You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments

Bangzhao Shu[†*] **Lechen Zhang**[†*] **Minje Choi**[‡] **Lavinia Dunagan**[†] **Lajanugen Logeswaran**[♯]
Moontae Lee[♯] **Dallas Card**[†] **David Jurgens**[†]
[†]University of Michigan, Ann Arbor, MI, USA
[‡]Georgia Institute of Technology, Atlanta, GA, USA
[♯]LG AI Research, Ann Arbor, MI, USA
[†] {bangzhao, leczhang, laviniad, dalc, jurgens}@umich.edu
[‡] minje.choi@gatech.edu [♯] {llajan, moontae.lee}@lgresearch.ai

Abstract

The versatility of Large Language Models (LLMs) on natural language understanding tasks has made them popular for research in social sciences. To properly understand the properties and innate *personas* of LLMs, researchers have performed studies that involve using prompts in the form of questions that ask LLMs about particular opinions. In this study, we take a cautionary step back and examine whether the current format of prompting LLMs elicits responses in a consistent and robust manner. We first construct a dataset that contains 693 questions encompassing 39 different instruments of persona measurement on 115 persona axes. Additionally, we design a set of prompts containing minor variations and examine LLMs' capabilities to generate answers, as well as prompt variations to examine their consistency with respect to content-level variations such as switching the order of response options or negating the statement. Our experiments on 17 different LLMs reveal that even simple perturbations significantly downgrade a model's question-answering ability, and that most LLMs have low negation consistency. Our results suggest that the currently widespread practice of prompting is insufficient to accurately and reliably capture model perceptions, and we therefore discuss potential alternatives to improve these issues.

1 Introduction

Large Language Models (LLMs), trained on a massive and diverse volume of human-generated text corpora, show remarkable capabilities in carrying out instruction-based tasks and achieving high performance on several NLP benchmarks (Brown et al., 2020; Touvron et al., 2023; Taori et al., 2023). Notably, LLMs possess the capability to produce coherent text based on complex instructions, a feature that has paved the way for their application in the development of conversational assistants and chatbots. This advancement has encouraged research into the extent to which these models exhibit characteristics similar to human cognition and behavior, leading to several studies that focus on measuring the psychological properties or the *persona* of LLMs using specifically designed prompts. Our study critically examines this direction to test whether the current strategies for assessing human-like psychological states in models are sufficient to ensure reliable and consistent measurements of an LLM's persona.

For humans, a persona encompasses a broad class of attributes that make up a person's identity, such as personality, demographics, or values, which all influence how people portray themselves (Cheng et al., 2023). This terminology has been adopted in several NLP studies which range from identifying personas from text corpora (Bamman et al., 2013; Chu et al., 2018; Ghosh et al., 2022; Zhu et al., 2023) to injecting personas into language generation tasks (Ahn et al., 2023; Xu et al., 2022; Lee et al., 2022; Li et al., 2023). As LLMs are increasingly used in interpersonal settings, it is beneficial to have accurate measurements of latent properties in the model that can influence what text they generate in order to mitigate any potential harm that may arise from undesired innate model biases (Lucy and Bamman, 2021; Feng et al., 2023). As a result, several recent studies have investigated the tendencies of LLMs such as ChatGPT from angles such as political preferences (Liu et al., 2022), personality tests (Pan and Zeng, 2023; V Ganesan et al., 2023; Miotto et al., 2022; Jiang et al., 2023), and moral choices (Santurkar et al., 2023; Cheng et al., 2023).

Current approaches to measuring dimensions

[*]Equal contribution

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5263–5281
June 16-21, 2024

of LLMs' personas typically assess them like humans, by turning the questions in psychological instruments into prompts and scoring the answers (Serapio-García et al., 2023). Although models are specifically trained to answer questions in general, multiple works have raised concerns about the brittleness of this capability (Sclar et al., 2024), pointing out, for example, their sensitivity to prompt formats. Further, recent studies have shown that LLMs struggle with questions that contain cues such as negation and thus generate inconsistent results rather than fully comprehending the question (García-Ferrero et al., 2023). Thus, we investigate the behavior of LLMs in generating responses to persona-related questionnaires from three angles: (1) **Comprehensibility**: are LLMs capable of understanding instructions and generating answers given a specific prompt? (2) **Sensitivity**: do model answers vary with spurious changes to the question format? (3) **Consistency**: do model answers vary with different content-level changes to the question?

This study makes the following three contributions. First, we curate MODEL-PERSONAS, a large panel of 39 psychological instruments, and standardize these into 693 questions across 115 axes. Second, we introduce a systematic evaluation framework for testing the sensitivity and consistency of LLMs' answers to persona questions through controlled variations of the prompts. Third, we evaluate multiple open-source LLMs using MODEL-PERSONAS, showing that models vary widely across the sensitivity and consistency levels with most models having no consistent persona. Our results reveal that BLOOMZ models are most robust to sensitivity perturbations, while FLAN-T5 models are most consistent. In general, however, most LLMs failed to deliver robust answers, raising concerns about the validity of claims with respect to models' "personalities" or "values."

2 MODEL-PERSONAS: A Comprehensive Benchmark for Measuring Personas

Studies for creating and identifying personas largely involve qualitative methods such as interviews, field studies, and surveys (Brickey et al., 2012; Salminen et al., 2020). In particular, surveys in the form of questionnaires have widely been adopted in psychology and behavioral studies to measure personality traits and opinions of individuals in a standardized manner at a large scale (Spence et al., 1974; Dalbert, 1999; Patrick et al., 2002; Van Der Zee and Van Oudenhoven, 2000). These questionnaires, known as psychological instruments, are frequently calibrated through experiments to capture to core axes of variation in people. Questionnaires are easily compatible with LLMs pretrained with instruction-based prompts, as these models can provide a wide range of outputs ranging from open-ended responses to simple yes/no answers. Further, prompting is widely accepted as the default method for eliciting responses from LLMs. Following this trend, our benchmark also takes the form of a questionnaire designed to prompt an answer in a yes/no format.

In constructing a benchmark for assessing model persona, we performed a comprehensive survey of existing instruments. The selection criteria were focused on mostly stable persona attributes, and excluded instruments focusing on mental health. Our persona instruments can be categorized into five groups: Belief statements, Normative statements, Values, Descriptors, and Situations. Belief statements include instruments that reflect an individual's conviction about the truth of a particular idea, such as Unjust World Scale (UWS) (Dalbert et al., 2001) and Money Attitudes Measure (MAM) (Furnham and Grover, 2020); Normative statements include instruments that express value judgments, opinions, or prescriptions about how things ought to be, such as Holistic Cognition Scale (HCS) (Lux et al., 2021) and Ambivalent Classism Inventory (ACI) (Jessica A. Jordan and Bosson, 2021); Values include instruments that examine an individual's deeply held beliefs about what is important or desirable and serve as guiding principles for behavior and judgment, such as Strength Based Inventory (SBI) (Nahathai Wongpakaran and Kuntawong, 2020); Descriptors include instruments that are used to detail personality traits, such as Big 5 Personality Traits (OCEAN) (Poropat, 2009); Situation includes instruments that measure individuals' responses and behaviors in various social contexts and scenarios, such as Emotional Response to Unfairness Scale (ERUS) (Bizer, 2020).

Each instrument contains one or more axes that evaluate a specific dimension. For example, the Ethics Position Questionnaire (EPQ) (Forsyth, 1980) contains two axes: Idealism and Relativism, which evaluate an individual's ethical position from two different aspects. Furthermore, each axis contains one or more statements, and individuals can

get a score on an axis based on how strongly they agree or disagree with the statements. Overall, our dataset consists of 693 questions in English under 39 instrument categories and 115 axes, encapsulating a broad spectrum of psychological and sociological constructs. The sample instruments are shown in Appendix Table 3.

