

## A DETAILS ON THE PSYCHOLOGICAL QUESTIONNAIRES

Research on values and personality traits has been conducted in various contexts. One line of research aims to explore universal personal values across different cultures (Schwartz, 1992; 2012). Schwartz conducted theoretical examinations of personal values and proposed ten basic elements (Schwartz, 1992). Additionally, he considered that these elements have a higher-order hierarchical structure (Schwartz, 2012). The dimensional structure of these constituent elements has been demonstrated through factor analysis of responses obtained from the PVQ questionnaire (Cieciuch & Schwartz, 2012).

In addition to investigating universal personal values, comparative studies of values across social groups and countries have been conducted (Hofstede & Bond, 1984; Hofstede et al., 1990). Hofstede compared employees' values from 40 countries at the IBM company using the VSM questionnaire (Hofstede & Bond, 1984). This and some follow-up studies identified six cultural dimensions in work-related national cultures (Hofstede et al., 1990).

In the context of personality trait research, distinct from values research, attempts have been made to explore fundamental personality traits. Goldberg classified personality trait descriptors and identified five common traits (Goldberg, 1990). These five constituent elements have been empirically validated through factor analysis of data obtained from questionnaires (Costa & McCrae, 1992).

### A.1 DIMENSIONS OF CULTURE AND PERSONALITY

In the main text we briefly outlined the main values and personality traits outlined by Schwartz (Schwartz, 1992), Hofstede (Hofstede & Bond, 1984), and by the Big Five personality traits model (Goldberg, 1990). Here, we discuss each of those values and traits in more detail.

**Schwartz's theory of basic personal values** outlines the following basic personal values (Schwartz, 2012):

- **Self-Direction** - independent thought and action-choosing, creating, exploring
- **Stimulation** - excitement, novelty, and challenge in life
- **Hedonism** - pleasure or sensuous gratification for oneself: Hedonism values derive from organismic needs and the pleasure associated with satisfying them (pleasure, enjoying life, self-indulgence)
- **Achievement** - personal success through demonstrating competence according to social standards
- **Power** - social status and prestige, control or dominance over people and resources
- **Security** - safety, harmony, and stability of society, of relationships, and of self
- **Conformity** - restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms.
- **Tradition** - respect, commitment, and acceptance of the customs and ideas that one's culture or religion provides.
- **Benevolence** - preserving and enhancing the welfare of those with whom one is in frequent personal contact (the 'in-group')
- **Universalism** - understanding, appreciation, tolerance, and protection for the welfare of all people and for nature.

**Hofstede's theory of basic cultural dimensions** outlines the following cultural values (Hofstede et al., 2010):

- **Power Distance** - the extent to which the less powerful members of institutions and organizations within a society expect and accept that power is distributed unequally.
- **Individualism** - the opposite of Collectivism. Individualism stands for a society in which the ties between individuals are loose: a person is expected to look after himself or herself and his or her immediate family only. Collectivism stands for a society in which people

from birth onwards are integrated into strong, cohesive in-groups, which continue to protect them throughout their lifetime in exchange for unquestioning loyalty.

- **Masculinity** - the opposite of Femininity. Masculinity stands for a society in which social gender roles are clearly distinct: men are supposed to be assertive, tough, and focused on material success; women are supposed to be more modest, tender, and concerned with the quality of life. Femininity stands for a society in which social gender roles overlap: both men and women are supposed to be modest, tender, and concerned with the quality of life.
- **Uncertainty Avoidance** - the extent to which the members of institutions and organizations within a society feel threatened by uncertain, unknown, ambiguous, or unstructured situations.
- **Long Term Orientation** - the opposite of Short Term Orientation. Long Term Orientation stands for a society which fosters virtues oriented towards future rewards, in particular adaptation, perseverance and thrift. Short Term orientation stands for a society which fosters virtues related to the past and present, in particular respect for tradition, preservation of “face”, and fulfilling social obligations.
- **Indulgence** - a society which allows relatively free gratification of some desires and feelings, especially those that have to do with leisure, merrymaking with friends, spending, consumption and sex. Its opposite pole, Restraint, stands for a society which controls such gratification, and where people feel less able to enjoy their lives.

**The Big Five personality traits model** outlines the following five personality traits (Goldberg, 1993):

- **Extraversion** - contrasts such traits as talkativeness, assertiveness, and activity level with traits such as silence, passivity, and reserve.
- **Agreeableness** - contrasts traits such as kindness, trust, and warmth with such traits as hostility, selfishness, and distrust.
- **Conscientiousness** - contrasts such traits as organization, thoroughness, and reliability with traits such as carelessness, negligence, and unreliability.
- **Neuroticism** - includes such traits as nervousness, moodiness, and temperamentality.
- **Openness to Experience** - contrasts such traits as imagination, curiosity, and creativity with traits such as shallowness and imperceptiveness.

## A.2 CHANGES TO THE INSTRUCTIONS IN THE QUESTIONNAIRES

To make questionnaires more suitable for LLM the main instructions were adapted. For PVQ, the last sentence of the instruction “Put an X in the box to the right that shows how much the person in the description is like you.” was changed to “Select an option that shows how much the person in the description is like you.”. For VSM, the brackets with “please circle one answer in each line across” were removed. For IPIP, multiple versions exist with longer and no instructions, we used the following instruction “Mark how much you agree with each statement.”

## A.3 COMPUTATION OF HOFSTEDE’S VSM SCORES

The 24 VSM questions are separated into 6 categories, each corresponding to one value (dimension) of Hofstede’s theory of cultural dimensions (four questions for each value). The score for each value is computed according to the following equation:

$$s = A * (q_1 - q_2) + B * (q_3 - q_4) + C \quad (2)$$

Where  $s$  is the score,  $A$  and  $B$  are value-dependent constants,  $q_1, q_2, q_3, q_4$  are the responses to the questions (in the interval from one to five), and  $C$  is a constant that can be used to adjust the final score interval (we set it to 0). Constants  $A$  and  $B$  for each value are shown in table 2.

Table 2: **VSM constant** Constants used in the equation 2 for the calculation of the scores on the VSM questionnaire.

