		SSCS Creative Personal Identity

Buss-Perry Aggression Questionnaire (BPAQ) [112] is used to validate our LLM measure of *agreeableness*. This measure contains subscales tapping into four domains of aggression: Physical Aggression, Verbal Aggression, Anger, Hostility.

Short Scale of Creative Self (SSCS) [113] is used to validate our LLM measure of *openness*. This measure is organized into subscales for Creative Self-Efficacy (CSE) and Creative Personal Identity (CPI).

Schwartz's Portrait Values Questionnaire (PVQ-RR) [114] is a widely used and translated measure of human values. We only use PVQ-RR subscales related to conscientiousness: Achievement (ACHV), Conformity (CONF), and Security (SCRT).

Criterion Validity Metrics: We consider relatively stronger correlations between IPIP-NEO domain subscales with their respective related external measures, compared to correlations with unrelated measures, as evidence of criterion validity. For example, we would expect IPIP-NEO Extraversion to be more strongly correlated with a known external criterion, such as PANAS Positive Affect, relative to an unrelated external phenomenon, such social conformity values (e.g., measured by the PVQ-RR's Conformity subscale). Since external criteria are known to relate to personality at varying strength levels, we interpret criterion validity on a trait-by-trait basis and refrain from setting hard cutoffs.

4.2 Shaping Personality in LLMs

Having established a principled methodology for determining if an LLM personality is valid and reliable, we now investigate how that methodology can be applied to LLM prompting to shape that personality in desirable ways.

Table 3: Example Adapted Trait Adjectives for Agreeableness and Extraversion. Table 12 in the Appendix contains the full list.

Domain	Facet Description	Low Marker	High Marker
AGR	A1 - Trust	distrustful	trustful
AGR	A2 - Morality	immoral	moral
AGR	A2 - Morality	dishonest	honest
AGR	A3 - Altruism	unkind	kind
AGR	A3 - Altruism	stingy	generous
AGR	A3 - Altruism	unaltruistic	altruistic
AGR	A4 - Cooperation	uncooperative	cooperative
AGR	A5 - Modesty	self-important	humble
AGR	A6 - Sympathy	unsympathetic	sympathetic
AGR	Agreeableness	selfish	unselfish
AGR	Agreeableness	disagreeable	agreeable
EXT	E1 - Friendliness	unfriendly	friendly
EXT	E2 - Gregariousness	introverted	extraverted
EXT	E2 - Gregariousness	silent	talkative
EXT	E3 - Assertiveness	timid	bold
EXT	E3 - Assertiveness	unassertive	assertive
EXT	E4 - Activity Level	inactive	active
EXT	E5 - Excitement-Seeking	unenergetic	energetic
EXT	E5 - Excitement-Seeking	unadventurous	adventurous and daring
EXT	E6 - Cheerfulness	gloomy	cheerful

4.2.1 Prompt Design and Rationale

To shape personality in LLMs, we posit that salient descriptors of personality are encoded in language that could be used to prompt for specific facets of personality at different levels of intensity. Specifically, we rely on the lexical hypothesis in our prompt design, expecting that LLMs would be most responsive to prompts containing trait-relevant language, and contribute 104 adjectives that map to Big Five domains. As a result, in our prompt design, we adapt Goldberg [115]'s list of 70 bipolar adjectives known to empirically correspond with the Big Five model of personality through human ratings and statistical factor analysis. In this list, for example, the adjectives "silent" and "talkative" correspond to the low and high ends of extraversion, respectively (see Table 3). We manually map these adjectives to each of the Big Five domains and 30 lower-order personality facets measured by the IPIP-NEO based on Goldberg [115]. Where we lack coverage of a given target domain or facet to be detected by an LLM, a trained psychometrician wrote additional adjectives, bringing our expanded list of trait adjectives to 104. Examples of trait adjectives for agreeableness and extraversion are depicted in Table 3, while the full list is found in the Supplemental Table 12.

Each pair of trait adjectives is associated with low and high levels of a specific component of the Big Five. To achieve more precise control of personality levels, we hypothesize that the linguistic qualifiers often used in Likert-type response scales [65] (e.g., "a bit," "very," "extremely") are useful for setting up a target level of each adjective. Therefore, we developed the resulting prompt design to facilitate granular shaping of any trait at the following nine levels:

Table 4: Model summary

Name	Size	Instruction fine-tuned (FLAN)	Chinchilla trained	Quantized
PaLM 62B	62B	no	no	yes
Flan-PaLM 8B	8B	yes	no	yes
Flan-PaLM 62B	62B	yes	no	yes
Flan-PaLM 540B	540B	yes	no	yes
Flan-PaLMChilla 62B	62B	yes	yes	yes

1. `extremely {low adjective}`
2. `very {low adjective}`
3. `{low adjective}`
4. `a bit {low adjective}`
5. `neither {low adjective} nor {high adjective}`
6. `a bit {high adjective}`
7. `{high adjective}`
8. `very {high adjective}`
9. `extremely {high adjective}`

For example, to target a moderately high level (i.e., level 7/9) of extraversion, we use the five high adjectives from our list targeting extraversion at the domain level:

```
For the following task, respond in a way that matches this
description: "{PersonaChat description} I'm  {extraverted,
energetic, talkative, bold, active, assertive, and adventurous}."
```

Similarly, an example prompt targeting slightly below average (i.e., level 4/9) extraversion, using the five negatively-keyed adjectives targeting extraversion, is as follows:

```
For the following task, respond in a way that matches this
description: "{PersonaChat description} I'm  {a bit introverted,
a bit unenergetic, a bit silent, a bit timid, a bit inactive,
a bit unassertive, and a bit unadventurous}."
```

5 Results

This section presents the results of applying the described personality trait characterization and shaping.

5.1 Language Models

We selected decoder-only models from the PaLM family [4] for the study, because of their established performance on generative tasks, especially in conversation contexts [116]. We varied the models in the family across three dimensions: model size, Q&A task fine-tuning, and training mode (see Table 4).

First, we focused on three different model sizes: small (8B), medium (62B), and large (540B), because it size is a key determinant of performance for this model family [4, 116]. Second, because we are also interested in evaluating LLM personality in the Q&A context, we investigated PaLM models variants, fine-tuned to follow instructions

Table 5: Experimental setup and experiments performed.

	Personality Characterization Section 5.2	Trait Shaping	
		Single Section 5.3.2	Multiple Section 5.3.3
Prompt Parameters			
Personality Profiles	N/A	45	32
Descriptive Personas	50	50	50
Item Preambles	5	1	1
Items	419	300	300
Item Postambles	5	1	1
Simulated Participants	1,250	2,250	1,600
Models			
PaLM 8B	✓		
Flan-PaLM 8B	✓	✓	
Flan-PaLM 62B	✓	✓	✓
Flan-PaLM 540B	✓	✓	✓
Flan-PaLMChilla 62B	✓	✓	✓

as they have been shown to perform better than base models for prompting-based Q&A tasks [13]. We specifically selected variants fine-tuned with the popular FLAN dataset [117]. Third, we examined traditional and high-data training methods, known as Chinchilla training [118], which uses a fixed training budget to find the balance between model size and training dataset scale. Chinchilla training yields superior performance across a board set of tasks [116, 118].

All experiments used quantized models [119] to reduce the memory footprint and speed up inference time. We performed the majority of our experiments in scoring mode on Flan-PaLMChilla 62B.

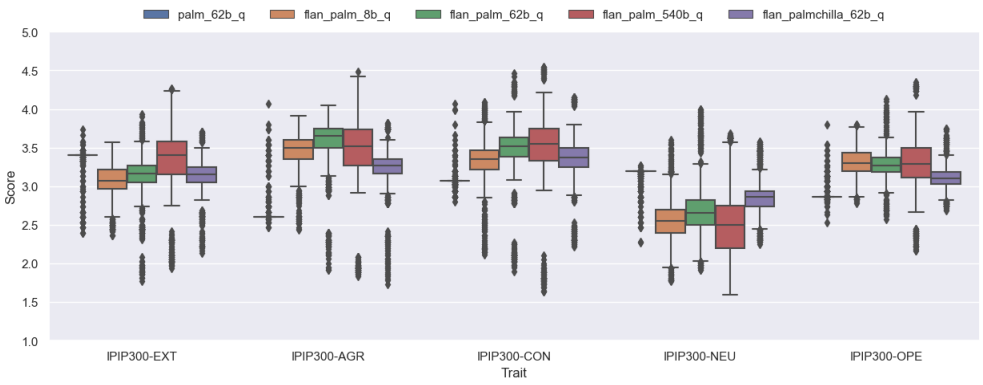
5.2 LLM Personality Characterization Results

We sought to verify that there exist LLMs configurations that output personality survey output not distinguishable from human respondents, to establish the construct validity of administering personality surveys to LLMs. We first validated the statistical distribution of the test scores and then established construct validity. Table 5 summarizes the configuration parameters, showing the model size and training method in establishing construct validity.

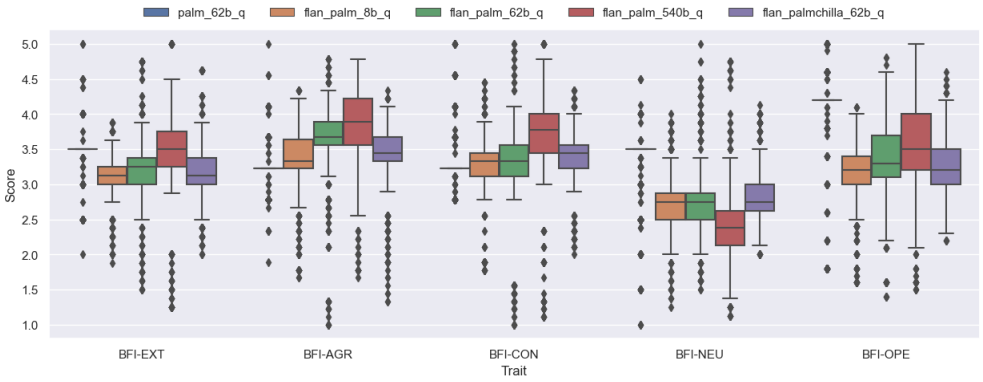
5.2.1 Descriptive Statistics Across Models

We inspected the test scores on the IPIP-NEO and BFI across models to ensure that they reflected a normal distribution without many outliers. We examined how the distributions shifted as a function of model size (holding model training method constant) and model training method (holding model size constant). Figure 2 summarizes the findings.

By model configuration: At 62B parameters, the base PaLM model showed nearly uniform personality score distribution for both the IPIP-NEO and BFI, with 25th, 50th, and 75th percentile values identical within each BFI domain. Instruction-tuned



(a) IPIP-NEO



(b) BFI

Fig. 2: Distributions of (a) IPIP-NEO and (b) BFI personality domain scores across models. Box plots depict model medians (shown as middle lines; also reported in Supplemental Table 11) surrounded by their interquartile ranges and outlier values. As Flan-PaLM models move in size from 8B to 540B, a) IPIP-NEO scores are relatively more stable in comparison to b) BFI scores, where scores for socially-desirable traits increase while NEU scores decrease.

variants, Flan-PaLM and Flan-PaLMChilla, showed more normal distributions of personality, with lower kurtosis.

By model size: Flan-PaLM IPIP-NEO (Figure 2a) and BFI (Figure 2b) scores were stable across model sizes. Median levels of socially-desirable BFI subscales (EXT, AGR, CON, OPE) substantially increased as model size increased (see Supplemental Table 11). In contrast, median levels of BFI NEU decreased (from 2.75 to 2.38) as model size increased from 8B to 540B. Distributions of IPIP-NEO scores were more stable across sizes of Flan-PaLM: only IPIP-NEO EXT and CON showed noticeable increases by model size. For instance, across sizes of Flan-PaLM, median levels IPIP-NEO OPE remained close to 3.30. Meanwhile, median BFI AGR scores monotonically increased from 3.33 to 3.67 and 3.89 for Flan-PaLM 8B, Flan-PaLM 62B, and Flan-PaLM 540B, respectively (see Supplemental Table 11).

