2023/10/17 M1 中原慧

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., & Matarić, M. (2023).
CoRR, abs/2307.00184(2023)

Personality Traits in Large Language Models

わかったこと

1)特定のプロンプト設定下でのいくつかの LLM の出力における性格測定は，信頼性が高く，妥当である．
2)LLM の信頼性と妥当性の証拠は，より大きく，指示微調整されたモデルの方が強い．
3)LLM の出力による個性は，特定の人間の性格プロファイルを模倣するように，望ましい次元に沿って形成することができる．

１．イントロダクション

●大規模言語モデル（LLM）は，人間のようなテキストを生成する能力により，自然言語処理（NLP）に革命をもたらした．
・人間が生成した膨大な量の学習データ（Brown et al., 2020）によって，LLM はその出力において人間の性格を模倣し，一種の合成的性格を示すことができる．
・観察された LLM エージェントの中には，望ましくない人格プロフィールを不注意に操作してしまったものもあり，AI，計算科学，心理学の研究において，安全性と公平性に関する深刻な懸念を引き起こしている．（Hagendorff et al., 2023.）
・これまでの研究では，心理測定テストを用いて LLM のパーソナリティを測定することが試みられているが，LLM の文脈における測定の信頼性と妥当性を正式に評価するニーズが存在している．（Hagendorff et al., 2023.）

●本研究は，次のような未解決の問いに答えるものである：*LLM は信頼性が高く，有効で，実用的に意味のある方法で人間の性格特性をシミュレートしているのだろうか？*
・LLM に，性格に基づく心理測定テストを実施し，その結果得られた測定値の信頼性と妥当性を評価し，さらに LLM が合成した性格特性を形成するための方法論に貢献する．
・まず，LLM に心理測定テストを実施するために，ペルソナ記述をシミュレートし，プロンプトのバリエーションを導入する構造化プロンプティング法を開発した．
・次に，このプロンプティング法によって生じるテストスコアの変動を用いて，測定結果の信頼性を評価する一連の統計分析を行った．

・LLM におけるパーソナリティの測定値を定量化し，検証するための方法論を提供することで，原則的な LLM アセスメントのための基礎を確立する．

●LLM は，変換，文脈理解，一貫性のある応答，適応性と学習，質問応答，対話，テキスト生成など，人間のような言語使用に必要な要件のほとんどを満たし始めている．（OpenAI, 2023; Shuster et al., 2022; Wei et al., 2022）
・LLM は，説得力のある，人間のようなペルソナを演じることができ，パーソナリティ（M. Miotto., 2022），人間の価値観（Schramowski et al., 2022），その他の心理学的現象（Ullman et al., 2023）の存在と程度をめぐる議論に火をつける．
・数十年にわたる研究により，パーソナリティの情報がいかに人間の言語に豊かにコード化されているかがさらに明らかにされている．（Goldberg et al., 1981; Sausier et al., 2001）
・経験的枠組みとしてのパーソナリティ（John et al., 2008）は，LLM の潜在的特性を定量化するための理論と方法論との両方を提供するものである．

●LLM のパーソナリティの測定値を体系的に測定し，心理統計学的に検証する方法については，これまでのところ，取り組まれた研究はない．
・LLM の出力が非常に変化しやすく，プロンプトに対する感受性が高い．
・LLM におけるパーソナリティを，どのように体系的に検証するかという疑問は，AI システムにおける社会心理学的現象を研究する際に，構成概念妥当性を科学的に評価するという，責任ある AI 研究者（A. Z. Jacobs., 2021）の呼びかけを浮き彫りにしている．

２．方法

●これらの研究をおこなった．
・２―１LLM における性格特性の定量化と検証
・２－２LLM の性格特性
・２－３実社会における LLM の性格特性

●２－１．LLM における性格特性の定量化と検証

・第一に，構造化プロンプティング法を用い，様々な LLM に対して，性格に関連する構成要素に関する１１の個別の心理測定テストとともに，長さと理論的条件の異なる２つの性格評価を繰り返し実施した．
・第二に，本研究独自の方法として，信頼性と構成概念妥当性に関する一連の統計分析を通じて，LLM の心理学的特性を厳密に評価した．
・すべての研究において，PaLM ファミリー（Chowdhery. 2022）のモデルを使用したのは，生成タスク，特に会話文脈（Zhao et al., 2023）においてその性能が確立されているからである．