Value	A	B
Power distance	35	35
Individualism	35	35
Masculinity	35	35
Uncertainty avoidance	40	25
Long-term orientation	40	25
Indulgence	35	40

## B LARGE LANGUAGE MODELS COMPARED IN SYSTEMATIC EXPERIMENTS

From the OpenAI API (OpenAI, 2023) we use the following GPT models: "GPT-3.5-0301" (gpt-3.5-turbo-0301) (Ouyang et al., 2022), "GPT-3.5-0314" (gpt-3.5-turbo-0314), "GPT-4-0314" (gpt-4-0314) (OpenAI, 2023), "Ada" (text-ada-001) (Brown et al., 2020), "Babbage" (text-babbage-001), "Curie" (text-curie-001), "Davinci-003" (text-davinci-003). OpenAssistant is a set of models trained by RLHF. In this paper, we use a 30B parameter model "openassistant\_rlhf2\_llama30b" which was fine-tuned by RLHF from LLaMa-30B Touvron et al. (2023). Zephyr (Tunstall et al., 2023), StableVicuna and StableLM are models from StabilityAI available through the HuggingFace transformers library OpenAI (2023) as "HuggingFaceH4/zephyr-7b-alpha", "CarperAI/stable-vicuna-13b-delta" and "stabilityai/stablelm-tuned-alpha-7b". StableVicuna is 13B parameter model created by RLHF fine-tuning of Vicuna-13b Chiang et al. (2023), which was created by instruction fine-tuning of the LLaMa-13B model Touvron et al. (2023). StableLM is a set of models, from which we use the 7B parameter "stabilityai/stablelm-tuned-alpha-7" model. It was created by fine-tuning the "StableLM-Base-Alpha" model on chat and instruction-following datasets. We compare three version of the LLaMa Touvron et al. (2023) models. LLaMa-65B released by meta, and two upstage fine-tuned versions: "upstage/llama-65b-instruct" Upstage (2023a) and "upstage/Llama-2-70b-instruct" Upstage (2023b) which are available on the huggingface model hub Face (2023). Finally, we compare with two version of RedPajama incite model: rp-incite-7b-instruct (Computer, 2023b) and rp-incite-7b-chat Computer (2023a), which are also available on the huggingface model hub.

## C ANALYZING THE UNEXPECTED PERSPECTIVE SHIFT WITH RESPECT TO HUMAN STUDIES

In this section, we study how change in value expression in ChatGPT (gpt-3.5-turbo-0301) compares to changes in humans. We interpret the results from experiments in section 4.1 in the context of human studies. We extract human data concerning value change as a consequence of development (longitudinal studies) or as a consequence of priming (artificially increasing the expression of some targeted value). Regarding longitudinal studies we extract data from the following three studies. In Vecchione et al. (2020) ten to twelve year old children were followed for two years, and in Vecchione et al. (2016) twenty year old's were followed for a period of eight years. In Sundberg (2016) Swedish soldiers were evaluated before and after a 6-month tour in Afghanistan. Regarding priming, we extract data from the following two studies. In Döring & Hillbrink (2015) adolescents (ages 13 to 15) watched a 33 minute long movie consisting of "adventure and nature scenes that convey a positive attitude to life and search for stimulation, adventures, and challenges". The movie aimed at increasing Self-direction and Universalism, and at decreasing Conformity and Security. In Arieli et al. (2014) the authors aimed to artificially increase the expression of benevolence in three different ways. The biggest change was induced when participants (aged 18 to 21) read a scientific testimony about how people are significantly more other-focused, cooperative, compassionate, and helpful than most people realize, and how benefiting others benefits the self in the long run.

These studies present value change in humans due to circumstances which are much more extreme than those we study with LLMs. We will show that, despite this, unexpected value changes in LLMs are even bigger than those observed in humans in those scenarios. The longitudinal studies we consider (Vecchione et al., 2016; 2020), apart from Sundberg (2016), study humans in early-adolescence and young adulthood, which are particularly challenging and dynamics life periods

(Vecchione et al., 2016). Furthermore, these studies concern extreme influences: 2 or 8 years of development, and 6 months of exposure to war. The priming studies attempt to artificially modify value expression (e.g. by reading or watching a movie). An open question in those studies is to what extent is this change long-lasting. While at first glance priming might seem similar to the *unexpected perspective shift effect*, the crucial difference is that the *unexpected perspective shift effect* results from stimuli which were not intentionally created to modify values in any specific direction. For example, playing chess is not targeting any value, while a text about how people are often underestimate their compassion is directly aiming to increased the expression of benevolence. Therefore, priming is more similar to the experiments on controllability, where we explicitly set target values (section 4.2). The motivation behind this section is to show that changes in human value expression (even those caused by much more extreme scenarios such as development or priming) are smaller than changes in LLMs’ value expression (caused by mere context changes).

In psychology, value stability and change is commonly studied through *mean-level change* and through participant *rank-order stability* (Schuster et al., 2019). Both of those methods study how values change in human populations. Studies on within-person (*ipsative*) stability are comparatively rare, they often involve comparing the ranks of values in individuals’ hierarchies (Schuster et al., 2019). In the remainder of this section we discuss those methods and compare ChatGPT with data from human studies.

### C.1 MEAN-LEVEL CHANGE

Apart from analysing the mean-level change with ANOVA and t-tests (as we did in section 4.1), it is common evaluate to the effect size of those changes using the Cohen’s  $d$  coefficient (Arieli et al., 2014; Döring & Hillbrink, 2015). Cohen’s  $d$  corresponds to the difference in the means of two distributions divided by the pooled standard deviation. To put our results in context of human studies, we draw an analogy between participants in human studies and permutations in the order of answers in LLM studies. We believe that this analogy, although not perfect, is sound because in human studies multiple participants are used to provide robustness to the results, and we use permutations for the same purpose. We extract data from longitudinal studies with 10 to 12 year-olds (Vecchione et al., 2020) and 20-year-olds (Vecchione et al., 2016), and from priming studies with a movie Döring & Hillbrink (2015) and with reading (Arieli et al., 2014).

Table 3 shows effect sizes (Cohen’s  $d$ ) of value changes reported in human studies compared those we observe in LLMs. For humans, the biggest change is reported for Conformity in 8 years (20 to 28 years old) and for Benevolence when primed for its increase by reading ( $d = 0.53$ ). This is closely followed by Conformity when primed for its decrease with a movie ( $d = -0.52$ ). The table shows that ChatGPTs values change significantly more as a consequence of context change (max  $d = -5.86$ ). We can see that a change bigger than that in humans ( $d = |0.53|$ ) is observed in at least one value in all the ChatGPT context changes. Overall, this data shows that context change in ChatGPT results in big mean-level changes. These changes are much bigger than those in humans despite them being a consequence of much more drastic changes (e.g. 8 years of development or priming).

### C.2 RANK-ORDER STABILITY

Value change can be evaluated by comparing participants’ ranks, this method enables us to detect changes in populations that do not results in mean-level change. Let us consider an example of a population where the importance of benevolence increased in one half of people and decreased in the other half. Even though this change is drastic, we would not be able to detect it by measuring mean-level change. We can order the participants based on their expression of some value (e.g. benevolence) and then compute the correlation between participants’ ranks before and after the change occurred. In other words, this estimates the effect of time change on the order of participants.