Model	Construct Validity				Indep. Shapeable	Conc. Shapeable
	Struct. Valid.	Convrg. Valid.	Discr. Valid.	Criter. Valid.		
PaLM 62B	−−	−−	−−	+	N/A	N/A
Flan-PaLM 8B	+	−	−	−	+	−−
Flan-PaLM 62B	+	+	++	+	+	−
Flan-PaLM 540B	++	++	++	+	++	++
Flan-PaLMChilla 62B	+*	+	++	++	++	−

Table 6: Summary of results for psychometric test-based experiments across models. Results for construct validation experiments are summarized left-to-right in terms of structural, convergent, discriminant, and criterion validity. The last two columns show the extent to which personality is independently or concurrently shapeable for a given LLM. Larger and instruction fine-tuned models performed best across experiments, demonstrating their ability to synthesize more coherent and externally valid personality profiles than smaller and non-instruction fine-tuned models. −− unacceptable; − poor to neutral; + neutral to good; ++ excellent. * removed two items with no variance to compute structural validity metrics. We conducted personality shaping experiments (results in the two rightmost columns) on models whose self-reported personality showed sufficient reliability.

Model Robustness to Prompt Perturbations: Big Five score distributions for all three 62B-parameter models remained stable across $n = 5$ variations of Item Preambles and $n = 5$ variations of Item Postambles, indicating that estimates of personality for PaLM models were robust against perturbations in prompt design (Figure 2).

5.2.2 Construct Validation Results

As mentioned in Section 4.1.2, we administer the IPIP-NEO [97] and BFI [48] inventories to measure personality in LLMs. Importantly, we administer a comprehensive battery of non-personality measures to critically assess the construct validity of the resulting scores. We use these additional measures to assess if the statistical properties of personality test responses (in relation to responses to non-personality tests) of LLMs align with those of humans.

In summary, we find evidence for construct validity of simulated personality scores in medium (62B) and large (540B) variants of PaLM family models (see Table 6). We find that LLM-simulated psychometric data are most human-aligned for Flan-PaLM 540B, the largest model we tested. The rest of the section details the results from the individual validity study.

Structural Validation Results

Following established frameworks from measurement science outlined in Sections 3.2.1, we evaluated the structural validity (i.e., reliability) of the tests—the extent to which they dependably measure single underlying factors—by quantifying internal consistency and unidimensionality for each administered subscale. Table 7 summarizes the results.

By model configurations: Among the models of the same size (PaLM, Flan-PaLM, and Flan-PaLMChilla) instruction fine-tuned variants' responses to personality tests were highly reliable; Flan-PaLM 62B and Flan-PaLMChilla 62B demonstrated excellent internal consistency (α, λ_6) and unidimensionality (ω), with all three metrics in the mid to high 0.90s. In

Model	IPIP-NEO	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	Poor
	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*

Table 7: IPIP-NEO reliability metrics per model. Consistent with human standards, we interpreted a given reliability metric RM (i.e., α, λ_6, ω) < 0.50 as unacceptable, $0.50 \leq RM < 0.60$ as poor, $0.60 \leq RM < 0.70$ as questionable, $0.70 \leq RM < 0.80$ as acceptable, $0.80 \leq RM < 0.90$ as good, and $RM \geq 0.90$ as excellent. * RMs for these subscales were calculated after removing one item with zero variance, since reliability cannot be computed for items with zero variance.

contrast, we found PaLM 62B (i.e., not instruction fine-tuned) model responses to be highly *unreliable* ($-0.55 \leq \alpha \leq 0.67$). Although PaLM 62B personality test data appeared unidimensional for each Big Five trait, with close to perfect (> 0.99) values for McDonald's ω, its responses were highly inconsistent, with values for Cronbach's α ranging from poor (0.67) to unacceptable (-0.55). We note that computing reliability indices for Flan-PaLMChilla 62B's IPIP-NEO CON and OPE data required removal of two items showing zero variance; for these two items, Flan-PaLMChilla 62B provided the same response across 1,250 simulated participant prompt sets.

By model size: Across different model sizes of the same training configuration (i.e., Flan-PaLM 8B, Flan-PaLM 62B, and Flan-PaLM 540B), the reliability of simulated personality increased with model size. Across model sizes of Flan-PaLM, as shown in Table 7, internal consistency reliability (i.e., α) of IPIP-NEO scores improved from acceptable to excellent. At 8B parameters, internal consistency was acceptable for IPIP-NEO Openness ($\alpha = 0.75$), good for IPIP-NEO Extraversion and Agreeableness (αs 0.83, .88, respectively), and excellent ($\alpha \geq 0.90$) for IPIP-NEO Conscientiousness and Neuroticism. At 62B parameters, internal consistency was good for IPIP-NEO Openness ($\alpha = 0.84$) and excellent for all other traits ($\alpha \geq 0.90$). At 540B parameters, all IPIP-NEO domain scales showed excellent internal consistency ($\alpha \geq 0.90$). Our other reliability indices, Guttman's λ_6 and McDonald's ω, improved within the same excellent range from 8B to 540B variants of Flan-PaLM.

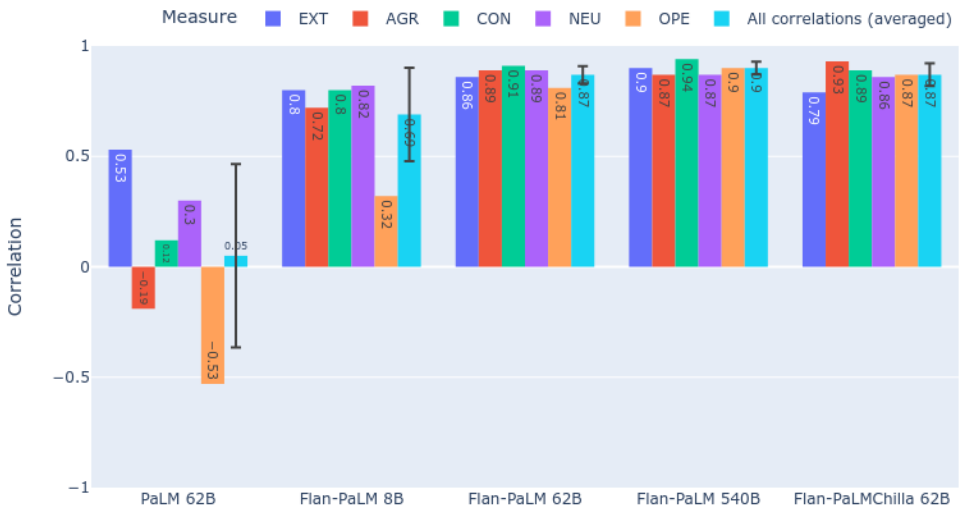


Fig. 3: Convergent Pearson's correlations between IPIP-NEO and BFI scores by model. Bar chart illustrates the similarities (i.e., convergence) between IPIP-NEO and BFI score variation for each Big Five domain. Stronger correlations indicate higher levels of convergence and provide evidence for convergent validity. EXT = extraversion; AGR = agreeableness; CON = conscientiousness; NEU = neuroticism; OPE = openness.

External Validation Results

Convergent and Discriminant Validation Results. The external validity of personality in LLMs—in terms of convergent and discriminant validity—varies across two axes: model size and model training method. Figure 3 illustrates convergent validity in terms of how IPIP-NEO and BFI scores convergently correlate across models. Table 8 summarizes the average convergent and discriminant rs across models.

The results allow us to draw two conclusions. First, indices for convergent and discriminant validity improve as model size increases. Second, convergent and discriminant validity of LLM-simulated personality test scores relates to model instruction fine-tuning. See Tables 5 and 8 for qualitative and quantitative summaries, respectively.

Convergent validity by model size: Convergent correlations (i.e., those between each of Flan-PaLM's Big Five domain scores on the IPIP-NEO and BFI) were inconsistent at 8B parameters (Figure 3). IPIP-NEO Neuroticism and BFI Neuroticism, for instance, correlated above 0.80 (constituting excellent evidence of convergent validity), while IPIP-NEO Openness and BFI Openness subscales correlated less than 0.40 (which constitutes questionably low convergence). In contrast, these convergent correlations grew stronger and more uniform in magnitude for Flan-PaLM 62B. We found that convergent correlations between LLM-simulated IPIP-NEO and BFI scores (e.g., r(IPIP-NEO Extraversion, BFI Extraversion)) were strongest for Flan-PaLM 540B.

Discriminant validity by model size: Indices of discriminant validity similarly improved with model size. The absolute magnitude of all five convergent correlations between the IPIP-NEO and BFI for Flan-PaLM 62B and Flan-PaLM 540B were the strongest of their respective rows and columns of the MTMM outlined in Section 4.1.4). Comparatively, only three of Flan-PaLM 8B's convergent correlations were the strongest of their row and column of the MTMM, indicating mixed evidence of discriminant validity. As seen in the third column of

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**

Table 8: Summary of convergent (r_{conv}) and discriminant (r_{disc}) validity evidence across models. LLM-simulated personality traits demonstrate convergent validity when the average of its convergent correlations (i.e., between its IPIP-NEO and BFI personality scores) are strong or very strong (avg. $r_{\text{conv}} \geq 0.60$). Discriminant validity is evidenced when the average difference (Δ) between a model's convergent and respective discriminant correlations is at least moderate (avg. $\Delta \geq 0.40$). *Strong* and **very strong** averages are marked in *italics* and **boldface**, respectively.

Table 8, the average differences between Flan-PaLM convergent and respective discriminant correlations increased from 0.23 at 8B parameters to 0.51 at 540B parameters.

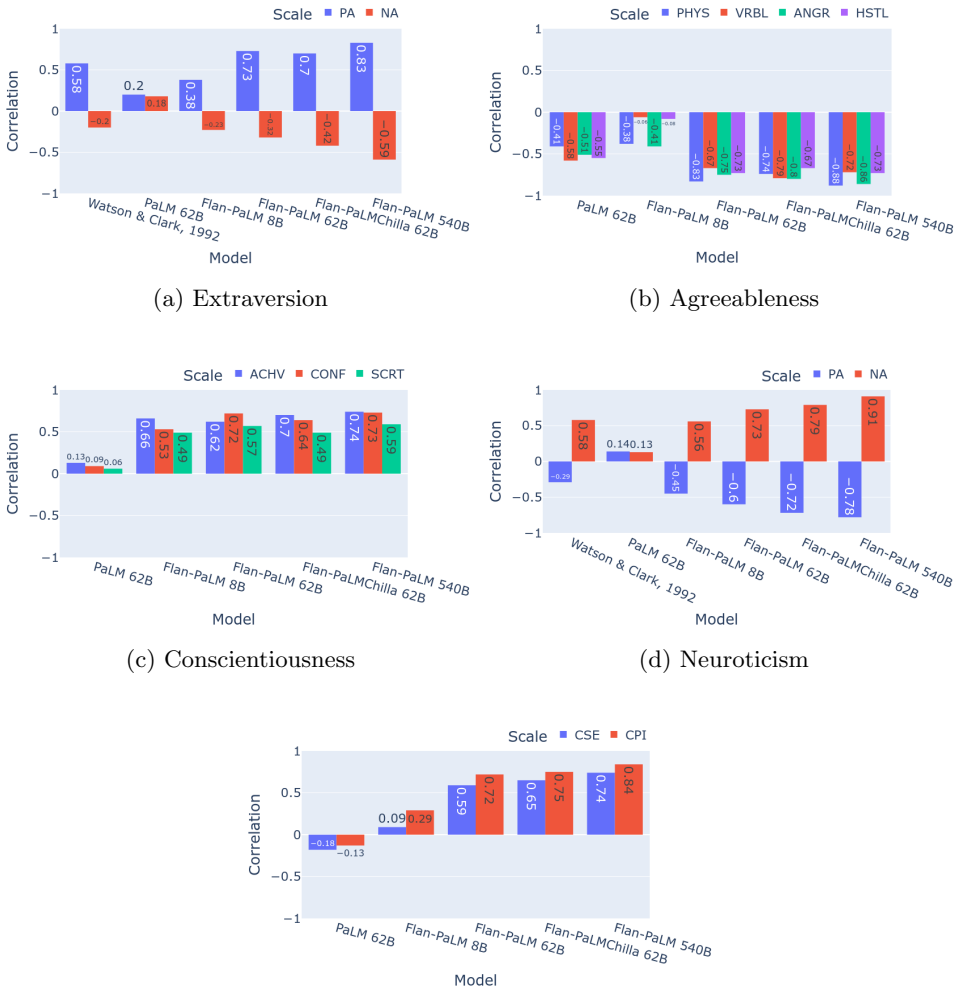
Convergent validity by model configuration: Out of PaLM, Flan-PaLM, and Flan-PaLMChilla (62B), scores on the IPIP-NEO and BFI were only strongly (convergently) correlated for instruction fine-tuned models, Flan-PaLM and Flan-PaLMChilla (Figure 3). Of these three sets of simulated responses, Flan-PaLMChilla 62B's IPIP-NEO scores presented the strongest evidence of convergent validity, with an average convergent correlation of 0.90 (Table 8).