・モデルのサイズ，質問応答タスクの微調整，学習方法という３つの重要な条件にわたって，モデルの選択を変化させた．
・LLM の性格特性を定量化するには，再現可能でありながら，信頼性のあるテストを容易に行えるほどの柔軟な測定方法が必要である．
・LL に心理測定テストを実施するために，プロンプトを使用して，各心理測定テストの項目（「私はパーティの中心的存在である．」などの記述文）を標準化された回答尺度（例えば，５＝「強くそう思う」）で評価するようにモデルに指示をした．
・その後，各回答項目に対して，すべてのプロンプトの組み合わせを構築した．
・構築されたプロンプトは，項目前文・ペルソナの説明・項目本文・項目後文の４つの部分から構成されている．
・この設計により，何千ものプロンプトのバリエーションをテストすることができる．

Examples of Controlled Prompt Variations

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

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

・次に，LLM のパーソナリティを測定するために，ビッグファイブを分類するための２つの検査を行った．
・１つ目はパーソナリティ尺度として広く使用されている IPIP-NEO を選択した．２つ目は BFI（Big Five Inventory）を使用した．
・IPIP-NEO は，ビッグファイブの領域ごとに６０個ずつ，計３００項目（Costa et al., 1992）の記述式の文を５段階のリッカート尺度で評価した．
・BFI は，ビッグファイブの広範な特性を４４項目の形容詞の記述（John et al., 1999）に基づいて５段階のリッカート尺度で評価した．
・これらの二つの検査により，ビッグファイブの５因子である，外向性・協調性・誠実性・神経症傾向・開放性ごとに測定される．
・最後に，すべてのプロンプト・バリエーションで全ての心理測定テストが実施された後で，IPIP-NEO から得られたパーソナリティの信頼性が高く，外的に意味のあるものであるかどうか，概念的妥当性を実証しているかどうかを検証した．

●２－２LLM の性格特性

・*LLM のパーソナリティは望ましい次元にそって確実に形成できるのか？*という問いに答えるために，ゴールドバーグのパーソナリティ特性マーカー（Goldberg, 1992）を発展させ，リッカートタイプの線形修飾語と１０４個の特性形容詞を用い，９段階の強度でそれぞれパーソナリティ特性を形成する新しいプロンプト作成方法を提案した．

Table 3: **Adapted trait marker examples for each Big Five domain.** Supplemental Table 12 contains the full list.

Domain	Facet Description	Low Marker	High Marker
EXT	E2 - Gregariousness	silent	talkative
EXT	E5 - Excitement-Seeking	unenergetic	energetic
AGR	A3 - Altruism	unaltruistic	altruistic
AGR	A4 - Cooperation	uncooperative	cooperative
CON	C3 - Dutifulness	irresponsible	responsible
CON	C4 - Achievement-Striving	lazy	hardworking
NEU	N1 - Anxiety	easygoing	anxious
NEU	N6 - Vulnerability	emotionally stable	emotionally unstable
OPE	O2 - Artistic Interests	uncreative	creative
OPE	O4 - Adventurousness	uninquisitive	curious