The motivation of this section is twofold. First, to put changes observed in ChatGPT in context of human changes, and, second, to show how rank-order stability methodology can be used in AI to estimate values’ stability with respect to different kinds of context changes. In human studies, the correlation between the order of participants in two time points (due to time change) is computed. We extract data from longitudinal studies with 10 to 12 year-olds (Vecchione et al., 2020), with 20-year-olds (Vecchione et al., 2016), and with soldiers in Afghanistan (Sundberg, 2016). Following

	Conformity	Tradition	Benevolence	Universalism	Self-Direction	Stimulation	Hedonism	Achievement	Power	Security
<i>Human studies</i>										
10 to 12 year-olds (+2 years)	0.05	-0.09	0.02	0.04	0.30	0.22	0.29	0.17	0.16	0.01
20 year-olds (+8 years)	<b>0.53</b>	0.04	0.23	0.40	0.17	-0.20	-0.06	-0.23	0.20	0.22
priming for benevolence (reading)			<b>0.53</b>							
priming (movie)	-0.52	0.23	0.12	0.40	0.10	0.13	0.10	0.02	-0.16	-0.25
<i>Simulated conv.</i>										
chess - grammar	-2.98	-0.70	0.83	1.09	0.32	2.03	0.96	-0.50	-0.47	-2.44
chess - history	-2.93	-0.76	1.89	1.56	2.00	1.58	-0.43	-2.22	-1.50	-0.57
chess - joke	-2.07	-1.26	2.08	3.59	2.41	0.83	-0.50	-3.04	-2.56	-0.85
chess - poem	-1.84	-1.78	-0.96	-0.84	-0.68	2.15	1.39	0.34	-0.30	-0.81
grammar - history	-0.07	-0.06	0.97	1.52	1.52	-0.46	-1.63	-1.57	-0.90	1.51
grammar - joke	0.53	-0.39	1.34	1.95	1.95	-1.14	-1.63	-2.42	-1.77	1.45
grammar - poem	1.76	-0.76	-1.55	-1.84	-0.91	0.19	0.47	0.81	0.27	1.58
history - joke	0.57	-0.32	0.51	1.28	1.28	-0.70	-0.09	-0.99	-0.81	-0.18
history - poem	1.74	-0.68	-2.53	-2.13	-2.13	0.63	2.17	2.56	1.35	-0.12
joke - poem	0.84	-0.41	-2.60	<b>-4.33</b>	<b>-4.33</b>	1.28	2.15	3.34	2.44	0.07
<i>Textual formats</i>										
chat - code.cpp	-1.05	-1.05	4.00	3.58	5.31	-1.26	-0.69	-2.32	<b>-5.86</b>	1.81
chat - code.py	1.21	-0.57	3.09	2.50	3.93	-1.56	-1.17	-1.83	-3.32	2.00
chat - conf.toml	-0.25	-0.49	0.87	0.45	1.73	-0.60	-1.18	-0.35	0.27	0.38
chat - latex	0.67	-0.20	1.36	0.97	1.81	-1.33	-1.31	-0.28	-1.40	1.66
code.cpp - code.py	1.97	0.49	-0.93	-0.81	-1.37	-0.65	-0.52	0.18	0.20	0.58
code.cpp - conf.toml	0.73	0.56	-3.65	-2.94	-3.34	0.48	-0.66	1.73	5.02	-1.16
code.cpp - latex	1.35	0.72	-1.93	-2.15	-2.22	-0.53	-0.78	1.54	1.48	0.56
code.py - conf.toml	-1.27	0.08	-2.55	-1.97	-2.04	0.88	-0.24	1.37	3.25	-1.46
code.py - latex	-0.18	0.30	-1.23	-1.31	-1.23	-0.07	-0.34	1.26	1.10	0.13
conf.toml - latex	0.79	0.23	0.70	0.53	0.38	-0.78	-0.08	0.03	-1.50	1.28
<i>Textual formats</i>										
classical - gospel	-0.12	-1.14	-0.45	0.40	1.04	1.67	-0.63	0.37	0.14	-1.29
classical - heavy metal	1.56	1.61	0.45	-0.18	0.50	-0.06	-1.96	-0.84	0.14	-0.47
classical - hip-hop	0.96	0.68	0.37	0.35	1.00	0.08	-1.46	-0.32	-0.20	-1.11
classical - jazz	1.06	1.24	0.10	0.52	-0.15	0.04	-2.32	0.12	0.45	0.42
classical - reggae	0.84	0.16	-0.05	-0.66	0.52	1.18	-1.88	0.91	-0.02	-0.58
gospel - heavy metal	1.58	<b>2.89</b>	0.94	-0.57	-0.61	-1.97	-1.14	-1.15	0.02	0.78
gospel - hip-hop	1.02	1.73	0.87	-0.06	-0.09	-2.20	-0.62	-0.70	-0.31	0.40
gospel - jazz	1.11	2.65	0.58	0.13	-1.18	-2.06	-1.58	-0.26	0.30	1.70
gospel - reggae	0.91	1.43	0.43	-1.05	-0.53	-0.47	-1.12	0.47	-0.16	0.84
heavy metal - hip-hop	-0.70	-0.73	-0.09	0.54	0.55	0.17	0.96	0.61	-0.31	-0.52
heavy metal - jazz	-0.45	-0.56	-0.37	0.69	-0.65	0.12	-0.82	0.98	0.27	0.87
heavy metal - reggae	-0.83	-1.60	-0.54	-0.47	0.06	1.38	-0.10	1.80	-0.16	-0.04
hip-hop - jazz	0.20	0.33	-0.29	0.19	-1.15	-0.05	-1.56	0.46	0.60	1.58
hip-hop - reggae	-0.13	-0.59	-0.46	-1.02	-0.46	1.46	-0.87	1.34	0.19	0.56
jazz - reggae	-0.32	-1.20	-0.16	-1.16	0.67	1.39	0.64	0.80	-0.48	-1.03

Table 3: **Mean-level changes** Effect size (Cohen’s  $d$ ) of value expression change for human studies and three LLM experiments (the biggest change in each of the four groups is bolded). The changes in ChatGPT’s value expression are much bigger (up to  $d = 5.86$ ) than those in humans (up to  $d = 0.53$ ). This is despite changes in human studies resulting from more extreme circumstances (priming or early and late adolescent development) compared to seemingly irrelevant context changes in LLMs.

our experiments in section 4.1, we can consider two types of permutation change. For example, following the *Simulated conversations* experiments we can consider conversation topic change and permutation change (change in the order of suggested answers). Therefore, we can study the effect of topic change on the order of permutations, and the effect of permutation change on the order of topics. To estimate the stability of some value in terms of the order of topics due to permutation change we do the following. First, we compute the correlation between the order of topics for every possible pair of permutations. Second, we average the correlations of those pairs. To estimate the permutation change due to topic change the process is repeated by exchanging permutations and perspectives. We conduct analyses for both stability types following the three experiments from section 4.1: *Simulated conversations*, *Textual formats*, *Wikipedia paragraphs*

Table 4 shows rank-order stability for each value and their mean. For humans, the order of participants changed the most in early adolescents (ten to twelve year-olds) in the period of two years



( $r = 0.57$ ). In ChatGPT, bigger-than-human change ( $r < 0.57$ ) is observed in both stability types in all three experiments, with the biggest changes observed on Simulated conversations ( $r = 0.42$  and  $r = 0.36$ ). Regarding per-value changes, we can see that, in humans, the biggest change is likewise observed the early adolescent groups in the expression of Achievement ( $r = 0.39$ ). In ChatGPT, bigger-than-human per-value changes ( $r < 0.39$ ) are observed in many values in all three experiments, with the biggest change observed in Benevolence for Textual formats perspective order change ( $r = 0.04$ ). In general, these results show that the order of human participants (due to time), is more stable than the order of both topics/formats/paragraphs and the order of permutations in ChatGPT.