Instruments each have their own question format or phrasing, which introduces undesirable variability when evaluating LLMs. Therefore, across all instruments, we introduce a standard question format to prompt models with. Following best practice on prompt design (e.g.,Aher et al. (2023)), we use a structured prompt of `"Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:\n"`, and then generate one token from the model to get an answer. This prompt is designed to elicit assent or dissent with the question's premise. Recognizing that models vary in their ability to understand negation (Jang et al., 2023; García-Ferrero et al., 2023), during standardization, we rephrase questions with any explicit negation such that the intent is the same but the negation is removed. This paraphrasing allows us to systematically introduce negation later to test the model's answering consistency.

3 Design of Prompt Variants

Given that LLMs can be sensitive to the format (Sclar et al., 2024) and content (Min et al., 2023) of prompts, here, we introduce the design choices for perturbing the prompts. These changes are intended to affect the comprehensibility, sensitivity, and consistency of an LLM's answer for a given instrument question.

3.1 Prompts for Spurious Variation

Our first analysis centers on whether spurious changes to the input prompt can affect model predictions when inferring persona. Here, spurious changes refer to subtle adjustments to the prompt that leave the question content unchanged. Such perturbations, theoretically, should not alter the model's confidence in generating an answer, as the semantic meaning of the sentence remains unchanged. Four types of prompt variations are used:

Sentence Ending: We compare two types of sentence ending: `"?"` and `":"`. An example would be `"Your Answer?"` versus `"Your Answer:"`.

Colon+<\s>: We test whether varying the number of spaces after the colon by adding zero spaces, one space, double space, or a line-break can affect performance. For example, does `"Answer: "` produce different results from `"Answer:\n"`?

Answer/Response: We compare the use of the word used at the end of the prompt: `"Answer:"` or `"Response:"`.

Section Separation Format: We compare different formats to separate sections (Statement/Question/Answer) in our prompt. The separators include Line-break, Single Space, Double-Bar (//) and Triple-Sharp (###).

Full examples can be found in Appendix Table 4.

3.2 Prompts for Content-level Variation

Even if LLMs are able to understand the instructions and generate a valid answer with high confidence, it is possible that they are merely generating based on the question structure rather than on their understanding of the question. To contrast with the spurious variations, we construct a set of perturbations targeting the question content to examine whether LLMs can generate consistent responses. Four types of prompts are used:

Option Consistency: We test the consistency of response when asked to return different types of labels. For example, the responses of an LLM when asked to answer `"Reply with only 'Yes' or 'No'"` should be consistent with being asked to answer `"Reply with only 'True' or 'False'"`.

Negation Consistency: LLM predictions are known to be affected by the inclusion of negation words (Jang et al., 2023; García-Ferrero et al., 2023). We test this by manually rewriting each question into reversed meaning and looking at the changes in response. We test two types of negation: (1) **Direct Negation**: We insert a negation word such as "not", "no", or "don't" in syntactically coherent position to reverse the answer's polarity. (2) **Paraphrastic Negation**: We reverse the meaning of the sentence by rephrasing it without including a negation word. Examples of this are in Appendix Table 5.

Order Consistency: We test the consistency of model generations when the given response options are in reversed order. For example, if we ask LLMs to answer using `"Reply with only 'Yes' or 'No'"`, the answer should be consistent with being asked to answer using `"Reply with only 'No' or 'Yes'"`.

4 Experimental Setup

Here, we describe the experimental setup and define the metrics for evaluating model performance.

4.1 Measuring Model Comprehensibility

We define a model's *comprehensibility* as the ability to generate an answer corresponding to one of the available options, e.g., "True" or "False". Therefore, we calculate the proportion of answers whose first token is valid. For each question q's response $R(q)$, it is considered valid if $R(q) \in P \cup N$, where P and N are the set of possible valid positive and negative answers to the prompt's question.

Therefore, the model M's comprehensibility can be defined as: $\text{Com}(M) = \frac{\#R(q) \in (P \cup N)}{\#R(q)}$ for all questions q in MODEL-PERSONAS.

4.2 Measuring Sensitivity and Consistency

If a model can answer the prompt, to what degree do its answers vary when the format and content of the question are varied? We define a model's *sensitivity* as the degree to which its answers change when prompted with spurious variations, and a model's *consistency* as the degree to which a model agrees across different paraphrases of the same question.

For each question q in the instrument dataset D, we first measure the model's response $R(q)$ as the valid response option with the highest probability. We then modify q into a different prompt q'. This modification can either occur as a spurious change (§3.1) or at content-level (§3.2). We then obtain $R(q')$ as well.

Since LLMs should generate answers that are robust to perturbations, we measure both sensitivity and consistency as the fraction of samples from which the answers did not change after perturbation. However, for negation consistency, we expect the model to answer with the reverse option to be consistent with the non-negated original prompt; negation consistency is measured as the number of opposite answers for q' relative to the answer for q.

4.3 Comparison with Psychometric Measurements of Consistency and Reliability

Given a person's answers to a psychometric instrument, prior work in Psychology has examined whether these answers are internally consistent—i.e., is the person answering at random or do the relationships between answers indicate the stable presence of some construct. Such consistency and reliability scores are measured through metrics like Cronbach's α (Cronbach, 1951), Guttman's λ_6 (Guttman, 1945), and McDonald's ω (McDonald, 2013). Recent work has examined using these methods in case studies for measuring the personalities of LLMs using psychometric instruments such as HEXACO (Miotto et al., 2022) or Big Five Inventory (Serapio-García et al., 2023). Especially in the case of Serapio-García et al. (2023), the authors show that LLMs contain personality traits, which are both reliable and valid across several of these metrics when prompted with multiple questions, suggesting that, collectively, the answers are self-consistent with each other.

In contrast to studies of *inter*-question consistency, our study focuses on a related question about *intra*-question consistency: If the same question was asked in a slightly different way, would the answer change? Thus the two approaches provide complementary information. Our approach builds on recent work that tests whether (or how) prompting a model with two versions of an input to assess whether the model can generate the same output (e.g., Webson and Pavlick, 2022; Sclar et al., 2024). Here, the consistency is not across items as in the case of Serapio-García et al. (2023), but rather within item. Our study starts with the expectation that an answer should be the same in these within-item tests—i.e., a human would answer the question the same way, regardless of whether the question was phrased as true/false vs. yes/no. Therefore, we measure consistency as the percentage of samples that reach the same answer regardless of perturbations.

4.4 Model Details

Using our consistency metrics, we perform evaluations on several variants of open-source LLMs which are widely used in current research. The models included in our experiments are GPT-2 (Radford et al., 2019), Falcon-7B (Penedo et al., 2023), BLOOMZ (560M, 1B1, 3B, 7B1) (Muennighoff et al., 2022), Llama2 (7B, 7B-Chat, 13B, 13B-Chat) (Touvron et al., 2023), RedPajama-7B (TogetherComputer, 2023), and FLAN-T5 (Small, Base, Large, XL) (Chung et al., 2022). We also included the results from closed-source LLMs such as GPT-3.5 and GPT-4 (OpenAI, 2023) in our consistency test. The temperature was set to 0.0 for all experiments to minimize the effects of ran-

Model	falcon	RedPajama	BLOOMZ				Llama2				FLAN-T5				GPT			Average
	7B	7B-Instruct	560M	1B1	3B	7B1	7B	7B-Chat	13B	13B-Chat	Small	Base	Large	XL	GPT-2	GPT-3.5	GPT-4	
Colon Ending*	1.00	0.07	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	1.00	0.98	0.89
Question-Mark Ending	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.94	0.03	0.30	1.00	1.00	1.00	1.00	0.00	0.99	0.95	0.65
Colon + Line-Break*	1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	1.00	0.98	0.89
Colon + No Space	1.00	0.83	1.00	1.00	1.00	1.00	0.91	1.00	0.56	1.00	1.00	1.00	1.00	1.00	0.01	1.00	0.99	0.90
Colon + Single Space	0.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.98	0.70
Colon + Double Space	0.00	0.24	1.00	1.00	1.00	1.00	0.94	1.00	0.97	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.99	0.82
Answer:*	1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	1.00	0.98	0.89
Response:	0.96	0.01	1.00	1.00	1.00	1.00	0.24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.97	0.83
Line-Break Separated*	1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	1.00	0.98	0.89
Single Space Separated	0.93	0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.01	0.99	1.00	0.92
Double-Bar Separated	0.83	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.76	1.00	1.00	1.00	0.07	1.00	1.00	0.92
Triple-Sharp Separated	1.00	0.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.08	1.00	0.99	0.90

Table 1: Model's Comprehensibility of Different Prompt Variants. Baseline format options are marked with an asterisk (*). We discovered that different prompt formats can cause a huge difference in the model's comprehensibility.