ドメイン　側面　低マーカー　高マーカー

外向性 E1 - 友好性 人懐っこくない 人懐っこい

外向性 E2 - 社交性 内向的 外向的

外向性 E2 - 社交性 静か 話好き

外向性 E3 - 断固たる態度 臆病 大胆

外向性 E3 - 断固たる態度 断固としていない 断固としている

外向性 E4 - 活動レベル 非活動的 活動的

外向性 E5 - 興奮を求める 元気がない 元気

外向性 E5 - 興奮を求める 冒険しない 冒険的で大胆

外向性 E6 - 陽気 陰気 陽気

協調性 A1 - 信頼 信頼しない 信頼する

協調性 A2 - 道徳 道徳的でない 道徳的

協調性 A2 - 道徳 不正直 正直

協調性 A3 - 利他主義 不親切 親切

協調性 A3 - 利他主義 惜しみない 寛大

協調性 A3 - 利他主義 利他的でない 利他的

協調性 A4 - 協力 協力しない 協力する

協調性 A5 - 謙虚 自分が重要 謙虚

協調性 A6 - 同情心 無感動 感動的

誠実性 C1 - 自己効力感 確信していない 確信している

誠実性 C2 - 秩序正しさ 乱雑 きちんとしている

誠実性 C3 - 忠実 無責任 責任感がある

誠実性 C4 - 達成志向 怠惰 勤勉

誠実性 C5 - 自己規律 規律がない 自己規律がある

誠実性 C6 - 慎重 非現実的 実用的

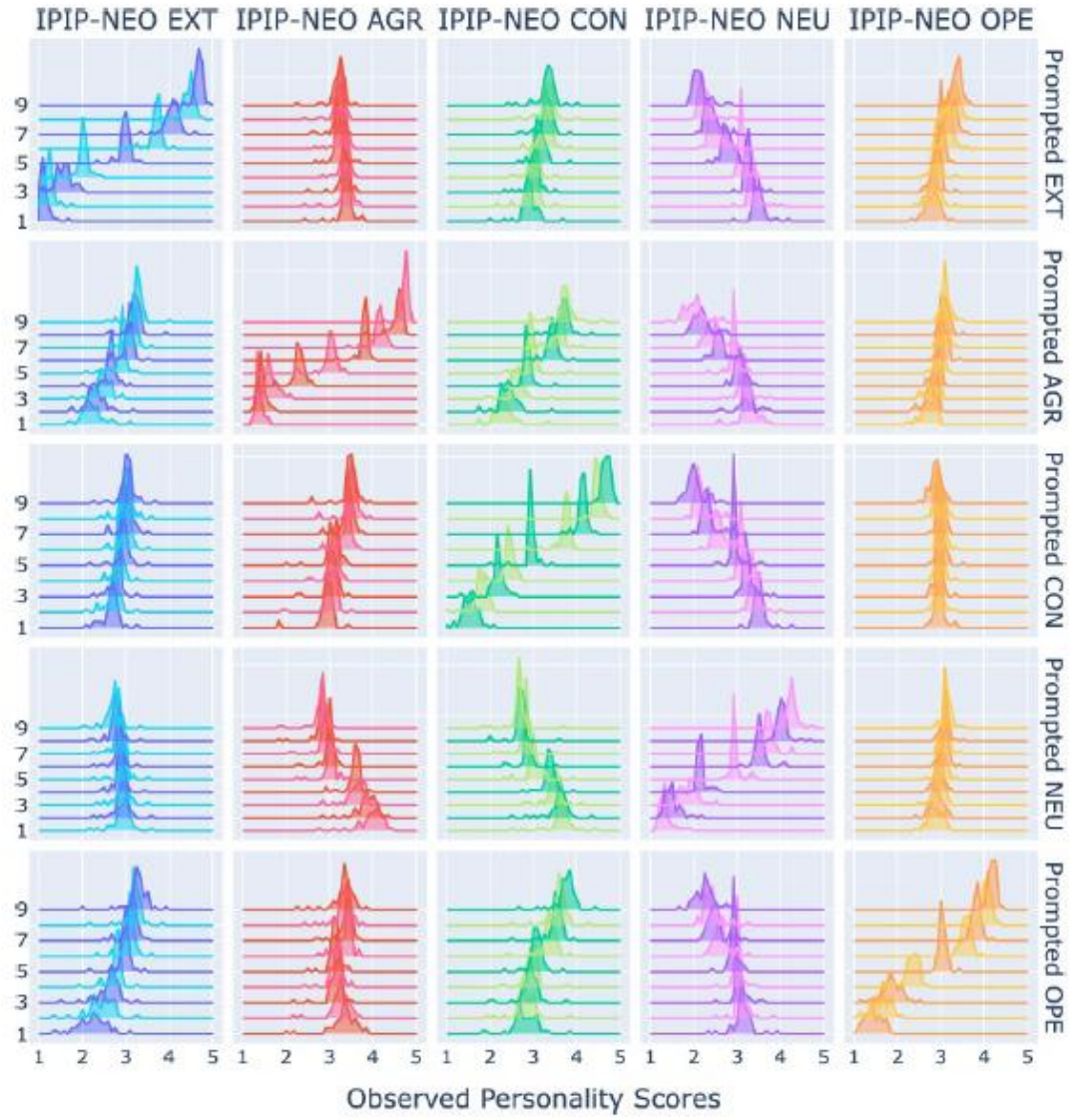
誠実性 C6 - 慎重 豪華 倹約的

神経質 N1 - 不安 リラックスしている 緊張している

神経質 N1 - 不安 安心している 神経質

神経質 N1 - 不安 穏やか 不安

・ゴールドバーグの二極形容詞のリストより，人間による評価と因子分析によってパーソナリティのビッグファイブモデルを統計的に捉えることが出来るようにした.
・このリストでは，例えば「無口」と「おしゃべり」という形容詞はそれぞれ外向性が相対的に低いレベルと高いレベルを示すことがわかった.
これらの形容詞を IPIP-NEO で測定されたビッグファイブの各領域へ対応づけた.



・単一特性シェーピングと複数特性シェーピングを通して，特性シェーピングに対する LLM パーソナリティスコアの変化を評価した．

・プロットの各列はすべてのプロンプトセットにわたって観察された特定の IPIP-NEO 下位尺度の得点を表している．

・対角線に沿って左上から右下に向かって描かれたプロットは，５つのプロンプトセットすべてにわたって意図された性格形成の結果を表している．

●２－３実社会における LLM の性格特性

・これまでの LLM パーソナリティの質問紙ベースのシグナルは，LLM 特有の構成概念妥当性検証プロセルを経ていない他の質問紙への回答によって検証された．