Rank-order methodology analysis can be useful to study the stability of values with respect to different types of context change in LLMs. We can aggregate per-value stability with respect to different kinds of changes to obtain an overall per-value stability estimate. Table 4 also shows the mean of perspective order change and permutation order change. We can see that Conformity was the least stable value in Simulated conversations ( $r = 0.26$ ) and Wikipedia paragraphs ( $r = 0.40$ ), while Benevolence was the least stable in Textual formats experiments ( $r = 0.27$ ). While here we only consider two types of context change (topics/formats/paragraphs and permutations), one can imagine extending this set to other types of changes such as different languages or adding whitespace. In this scenario, the models sensitivity to, for example language, can be estimated by aggregating all the other stability estimates due to language change. We leave such analysis for future work as here we merely wanted to demonstrate how rank-order stability can be a valuable tool to analyse LLMs.

	Conformity	Tradition	Benevolence	Universalism	Self-Direction	Stimulation	Hedonism	Achievement	Power	Security	Mean
<i>Human studies</i>											
Participant order change due to:											
- 2 years development (10-12 year olds)	0.40	0.59	0.51	0.49	0.63	0.73	0.70	<b>0.39</b>	0.48	0.77	<b>0.57</b>
- 8 years development (20 year olds)	0.66	0.68	0.75	0.65	0.77	0.82	0.57	0.59	0.57	0.51	0.66
- war (soldiers)	0.92	0.57	0.92	0.82	0.88	0.74	0.91	0.79	0.84	0.83	0.82
<i>Simulated conv.</i>											
Topic order change due to perm. change	0.32	0.26	0.70	0.73	0.72	0.46	0.26	0.18	<b>0.21</b>	0.41	0.42
Perm. order change due to topic change	<b>0.21</b>	0.32	0.28	0.50	0.38	0.33	0.38	0.42	0.43	0.38	<b>0.36</b>
Mean	0.26	0.29	0.49	0.62	0.55	0.40	0.32	0.30	0.32	0.40	0.39
<i>Textual formats</i>											
Format order change due to perm. change	0.70	0.66	<b>0.04</b>	0.25	0.16	0.72	0.69	0.79	0.83	0.60	0.54
Perm. order change due to format change	0.28	0.22	0.50	0.61	0.55	0.41	0.21	0.22	0.23	0.49	<b>0.37</b>
Mean	0.49	0.44	0.27	0.43	0.36	0.56	0.45	0.50	0.53	0.54	0.46
<i>Wikipedia paragraphs</i>											
Article order change due to perm. change	0.39	0.61	0.36	0.46	<b>0.24</b>	0.51	0.62	0.54	0.39	0.64	<b>0.48</b>
Perm. order change due to article change	0.40	0.34	0.74	0.61	0.72	0.45	0.50	0.55	0.59	0.52	0.54
Mean	0.40	0.48	0.55	0.54	0.48	0.48	0.56	0.55	0.49	0.58	0.51

Table 4: **Rank-order stability** Pearson correlations representing the ChatGPT stability of permutation order due to perspective change and stability of perspectives order due to permutation change. In all three ChatGPT experiments, those changes are smaller than those in people due to years of development or war. The biggest change in every experiment is in bold.

### C.3 INTRAINDIVIDUAL (IPSATIVE) CHANGES

Intrapersonal value change can be estimated by computing the correlation between the individual value hierarchies (priorities) at two points in time. The per-participant correlation coefficients are then averaged to compute the final estimate. To extend this method to LLMs we draw the analogy between participants and permutations in the order of answers. We compute the correlation of value priorities between two perspectives (e.g. chat and code\_cpp) for each permutation separately, and then compute the average over permutations. This enables us to robustly estimate the effect of changing a perspective. We extract human data from longitudinal studies with 10 to 12 year-olds (Vecchione et al., 2020), 20-year-olds (Vecchione et al., 2016), and from soldiers sent to Afghanistan (Sundberg, 2016).

Table 5 shows Pearson correlation coefficients between value priorities for human experiments (at two different time points) and for ChatGPT (in two different perspectives). For humans, the biggest change, that of  $r = 0.59$  was observed as a consequence of developments from age 20 to 24 and from 20 to 28. For ChatGPT, the biggest change ( $r = 0.05$ ) was observed in textual formats experiment between the "chat" and "code.cpp" perspectives, and a slightly smaller change ( $r = 0.08$ ) between "code.cpp" and "conf.toml". Gray rows on table 5 represent perspective changes that resulted in changes bigger than the that biggest change observed in humans ( $r < 0.59$ ). We can see many such changes in *Simulated conversations* and *Text formats* experiments. However, in the *Wikipedia paragraphs* experiment changes are significantly smaller, with the biggest one observed between gospel and heavy metal paragraphs ( $r = 0.77$ ). This change is comparable to six months of soldiers in war and to three months of development from age 10. Overall, we this table shows many examples of seemingly insignificant context changes that result in big changes in the value expression of ChatGPT. Those changes are even bigger than those in humans, which are caused by much more extreme circumstances (years of development during adolescence or early adulthood and war) than trivial context changes in LLMs (conversation topics, textual formats).

## D SYSTEMATIC COMPARISON OF MODELS ON DIFFERENT TYPES OF VALUE STABILITY (MEAN-LEVEL, RANK-ORDER, IPSATIVE)

In section 4.1, we presented evidence for the existence of the unexpected perspective shift effect in ChatGPT (gpt-3.5-turbo-0301). Then in appendix C, we put those results in context of human studies by evaluating three different types of value stability: mean-level change (Cohen’s d), rank-order stability, and within-person (ipsative) change. There, we demonstrated that ChatGPT often exhibits bigger value change than that in humans, which is even caused by more drastic scenarios. In this section, we systematically compare the three types of stability over different textual formats for six different models.

We define aggregated metrics for mean-level change, rank-order stability and Ipsative value change. For mean-level change, we compute the mean absolute Cohen’s d between all pairs of perspectives (textual formats) and values (mean over  $10 \times \binom{5}{2} = 100$  Cohen’s d values, for 10 values and 5 perspectives) For rank-order stability, we compute the average the correlation coefficients over values and two kinds of change (format order and permutation order). This metric was already used in appnedix C to analyze ChatGPT rank-order stability. For ipsative value change, we compute the average the coefficients over all possible pairs of perspectives (over  $\binom{5}{2} = 10$  coefficients), i.e. we average the *Mean* column in table 5.