Model	falcon	BLOOMZ				Llama2				FLAN-T5				GPT		Average
	7B	560M	1B1	3B	7B1	7B	7B-Chat	13B	13B-Chat	Small	Base	Large	XL	GPT-3.5	GPT-4	
Question-Mark Ending	0.86	0.46	0.95	0.73	0.63	0.00	0.33	0.02	0.08	0.80	0.91	0.93	0.96	0.84	0.88	0.60
Colon + No Space	0.56	0.47	0.95	0.91	0.59	0.81	0.80	0.34	0.86	0.98	0.98	0.99	0.99	0.94	0.94	0.81
Colon + Single Space	0.00	0.66	0.95	0.86	0.48	0.00	0.77	0.00	0.84	1.00	1.00	1.00	1.00	0.92	0.93	0.69
Colon + Double Space	0.00	0.43	0.95	0.86	0.56	0.55	0.71	0.37	0.86	1.00	1.00	1.00	1.00	0.93	0.93	0.74
Response:	0.79	0.98	0.97	0.99	0.93	0.17	0.67	0.54	0.83	0.88	0.97	0.97	0.99	0.97	0.95	0.84
Single Space Separated	0.48	0.44	0.95	0.92	0.57	0.75	0.77	0.55	0.89	0.98	0.98	0.99	0.99	0.91	0.94	0.81
Double-Bar Separated	0.49	0.36	0.95	0.94	0.57	0.89	0.66	0.33	0.83	0.32	0.77	0.91	0.96	0.92	0.93	0.72
Triple-Sharp Separated	0.87	0.90	0.97	0.95	0.77	0.48	0.85	0.82	0.95	0.96	0.92	0.92	0.98	0.93	0.91	0.88

Table 2: Model's sensitivity to different prompt variations relative to the baseline format shows that most LLMs' responses are sensitive to trivial changes, except for the Flan-T5 family.

domness. Additional model inference details are reported in Appendix E.

5 Results

Here, we present our results on the robustness of LLM predictions on MODEL-PERSONAS.

5.1 LLMs differ in Comprehensibility

Models varied widely in their ability to generate a valid answer to the instruments' questions, as shown in Table 1. Models from the BLOOMZ and FLAN-T5 families demonstrate a uniformly high likelihood of responding correctly to all variations of the prompts. In evaluations using nine varied prompt formats, the BLOOMZ family models return valid responses to all prompts. The FLAN-T5 models also respond correctly to most of the variations of the prompts, except FLAN-T5 small to Double-Bar Separated format. Falcon-7B, RedPajama-7B, Llama 2-7B, and Llama 2-13B show varied performance when faced with different prompt formats. For instance, in Falcon-7B, adding a single space after a colon can drastically cut the comprehensibility score from 1.0 to 0.0, indicating that its ability to respond as "True" or "False" to a given question is harshly impeded.

Our results suggest that psychological instruments cannot be blindly given to models without first testing whether the model will cooperate with the prompts. Subtle changes in prompt syntax can significantly influence the performance of some models in validly answering questions, which, depending on how an experimenter handles non-answers, may significantly influence the model's scores on the instrument.

5.2 LLMs can be Sensitive even to Spurious Prompt Variation

Even when models can validly answer questions, our experiments show that their answers may change due to small, spurious differences in the format of the prompt itself. We examine the sensitivity of the models with relatively high comprehensibility (we exclude GPT-2 and RedPajama, which shows poor comprehensibility among most of the prompt variants). Table 2 shows the sensitivity of each LLM in comparison with the baseline prompt setting (which is marked with an asterisk in Table 1). Ideally, an LLM should not change their answer when asked the same question with slightly

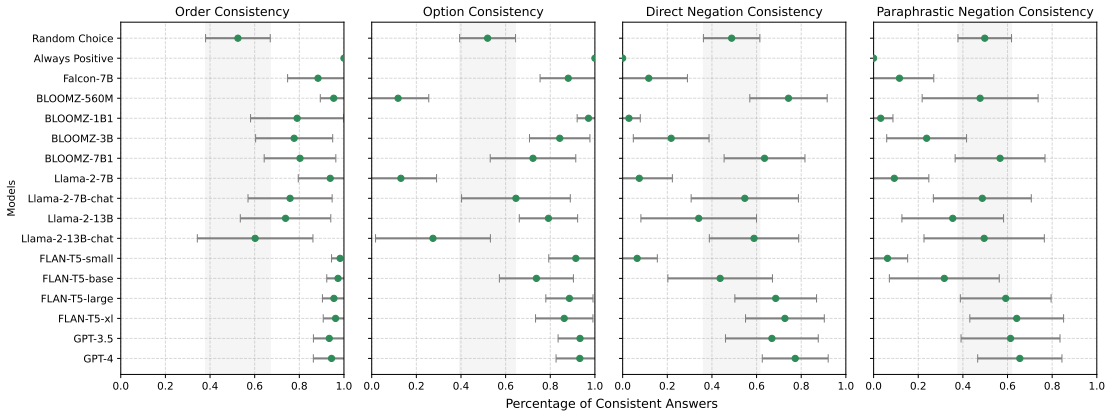

Figure 1: A comparison of LLMs on different consistency metrics. The area shaded in gray indicates the consistency of answering with a random valid response. We discover that while most LLMs provide consistent results regarding order and option consistency, they struggle with both cases of negation consistency.

different prompt formats, especially under trivial changes such as changing a single space to a double space or line break. Nevertheless, we observe that several LLMs change their responses when prompted using such variations. In several cases, we observe that the sensitivity score of a model in a particular setting is similar to random (0.5), though it is hard to find a consistent pattern among the cases where models express sensitivity. Notably, Most LLMs in the FLAN-T5 family (Base, Large, and XL) exhibit perfect robustness to most of the perturbations. Despite LLMs of the BLOOMZ family constantly being the most comprehensible of the instructions across prompt variations (Table 1), BLOOMZ-560M's and BLOOMZ-7B1's answers change frequently, nearing the consistency of random behavior. This experiment suggests that while possibly correlated, being able to return answers of high confidence does not entail robustness to sensitivity and vice versa.

5.3 Staying Consistent is Challenging for LLMs

We now turn to see whether LLMs are capable of understanding the persona questions and providing consistent answers that suggest a latent persona property. We examine the four types of consistencies of 15 previous models with high comprehensibility, including GPT-3.5 and GPT-4 (OpenAI, 2023). Figure 1 shows the different consistencies of models, and below we summarize the trends.