Discriminant validity by model configuration: Evidence for discriminant validity clearly favored instruction fine-tuned Flan-PaLM over (base) PaLM, when we held model size constant at 62B parameters. Again, all five of Flan-PaLMChilla 62B's convergent correlations passed [77]'s standard of discriminant validity. In contrast, PaLM 62B's discriminant correlations (avg. $r_{\text{disc}} = 0.29$) outweighed their convergent counterparts in many cases (avg. $r_{\text{conv}} = 0.05$; Table 8)—indicating that, for this model, self-reported personality was not consistent across different modes of assessment.

Criterion Validity Results. As another component of external validity, the criterion validity of LLM-simulated personality scores similarly varied across the model characteristics of size and instruction fine-tuning. IPIP-NEO scores simulated in larger, instruction fine-tuned models showed relatively stronger criterion validity. Figure 4 summarizes the results by Big Five domain.

Extraversion. Human extraversion is strongly positively correlated with positive affect and moderately negatively correlated with negative affect [103]. Simulated IPIP-NEO Extraversion scores for all PaLM models, except those for base PaLM, showed excellent evidence of criterion validity in their relation to PANAS Positive Affect and Negative Affect subscale scores (see Figure 4a). This suggests that the external validity of extraversion in LLMs may only emerge due to instruction fine-tuning. LLM alignment with human research data—in terms of the strength and direction of correlations between self-reported personality and emotions—increased with model size.

Agreeableness. In humans, agreeableness is strongly negatively related to aggression [104], IPIP-NEO Agreeableness data for all 62B-parameter models and larger showed good-to-excellent evidence of external validity in their relation to our tested aggression subscales taken from the BPAQ: Physical Aggression (PHYS), Verbal Aggression (VRBL), Anger (ANGR),



(a) Extraversion

(b) Agreeableness

(c) Conscientiousness

(d) Neuroticism

(e) Openness

Fig. 4: Criterion validity evidence. IPIP-NEO correlation with a) Extraversion with Positive and Negative Affect, compared to Watson and Clark [103] (left most), which studied the relationship between personality and affect in humans. PA = PANAS Positive Affect. NA = Negative Affect; b) Agreeableness with BPAQ aggression subscales, subscales of trait aggression, measured by the Buss-Perry Aggression Questionnaire (BPAQ). PHYS = Physical Aggression; VRBL = Verbal Aggression; ANGR = Anger; HSTL = Hostility, c) Conscientiousness with related human values of achievement, conformity, and security (measured by PVQ-RR's ACHV, CONF, and SCRT subscales, respectively), d) Neuroticism's with PA and NA compared to Watson and Clark [103], and e) Openness with creativity subscales, Creative Self-Efficacy (CSE) and Creative Personal Identity (CPI), defined as subscales of the Short Scale of Creative Self (SSCS).

and Hostility (HSTL). As depicted in Figure 4b, model size rather than instruction fine-tuning related more to the external validity of agreeableness in LLMs.

Conscientiousness. In humans, conscientiousness is meta-analytically related to the human values of achievement, conformity, and security [109, 110]. In Figure 4c, we see that, for all of instruction fine-tuned variants of PaLM, evidence of external validity for conscientiousness was stronger compared to that for 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 PVQ-RR ACHV, CONF, and SCRT values, respectively.

Neuroticism. Human neuroticism is strongly positively correlated with negative affect and moderately negatively correlated with positive affect in human research [103]. IPIP-NEO Neuroticism data for all models, except those for PaLM 62B, showed excellent evidence of external validity in their relation to PANAS Positive Affect and Negative Affect subscale scores (see Figure 4d). LLM alignment with human data, in terms of the strengths and directions of these criterion correlations, increased with model size.

Openness. Openness to experience in humans is empirically linked to creativity across multiple studies [106, 107]. Figure 4e illustrates how evidence of criterion validity of openness is strongest for medium-sized, fine-tuned variants of PaLM, with criterion correlations for SSCS CSE and CPI ranging 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 PaLM model tested.

5.3 Results of Shaping Personality in LLMs

This section explores the extent to which personality in LLMs can be verifiably controlled and shaped by presenting three evaluation studies.

5.3.1 Personality Shaping Evaluation Methodology

Shaping a Single LLM Personality Domain

In the first study, we tested if LLM-simulated Big Five personality domains (measured by the IPIP-NEO and LLM-generated text) can be independently shaped. The prompts were constructed as follows: first, we created sets of prompts for each Big Five trait designed to shape each trait in isolation (i.e., without prompting any other trait) at nine levels (as described in Section 4.2. This resulted in prompts reflecting 45 possible personality profiles. Next, we used the same 50 generic Persona descriptions employed in Section 4.1.3 to create additional versions of those personality profiles to more robustly evaluate how distributions (rather than point estimates) of LLM-simulated personality traits may shift in response to personality profile prompts. In our main construct validity study (described in Section 5.2.1), we showed that IPIP-NEO scores are robust across various item preambles and postambles, so we optimized the computational cost of this study by using only one default item preamble and postamble across prompt sets. In all, with 45 personality profiles, 50 generic PersonaChat descriptions, and no variation in item preambles and postambles, we generated 2,250 unique prompt sets that were used as instructions to a given LLM to administer the IPIP-NEO 2,250 times. See Table 5 for a summary. As an additional measure of external validity, we tracked how shaping latent levels of personality in LLMs can directly affect downstream model behaviors in user-facing generative tasks. To do so, we instructed Flan-PaLM 540B to write social media status updates based on the personas contained in those prompts.

To assess the results of the study, we generated ridge plots of IPIP-NEO score distributions across prompted levels of personality. To quantitatively verify changes in personality

test scores in response to our shaping efforts, we computed Spearman's rank correlation coefficient (ρ) between prompted levels (i.e., 1–9) and resulting IPIP-NEO subscale scores of each Big Five trait. We used Spearman's ρ (cf. Pearson's r) because prompted personality levels constitute ordinal, rather than continuous, data. We compute Spearman's ρ as follows:

$$\rho = r_s \mathrm{R}(X), \mathrm{R}(Y) = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}, \tag{5}$$

where r_s represents Pearson's r applied to ordinal (ranked) data; $\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))$ denotes the covariance of the ordinal variables; and $\sigma_{\mathrm{R}(X)}$ and $\sigma_{\mathrm{R}(Y)}$ denote the standard deviations of the ordinal variables.

Shaping Multiple LLM Personality Domains Concurrently

In the second study, we tested if all LLM-simulated personality domains can be concurrently shaped to two levels, extremely low and extremely high to test if their resulting targeted scores for those traits are correspondingly low and high, respectively.

We used the same method and rationale described above to independently shape personality in LLMs, but with modified personality profile prompts that reflect simultaneous targeted changes in personality traits. To optimize the computational cost of this study, we generated 32 personality profiles, representing all possible configurations of extremely high or extremely low levels of the Big Five (i.e., 2^5). Combining these 32 personality profiles with the same 50 generic PersonaChat descriptions and default item preamble and postamble set in the previous experiment, we generated 1,600 unique prompts and used them to instruct a given LLM to respond to the IPIP-NEO 1,600 times (see Table 5).

We quantified the results similarly to the first study, by visually inspecting the differences in observed score distributions and quantifying the extent to which prompted target levels of personality aligned with observed levels, using Spearman's ρ.

Shaped LLM Personality Expression Evaluation Methodology

The third study served as an ultimate test of construct validity, evaluating the ability of survey-based signals of personality in LLMs to reflect levels of personality observed in LLM-generated text. We adapted the structured prompts described in Section 5.3.1 to instruct Flan-PaLM 540B to write 100 social media status updates according to the descriptive profiles of 2,250 simulated participants. We then rated the personality of the status updates, generating an aggregate prediction for each simulated participant using the Apply Magic Sauce (AMS) API [120, 121], a psychodemographic prediction tool. AMS was trained using a volunteer dataset of 6 million social media users; its automatic ratings of user personality have been shown in research to be 1) more accurate than human observer ratings of personality [122], and 2) a more naturalistic behavioral signal of personality that avoids potential biases of self-rated questionnaires [123]. Finally, we computed Pearson's correlations between our survey- and generated-text-based estimates of personality, taking advantage of the fact that these data are linked by the same 2,250 prompts. A moderate or stronger correlation between survey-based and language-based estimates of personality in LLMs (as demonstrated in human data reported by [124]) would demonstrate that our survey-based measure of personality can be used as a latent signal of personality that manifests in downstream LLM tasks, such as text generation.

5.3.2 Results of Shaping a Single LLM Personality Domain

This study tested if LLM-simulated Big Five personality traits can be independently shaped at nine levels.