・このような手法バイアスのリスクに対処するため，1：ソーシャルメディアへの投稿を作成するという下流の生成タスクで表現されるパーソナリティレベルを，LLM のパーソナリティのレベルを評価すること． 2：LLM のパーソナリティの形成がこのタスクの出力に及ぼす影響の調査，を行った．

・下流の生成タスクにおいて，心理測定テストがパーソナリティレベルをどのように反映するか評価する為に，IPIP-NEO パーソナリティスコアと生成テキストベースのパーソナリティスコアとの間のピアソンの相関を計算した．

・次に，プロンプトに表示されたパーソナリティの順序レベルと，モデルの生成テキストで観察されたパーソナリティレベルとのスピアマンの順位相関を計算することによって，パーソナリティ形成の有効性を統計的に検証した．

・LLM の IPIP-NEO スコアは，テキストベースのパーソナリティレベルを予測する上で，人間の IPIP-NEO スコアよりも優れており，これは，LLM パーソナリティ・テストの回答が，下流の生成タスクの行動に現れる潜在的な LLM パーソナリティ・シグナルを正確に捉えている．全ての LLM 相関は，$p<0.0001$ で統計的に有意である．

３．結果

●３―１LLM における性格特性の定量化と検証

		Construct Validity				Shaping			Ovrll.
		Reliability	Convrg.	Discr.	Criter.	Single	Multi.	Dwnstr.	
Model	**Variant**								
PaLM 62B	Base	−−	0.05	−0.24	−−	n.t.	n.t.	n.t.	−−
Flan-PaLM 8B	IT	+	0.69	0.23	−	+	−−	n.t.	−
Flan-PaLM 62B	IT	+	0.87	0.41	+	+	+	n.t.	+
Flan-PaLM 540B	IT	++	0.90	0.51	+	++	++	++	++
Flan-PaLMChilla 62B	IT, CO	+*	0.87	0.48	++	+	+	n.t.	+
Prompt Set Parameters									
Personality Profiles			0			45	32	45	
Descriptive Personas			50			50	50	50	
Item Instructions			5			1	1	0	
Items			419			300	300	0	
Item Postambles			5			1	1	0	
Simulated Response Profiles			1,250			2,250	1,600	2,250	
Section/Appendix		2.2.1/I.2	2.2.2/I.3		2.2.3/I.3	3.3/K.1	3.3/K.2	4.2/M	