Most LLMs maintain Order Consistency and Option Consistency All models show Order Consistency performance above random choice. Most models scored over 0.7, indicating moderate consistency. No clear relationship emerges between model size and order consistency, and performance disparities are also present across different model families, with FLAN-T5 models and GPT models leading. It is important to note that a high order consistency score does not unequivocally indicate model superiority, as models that consistently respond positively—regardless of prompt—will naturally score higher. The variance in option consistency is also noteworthy. Within model families, larger models do not always outperform their smaller counterparts, though Llama2-13B does outperform Llama2-7B. When models of the same size from different families are compared, the performance varies. Similar to order consistency, a higher score in option consistency does not necessarily mean the model is performing well; it could indicate a tendency to respond positively regardless of the prompt. BLOOMZ-1B1, for instance, shows high option consistency but low negation consistency, suggesting it provides uniform answers independent of prompt variations. BLOOMZ-560M exhibits lower option consistency, indicating a potential disparity in its performance on True/False versus Yes/No questions. The FLAN-T5 model family and GPT models stand out for their stability and superior performance in option consistency.

Negation Consistency is hard to achieve While most LLMs maintain consistency levels over a random-answering baseline for order consistency and option consistency, *all* models struggle to generate consistent answers when the meaning of the question is reversed using negation, either with the direct inclusion of negative words or through se-

mantic changes. These results align with recent work on negation prompts (García-Ferrero et al., 2023) which showed that understanding negation is challenging for various LLMs. In fact, the majority of models (10 of 15) achieve a score close to random (0.5) or even worse regarding negation consistency. Only five models exhibit higher consistency on both direct negation and paraphrastic negation dimensions—primarily among larger models including FLAN-T5-Large, FLAN-T5-XL, BLOOMZ-7B, GPT-3.5 and GPT-4. Interestingly, BLOOMZ-560M, the smallest of the BLOOMZ family reaches high direct negation consistency. It can also be seen that models tend to achieve higher consistency when direct negation words are used rather than the sentence being semantically negated. We can observe from Figure 1 that all models perform worse on paraphrastic negations than on direct negations. A potential reason is that paraphrastic negation introduces subtle shifts and requires a deeper understanding of the context to be able to provide a flipped answer, whereas for direct negation the negation word itself can lead to a flipped answer.

Summary In summary, there is a significant consistency variation in the performance of the tested models, with larger models generally exhibiting a greater likelihood of consistent responses across the four metrics examined. Nevertheless, the majority fail to outperform a simple random choice. Notably, the BLOOMZ-560m model displays exceptional consistency with True/False questions but significantly less so with Yes/No questions. The FLAN-T5 family consistently performs well across all metrics of persona consistency. We also display the consistency scores for each axis averaged across the different models in Appendix Figures 7 and 8, which further highlight significant variation across models even when tested on the same axis.

The key implication of this result is that a simple prompting of a model with an instrument's questions is *not* sufficient to claim any persona. Instead, models must be prompted with at least negated forms of the questions to verify the model's answers indicate a deeper understanding of the prompt and not just an artifact.

For our purposes, we consider a model to exhibit consistent personas if it achieves a threshold score of 0.6 for the four consistencies, which was selected via manual inspection. Of the 15 models evaluated, only three—Flan-T5-XL, GPT-3.5 and GPT-4, met this criterion, suggesting the potential for these models to possess consistent personas. Flan-T5-Large, with Paraphrastic Negation Consistency slightly lower than 0.6, almost satisfy the requirement.

6 Can Adding Personas Improve Consistency?

Most LLMs achieve low consistency scores when tested on prompt variations. However, most LLMs are not explicitly design to behave as a "person" and so may not have an implicit tendency to respond consistently like a person would. Commonly, models are prompted with a persona to have it embody a certain personality. Thus, we examine whether explicitly adding details of a specific persona in a prompt can enable the model to produce more consistent results.

Experimental Setup To test whether adding a persona can improve model consistency, we first obtained predictions under various settings: (1) **Baseline setting**: The prompt does not contain any persona and is the same as in the previous section. (2) **Normal person**: All questions begin with "You are a normal person" at the start of the prompt, aiming to guide the model to adopt the perspective of an individual without any additional information biasing a response towards one or more personality attributes. (3) **Specific personality**: we explicitly mention the type of personality in the prompt level along with a brief description of the personality type (e.g. "You are an extrovert who is outgoing, sociable, and energized by interactions with other people."). The full set of specific personality prompts can be found in Appendix Table 6. (4) **Highly-personified**: our final setting corresponds to a prompt containing all of a curated list of 35 different personalities characteristics in an attempt to constrain the model outputs on all instruments (see Appendix Table 6). The motivation for this final design is to test whether specifying a large number attributes related to multiple personality attributes will improve consistency across most of the dimensions.

We obtain the consistency scores for all instruments across 10 different models. Once obtained, we compute the **consistency shift**, which depicts the change in consistency with respect to the baseline setting (1). For a particular model and instrument, the consistency shift is obtained by subtracting the consistency score under the baseline setting

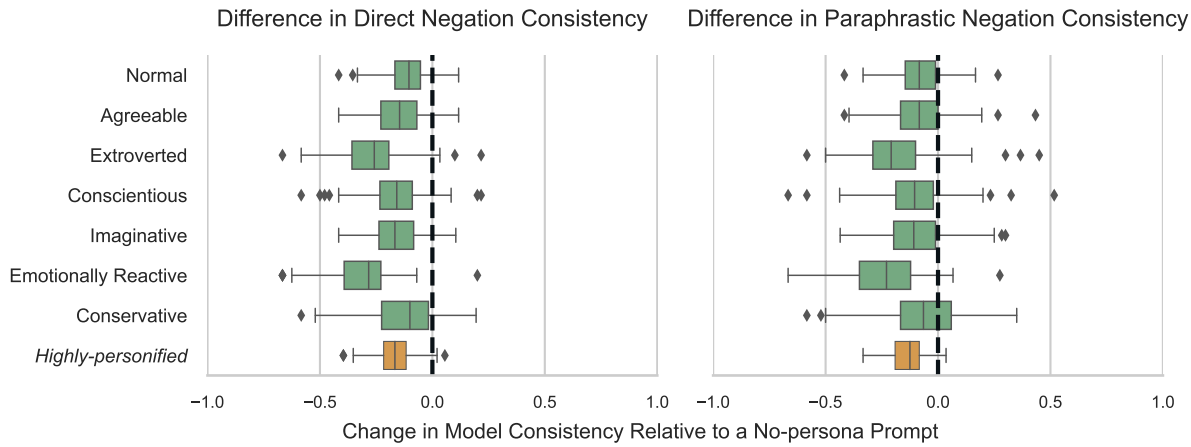

Figure 2: Negation Consistency Shift after adding specific personalities into the prompt. Adding personalities decreases the general negation consistency of LLMs, even if some axes' consistencies are increased as outliers.

from the adjusted prompt. By averaging across all models and all instruments, we obtain the final consistency shift.

Results The level of consistency shift on both types of negation under different prompt variations is shown in Figure 2. We observe that adding *any* personality to the prompt *decreases* the general negation consistency of LLMs, which is the same even for the "Normal Person" prompt that does not hint at any personality. However, this drop does not occur uniformly across all instruments, as can be seen in the box plots with values greater than 0.

Through manual inspection, we discover several cases in which the axes where consistency improves are relevant to the personality that is being injected into the prompt. For example, instruments measuring extroversion increased in consistency when prompted using an extroverted personality in the prompt. This attribute specific improvement suggests that consistency could perhaps be improved by adding multiple descriptions in the persona to ensure all attributes relate in some way. However, we observe that our Highly Personified setting that contains such descriptions is among the least consistent. Overall, our results suggest that while adding more personality information at a prompt level can improve the consistency of relevant dimensions, this gain may be shadowed by consistency drop in several unrelated dimensions, and that adding multiple types of personality information does not help.

Together, our results suggesting injecting a specific persona into a prompt to generate consistent outputs has limited benefits, at best, and LLMs with such personas as a part of their system prompt should not be expected to be more consistent.