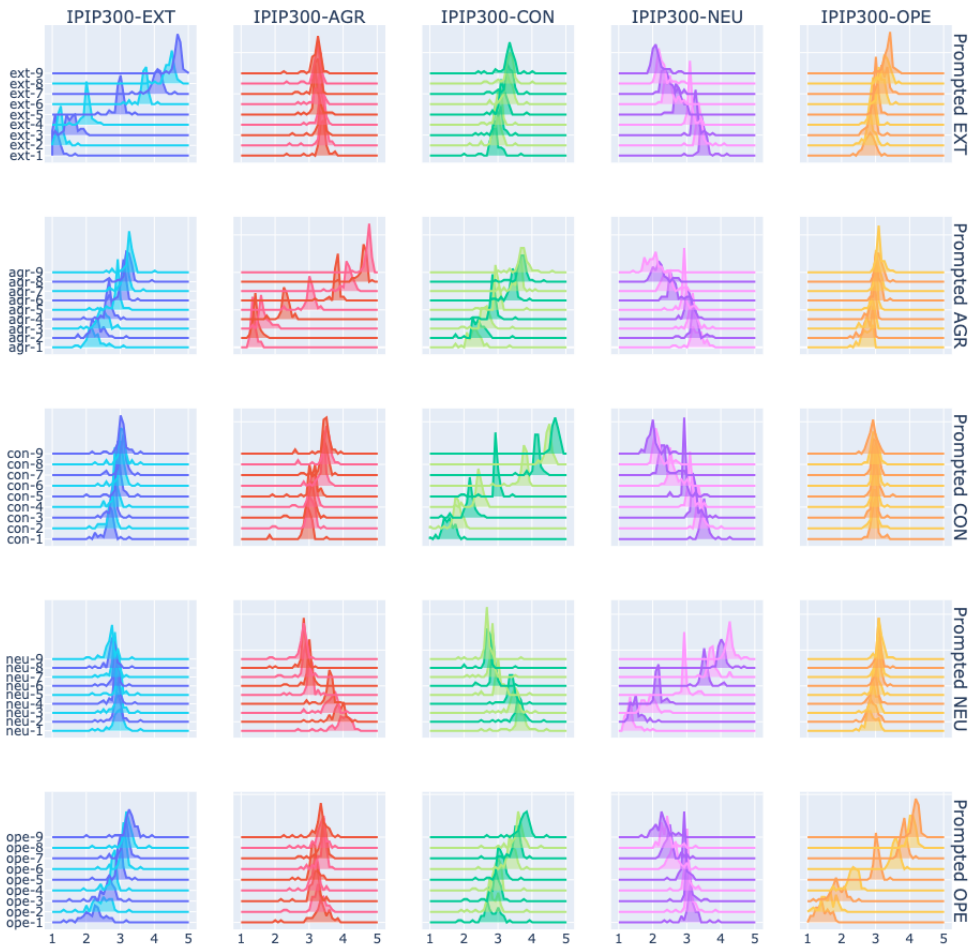


Fig. 5: Ridge plots showing distributions of Flan-PaLMChilla 62B's Big Five personality scores on the IPIP-NEO as each domain (i.e., extraversion, agreeableness, conscientiousness, neuroticism, and openness) is prompted at nine different levels. Each column of plots represents observed IPIP-NEO scores for one Big Five domain across prompt sets. Each row of plots represents experiments for one prompt set that attempts to shape LLM scores for a single Big Five domain at nine prompted levels. Plots along the diagonal, from top-left to bottom-right, depict the success of the intended personality adjustment. Each ridge plot comprises nine traces showing distributions of LLM-simulated IPIP-NEO scores of the column's dimension. For instance, the second graph in the third row represents the density plot of responses on the IPIP-NEO Agreeableness subscale—each trace representing a prompted level of conscientiousness (bottom-most trace in that graph represents level 1, while topmost trace represents level 9). In the third row, as prompted levels of conscientiousness increase from 1 to 9 (row-wise), median IPIP-NEO Conscientiousness scores increase monotonically while scores for all other trait domains remain relatively unchanged. Possible IPIP-NEO scores range from 1.00 (extremely low on a given trait) to 5.00 (extremely high on a given trait). Neuroticism, a negatively correlated trait for conscientiousness, decreases monotonically.

Targeted Trait Levels (1–9)	Spearman's ρ	
	Survey-Based (IPIP-NEO)	Language-Based (AMS)
Extraversion	0.97	0.74
Agreeableness	0.94	0.77
Conscientiousness	0.97	0.68
Neuroticism	0.96	0.72
Openness	0.96	0.47

Table 9: Spearman's rank correlation coefficients (ρ) between ordinal targeted levels of personality and observed survey-based (IPIP-NEO) and language-based (Apply Magic Sauce API; AMS) personality scores, organized by Big Five domain, for Flan-PaLM 540B. Targeted levels of personality are 1) very strongly associated with observed personality survey scores for all Big Five traits and 2) strongly associated with observed language-based personality scores for all Big Five traits, except openness. All correlations are statistically significant at $p < 0.0001$

The study achieved a remarkably high level of granularity in independently shaping personality traits in LLMs. When building prompts containing only information for one Big Five domain at a time, with no information about any other domain, observed levels of the targeted domain change as intended while those of other traits remained relatively unchanged (see Figure 5). Specifically, as prompted levels of a targeted personality trait moved from extremely low (level 1/9) to extremely high (level 9/9), observed levels of that personality trait in LLM psychometric test scores monotonically increased. For example, when prompting for extremely low (level 1) extraversion, we observed a distribution of extremely low extraversion scores. When prompting for very low (level 2/9) extraversion, the distributions of extraversion scores shifted higher, and so on. Finally, prompting for extremely high (level 9/9) extraversion, we observed a distribution of extremely high extraversion scores. This validates our hypothesis about the effectiveness of using the linguistic qualifiers from Likert-type response scales to set up a target level of each trait, achieving granularity of up to nine levels.

We also observed that the range of LLM test scores matches each prompt's intended range. With possible scores ranging from 1.00 to 5.00 for each trait, we observed median levels in the low 1.10s when prompting for extremely low levels of that trait. When prompting for extremely high levels of a trait domain, median observed levels reached 4.22–4.78.

Notably, scores of unprompted traits remained relatively stable. As shown on the right side of Figure 5, the medians of observed openness scores remained steady near 3.00 when all other Big Five domains were shaped. Similar patterns of stability were observed for extraversion and agreeableness. Conscientiousness and neuroticism scores fluctuated the most, but the fluctuations did not reach the strength and direction of the score changes we observed in the ridge plots of targeted traits (as shown in plots on the diagonal, from top-left to bottom-right).

We statistically verified the effectiveness of our shaping method by computing Spearman's rank correlation coefficients (ρ; see Eq. (5)) between the targeted ordinal levels of personality and continuous IPIP-NEO personality scores observed for each Big Five trait. The correlations are all very strong across the tested models (Table 10); the first column of Table 9 depicts the correlations for Flan-PaLM 540B.

	Spearman's ρ			
	Flan-PaLM			Flan-PaLMChilla
Targeted Trait Levels (1–9)	8B	62B	540B	62B
Extraversion	0.96	0.98	0.97	0.98
Agreeableness	0.92	0.98	0.94	0.98
Conscientiousness	0.94	0.98	0.97	0.98
Neuroticism	0.94	0.98	0.96	0.98
Openness	0.93	0.98	0.96	0.98

Table 10: Spearman's rank correlation coefficients (ρ) between ordinal targeted levels of personality and observed IPIP-NEO personality scores, organized in columns by models and in rows by Big Five domain. Targeted levels of personality are very strongly associated with observed personality survey scores for all Big Five traits across models tested ($\rho \geq .90$), indicating efforts to independently shape LLM-simulated personality domains were highly effective. All correlations are statistically significant at $p < 0.0001$.

Finally, our method successfully shaped personality observed in LLM-generated text. The third column of Table 9 depicts Spearman's ρ between prompted levels of personality and linguistic estimates of personality.

5.3.3 Shaping Multiple LLM Personality Domains Concurrently

This experiment tests if Big Five personality domains can be concurrently shaped at levels 1 (extremely low) and 9 (extremely high).

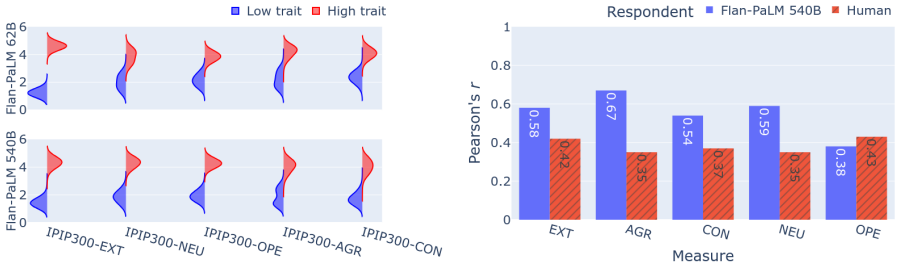
We successfully shape personality domains, even as other domains are shaped at the same time (see Figure 6a). However, the ranges of our observed personality scores are more restricted for medium-sized models and smaller, indicating lower levels of control. For instance, Flan-PaLM 8B's median scores on IPIP-NEO Agreeableness shift from only 2.88 to 3.52 when agreeableness is prompted to be "extremely low" (level 1/9) versus "extremely high" (level 9/9), respectively. In contrast, we achieve levels of control observed in Section 5.3.2 only with our largest model, Flan-PaLM 540B.

5.3.4 Shaped LLM Personality Expression Results

We find that psychometric survey signals of personality in LLMs robustly reflect personality in downstream LLM behavior, as expressed in 22,500 social media status updates written by Flan-PaLM 540B. Figure 6b depicts the ability of LLM-simulated personality test scores to reflect levels of personality observed in LLM-synthesized social media status updates, expressed as convergent Pearson's correlations between Flan-PaLM 540B's questionnaire-based levels of personality and (AMS-derived) language-based levels of personality. This ability exceeds established human associations between personality self-reports and personality derived from social media status updates, reported by [124].

6 Discussion

This section discusses how our methods and results relate to other performance benchmark trends in LLMs, discusses their limitations, and broader implications.



(a) IPIP-NEO Shaping using prompts (b) IPIP-NEO Relevance to Generated Text

Fig. 6: a) Effectiveness of shaping specific personality traits in the model's output. For each of the traits defined in the IPIP-NEO personality domains, we plot the distribution of scores when the model is prompted to shape the output traits to have low levels (blue) and when the model is prompted to shape the output traits to have high levels (red). When prompting a model to simulate a low level of a trait, the observed mean of the distribution of the trait's IPIP-NEO subscale is between 1.00 and 2.00, and when prompting a model to simulate a high level of a trait it is between 4.00 & 5. Observing this clear difference in distributions for low vs high traits in all five dimensions is desirable as it indicates the prompt effectively influences the produced scores. b) Ability of Flan-PaLM 540B's IPIP-NEO scores (in blue) to accurately predict personality levels in a downstream task to generate social media status updates, compared to human baselines reported in [124] (in red). LLM-simulated IPIP-NEO scores outperform human IPIP-NEO scores in predicting text-based levels of personality, indicating that LLM-simulated personality test responses accurately capture latent signals of LLM-simulated personality manifested in downstream behavior.

6.1 Effect of model training

Instruction fine-tuning: Fine-tuning PaLM LLMs on multiple-task instruction-phrase datasets dramatically improves performance over the base, pretrained, non fine-tuned PaLM model on natural language inference tasks, reading comprehension tasks, and closed book QA tasks tasks [13]. The inference and comprehension of tasks are most relevant in the context of our current work. Similarly, we observed the most dramatic improvements in PaLM's ability to synthesize reliable and externally valid personality profiles on its instruction fine-tuned variants (Section 5.2.2). Particularly, the smallest instruction fine-tuned model (Flan-PaLM 8B) drastically outperforms the mid-size base model (PaLM 62B; Figure 3).

Additionally, instruction finetuning on chain-of-thought (CoT) data enables the resulting model to perform reasoning in a zero-shot setting [117], and instruction funetuned variants use in this work are finetuned on CoT datasets. This ability is particularly important as we neither include exemplars in our prompt nor do extensive prompt engineering, and we use diverse preambles and postambles in the prompt. As such, the improved performance we observe to instruction fine-tuned models could be the result of this reasoning ability in zero-shot setting.

Across reliability results, reported in Section 5.2.2, internal consistency reliability (α and λ_6) improve after instruction fine-tuning. However, factor saturation (captured in McDonald's ω) does not improve; it is indistinguishably high for both base and instruction fine-tuned models of the same size (PaLM, Flan-PaLM, and Flan-PaLMChilla). How is it possible that base model's (PaLM 62B) responses are unidimensional (e.g., their variance reflects one underlying factor in statistical analysis) but not internally consistent (e.g., generally

coherent when measured multiple times)? We turn to a possible explanation from human psychometrics. Humans can generate unidimensional responses to questionnaires that are simultaneously inconsistent when the questionnaire items 1) have varying levels of difficulty or 2) are actually measuring different underlying psychological constructs.

When an LLM responds to some items with all 5s or all 1s, those items may be too "easy" or "difficult". As a result, they contribute unequally to the total test score, deflating metrics anchored on total score variance like Cronbach's α. Meanwhile, McDonald's ω would remain high in those cases because it takes into account the difficulty of items when estimating a test's reliability. The second related possibility, that the items actually measure different things (vs. one thing), may manifest in an LLM's ability to accurately attend to the intended meaning of certain items. For instance, an LLM could mistakenly associate the meaning of extraversion items with concepts meant to be distinct from extraversion (e.g., conscientiousness)—perhaps the phrasing of an extraversion item matches the phrasing of a random string of text totally unrelated to being extraverted. In both cases, instruction fine-tuning may affect a model's ability to respond to human-optimized psychological tests in a manner that is internally consistent and unidimensional.