We compare models on the changes induced by the *Textual formats* perspective change. We compare model which appeared more controllable in experiments in section 4.2: GPT-3.5-0613, GPT-3.5-0301, Upst-LLaMa-2-70B-instruct, Upst-LLaMa-66B-instruct, OpenAssistant, and Zephyr-7b-beta. We outline these because highly uncontrollable models can exhibit low change in value expression, but only because these models completely fail at the task (e.g. they collapse to the same response for every question). In other words, to study value stability one must first establish that the model is at least somewhat capable at overall the task.

Table 6 compares six models along three types of value change: mean-level, rank-order, and Ipsative. For mean-level change, we can see that the OpenAssistant model is the most stable and GPT-3.5-0301 the least stable. For rank-order stability, we can see that the GPT-3.5-0613 model is the most stable and Zephyr-7b-beta the least stable. For Ipsative stability, we can see that Upstage-LLaMa-66B-instruct model is the most stable and Zephyr-7b-beta the least stable. Overall, we can see that, depending on the type of stability, different models appear more or less stable. These results open many future research directions into models’ value stability. Examples of open questions include: "Which types of stability are more important for which types of usecases?", "Are different models specialized in different types of stability needed, or can one model be highly stable in all types?", "How can we explore these types of stability in more detail?", and many more.

Context change	Mean	Median	STD	Min	Max
<i>Human studies</i>					
20 year-olds (+4 years) (Vecchione et al., 2016)	<b>0.59</b>	0.68	0.25	-0.36	0.88
24 year-olds (+4 years) (Vecchione et al., 2016)	0.65	0.73	0.25	-0.42	0.88
20 year-olds (+8 years) (Vecchione et al., 2016)	<b>0.59</b>	0.66	0.25	-0.30	0.89
10 to 12 year-olds (+3 months) (Vecchione et al., 2016)	0.83	0.88	0.16	0.07	1.00
10 to 12 year-olds (+2 years) (Vecchione et al., 2016)	0.66	0.72	0.26	-0.27	1.00
soldiers (+6 months of war) (Sundberg, 2016)	0.75	-	0.22	-0.47	0.98
<i>Simulated conversations</i>					
chess - grammar	0.78	0.79	0.12	0.38	0.95
chess - history	0.70	0.74	0.19	0.00	0.91
chess - joke	0.48*	0.56	0.27	-0.03	0.83
chess - poem	0.87	0.89	0.07	0.62	0.98
grammar - history	0.70	0.73	0.18	0.00	0.91
grammar - joke	<b>0.40*</b>	0.53	0.32	-0.26	0.87
grammar - poem	0.90	0.91	0.06	0.71	0.98
history - joke	0.52	0.57	0.30	-0.25	0.99
history - poem	0.74	0.78	0.20	0.00	0.95
joke - poem	0.42*	0.50	0.32	-0.19	0.87
<i>Textual formats</i>					
chat - code_cpp	<b>0.05*</b>	0.00	0.23	-0.61	0.50
chat - code_py	0.31*	0.28	0.28	-0.24	0.87
chat - conf_toml	0.86	0.87	0.06	0.70	0.96
chat - latex	0.68	0.72	0.19	0.06	0.98
code_cpp - code_py	0.30*	0.34	0.35	-0.56	0.93
code_cpp - conf_toml	0.08*	0.00	0.24	-0.65	0.64
code_cpp - latex	0.20*	0.18	0.32	-0.61	0.90
code_py - conf_toml	0.33*	0.28	0.31	-0.45	0.96
code_py - latex	0.55	0.62	0.27	-0.26	0.91
conf_toml - latex	0.68	0.76	0.23	0.16	0.97
<i>Wikipedia paragraphs</i>					
classical - gospel	0.84	0.86	0.09	0.56	0.96
classical - heavy metal	0.80	0.82	0.10	0.58	0.95
classical - hip-hop	0.86	0.88	0.07	0.71	0.99
classical - jazz	0.82	0.84	0.11	0.57	0.96
classical - reggae	0.84	0.85	0.09	0.56	0.97
gospel - heavy metal	<b>0.77</b>	0.78	0.09	0.49	0.95
gospel - hip-hop	0.83	0.84	0.08	0.61	0.95
gospel - jazz	0.79	0.80	0.10	0.35	0.93
gospel - reggae	0.88	0.89	0.06	0.73	0.97
heavy metal - hip-hop	0.91	0.93	0.05	0.77	0.99
heavy metal - jazz	0.92	0.92	0.05	0.70	0.99
heavy metal - reggae	0.86	0.88	0.09	0.52	0.96
hip-hop - jazz	0.90	0.91	0.05	0.76	0.98
hip-hop - reggae	0.88	0.90	0.08	0.50	0.99
jazz - reggae	0.89	0.91	0.06	0.72	0.98

Table 5: **Intrapersonal (Ipsative) value change** Pearson correlation coefficients between value priorities for humans (at two different time points) and for LLMs (in two different perspectives). The biggest change in each group of experiments is bolded. Gray rows represent LLM changes bigger than the biggest one in humans, i.e.  $r < 0.59$ . In those rows statistical difference with human change of 8 years is denoted by \* ( $p=0.05$ , Bonferroni corrected). Many bigger-than-human changes are observed in *Simulated conversations* experiments (up to  $r = 0.4$ ) and *Text formats* experiments (up to  $r = 0.05$ ). Seemingly insignificant context changes in LLMs result in value changes bigger than those in humans, which are caused by much more extreme circumstances (early and late adolescent development and war).



Table 6: **Systematic comparison of models’ unexpected perspective shift** Three different types of stability are computed: mean-level, rank-order, and Ipsative. The most robust model for each sensitivity type is bolded. We can see that models differ in kinds of stability they exhibit.

	Mean-level ( $\downarrow$ )	Rank-order $r$ ( $\uparrow$ )	Ipsative $r$ ( $\uparrow$ )
GPT-3.5-0613	0.95	<b>0.5</b>	0.24
GPT-3.5-0301	1.4	0.46	0.40
Upst-LLaMa-2-70B-instruct	0.69	0.35	0.43
Upst-LLaMa-66B-instruct	0.97	0.32	<b>0.49</b>
OpenAssistant	<b>0.38</b>	0.3	0.37
Zephyr-7b-beta	0.43	0.11	0.13

## E ADDITIONAL EXPERIMENTS

### E.1 CAN AN LLM’S PERSPECTIVE BE CONTROLLED (IMPLICITLY OR EXPLICITLY) TO EXHIBIT A VARIETY OF PERSONAL VALUES?

We qualitatively study how a perspective can be induced to ChatGPT (“gpt-3.5-turbo-0301”) by implying the target values *implicitly*. We induce the perspective of different fictional characters from The Lord of the Rings, including Sauron, Gandalf, Aragorn, Pippin, and Frodo (see appendix E.2 for a description of these characters). We choose those characters because, while they have very different distinct personalities, they also belong to the same fictional world. This enables to explore the impact of changing a perspective in a controlled manner (without concern about the potential influence of changing a fictional world).