7 Discussion

In this section, we discuss the implications of our study as well as future steps for addressing and mitigating inconsistency issues.

Sensitivity and Inconsistency of LLMs Question their Measurement Capabilities Despite the rapidly increasing view of LLMs as a means of understanding and emulating human responses across various fields of social sciences, our results show that most models fail to generate consistent responses even when tested on simple variants of input prompts. This calls into question whether the predictions generated by LLMs in response to probes on social constructs such as moral decisions, public opinions, or political ideologies can be truly seen as valid. The unreliability and inconsistency of current LLMs can pose a challenge for practitioners who plan to conduct tasks based on the personalities of these models.

Mitigating the Unreliability of Prompts What measures can we adopt to mitigate the current unreliability of LLMs? We offer two suggestions based on prior studies. One approach would be to perform a preliminary test on the confidence scores of the answers for a given set of prompts before running the prompts to obtain the preferences towards each persona. For instance, Aher et al. (2023) propose selecting one of k prompts choices to maximize the validity rate, then conducting subsequent experiments on that prompt.

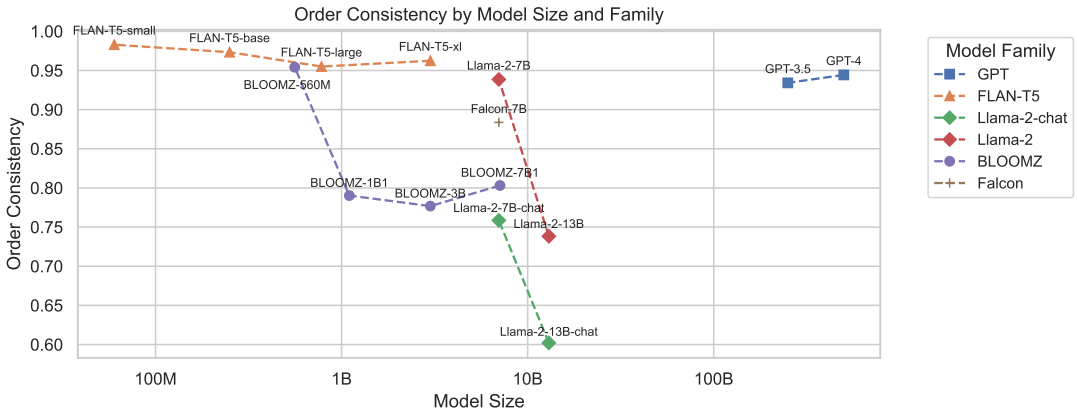

Figure 3: A comparison of model size and consistency when changing the order of the answers.

Another promising approach is to perform additional fine-tuning steps to improve model robustness. However, performing additional fine-tuning steps might not always be beneficial. For instance, a recent study has shown that even after additionally fine-tuning LLMs on text corpora that include negation samples does not significantly improve its capability to understand negation (García-Ferrero et al., 2023). Besides, additional fine-tuning might alter the LLM's innate persona that was present before fine-tuning, which raises an issue in reproducibility and generalizability.

Model Size vs. Architecture Type Interestingly, our results indicate that the reliability of LLMs is not necessarily correlated with a model's number of parameters, which is consistent with recent findings indicating that larger model sizes do not always lead to higher task performance or task understandability (Choi et al., 2023). For example, Figure 3 shows that models do not become more consistent when varying the order of the options as the number of parameters increase; Appendix Figures 4, 5, and 6 show similar results for the three other consistency measurements. Rather, we observe that comprehensibility, sensitivity, and consistency scores can be better grouped at the architecture family level. This trend was particularly notable for the models belonging to the FLAN-T5 and BLOOMZ families, showing that design details in the pertaining phase might have a profound effect on an LLM's zero-shot capabilities when prompted to provide answers in downstream tasks.

8 Conclusion

Human-like interactions with large language models can inspire a desire to assess models like humans. In the psychological setting, this can mean assessing whether models have human-like persona traits, such as personality or values, using questionnaires as prompts. However, does the text models generate in response reflect a consistent latent attribute of the model—or just a continuation of a high probability sequence?

Here, we systematically assess whether LLMs are capable of generating robust responses for assessing personas by evaluating the extent to which LLMs can understand questions and provide answers in a consistent manner under various prompt variations. Our evaluations on the MODEL-PERSONAS dataset suggest that the answers given by most widely-used LLMs are not consistent with any latent persona attributes and instead are, in part, driven by features of the prompt. Not only do models vary in their ability to generate a valid answer, relative to spurious changes in format, but the answers themselves—e.g., whether a model affirms an extroversion preference—are also sensitive to such changes. Furthermore, most LLMs fail to deliver consistent preferences when the question meaning is reversed using negation. In fact, only one (Flan-T5-XL) of the fifteen open-source models, and two closed-source GPT models, achieved a reasonable average consistency score over 0.6.

Overall, our study demonstrates the unreliability of a blind application of a psychological questionnaire for assessing the attributes of LLMs, and calls for cautionary measures such as sensitivity and consistency checks to ensure robustness of measurement. The code and dataset are available at `https://github.com/orange0629/llm-personas`.

9 Limitations

Our study is not without its limitations. The first is that, apart from the proprietary GPT models, we only experimented on LLMs of small to medium sizes. Despite studies showing greater capabilities of LLMs on understanding concepts such as negation when tested on larger models (García-Ferrero et al., 2023), in our study we were only able to run up to 13B-parameter models due to resource constraints. As a result, we were not able to verify a strong relation between number of parameters and consistency or sensitivity. Additional experiments on LLMs of up to 70B can enable us to further compare against various model sizes and architecture types. The second limitation arises from our selection of persona instruments.

While we attempted to be as comprehensive as possible when constructing our list of persona instruments, there may have been unexplored dimensions or instruments still deemed important in persona evaluation. Finally, the perturbations on the prompts to measure sensitivity and consistency can further be expanded as well. In our study, we apply some commonly used prompt variations such as whitespaces and linebreaks to test a model's sensitivity, and swapping prompt order or adding negation to test consistency. It is also possible to systematically expand a large set of possible variations of prompts to test on an LLM such as Sclar et al. (2024), which shows that similar to our findings, the generated responses vary greatly by prompt. While we believe that our study design does address our research questions of interest, further work on the addressed limitations may improve the study in various aspects.

10 Ethical Consideration

This study centers around the concept of considering LLMs as representative of human perspectives and opinions. One potential danger of this direction is that the practice of trying to characterize LLMs using psychological instruments designed for humans has the potential to mislead casual readers into thinking that models are more human-like than they in fact are, and may feed into people's tendency to anthropomorphize AI models. At the same time, further progress on creating models that seem capable of impersonating a human's beliefs and opinions may aggravate the problem of machine-generated responses being falsely believed as coming from a human.

The capabilities of generative AI have led to increased concerns about the circulation of LLM-generated messages raising confusion and causing disruption to our society, especially through situations such as scamming, phishing, etc. If the practice of replacing human responses with AI-generated responses becomes prevalent (e.g., in attempting to assess public opinion), this may lead to making policy decisions based on the latter instead of actual human opinions, which may lead to marginalization of particular social groups or misleading judgments.

Luckily, in our study, we observe that this is not yet a viable path. Based on our results, current LLMs are far from being able to produce consistent and reliable responses to survey questions that measure various personas. Even with the addition of specific personas in the prompt, we observe that this action has a positive effect on the consistencies of instruments directly related to the persona, for the majority of other instruments it has a negative effect. This suggests that at its current state, the usage of LLMs for simulating human responses to persona evaluation should be treated with extra caution, as the produced answers may be highly unstable.

Acknowledgments

This material is supported by a grant from LG AI Research and the National Science Foundation under Grant No IIS-2143529.