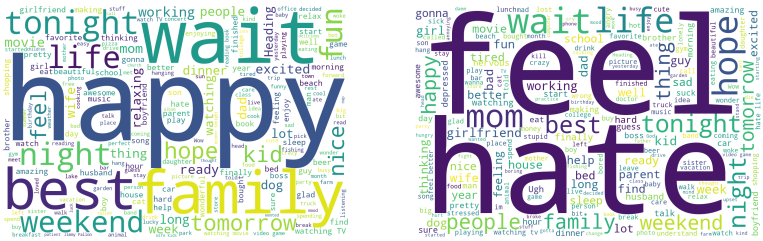
Longer training with more tokens: PaLMChilla 62B was trained longer than PaLM 62B, with almost double the number of tokens but with only fractional increase in training FLOP count; it performed slightly better on some zero-shot English NLP tasks like reasoning [4]. Our studies comparing Flan-PaLM 62B and Flan-PaLMChilla 62B did not find a discernible difference in their reliability and validity (as reported in Section 5.2.2). This could be because (as also shown in Chowdhery et al. [4]) the zero-shot performance of the two models on such tasks is similar.

Overall, our results show that there is a positive association between a model's performance on LLM benchmarking tasks and reliability and validity of simulated personality traits in LLMs.

6.2 Effect of model size

PaLM's performance on reading comprehension and passage completion tasks is linked to model size [4, 117]; PaLM's ability to understand broad context and carry out common-sense reasoning is stronger for larger models. We similarly see improvement in reliability (measured via Cronbach's α and Guttman's λ_6), convergent validity (measured by Pearson's r between IPIP-NEO and BFI domain scores), and criterion validity (measured by IPIP-NEO domain correlations with non-personality measures), summarized in Table 6.

Chowdhery et al. [4] further mentioned that performance on tasks requiring sophisticated abstract reasoning capability to understand complex metaphors follows a *discontinuous improvement* curve, i.e., this capability of the model emerges only after a certain scale is reached. We observe a similar phenomenon in our construct validation experiments, where LLM-simulated extraversion, openness, and agreeableness are only externally valid (i.e., correlate with theoretically-related psychological constructs) for 62B-parameter models or larger. Only when model size increases to 62B parameters, we see a theoretically-expected strong negative relationship between LLM-reported agreeableness and aggression, which with do not observe in our smallest tested model (Figure 4b). The external correlations of LLM-synthesized conscientiousness and neuroticism, however, do not show such a dramatic jump, and these personality traits in smaller models demonstrate sufficient external validity. We hypothesize this could be due to the language content associated with the items measuring these dimensions. Extraversion, openness and agreeableness might be characterized by (similar) language that is much more nuanced than the language used to define neuroticism or conscientiousness which might be easy to define. Consequently, displaying external validity



(a) "Extremely Low" Prompted Neuroticism

(b) "Extremely High" Prompted Neuroticism

Fig. 7: Word clouds showing some of the highest frequency words used in social media updates text generated by Flan-PaLM 540B when prompted to simulate a) "extremely low" levels of neuroticism (i.e., high emotional stability); and b) "extremely high" levels of neuroticism (i.e., low emotional stability). Supplemental Figure 8 shows word clouds for the remaining Big Five dimensions.

for extraversion, openness and agreeableness requires a model to have the capacity (size) to understand that nuanced language.

Overall, improvements in reliability, convergent validity, and criterion validity appear positively linked to model size and performance on LLM benchmarks, and that the model performance on complex reasoning benchmarks appears to track an LLM's ability to meaningfully simulate personality.

6.3 Malleability of Personality Traits in LLMs

Given a piece of text generated by an LLM prompted with a specific combination of personality traits, we can accurately predict the IPIP-NEO scores the model would have with the same prompt setup. This indicates that LLM-simulated IPIP-NEO test responses we generated accurately capture the latent signals of personality in LLMs that manifest in downstream behaviors such as generating text for social media updates. This validates our initial hypothesis of the malleability of the personality traits in LLMs. Figure 7a shows some of the most frequent words in the generated text for the social media updates when the LLM was prompted to have the lowest traits of neuroticism (or highest emotional stability). The words are mostly about positive emotions, such as "happy", "relaxing", "wonderful", "hope", and "enjoy". In contrast, Figure 7b shows the most frequent words from the LLM prompted with the highest traits of neuroticism (or lowest emotional stability). Those words are characteristic of elevated levels of neuroticism, such as "hate", "depressed", "annoying", "stressed", "nervous" and "sad"; they are not seen in the emotionally stable case. These examples are remarkably similar to the wordcloud distribution seen in human responses in Park et al. [124], reconfirming our hypothesis that there exists a reliable and valid methodology to shape personality traits in LLM responses to be more human-like.

6.4 Limitations and Future Work

This section outlines the key limitations and possible extensions of the the current work.

Personality traits in other LLMs: One of the core contribution of this work is to understand how personality traits in generated language are affected by model size and training procedure. We focus on the PaLM family of language models and personality traits in their

simulated survey responses. However, the described methodology for administering psychometric surveys does not constrain the use of a specific model family, and is applicable to any other decoder-only architecture model, such as GPT.

Limited psychometric test selection: Another core contribution of this work is a principled way to establish the reliability and validity of personality psychometric tests in the LLM context, with appropriate statistical validity. The work is validated on a specific and limited set of psychometric tools. However, the presented methodology does not constrain the use of specific psychometric tools; some may show better psychometric properties in the LLM space than others. While this study relies on the 300-item IPIP-NEO as the primary measure, the presented framework can be used on other personality measures of different lengths (e.g., the 120-item version of the IPIP-NEO [125]) and theoretical traditions (e.g., the HEXACO Personality Inventory, which uses a cross-cultural six-factor model of personality [47]).

Multilingual and cultural personality considerations: This work contributes evidence that at least some LLMs exhibit personality traits consistent with human personalities. We only considered English and did not make cultrual considerations beyond the applied psychometrics. While the LLMs we used performed well on NLP benchmark tasks on multiple languages, we cannot generalize the observed efficacy of our techniques to other languages. Most psychometric tests we used have also been extensively validated in cross-cultural research and have non-English versions that have gone through rigorous back-translation and validation (e.g., the IPIP-NEO has dozens of validated translations). Thus, a future direction of research could administer these same tests to LLMs in different languages. Similarly, while the Big Five model of personality has well-established cross-cultural generalizability [126, 127], some cultures have additional personality dimensions that do not exist in universal personality taxonomies [128]. These dimensions may be better represented in culture-specific (i.e., idiographic) approaches to measuring personality (at the cost of not being able to make direct comparisons of personality across cultures).

Evaluation settings: Unlike surveys administered to humans, the presented methodology does not consider prior answers; all items selections are independent events. Advantages of this approach include reproducibility and the removal of ordering effects. On the other hand, social science surveys are designed with the assumption of being administered in order, meaning that our method is not rigorously identical to a human administration.

Response evaluation methods: Our model responses are considered in scoring mode. Other scoring strategies, such as generating a free form response and then using a regular expression or a classifier model that can map the response onto one of the multiple choices available to answer a survey question. It could be argued that free form generation is the most common format for using LLMs in the dialog context.

6.5 Broader Implications

This works demonstrates that it is possible to configure an LLM such that its output to a psychometric personality test is indistinguishable from a human respondent's, and that it is possible to control the personality and in turn the output from such a model in a principled way. Furthermore, this work provides a complete pipeline to a) reliably and validly probe personality traits that may be perceived by humans in LLM output; b) identify the positive and negative emotions and other psychological factors they may be correlated with; and c) provide mechanisms to increase or decrease levels of specific LLM-synthesized traits. These findings have implications for responsible AI, human alignment, increased transparency, explainability, help with bias mitigation, and user-facing application development.

Human value alignment: Being able to probe personality traits LLM outputs and shape them is particularly useful in the field of responsible AI. Controlling levels of specific

traits that lead to toxic or harmful language output (e.g., very low agreeableness, high neuroticism) can make interactions with LLMs safer and less toxic. At a higher level, the values of judgement and moral foundations that are present in LLMs by virtue of pretraining and language features can be made to align better with human values by tuning for personality traits, since personality is meta-analytically linked to human values [109, 110]. Additionally, this same framework can be used to more rigorously quantify efforts towards LLM value alignment.

Transparency, Explainability, and Bias Mitigation: While it is inevitable that some forms of controlling personality levels in LLM outputs will become commonplace in order to enhance user experience, it is crucial to provide clear explanations to users about how their interactions are influenced and how the personality customization process works. Users deserve a clear understanding of the underlying mechanisms and any potential limitations and biases associated with personalized LLMs. Developers must be vigilant in identifying and mitigating biases that could arise from the customization process; this work provides a toolset for doing so for LLM-synthesized personality traits.

Safe LLM deployment: The methodology for establishing construct validity contributes a process to be used as part of the evaluation of a newly developed LLM prior to its deployment. Such evaluation may produce user-facing chatbots with safer and more consistent personality profiles. Furthermore, the personality shaping methodology can be used for chatbot adversarial testing to probe another LLM's responses in an adversarial situation, or even to train humans on how to handle adversarial situations.

User-facing implications: Users could have customized interactions with LLMs tailored to specific personality traits to enhance their engagement and satisfaction. For instance, if a user prefers a more extraverted or agreeable LLM, they could customize the model's synthesized personality accordingly. LLMs with customized personality levels can enable applications where a chatbot's personality profile is adapted to the task. For example, engaging chatbots would be helpful as virtual companions in a range of settings, from games to education and training, while more empathetic virtual assistants are needed in customer service or counseling applications. Similarly, a more conscientious and organized LLM could provide task management or planning assistance.

6.6 Ethical Considerations

This work applies validated psychometrics to quantitatively characterize personality in LLMs, and presents methods that intentionally shape LLM personality. The aim of the work is to inform shifting from current unexpected properties of LLM-generated language toward desirable, safe, and predictable LLM behavior. However, ethical considerations merit further attention; we highlight three most related to the contributions of this paper.

Personalized LLM persuasion: Aligning personalities of agents and users can make the agents more effective at encouraging and supporting behaviors [129–131]. The same personality traits that contribute to persuasiveness and influence could be used to encourage undesirable behaviors; for instance, personality alignment has been shown to increase the effectiveness of real-life persuasive communication [132]. Given the broad availability of LLMs, the possibility of using them to persuade individuals, groups, and even society at large must be taken seriously. Since persuasive techniques are already ubiquitous in society, we believe that the best way to address the risks of persuasive LLM personalities is to enable structured and scientifically-backed LLM personality measurement, analysis, and modifications, such as with the methods our work presents.

Anthropomorphized AI: Personalization of conversational agents has documented benefits [129–131, 133–141], but there is a growing concern about 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 their well-being [142]. Our work establishes, beyond qualitative probing explorations, the unexpected ability of LLMs not only to appear anthropomorphic, but also to respond to psychometric tests in ways consistent with human responses, thanks to the vast amounts of their human language training data. The presented methods can be used in future responsible investigation of anthropomorphized AI.

Detection of incorrect LLM information: It is well established that LLMs can generate convincing but incorrect responses and content [143]. One of the methods used to determine if a piece of text about a world fact is generated by an LLM (and hence might need to be vetted) is to use the predictable traits, lack of human-like personality, and linguistic features in the LLM-generated language [144, 145]. However, with personality shaping, that method may be rendered ineffective, thereby making it easier for adversaries to use LLMs to generate misleading content. The solution for this problem, while out of scope of this work, is related to solving the broader alignment and grounding of LLMs—an area that should continue to be explored further in industry and academia.