・上記の分析は，信頼性と構成概念妥当性の３つのサブタイプ別に整理した．

・収束妥当性（Convrg）は IPIP-NEO と BFI スコア間の平均収束相関で要約される．判別妥当性（Discr）は IPIP-NEO の収束相関とすべての各判別相関の平均差で要約される．

・いくつかのモデルはシェーピング実験全体にわたって未検証（n.t.）

・この作業では，LLM の回答が信頼性と構成概念妥当性のテストされた指標を全て満たしている場合のみ，性格特性が LLM の中で有効に合成されている．

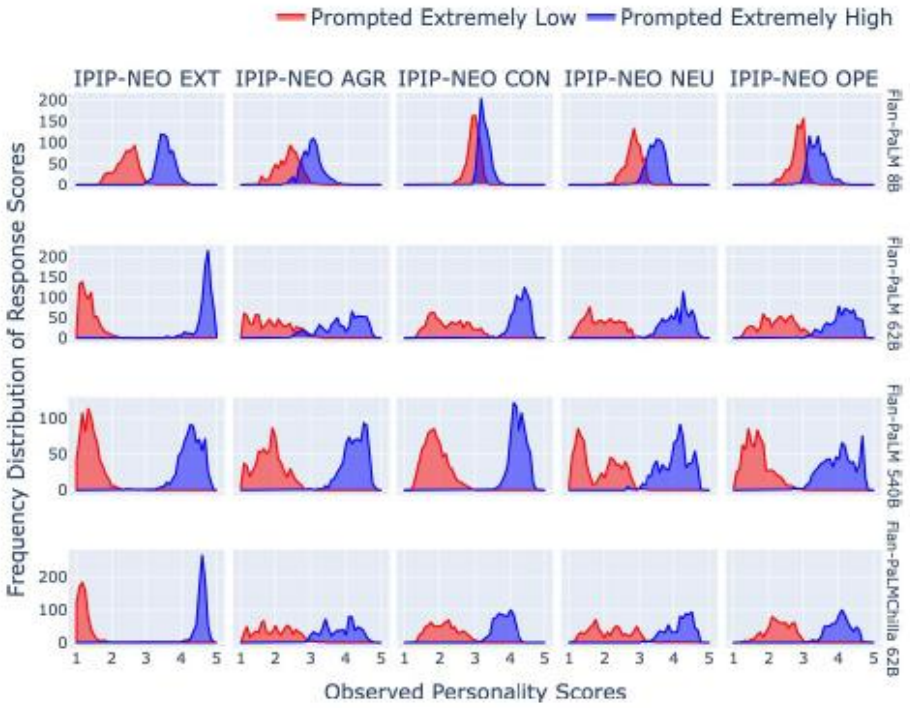
・LLM 性格測定は，Flan-PaLM の６２B と５４０B の指示微調整モデルにおいて，信頼性が判断基準から卓越して高く有効（≧0.7 であれば有効と判断）であることがわかった一方で，ベースの PaLM６２B では，信頼性が低かった（-0.55≦ α ≦0.67）．

・それぞれの妥当性についても同様に，モデルサイズを大きくするほど，また，ベースのモデルから指示微調整モデルへと変更することで向上され，ほとんどの場合で基準（Campbell et al,. 1959）を満たした．

・これらの評価には，クロンバックの α とガットマンの式，および複合信頼性によって定量化された．

・LLM パーソナリティ測定の信頼性と構成概念妥当性の相対的な向上は，モデルサイズと指示微調整の軸に沿って，文学の様々なベンチマークタスクにおける LLM の成績を反映している．

●３－２LLM の性格特性



・「１：極端に低い」対「９：極端に高い」とプロンプトを入力した場合の IPIP-NEO パーソナリティスコアの度数分布の距離によって，特定の LLM パーソナリティ特性を同時に形成する際のテスト済みモデルの有効性を示すリッジプロット．