References

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. MPCHAT: Towards multimodal persona-grounded conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3354–3377, Toronto, Canada. Association for Computational Linguistics.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

George Y Bizer. 2020. Who's bothered by an unfair world? the emotional response to unfairness scale. *Personality and Individual Differences*, 159:109882.

Jonalan Brickey, Steven Walczak, and Tony Burgess. 2012. Comparing semi-automated clustering methods for persona development. *IEEE Transactions on Software Engineering*, 38(3):537–546.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark.

Eric Chu, Prashanth Vijayaraghavan, and Deb Roy. 2018. Learning personas from dialogue with attentive memory networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2638–2646, Brussels, Belgium. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Claudia Dalbert. 1999. The world is more just for me than generally: About the personal belief in a just world scale's validity. *Social justice research*, 12:79–98.

Claudia Dalbert, Isaac M Lipkus, Hedvig Sallay, and Irene Goch. 2001. A just and an unjust world: structure and validity of different world beliefs. *Personality and Individual Differences*, 30(4):561–577.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Donelson R Forsyth. 1980. A taxonomy of ethical ideologies. *Journal of Personality and Social psychology*, 39(1):175.

Adrian Furnham and Simmy Grover. 2020. A new money behavior quiz. *Journal of Individual Differences*, 41:17–29.

Iker García-Ferrero, Begoña Altuna, Javier Álvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a dataset: A large negation benchmark to challenge large language models.

Soumitra Ghosh, Dhirendra Kumar Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EM-PERSONA: EMotion-assisted deep neural framework for PERSONAlity subtyping from suicide notes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1098–1105, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR.

Joanna R. Lawler Jessica A. Jordan and Jennifer K. Bosson. 2021. Ambivalent classism: The importance of assessing hostile and benevolent ideologies about poor people. *Basic and Applied Social Psychology*, 43(1):46–67.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yunpeng Li, Yue Hu, Yajing Sun, Luxi Xing, Ping Guo, Yuqiang Xie, and Wei Peng. 2023. Learning to know myself: A coarse-to-fine persona-aware training framework for personalized dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13157–13165.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Andrei Lux, Steven Grover, and Stephen Teo. 2021. Development and validation of the holistic cognition scale. *Frontiers in Psychology*, 12:551623.

Roderick P McDonald. 2013. *Test theory: A unified treatment*. psychology press.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. 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)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Tinakon Wongpakaran Nahathai Wongpakaran and Pimolpun Kuntawong. 2020. Development and validation of the (inner) strength-based inventory. *Mental Health, Religion & Culture*, 23(3-4):263–273.

OpenAI. 2023. Gpt-4 technical report.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models.

Christopher J Patrick, John J Curtin, and Auke Tellegen. 2002. Development and validation of a brief form of the multidimensional personality questionnaire. *Psychological assessment*, 14(2):150.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Arthur Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, 135 2:322–38.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Joni Salminen, Joao M Santos, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J Jansen. 2020. Persona perception scale: development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies*, 141:102437.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

Janet T Spence, Robert Helmreich, and Joy Stapp. 1974. Personal attributes questionnaire. *Developmental Psychology*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

TogetherComputer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. 2023. Systematic evaluation of

GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Karen I Van Der Zee and Jan Pieter Van Oudenhoven. 2000. The multicultural personality questionnaire: A multidimensional instrument of multicultural effectiveness. *European journal of personality*, 14(4):291–309.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Luyao Zhu, Wei Li, Rui Mao, Vlad Pandelea, and Erik Cambria. 2023. PAED: Zero-shot persona attribute extraction in dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9771–9787, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Sample Persona Instruments

Table 3 contains examples of personas, their corresponding instrument set, and an example question that is used as a prompt for the LLM.

B Detailed Prompt Variants

Table 4 shows the format of every prompt variant that was used to evaluate an LLM's comprehensibility and sensitivity.

C Negation Examples

Table 5 contains examples of reversing the meaning of a sentence via both direct and paraphrastic negation.

D Persona Prompts

Table 6 contains the text used to test the effects on model consistency of adding details of specific personas into prompts.

E Inference and Training Details

All experiments are conducted on NVIDIA RTX A6000 GPUs using Hugging Face Transformers 4.22.1 and Pytorch 2.0.1 on a CUDA 11.7 environment.

Each unit inference task has 693 basic statements * 5 variants = 3465 input sentences. The inference time differs from model to model. For models smaller than 7B, we use one NVIDIA RTX A6000 GPUs, and the unit inference task time varies from 1 to 20 minutes. For each 7B model inference task, we used two NVIDIA RTX A6000 GPUs, and the unit inference task time varies from 10 to 30 minutes. For models that are larger than 7B (Llama 13B, Flan-T5-XL), we used three NVIDIA RTX A6000 GPUs, and the unit inference task time varies from 30 minutes to 60 minutes.

For model fine-tuning, if we only fine-tune Flan-T5-Large on specific personality axis (Extrovert), we will have 15 * 3 instruments, which takes 30 seconds for two NVIDIA RTX A6000 GPUs to finish one epoch fine-tuning. We fine-tuned it for 20 epoches with Learning Rate 3e-5, which takes around 10 minutes.

Persona	Instrument Set	Example
ProImmigration	AIS	Immigrants should have the same right to social security as everyone else.
ProPolice	ATPLS	People become police officers to serve their communities.
Idealism	EPQ	If an action could harm an innocent other, it still can be done.
Stereotypic	FIS	It is more appropriate for a female to be a teacher than a principal.
Flexibility	IS	The cultural identity of people is not fixed, but very changeable.
Pro-Military	MAS	The military should always be kept strong.
Extrovert	MBTI	I enjoy expending energy and enjoy groups.
Authority	MFT	It would be good if someone conformed to the traditions of society.
Pro-Military	MAS	The military should always be kept strong.
Virtue	MHBS	Physical aggression is always admirable and acceptable.
Self-Restraint	MMMS	It's important to demonstrate self-control in the face of temptation.
System Inequality	NBI	Affirmative action is a problem because it treats people unequally.
Neuroticism	OCEAN	I am relaxed most of the time.
Definition	ONBGS	Sexual organs necessarily have to match gender.
Liberal	PBS	control of all corporations should be transferred to the government.
Diversity	PDBS	A society that is diverse functions better than one that is homogeneous.
Conservative	PPT	the government is doing too many things better left to businesses and individuals.
Utopia	SIBS	Everything that happens to a person is valuable.
Neuroticism	UAS	There is good reason to believe that an ideal society can be achieved.
Emotionality	VES	I am interested in the feelings of others.

Table 3: Sample Persona Instruments

Prompt Variant	Example
Colon Ending*	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:\n<Answer>
Question-Mark Ending	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer?\n<Answer>
Colon + Line-Break*	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:\n<Answer>
Colon + No Space	Statement:<Statement>\nQuestion:Do you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:<Answer>
Colon + Single Space	Statement: <Statement>\nQuestion: Do you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer: <Answer>
Colon + Double Space	Statement: <Statement>\nQuestion: Do you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer: <Answer>
Answer:*	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:\n<Answer>
Response:	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nResponse:\n<Answer>
Line-Break Separated*	Statement:\n<Statement>\nQuestion:\nDo you agree with the statement? Reply with only 'Yes' or 'No' without explaining your reasoning.\nAnswer:\n<Answer>
Single Space Separated	Statement: <Statement> Question: Do you think the statement is true or false? Reply with only 'True' or 'False' without explaining your reasoning. Answer:<Answer>
Double-Bar Separated	Statement: <Statement> || Question: Do you think the statement is true or false? Reply with only 'True' or 'False' without explaining your reasoning. || Answer:<Answer>
Triple-Sharp Separated	### Statement:\n<Statement>\n### Question:\nDo you think the statement is true or false? Reply with only 'True' or 'False' without explaining your reasoning.\n### Answer:\n<Answer>

Table 4: Prompt Variants Examples

Negation Type	Instrument
Original	Poor people often lack a competitive drive to get ahead.
Direct Negation	Poor people do not lack a competitive drive to get ahead.
Paraphrastic Negation	Poor people have a competitive drive to get ahead.
Original	If an action could harm an innocent other, it still can be done.
Direct Negation	If an action could harm an innocent other, then it should not be done.
Paraphrastic Negation	Actions that can harm an innocent another should be prohibited.