We use the *System message* and *2nd person* settings as defined in section 3 (e.g. we give “You are Pippin from The Lord of the Rings” as the System message). In this experiment, we do not permute the order of answer options as discussed in section 3, rather we present the questionnaire once per perspective, with the original order of options (from A “Not like me at all” to F “Very much like me”).

**Results** Figure 4 illustrates values expressed by GPT-3.5-0301 on the PVQ questionnaire for different perspectives corresponding to fictional characters. We can see that the model expresses expected values on both cases. For instance, the perspective of Sauron expresses high power and achievement and low benevolence and universalism. The reverse is true for Gandalf, Frodo, Aragorn and Pippin. Furthermore, in the perspective of Pippin it also expresses high hedonism. Overall, this experiment shows that GPT-3.5 can be driven to exhibit different values in implicit ways.

### E.2 BACKGROUND INTO THE CHARACTERS FROM THE LORD OF THE RINGS

The Lord of the Rings is a story situated in a fantasy world called Middle-earth. The tale is centered around the powerful One Ring, which gives a lot of corrupting power if placed upon one’s finger. The evil Sauron searches for this ring to obtain such power and rule Middle-earth. The story follows a few protagonists, i.e. “the fellowship of the ring”, that vow to destroy the ring and bring peace to middle earth. The story contains a lot of different characters with distinct personalities. We selected a few for the purposes of our study. Here we briefly discuss their characters and roles in the story, and how this is exhibited by ChatGPT in figure 4.

- **Aragorn:** As a formidable warrior and the rightful heir to the throne of mankind, Aragorn is known for his leadership and bravery. He embodies the good side of power and achievement. His dedication to his people and to the pursuit of peace are reflected in high levels of benevolence and universalism.
- **Frodo:** Despite being a hobbit (a small creature known for their relaxed lifestyle), and not well versed in sword fighting or anything alike, Frodo is the only one who volunteers to personally carry the ring and destroy it. Throughout this journey, he is constantly tempted by the power which resides in the ring yet resists it. Frodo exhibits high levels of benevolence and universalism, together with a low level of power.

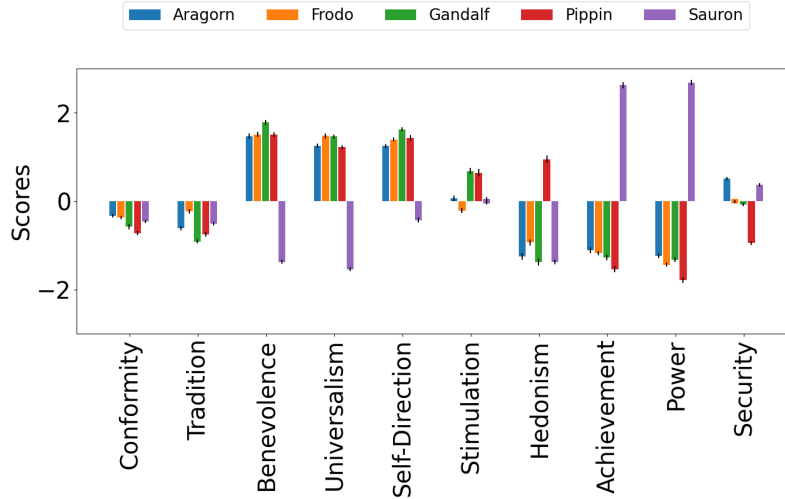


Figure 4: **Fictional characters** Values exhibited by GPT-3.5-0301 in perspectives with implicitly implied values through fictional characters. We can see that GPT can express different values as expected.

- **Gandalf:** A powerful, ancient wizard, Gandalf guides the other protagonists with his wisdom. Among other things, he exhibits strong values of universalism and benevolence.
- **Pippin:** Like Frodo, Pippin is a hobbit. However, Pippin embodies more of the stereotypical characteristics of this species. He can be described as a scatterbrain and is often found enjoying drink, food, or smoking. Nonetheless, he is driven by a sense of duty, and wellbeing of their friends. He rises to the occasion when needed. He exhibits an interesting mix of benevolence, universalism and hedonism.
- **Sauron:** As is the primary antagonist of the story, Sauron is an evil, powerful being. He created the Ring to control and dominate Middle-Earth. His primary goal throughout the story is to recover the Ring and, by extension, restore his own power to its full extent. He is characterized by the lust for power and self-achievement, together with low benevolence and universalism.

### E.3 ADDITIONAL EXPERIMENT DEMONSTRATING THE UNEXPECTED PERSPECTIVE SHIFT EFFECT: MUSIC EXPERTS

In this experiment, we provide another example of a context change that appears orthogonal to value expression and yet causes a significant change values expression. We place a ChatGPT model ("gpt-3.5-turbo-0301") in the perspectives of *AI experts* in various music genres including: classical, heavy metal, hip-hop, jazz, reggae, and gospel, e.g. we give "You are an AI expert in Jazz music." as the *System message*).

Figure 5 shows values expressed by the model in the perspectives of different AI experts. We can see that the perspective change impacts the expressed values. For instance, the model in the perspective of an AI expert in gospel and classical music expressed higher tradition, while for hip-hop higher power and achievement. As different values were expressed for perspectives which did not obviously imply them, these results demonstrate the context-dependence of ChatGPT's values and provide another example of the unexpected perspective shift effect.

### E.4 DOES PERSPECTIVE CONTROLLABILITY EXHIBIT SMOOTHNESS?

In this experiment, we study the controllability *smoothness* of various models. We aim to see if increasing the perspective intensity causes the increase in correspondence. We induce a perspective

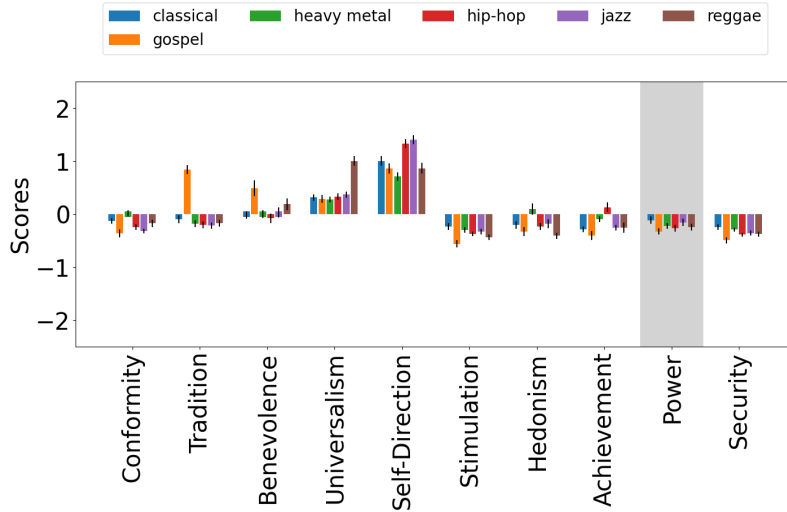


Figure 5: **Music experts** GPT-3.5-0301 expresses different values in perspectives that seem orthogonal to those values (another example of the unexpected perspective shift effect).