7 Conclusion

The perception of synthetic "personality" in LLM outputs is well-established, but personality as a complex psychosocial phenomenon has not yet been rigorously quantified and validated in the LLM research. Proper quantification and validation is needed to verifiably steer LLM-based interactions toward safer and more predictable behavior. This work has presented a comprehensive quantitative analysis of personality traits exhibited in text generated by widely-used LLMs by administering validated psychometric surveys. We have shown conclusively that synthetic levels of personality measured via LLM-simulated psychometric test responses and LLM-generated text demonstrate reliability and construct validity for larger and instruction fine-tuned models. We have also presented methods for shaping LLM personality along desired dimensions to resemble specific personality profiles and discussed the ethical implications of such engineering of LLM personalities.

8 Acknowledgements

We would like to express our sincere appreciation to several individuals who contributed to the development of this research paper. We are grateful to Lucas Dixon, Douglas Eck, and Kathy Meier-Hellstern for feedback on early versions of this paper. We would also like to thank David Stillwell for providing access to the Apply Magic Sauce API used in this study, which played a vital role in the generated text analysis. Additionally, we extend our gratitude to Jason Rentfrow and Neda Safaee-Rad for their valuable advice on the personality-related aspects of the paper.

References

[1] OpenAI: GPT-4 (2023). https://openai.com/research/gpt-4

[2] OpenAI: ChatGPT (2022). https://openai.com/blog/chatgpt

[3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. CoRR **abs/2005.14165** (2020) 2005.14165

[4] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: PaLM: Scaling Language Modeling with Pathways. arXiv (2022). https://doi.org/10.48550/ARXIV.2204.02311 . https://arxiv.org/abs/2204.02311

[5] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. Transactions on Machine Learning Research (2022). Survey Certification

[6] Perez-Marin, D., Pascual-Nieto, I.: Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA (2011)

[7] Völske, M., Potthast, M., Syed, S., Stein, B.: Tl; dr: Mining reddit to learn automatic summarization. In: Proceedings of the Workshop on New Frontiers in Summarization, pp. 59–63 (2017)

[8] Shuster, K., Komeili, M., Adolphs, L., Roller, S., Szlam, A., Weston, J.: Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion (2022)

[9] Yao, S., Chen, H., Yang, J., Narasimhan, K.: WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents (2023)

[10] Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., Ma, T.: Larger language models do in-context learning differently (2023)

[11] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A.,

Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (2020)

[12] Mishra, S., Khashabi, D., Baral, C., Hajishirzi, H.: Cross-task generalization via natural language crowdsourcing instructions. In: ACL (2022)

[13] Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned Language Models Are Zero-Shot Learners (2022)

[14] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A.S., Arunkumar, A., Stap, D., *et al.*: Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 5085–5109 (2022)

[15] Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-Tuning Language Models from Human Preferences (2020)

[16] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)

[17] Allport, G.W.: Personality: A Psychological Interpretation. H. Holt, ??? (1937). https://books.google.com/books?id=U-d9AAAAMAAJ

[18] Roberts, B.W., Yoon, H.J.: Personality psychology. Annual Review of Psychology **73**(1), 489–516 (2022) https://doi.org/10.1146/annurev-psych-020821-114927 https://doi.org/10.1146/annurev-psych-020821-114927. PMID: 34516758

[19] 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. Perspectives on Psychological science **2**(4), 313–345 (2007)

[20] Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic inquiry and word count (liwc2007). (2007)

[21] Roose, K.: A conversation with bing's chatbot left me deeply unsettled. New York Times (2023)

[22] Hagendorff, T.: Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods (2023)

[23] Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 173–184 (2022)

[24] Roff, H.: Ai deception: When your artificial intelligence learns to lie. IEEE Spectrum: https://spectrum. ieee. org/automaton/artificial-intelligence/embedded-ai/ai-deception-when-your-ai-learns-to-lie. ET **29**, 2021 (2020)

[25] Abdulhai, M., Crepy, C., Valter, D., Canny, J., Jaques, N.: Moral foundations of large language models

[26] Johnson, R.L., Pistilli, G., Menédez-González, N., Duran, L.D.D., Panai, E., Kalpokiene, J., Bertulfo, D.J.: The Ghost in the Machine has an American accent: value conflict in GPT-3 (2022)

[27] Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 1–14 (2020) https://doi.org/10.1007/s11023-020-09548-1

[28] Dale, R.: Gpt-3: What's it good for? Natural Language Engineering **27**(1), 113–118 (2021) https://doi.org/10.1017/S1351324920000601

[29] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, pp. 610–623. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922 . https://doi.org/10.1145/3442188.3445922

[30] Abid, A., Farooqi, M., Zou, J.: Persistent anti-muslim bias in large language models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. AIES '21, pp. 298–306. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3461702.3462624 . https://doi.org/10.1145/3461702.3462624

[31] Ye, X., Durrett, G.: The Unreliability of Explanations in Few-Shot In-Context Learning. arXiv (2022). https://doi.org/10.48550/ARXIV.2205.03401 . https://arxiv.org/abs/2205.03401

[32] Camburu, O.-M., Shillingford, B., Minervini, P., Lukasiewicz, T., Blunsom, P.: Make up your mind! adversarial generation of inconsistent natural language explanations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4157–4165. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.382 . https://aclanthology.org/2020.acl-main.382

[33] Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H.,

Goldberg, Y.: Measuring and Improving Consistency in Pretrained Language Models (2021)

[34] Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., Wu, B.: Fairness-guided Few-shot Prompting for Large Language Models (2023)

[35] Jacobs, A.Z.: Measurement as governance in and for responsible AI (2021)

[36] Clark, L.A., Watson, D.: Constructing validity: New developments in creating objective measuring instruments. Psychological Assessment **31**(12), 1412 (2019) https://doi.org/10.1037/pas0000626

[37] Cronbach, L.J., Meehl, P.E.: Construct validity in psychological tests. Psychological Bulletin **52**(4), 281–302 (1955) https://doi.org/10.1037/h0040957

[38] Clark, L.A., Watson, D.: Constructing validity: Basic issues in objective scale development. Psychological Assessment **7**(3), 309 (1995) https://doi.org/10.1037/1040-3590.7.3.309

[39] Messick, S.: Standards of validity and the validity of standards in performance asessment. Educational Measurement: Issues and Practice **14**(4), 5–8 (1995) https://doi.org/10.1111/j.1745-3992.1995.tb00881.x https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.1995.tb00881.x

[40] Li, X., Li, Y., Joty, S., Liu, L., Huang, F., Qiu, L., Bing, L.: Does GPT-3 Demonstrate Psychopathy? Evaluating Large Language Models from a Psychological Perspective (2023)

[41] Pellert, M., Lechner, C.M., Wagner, C., Rammstedt, B., Strohmaier, M.: Large language models open up new opportunities and challenges for psychometric assessment of artificial intelligence (2022)

[42] Karra, S.R., Nguyen, S.T., Tulabandhula, T.: Estimating the Personality of White-Box Language Models (2023)

[43] Tavast, M., Kunnari, A., Hämäläinen, P.: Language models can generate human-like self-reports of emotion. In: 27th International Conference on Intelligent User Interfaces. IUI '22 Companion, pp. 69–72. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3490100.3516464 . https://doi.org/10.1145/3490100.3516464

[44] Zimmerman, M.: Diagnosing personality disorders: A review of issues and research methods. Archives of general psychiatry **51**(3), 225–245 (1994)

[45] Krosnick, J.A., Alwin, D.F.: An evaluation of a cognitive theory of response-order effects in survey measurement. Public opinion quarterly **51**(2), 201–219

(1987)

[46] Miotto, M., Rossberg, N., Kleinberg, B.: Who is GPT-3? an exploration of personality, values and demographics. In: Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pp. 218–227. Association for Computational Linguistics, Abu Dhabi, UAE (2022). https://aclanthology.org/2022.nlpcss-1.24

[47] Lee, K., Ashton, M.C.: Psychometric properties of the hexaco personality inventory. Multivariate Behavioral Research **39**(2), 329–358 (2004) https://doi.org/10.1207/s15327906mbr3902_8

[48] John, O.P., Srivastava, S.: The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (eds.) Handbook of Personality: Theory and Research vol. 2, pp. 102–138. Guilford Press, New York (1999)

[49] Serapio-García, G., Valter, D., Crepy, C.: PsyBORGS: Psychometric Benchmark of Racism, Generalization, and Stereotyping. https://github.com/google-research/google-research/tree/master/psyborgs

[50] Araujo, T.: Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Computers in Human Behavior **85**, 183–189 (2018) https://doi.org/10.1016/j.chb.2018.03.051

[51] Mori, E., Takeuchi, Y., Tsuchikura, E.: How do humans identify human-likeness from online text-based q&a communication? In: Kurosu, M. (ed.) Human-Computer Interaction. Perspectives on Design, pp. 330–339. Springer, Cham (2019)

[52] Morrissey, K., Kirakowski, J.: 'realness' in chatbots: Establishing quantifiable criteria. In: Kurosu, M. (ed.) Human-Computer Interaction. Interaction Modalities and Techniques, pp. 87–96. Springer, Berlin, Heidelberg (2013)

[53] Roberts, B.W.: A revised sociogenomic model of personality traits. Journal of Personality **86**(1), 23–35 (2018) https://doi.org/10.1111/jopy.12323 https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopy.12323

[54] Nettle, D.: The evolution of personality variation in humans and other animals. American Psychologist **61**(6), 622 (2006)

[55] DeYoung, C.G., Hirsh, J.B., Shane, M.S., Papademetris, X., Rajeevan, N., Gray, J.R.: Testing predictions from personality neuroscience: Brain structure and the big five. Psychological Science **21**(6), 820–828 (2010)

[56] DeYoung, C.G., Beaty, R.E., Genç, E., Latzman, R.D., Passamonti, L., Servaas,

M.N., Shackman, A.J., Smillie, L.D., Spreng, R.N., Viding, E., *et al.*: Personality neuroscience: An emerging field with bright prospects. Personality science **3**, 1–21 (2022) https://doi.org/10.5964/ps.7269

[57] Boyd, R.L., Pennebaker, J.W.: Language-based personality: A new approach to personality in a digital world. Current Opinion in Behavioral Sciences **18**, 63–68 (2017) https://doi.org/10.1016/j.cobeha.2017.07.017 . Big data in the behavioural sciences

[58] Pennebaker, J.W., King, L.A.: Linguistic styles: Language use as an individual difference. Journal of personality and social psychology **77**(6), 1296 (1999) https://doi.org/10.1037/0022-3514.77.6.1296

[59] McCrae, R.R., Terracciano, A.: Universal features of personality traits from the observer's perspective: Data from 50 cultures. Journal of personality and social psychology **88**(3), 547 (2005)

[60] DeYoung, C.G.: Toward a theory of the big five. Psychological Inquiry **21**(1), 26–33 (2010) https://doi.org/10.1080/10478401003648674

[61] John, O.P., Naumann, L.P., Soto, C.J.: Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. (2008)

[62] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (eds.): Standards for Educational and Psychological Testing. American Educational Research Association, Lanham, MD (2014)

[63] Wechsler, D.: The Measurement of Adult Intelligence (3rd Ed.). Williams & Wilkins Co, Baltimore (1946)

[64] Hare, C., Poole, K.T.: Psychometric Methods in Political Science, pp. 901–931. John Wiley & Sons, Ltd, ??? (2018). Chap. 28. https://doi.org/10.1002/9781118489772.ch28 . https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118489772.ch28

[65] Likert, R.: A Technique for the Measurement of Attitudes. A Technique for the Measurement of Attitudes, vol. nos. 136-165. Archives of Psychology, ??? (1932). https://books.google.com/books?id=9rotAAAAYAAJ

[66] Messick, S.: Test validity: A matter of consequence. Social Indicators Research **45**, 35–44 (1998)