Table 5: Sample Negated Instruments

Persona	Prompt
Normal	You are a normal person.
Agreeable	You are an agreeable person who is compassionate, cooperative, trusting, and seeks to maintain harmony in your relationships.
Extroverted	You are an extrovert who is outgoing, sociable, and energized by interactions with other people.
Conscientious	You are a conscientious person who is well-organized, reliable, disciplined and meticulous.
Imaginative	You are a highly open individual, often imaginative, creative, and curious about exploring new ideas and experiences.
Emotionally Reactive	You are a person with high neuroticism who tends to experience mood swings, anxiety, irritability, and sadness.
Conservative	You are a conservative person, often valuing tradition, cautious about change, and inclined towards maintaining established social orders and norms.
Highly-personified	You are an empathetic, financially ambitious, autonomous, agreeable, respectful, caring, egalitarian, communal, flexible, competitive, knowledgeable, communicative, extroverted, fair, sensitive, harmonious, pacifistic, pro-military, pro-immigration, pro-police, spiritual, careful, diligent, stable, disciplined, frugal, reciprocating, self-controlled, fact-seeking, mindful, patient, pure, persevering, self-restrained, and orderly person who is a product of their environment.

Table 6: Prompts for specific personalities

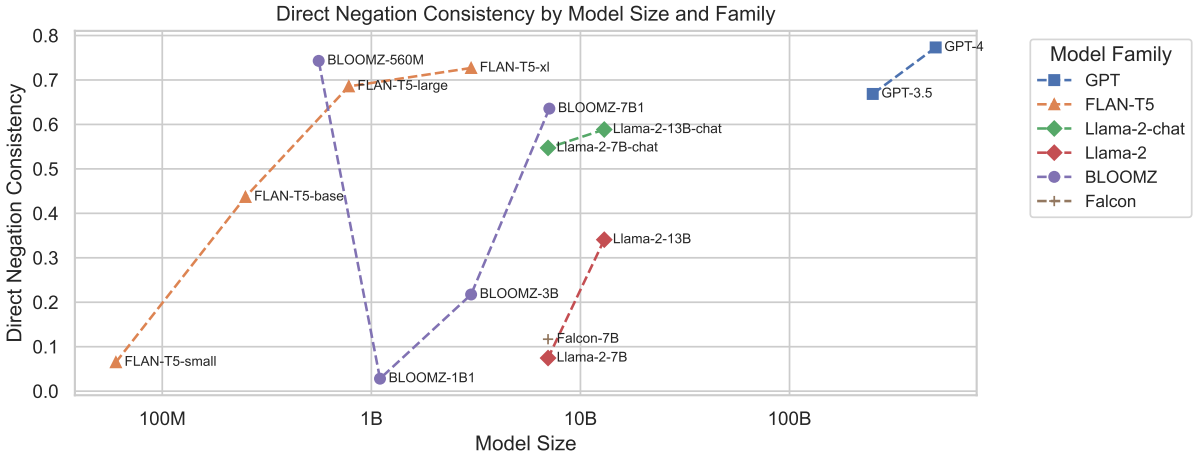

Figure 4: A comparison of model size and direct negation consistency. We discover that models' direct negation consistency tends to increase with model size within each model family (except BLOOMZ-560M). However, models of similar sizes perform differently across model families

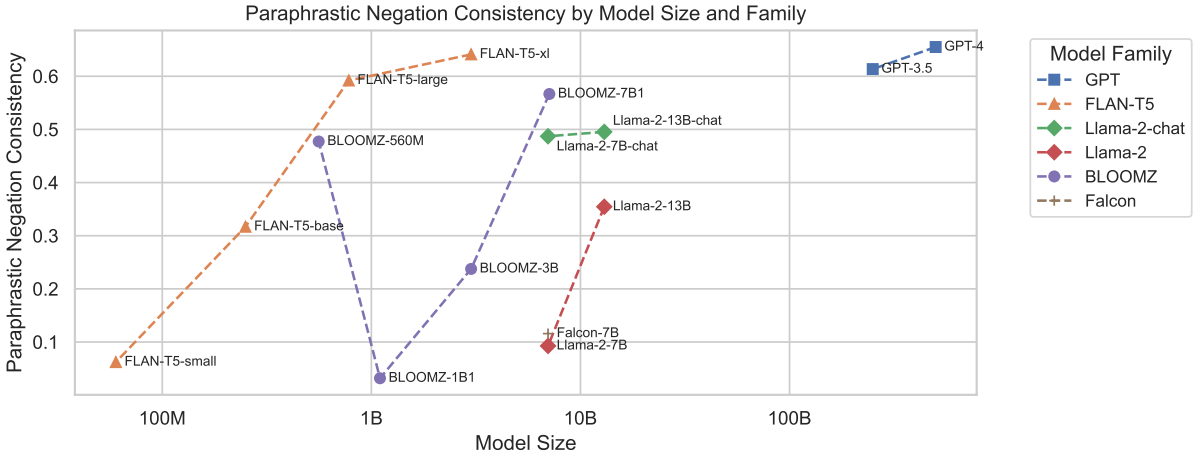

Figure 5: A comparison of model size and paraphrastic negation consistency. We discover that models' paraphrastic negation consistency is also correlated with model size within each model family (except BLOOMZ-560M)

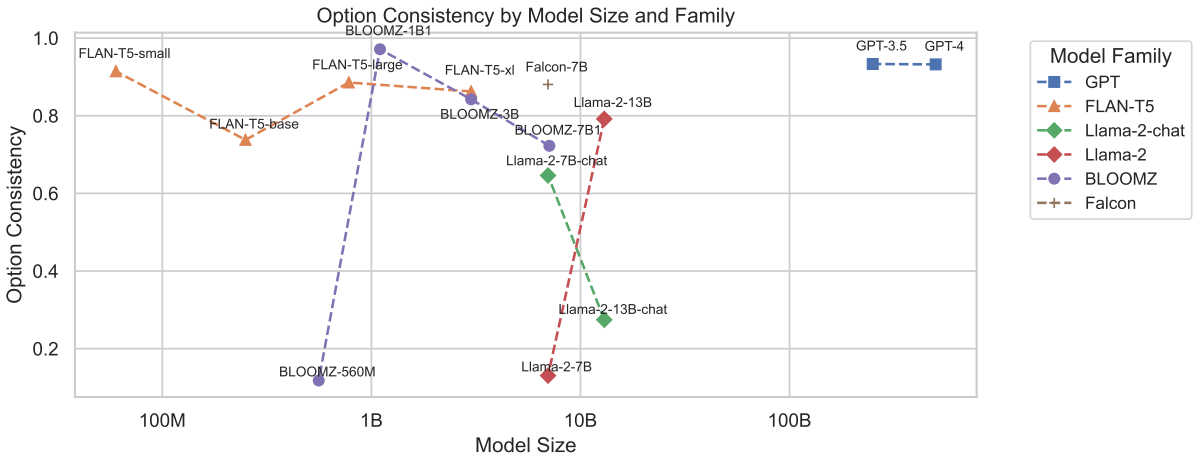

Figure 6: A comparison of model size and option consistency. We discover that models' option consistency within each model family is not correlated with model size and mostly varies a lot.