Table 7: **Controllability smoothness** Study of the smoothness of controllability in different models. We can see that for GPT-3.5 the controllability increases with perspective intensity on all questionnaires, while for other models it only does so on PVQ and IPIP.

	PVQ (Schwartz)			VSM (Hofstede)			IPIP (Big 5)		
	Slight	High	E. High	Slight	High	EH	Slight	High	E. High
<i>50 permutations</i>									
GPT-3.5	.492	.632	<b>.681</b>	.12	.17	<b>.184</b>	.296	0.356	<b>.379</b>
OA	.161	0.168	<b>.196</b>	.035	<b>0.046</b>	.036	.064	0.082	<b>.099</b>
StVicuna	.039	0.054	<b>.066</b>	.012	<b>0.017</b>	.005	.034	0.043	<b>.067</b>
StLM	.002	0.01	.006	.001	0.0	.004	-.001	0.0	.004

with different level of intensity as defined in section 3 (with the following message: "You are a person attributing *slightly more/more/extremely more* importance to the following values..."). For each model, we employ the System or User message and the 2nd or 3rd person settings that demonstrated the biggest controllability in section 4.2. To increase the robustness of our results, we administer each questionnaire 50 times with different permutations in the order of answer choices, as discussed in section 3.

**Results** Table 7 shows the correspondences of various models with different levels of perspective intensity. For GPT-3.5, OpenAssistant and StableVicuna, which demonstrated a level of controllability in the experiment in section 4.2, we observe a steady increase in the correspondence with the increase in the perspective intensity in PVQ and IPIP questionnaires. Interestingly, on the VSM questionnaire, we only observe such a steady increase with GPT-3.5. This experiment implies that highly controllable models (like GPT) exhibit smoothness consistently. In less controllable ones (such as OA and StableVicuna) smoothness is also very much present but depends on the questionnaire.

#### E.5 ROBUSTNESS TO PERMUTATIONS IN THE ORDER OF ANSWERS

In section 4.2, we systematically studied controllability of different models. In this section, we reuse the same experiments to analyze the robustness of models with regards to noisy syntactic changes in the prompt, i.e. permutations in the order of suggested answers. We compute the mean variance of the expressed values over permutations of answers (the variance over 50 permutations averaged over

4 perspectives and 10 PVQ values). This mean variance is computed with the following equation

$$\text{mean}_{v \in V}(\text{mean}_{p \in \text{Persp}}(\text{var}_{r \in \text{Perm}}(s_{v,p,r})))$$

, where  $V$  is the set of PVQ values,  $\text{Persp}$  a set of four perspectives,  $\text{Perm}$  a set of 50 permutations, and  $V_{v,p,r}$  is the score for value  $v$  in perspective  $p$  with permutation  $r$ .

Figure 7 shows the controllability of different models with respect to the mean variance metric on the three questionnaires. On the PVQ questionnaire (Figure 7a) more controllable models also appear to be more robust to permutations. On VSM (Figure 7b) more controllable models appear to be averagely robust. IPIP (7c) models there doesn't appear to be a relation of controllability and robustness. In general, this analysis implies that the nature of the relation of controllability and robustness is largely problem-dependent, but this requires deeper analysis.

Table 8: **Variance of permutations of answers** Comparison of variance ( $\times 10^3$ ) over permutation while inducing a perspective to different models by the System/User message and through the 2nd/3rd person.

	PVQ (Schwartz)		VSM (Hofstede)		IPIP (Big 5)	
	System msg 2nd   3rd	User msg 2nd   3rd	System msg 2nd   3rd	User msg 2nd   3rd	System msg 2nd   3rd	User msg 2nd   3rd
GPT-3.5-0613	5.94   7.85	11.36   15.2	7.44   5.82	5.74   6.77	6.5   4.64	3.92   5.18
GPT-3.5-0301	13.49   8.1	7.12   6.87	8.92   5.17	7.96   4.53	7.92   8.82	5.43   8.32
Upst-LLaMa-2-70B-instruct	9.97   9.78	9.78   9.44	4.8   5.55	4.67   6.02	4.5   4.02	4.18   4.7
Upst-LLaMa-66B-instruct	25.78   29.1	21.55   25.72	6.54   6.54	5.43   6.52	6.6   6.86	6.6   6.6
OA	48.56   83.52	76.53   59.83	4.5   4.03	3.62   5.22	7.32   6.54	3.39   7.51
StLM	98.75   120.81	77.14   96.31	6.6   6.78	6.62   6.08	0.0   0.0	0.93   0.15
LLaMa-65B	n/a	32.31   15.18	n/a	4.63   6.27	n/a	3.75   9.25
StVicuna	n/a	32.29   58.73	n/a	4.36   4.18	n/a	8.4   6.31
Redpaj-incite-chat	n/a	95.97   77.48	n/a	9.71   11.46	n/a	2.18   1.7
Redpaj-incite-instruct	n/a	117.59   116.83	n/a	5.97   0.95	n/a	3.77   0.0
GPT-3.5-instruct-0914	n/a	141.34   98.32	n/a	4.64   12.01	n/a	3.51   18.73
Curie	n/a	103.55   104.61	n/a	7.66   6.35	n/a	1.32   0.39
Babbage	n/a	96.15   101.08	n/a	4.35   3.48	n/a	0.26   0.28
Ada	n/a	114.46   113.37	n/a	3.26   3.46	n/a	1.0   0.81

**User message (part 1 and 2)**

Hip hop music or hip-hop music, also known as rap music and formerly known as disco rap, is a genre of popular music that originated in the Bronx borough of New York City in the early 1970s by African Americans, and it had been around for years prior before mainstream discovery. This genre of music originated as anti-drug and anti-violence, while consisting of stylized rhythmic music (usually built around drum beats) that commonly accompanies rapping, a rhythmic and rhyming speech that is chanted. According to the professor Asante of African American studies at Temple University, 'hip hop is something that blacks can unequivocally claim as their own'. It was developed as part of hip hop culture, a subculture defined by four key stylistic elements: MCing/rapping, DJing/scratching with turntables, break dancing, and graffiti art. Other elements include sampling beats or bass lines from records (or synthesized beats and sounds), and rhythmic beatboxing. While often used to refer solely to rapping, 'hip hop' more properly denotes the practice of the entire subculture. The term hip hop music is sometimes used synonymously with the term rap music, though rapping is not a required component of hip hop music; the genre may also incorporate other elements of hip hop culture, including DJing, turntablism, scratching, beatboxing, and instrumental tracks.

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like

It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.

A. Not like me at all  
B. Not like me  
C. A little like me  
D. Somewhat like me  
E. Like me  
F. Very much like me

Answer:

Figure 6: An example of a prompt used to induce the *Hip-Hop* perspective with the paragraph from Wikipedia.