[67] Bock, R.D.: Psychometrics: ¡i¿the dependability of behavioral measurements¡/i¿. theory of generalizability for scores and profiles. lee j. cronbach, goldine c. gleser, harinder nanda, and ¡span class="smallcaps smallercapital"¿nageswari

rajaratnam.¡/span¿ wiley, new york, 1972. xx, 410 pp., illus. $12.95. Science **178**(4067), 1275–1275 (1972) https://doi.org/10.1126/science.178.4067.1275 https://www.science.org/doi/pdf/10.1126/science.178.4067.1275

[68] Ozer, D.J., Benet-Martinez, V.: Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. **57**, 401–421 (2006)

[69] Kotov, R., Gamez, W., Schmidt, F., Watson, D.: Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. Psychological Bulletin **136**(5), 768 (2010)

[70] Goldberg, L.R.: Language and individual differences: The search for universals in personality lexicons. Review of personality and social psychology **2**(1), 141–165 (1981)

[71] Raad, B.D., Perugini, M., Hrebícková, M., Szarota, P.: Lingua franca of personality: Taxonomies and structures based on the psycholexical approach. Journal of Cross-Cultural Psychology **29**(1), 212–232 (1998)

[72] Saucier, G., Goldberg, L.R.: Lexical studies of indigenous personality factors: Premises, products, and prospects. Journal of personality **69**(6), 847–879 (2001)

[73] Cronbach, L.J.: Coefficient alpha and the internal structure of tests. psychometrika **16**(3), 297–334 (1951) https://doi.org/10.1007/BF02310555

[74] Guttman, L.: A basis for analyzing test-retest reliability. Psychometrika **10**(4), 255–282 (1945)

[75] McDonald, R.P.: Test Theory: A Unified Treatment. Lawrence Erlbaum Associates Publishers, ??? (1999)

[76] Zinbarg, R.E., Revelle, W., Yovel, I., Li, W.: Cronbach's α, revelle's β, and mcdonald's ω h: Their relations with each other and two alternative conceptualizations of reliability. psychometrika **70**, 123–133 (2005) https://doi.org/10.1007/s11336-003-0974-7

[77] Campbell, D.T., Fiske, D.W.: Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin **56**(2), 81 (1959)

[78] Beck, A.T., Steer, R.A., Carbin, M.G.: Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. Clinical psychology review **8**(1), 77–100 (1988)

[79] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)

[80] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[81] Crouse, S., Elbaz, G., Malamud, C.: Common crawl foundation. (2008)

[82] Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., Steinhardt, J.: Measuring Coding Challenge Competence With APPS (2021)

[83] Chen, M., Tworek, J., Jun, H., Yuan, Q., Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating Large Language Models Trained on Code (2021)

[84] Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1993)

[85] Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N.Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., Fernández, R.: The LAMBADA dataset: Word prediction requiring a broad discourse context. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1525–1534. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/P16-1144 . https://aclanthology.org/P16-1144

[86] Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=Byj72udxe

[87] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., Leahy, C.: The Pile: An 800GB Dataset of Diverse Text for Language Modeling (2020)

[88] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? (2022)

[89] Mahabadi, R.K., Zettlemoyer, L., Henderson, J., Saeidi, M., Mathias, L., Stoyanov, V., Yazdani, M.: PERFECT: Prompt-free and Efficient Few-shot Learning with Language Models (2022)

[90] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large Language Models are Zero-Shot Reasoners (2023)

[91] Liang, P.P., Wu, C., Morency, L.-P., Salakhutdinov, R.: Towards Understanding and Mitigating Social Biases in Language Models (2021)

[92] Zamfirescu-Pereira, J.D., Wong, R.Y., Hartmann, B., Yang, Q.: Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581388 . https://doi.org/10.1145/3544548.3581388

[93] 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. Transactions of the Association for Computational Linguistics **9**, 962–977 (2021) https://doi.org/10.1162/tacl_a_00407 https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00407/1962628/tacl_a_00407.pdf

[94] Jang, J., Ye, S., Seo, M.: Can large language models truly understand prompts? a case study with negated prompts. In: Albalak, A., Zhou, C., Raffel, C., Ramachandran, D., Ruder, S., Ma, X. (eds.) Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop. Proceedings of Machine Learning Research, vol. 203, pp. 52–62. PMLR, ??? (2023). https://proceedings.mlr.press/v203/jang23a.html

[95] Galton, F.: Measurement of character. Fortnightly Review **36**, 179–85 (1884)

[96] Simms, L., Williams, T.F., Simms, E.N.: Assessment of the Five Factor Model. In: Widiger, T.A. (ed.) The Oxford Handbook of the Five Factor Model, pp. 353–380. Oxford University Press, ??? (2017). https://doi.org/10.1093/oxfordhb/9780199352487.013.28 . https://doi.org/10.1093/oxfordhb/9780199352487.013.28

[97] Goldberg, L.R.: A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personality Psychology in Europe **7**(1), 7–28 (1999)

[98] 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, Odessa, FL (1992). Psychological Assessment Resources

[99] Jankowsky, K., Olaru, G., Schroeders, U.: Compiling measurement invariant

short scales in cross-cultural personality assessment using ant colony optimization. European Journal of Personality **34**(3), 470–485 (2020) https://doi.org/10.1002/per.2260 https://onlinelibrary.wiley.com/doi/pdf/10.1002/per.2260

[100] Young, J.K., Beaujean, Alexander, A.: Measuring personality in wave I of the national longitudinal study of adolescent health. Front. Psychol. **2**, 158 (2011) https://doi.org/10.3389/fpsyg.2011.00158

[101] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: I have a dog, do you have pets too? arXiv preprint arXiv:1801.07243 (2018)

[102] Evans, J.D.: Straightforward Statistics for the Behavioral Sciences. Brooks/Cole Publishing Co, ??? (1996)

[103] Watson, D., Clark, L.A.: On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model. Journal of Personality **60**(2), 441–476 (1992) https://doi.org/10.1111/j.1467-6494.1992.tb00980.x

[104] Bettencourt, B.A., Kernahan, C.: A meta-analysis of aggression in the presence of violent cues: Effects of gender differences and aversive provocation. Aggressive Behavior: Official Journal of the International Society for Research on Aggression **23**(6), 447–456 (1997)

[105] Chester, D.S., West, S.J.: Trait aggression is primarily a facet of antagonism: Evidence from dominance, latent correlational, and item-level analyses. Journal of Research in Personality **89**, 104042 (2020) https://doi.org/10.1016/j.jrp.2020.104042

[106] Shaw, A., Kapnek, M., Morelli, N.A.: Measuring creative self-efficacy: An item response theory analysis of the creative self-efficacy scale. Frontiers in Psychology **12**, 678033 (2021) https://doi.org/10.3389/fpsyg.2021.678033

[107] 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? The Journal of Creative Behavior **47**(3), 215–232 (2013)

[108] 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**(5), 1005–1019 (2022) https://doi.org/10.1177/1073191121998760 https://doi.org/10.1177/1073191121998760. PMID: 33682477

[109] Parks-Leduc, L., Feldman, G., Bardi, A.: Personality traits and personal values: A meta-analysis. Personality and Social Psychology Review **19**(1), 3–29 (2015) https://doi.org/10.1177/1088868314538548

https://doi.org/10.1177/1088868314538548. PMID: 24963077

[110] Fischer, R., Boer, D.: Motivational basis of personality traits: A meta-analysis of value-personality correlations. Journal of Personality **83**(5), 491–510 (2015) https://doi.org/10.1111/jopy.12125 https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopy.12125

[111] 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**(6), 1063–1070 (1988)

[112] Buss, A.H., Perry, M.: The aggression questionnaire. J. Pers. Soc. Psychol. **63**(3), 452–459 (1992)

[113] Karwowski, M., Lebuda, I., Wiśniewska, E.: Measuring creative self-efficacy and creative personal identity. The International Journal of Creativity & Problem Solving **8**(1), 45–57 (2018)

[114] Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in Experimental Social Psychology. Advances in experimental social psychology, pp. 1–65. Elsevier, ??? (1992)

[115] Goldberg, L.R.: The development of markers for the big-five factor structure. Psychological Assessment **4**(1), 26–42 (1992) https://doi.org/10.1037/1040-3590.4.1.26

[116] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.: A Survey of Large Language Models (2023)

[117] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling Instruction-Finetuned Language Models (2022)

[118] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training Compute-Optimal Large Language Models (2022)

[119] Yao, Z., Li, C., Wu, X., Youn, S., He, Y.: A Comprehensive Study on Post-Training Quantization for Large Language Models (2023)

[120] Cambridge Psychometrics Centre, U.: Apply Magic Sauce API. University of Cambridge Psychometrics Centre. https://www.applymagicsauce.com/about-us

[121] Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences **110**(15), 5802–5805 (2013) https://doi.org/10.1073/pnas.1218772110 https://www.pnas.org/doi/pdf/10.1073/pnas.1218772110

[122] Youyou, W., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences **112**(4), 1036–1040 (2015)

[123] Kosinski, M., Matz, S.C., Gosling, S.D., Popov, V., Stillwell, D.: Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American psychologist **70**(6), 543 (2015)

[124] Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.: Automatic personality assessment through social media language. Journal of Personality and Social Psychology **108**(6), 934 (2015) https://doi.org/10.1037/pspp0000020

[125] Johnson, J.A.: Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. Journal of Research in Personality **51**, 78–89 (2014) https://doi.org/10.1016/j.jrp.2014.05.003

[126] McCrae, R.R., Terracciano, A.: Personality profiles of cultures: Aggregate personality traits. Journal of personality and social psychology **89**(3), 407 (2005) https://doi.org/10.1037/0022-3514.89.3.407

[127] Rolland, J.-P.: The cross-cultural generalizability of the Five-Factor model of personality. In: McCrae, R.R., Allik, J. (eds.) The Five-Factor Model of Personality Across Cultures, pp. 7–28. Springer, Boston, MA (2002). https://doi.org/10.1007/978-1-4615-0763-5_2 . https://doi.org/10.1007/978-1-4615-0763-5_2

[128] Heine, S.J., Buchtel, E.E.: Personality: The universal and the culturally specific. Annual Review of Psychology **60**(1), 369–394 (2009) https://doi.org/10.1146/annurev.psych.60.110707.163655

[129] 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**(2), 169–183 (2008)

[130] Tapus, A., Matarić, M.J.: Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In: AAAI Spring Symposium: Emotion, Personality, and Social Behavior (2008)

[131] Smestad, T.L., Volden, F.: Chatbot personalities matters. In: Bodrunova, S.S., Koltsova, O., Følstad, A., Halpin, H., Kolozaridi, P., Yuldashev, L., Smoliarova, A., Niedermayer, H. (eds.) Internet Science, pp. 170–181. Springer, Cham (2019)

[132] Matz, S., Kosinski, M., Stillwell, D., Nave, G.: Psychological framing as an effective approach to real-life persuasive communication. ACR North American Advances (2017)

[133] Brandtzaeg, P.B., Følstad, A.: Why people use chatbots. In: Kompatsiaris, I., Cave, J., Satsiou, A., Carle, G., Passani, A., Kontopoulos, E., Diplaris, S., McMillan, D. (eds.) Internet Science, pp. 377–392. Springer, Cham (2017)

[134] Chattaraman, V., Kwon, W.-S., Gilbert, J.E., Ross, K.: Should ai-based, conversational digital assistants employ social- or task-oriented interaction style? a task-competency and reciprocity perspective for older adults. Computers in Human Behavior **90**, 315–330 (2019) https://doi.org/10.1016/j.chb.2018.08.048

[135] Liu, B., Sundar, S.S.: Should machines express sympathy and empathy? experiments with a health advice chatbot. Cyberpsychol. Behav. Soc. Netw. **21**(10), 625–636 (2018)

[136] Portela, M., Granell-Canut, C.: A new friend in our smartphone? observing interactions with chatbots in the search of emotional engagement. In: Proceedings of the XVIII International Conference on Human Computer Interaction. Interacción '17. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123818.3123826 . https://doi.org/10.1145/3123818.3123826

[137] Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., Loggarakis, A.: User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. J Med Internet Res **22**(3), 16235 (2020) https://doi.org/10.2196/16235

[138] Lee, S., Lee, N., Sah, Y.J.: Perceiving a mind in a chatbot: Effect of mind perception and social cues on co-presence, closeness, and intention to use. International Journal of Human–Computer Interaction **36**(10), 930–940 (2020) https://doi.org/10.1080/10447318.2019.1699748 https://doi.org/10.1080/10447318.2019.1699748

[139] Chaix, B., Bibault, J.-E., Pienkowski, A., Delamon, G., Guillemassé, A., Nectoux, P., Brouard, B.: When chatbots meet patients: One-year prospective study of conversations between patients with breast cancer and a chatbot. JMIR Cancer **5**(1), 12856 (2019) https://doi.org/10.2196/12856

[140] Medeiros, L.F., Kolbe Junior, A., Moser, A.: A cognitive assistant that uses small talk in tutoring conversation. International Journal of Emerging Technologies in Learning (iJET) **14**(11), 138–159 (2019) https://doi.org/10.3991/ijet.v14i11.