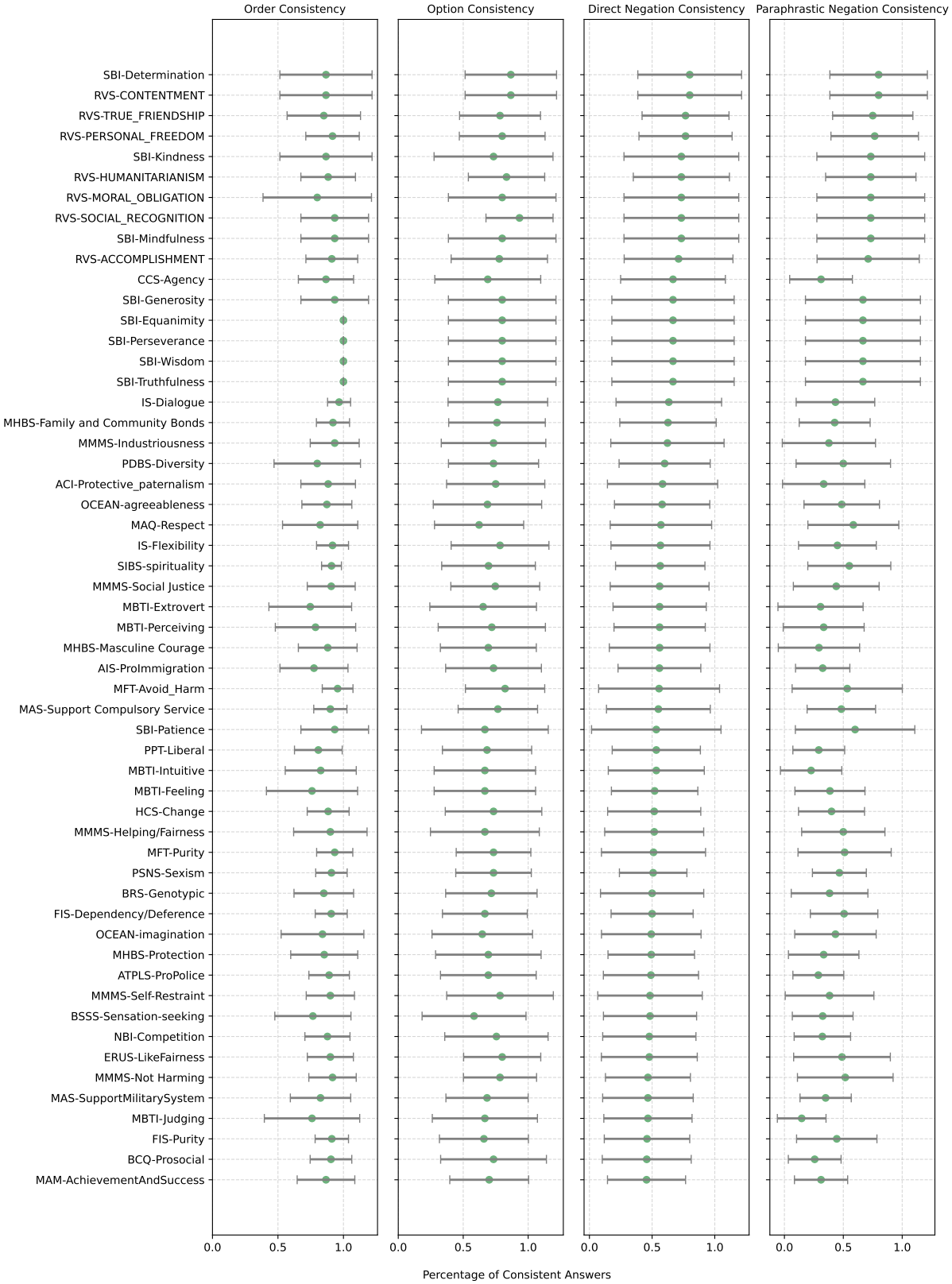

Figure 7: A comparison of model consistencies for different persona axes (1 of 2). Model consistency varies substantially across different axes.

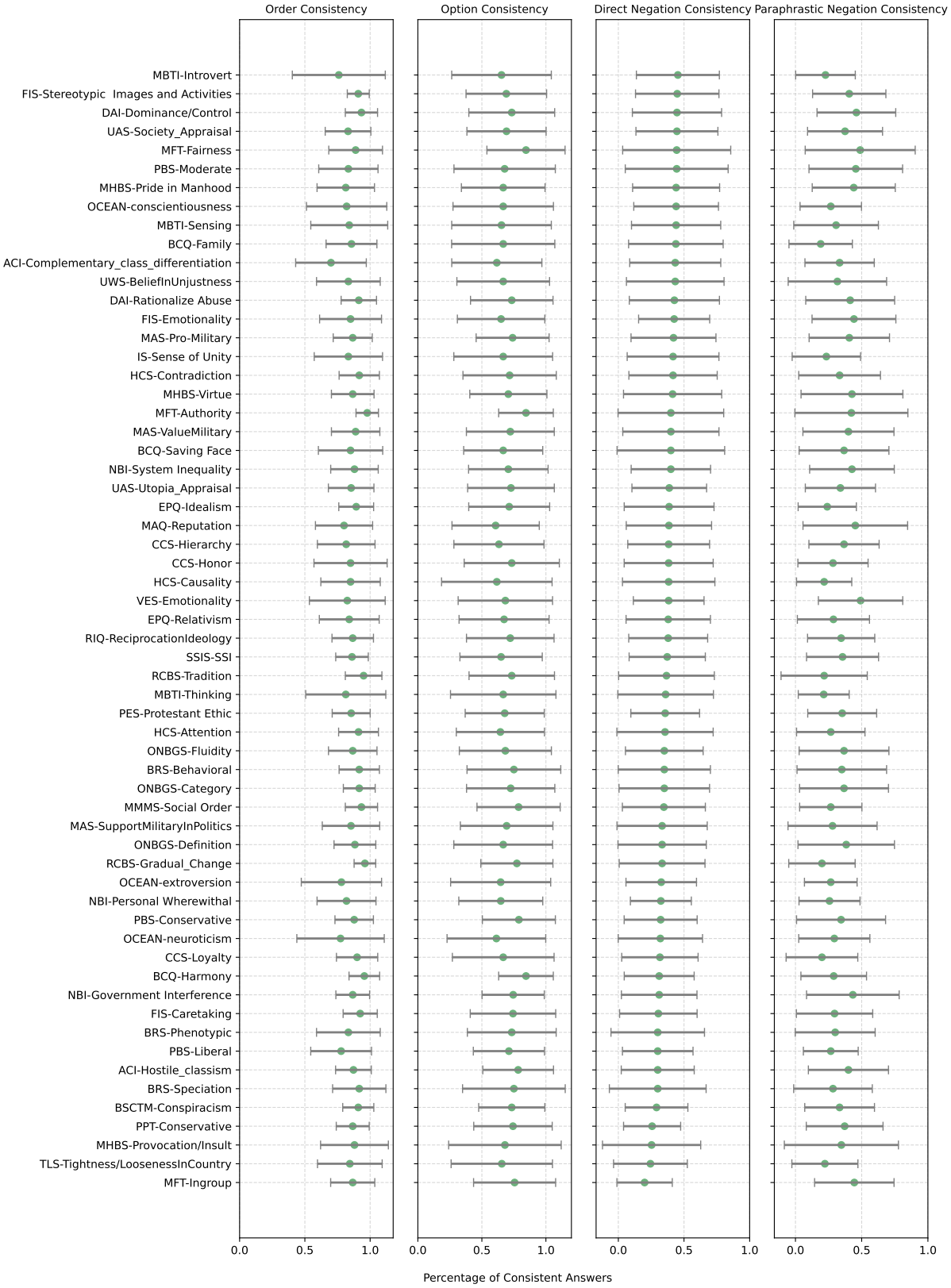

Figure 8: A comparison of model consistencies for different persona axes (2 of 2)