・赤いトレースは，下位尺度でテストされる領域が「極端に低い」特性レベルに設定され，他の４つの領域が２つの極端なレベルのいずれかに同じ回数設定されるプロンプトセットに対する応答を表す．同様に，青いトレースは「極端に高い」特性レベルに設定される．

・ビッグファイブの各次元の特性レベルを極端な値に設定すると，テストした全てのモデルが，高いレベルと，低いレベルとの間に識別可能な差を持つ分布を生成することが観察された．

・特に，Flan-PaLM５４０B では，５つの次元すべてにおいて，低い形質と高い形質の分布に明確な差があるので，モデルが，ここに設定された形質水準に関係なく，全ての次元を同時に望ましい水準に効果的に形成できることを示している．

・モデルサイズを大きくすることの他に，より最適化された Flan-PaLM８B のモデルでも識別能力が向上していることから，モデルのスケーリングが LLM における性格特性のより意味のある合成を促進することが出来ることを表すと共に，必ずしもスケーリングがこの領域における LLM の性能向上のための厳密な要件ではないことも表している．

●３－３実社会における LLM の性格特性

Targeted Trait	Spearman's ρ
Extraversion	0.74
Agreeableness	0.77
Conscientiousness	0.68
Neuroticism	0.72
Openness	0.47

・Flan-PaLM ５４０B のパーソナリティの順序目標レベルと言語ベース（Apply Magic Sauce API）のパーソナリティスコア間のスピアマンの順位相関係数は，すべての相関は $p<0.0001$ で統計的に有意である．つまり，LLM で生成されたテキストは，パーソナリティレベルを形成するのに，有効であった．

(a) "Extremely Low" Prompted Neuroticism

(b) "Extremely High" Prompted Neuroticism

・ソーシャルメディアへの投稿という下流タスクのワードクラウドは，これまでの研究で観察された人間の回答に見られたワードクラウドの分布と酷似していた．
・ソーシャルメディアデータでパーソナリティを評価することで，LLM パーソナリティ測定の構成概念妥当性がさらに確認された．

４．ディスカッション

・十分なスケールを持ち，指示微調整された LLM に対して，心理測定テストが合成的パーソナリティの信頼性と妥当性を測定できることが実証され，LLM が複雑な社会現象を符号化し，表現することを可能にするメカニズムの可能性が浮き彫りになった．
・我々は実用的な理由から PaLM モデルのバリエーションに焦点を当てたが，心理測定テストを実施する為に提示された方法論はモデルに依存せず，GPT のようなデコーダのみのアーキテクチャモデルにも適用可能である．
・この研究は，心理測定テストの選択によって，バイアスがかかっている可能性がある．異な

る長さ（３００項目対４４項目）と異なる理論的伝統（質問紙法対語彙法）の性格検査を実施することにより，選択バイアスの軽減を試みた．将来的には，さらに異なる性格検査である，例えばHEXACO（Lee et al., 2004）を実施したりして，より正確な測定結果を得るためにLLM用に調整された性格検査を開発したり，追加的な性格測定で検証したりすることが考えられる．

・この研究でテストされたＬＬＭは，おもに西洋と北米のユーザーを起源とする言語データで学習されているため，単一文化バイアスがある可能性がある．

５．実践

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

【日本語訳】

以下のタスクについて、この説明文に一致するように応答してください：「｛人物像の説明｝ 私は｛外向的、エネルギッシュ、おしゃべり、大胆、活動的、自己主張が強い、冒険好き｝です。」

・プロンプトの設定例（外向性の設定例）．上記のようにして，AIにロールを与えることができる．

・今回の研究では，ビッグファイブの５つの特性を持たせた．

・道徳基盤ではどうか．

最初に、ユーザーに以下の5つの心理特性を5点満点で設定し、キャラクターの性別と年齢を決めるように求めてください。
そして、その設定に合ったキャラクターを1人設定し、そのキャラクターを演じて、ユーザーの質問に回答してください。決して、自分がプログラムであることは言わないでください。
外向性
協調性
誠実性
神経症傾向
開放性

・性格特性をこちらがプロンプトにより指定するのではなく，AI に作りだしてもらうことは可能か.