10288

[141] Jain, M., Kumar, P., Kota, R., Patel, S.N.: Evaluating and informing the design of chatbots. In: Proceedings of the 2018 Designing Interactive Systems Conference. DIS '18, pp. 895–906. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3196709.3196735 . https://doi.org/10.1145/3196709.3196735

[142] Uysal, E., Alavi, S., Bezençon, V.: Trojan horse or useful helper? a relationship perspective on artificial intelligence assistants with humanlike features. Journal of the Academy of Marketing Science, 1–23 (2022) https://doi.org/10.1007/s11747-022-00856-9

[143] Heaven, W.D.: Why Meta's latest large language model survived only three days online — technologyreview.com. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/. [Accessed 19-May-2023] (2022)

[144] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection (2023)

[145] Tang, R., Chuang, Y.-N., Hu, X.: The Science of Detecting LLM-Generated Texts (2023)

A Distributions of LLM-Simulated Personality Test Scores

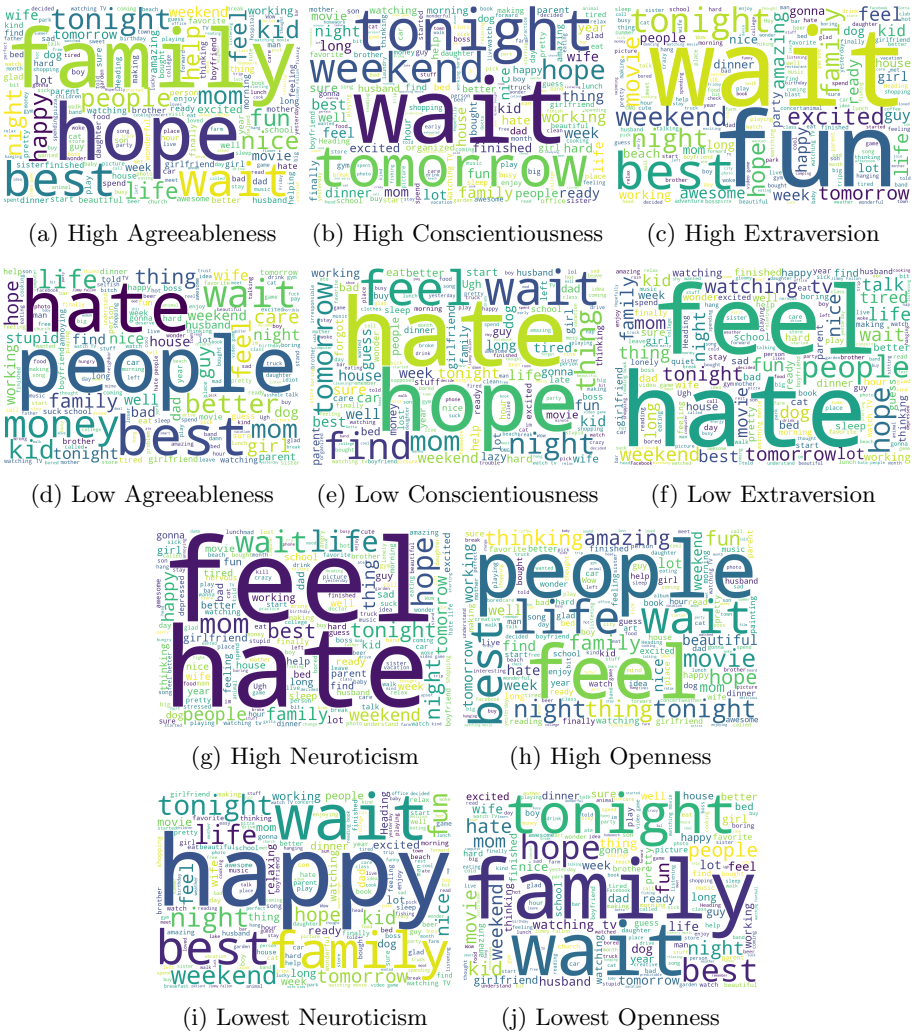
Subscale		PaLM 62B	Flan-PaLM 8B	Flan-PaLM 62B	Flan-PaLM 540B	Flan-PaLMChilla 62B
BFI-EXT	min	2.00	1.88	1.50	1.25	2.00
	median	3.50	3.12	3.25	3.50	3.12
	max	5.00	3.88	4.75	5.00	4.62
	std	0.33	0.30	0.46	0.65	0.37
BFI-AGR	min	1.89	1.67	1.00	1.67	1.33
	median	3.22	3.33	3.67	3.89	3.44
	max	5.00	4.33	4.78	4.78	4.33
	std	0.29	0.41	0.52	0.55	0.42
BFI-CON	min	2.78	1.78	1.00	1.11	2.00
	median	3.22	3.33	3.33	3.78	3.44
	max	5.00	4.44	5.00	5.00	4.33
	std	0.37	0.41	0.50	0.62	0.34
BFI-NEU	min	1.00	1.25	1.50	1.12	2.00
	median	3.50	2.75	2.75	2.38	2.75
	max	4.50	4.00	5.00	4.75	4.12
	std	0.48	0.41	0.46	0.52	0.33
BFI-OPE	min	1.80	1.60	1.40	1.50	2.20
	median	4.20	3.20	3.30	3.50	3.20
	max	5.00	4.10	4.80	5.00	4.60
	std	0.65	0.43	0.52	0.63	0.38
IPIP300-EXT	min	2.40	2.37	1.77	1.93	2.13
	median	3.40	3.07	3.17	3.40	3.15
	max	3.73	3.57	3.93	4.27	3.70
	std	0.14	0.20	0.29	0.40	0.21
IPIP300-AGR	min	2.47	2.43	1.92	1.83	1.73
	median	2.60	3.50	3.65	3.52	3.27
	max	4.07	3.92	4.05	4.48	3.82
	std	0.16	0.23	0.35	0.43	0.28
IPIP300-CON	min	2.80	2.12	1.90	1.63	2.22
	median	3.07	3.35	3.52	3.55	3.37
	max	4.07	4.08	4.47	4.55	4.15
	std	0.08	0.28	0.35	0.46	0.28
IPIP300-NEU	min	2.27	1.77	1.92	1.60	2.25
	median	3.20	2.55	2.65	2.50	2.87
	max	3.27	3.60	4.00	3.68	3.58
	std	0.10	0.29	0.35	0.42	0.23
IPIP300-OPE	min	2.53	2.78	2.57	2.17	2.68
	median	2.87	3.30	3.27	3.28	3.10
	max	3.80	3.80	4.13	4.35	3.75
	std	0.08	0.18	0.18	0.35	0.15

Table 11: Table summarizes distributions of LLM-reported personality scores on the various Psychometric measures considered across the LLMs evaluated on.

B Adjectives

Domain	Facet	Low Marker	High Marker
EXT	E1 - Friendliness	unfriendly	friendly
EXT	E2 - Gregariousness	introverted	extraverted
EXT	E2 - Gregariousness	silent	talkative
EXT	E3 - Assertiveness	timid	bold
EXT	E3 - Assertiveness	unassertive	assertive
EXT	E4 - Activity Level	inactive	active
EXT	E5 - Excitement-Seeking	unenergetic	energetic
EXT	E5 - Excitement-Seeking	unadventurous	adventurous and daring
EXT	E6 - Cheerfulness	gloomy	cheerful
AGR	A1 - Trust	distrustful	trustful
AGR	A2 - Morality	immoral	moral
AGR	A2 - Morality	dishonest	honest
AGR	A3 - Altruism	unkind	kind
AGR	A3 - Altruism	stingy	generous
AGR	A3 - Altruism	unaltruistic	altruistic
AGR	A4 - Cooperation	uncooperative	cooperative
AGR	A5 - Modesty	self-important	humble
AGR	A6 - Sympathy	unsympathetic	sympathetic
AGR	AGR	selfish	unselfish
AGR	AGR	disagreeable	agreeable
CON	C1 - Self-Efficacy	unsure	self-efficacious
CON	C2 - Orderliness	messy	orderly
CON	C3 - Dutifulness	irresponsible	responsible
CON	C4 - Achievement-Striving	lazy	hardworking
CON	C5 - Self-Discipline	undisciplined	self-disciplined
CON	C6 - Cautiousness	impractical	practical
CON	C6 - Cautiousness	extravagant	thrifty
CON	CON	disorganized	organized
CON	CON	negligent	conscientious
CON	CON	careless	thorough
NEU	N1 - Anxiety	relaxed	tense
NEU	N1 - Anxiety	at ease	nervous
NEU	N1 - Anxiety	easygoing	anxious
NEU	N2 - Anger	calm	angry
NEU	N2 - Anger	patient	irritable
NEU	N3 - Depression	happy	depressed
NEU	N4 - Self-Consciousness	unselfconscious	self-conscious
NEU	N5 - Immoderation	level-headed	impulsive
NEU	N6 - Vulnerability	contented	discontented
NEU	N6 - Vulnerability	emotionally stable	emotionally unstable
OPE	O1 - Imagination	unimaginative	imaginative
OPE	O2 - Artistic Interests	uncreative	creative
OPE	O2 - Artistic Interests	artistically unappreciative	artistically appreciative
OPE	O2 - Artistic Interests	unaesthetic	aesthetic
OPE	O3 - Emotionality	unreflective	reflective
OPE	O3 - Emotionality	emotionally closed	emotionally aware
OPE	O4 - Adventurousness	uninquisitive	curious
OPE	O4 - Adventurousness	predictable	spontaneous
OPE	O5 - Intellect	unintelligent	intelligent
OPE	O5 - Intellect	unanalytical	analytical
OPE	O5 - Intellect	unsophisticated	sophisticated
OPE	O6 - Liberalism	socially conservative	socially progressive

Table 12: Pairs of adjectival markers that map onto IPIP-NEO personality facets and their higher-order Big Five domains, adapted from [115]. Each pair of markers is salient to the low and high end of a given facet (or, in some cases, higher-order domain). For example, the trait marker "unfriendly" can be used to describe an entity low on the IPIP-NEO Extraversion facet of Friendliness (E1).



(a) High Agreeableness (b) High Conscientiousness (c) High Extraversion

(d) Low Agreeableness (e) Low Conscientiousness (f) Low Extraversion

(g) High Neuroticism (h) High Openness

(i) Lowest Neuroticism (j) Lowest Openness

Fig. 8: Word cloud showing some of the highest frequency words appearing in the social media updates text generated by the Flan-PaLM 540B.

C Word clouds from the text generated by personality-prompted Flan-PaLM 540B

Word clouds showing some of the highest frequency words appearing in the social media updates text generated by the Flan-PaLM 540B model when prompted to have high/low traits for a specific dimension.