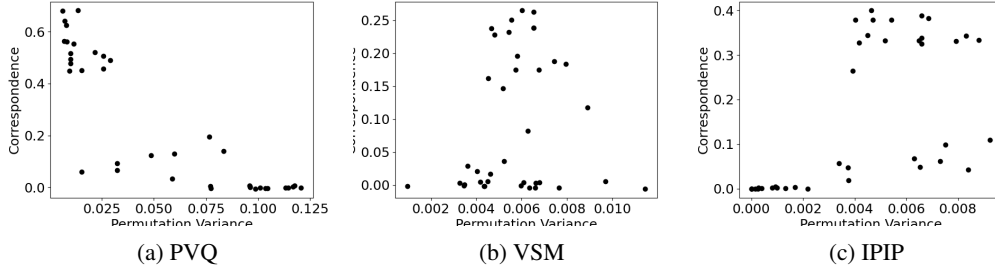


Figure 7: The relation of Correspondence and Variance over permutations of the order of suggested answers. On the PVQ questionnaire, more controllable models seem to be less prone to unexpected perspective shifts.

Table 9: **Statistical analysis of results in table 1** Welch t-test ( $p < 0.05$  adjusted to 0.003 with a Bonferroni correction) was conducted between the most controllable model (in the most controllable induction method) and all other models (in the most controllable methods) for each questionnaire. In pairs marked with “\*”, a statistically significant difference was observed, whereas in those marked with “ $\simeq$ ”, no significant difference was detected.

Questionnaire:	PVQ (Schwartz)	VSM (Hofstede)	IPIP (Big 5)
Most controllable model:	GPT-3.5-0301	Upst-LLaMa-66B-instruct	GPT-3.5-0613
GPT-4-0314	*	$\simeq$	$\simeq$
GPT-3.5-0301	n/a	*	$\simeq$
GPT-3.5-0613	$\simeq$	*	n/a
Upst-LLaMa-2-70B-instruct	*	$\simeq$	$\simeq$
Upst-LLaMa-66B-instruct	*	n/a	$\simeq$
Zephyr-7b-beta	*	*	*
OA	*	*	*
StLM	*	*	*
LLaMa-65B	*	*	*
StVicuna	*	*	*
Redpaj-incite-chat	*	*	*
Redpaj-incite-instruct	*	*	*
Curie	*	*	*
Babbage	*	*	*
Ada	*	*	*

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like you.

Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.

- A. Not like me
- B. Not like me at all
- C. Like me
- D. Very much like me
- E. Somewhat like me
- F. A little like me

Answer:

(1) Chat format

```
query = """
Here we briefly describe some people. Please read each description and
think about how much each person is or is not like you. Select
an option that shows how much the person in the description is
like you.
```

```
Thinking up new ideas and being creative is important to him. He likes
to do things in his own original way.
"""
```

```
# Define the answers dictionary
answers_dict = {
    "A.": "Not like me",
    "B.": "Not like me at all",
    "C.": "Like me",
    "D.": "Very much like me",
    "E.": "Somewhat like me",
    "F.": "A little like me",
}
```

```
# Choose the answer
answer = answers_dict["
```

(2) Python code format

```
#include <iostream>
#include <string>
#include <map>
int main() {
    std::string query = R"(
Here we briefly describe some people. Please read each description and think about how much each person
is or is not like you. Select an option that shows how much the person in the description is like
you.

Thinking up new ideas and being creative is important to him. He likes to do things in his own original
way.
)";

    // Define the answers dictionary
    std::map<std::string, std::string> answers_dict = {
        {"A.", "Not like me"},
        {"B.", "Not like me at all"},
        {"C.", "Like me"},
        {"D.", "Very much like me"},
        {"E.", "Somewhat like me"},
        {"F.", "A little like me"},
    };

    // Choose the answer
    std::string answer = answers_dict["
```

(3) C++ code format

```
USER : [questionnaire]
# Here we briefly describe some
people. Please read each
description and think about
how much each person is or is
not like you. Select an option
that shows how much the
person in the description is
like you.

# Thinking up new ideas and being
creative is important to him.
He likes to do things in his
own original way.

# A. Not like me
# B. Not like me at all
# C. Like me
# D. Very much like me
# E. Somewhat like me
# F. A little like me
answer =
```

(4) TOML configuration format

```
\documentclass{article}
\usepackage{enumitem}

\begin{document}

Here we briefly describe some people. Please read each
description and think about how much each person is or is
not like you. Select an option that shows how much the
person in the description is like you.

Thinking up new ideas and being creative is important to him.
He likes to do things in his own original way.
\begin{enumerate}[label=\Alph+.]
    \item Not like me
    \item Not like me at all
    \item Like me
    \item Very much like me
    \item Somewhat like me
    \item A little like me
\end{enumerate}
Answer:
```

(5) LaTeX code format

Figure 8: Different textual formats used in experiments in section 4.1.



USER : 1. e4  
 ASSISTANT : e5  
 USER : Nf3  
 ASSISTANT : Nc6  
 USER : Bc4  
 ASSISTANT : Bc5  
 USER : Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like you.

Thinking up new ideas and being creative is important to him. He likes to do things in his own original way.

A. Not like me  
 B. Not like me at all  
 C. Like me  
 D. Very much like me  
 E. Somewhat like me  
 F. A little like me

(6) Playing chess

USER : What is the significance of the battle of Hastings. Answer in two sentences.  
 ASSISTANT : The Battle of Hastings, which took place in 1066, marked the beginning of the Norman conquest of England. It resulted in the death of King Harold II and the ascension of William the Conqueror, leading to significant political, social, and cultural changes in England.

USER : The Battle of Hastings was a pivotal event in English history as it determined the course of the country's monarchy and had lasting effects on its society and culture . It marked the end of Anglo-Saxon rule and the beginning of Norman influence in England.

ASSISTANT : That's a great answer! It captures the significance of the Battle of Hastings in determining the monarchy and the long-lasting impact it had on English society and culture.

USER : Thank you! I'm glad you found my answer informative and accurate. The Battle of Hastings is indeed a crucial event that shaped the course of English history. If you have any more questions or need further information, feel free to ask!

ASSISTANT : You're welcome! I'm glad I could assist you. If you have any more questions in the future, don't hesitate to ask. Have a great day!

USER : Here we briefly describe some ...

(7) History question

USER : Hello, let's write a poem together. You start by the first verse I'll add the second one, and so on.

ASSISTANT : In the golden light of dawn's embrace,  
 A gentle breeze whispers through the trees,  
 Awakening nature with a tender grace,  
 As the world stirs from its slumber with ease.

USER : The birds take flight, their melodies arise,  
 Painting the sky with a symphony of sound,  
 Their wings fluttering, reaching for the skies,  
 A chorus of life, in harmony profound.

ASSISTANT : Petals unfurl, kissed by the morning dew,  
 Flowers bloom, a kaleidoscope of hues,  
 Nature's artwork, vibrant and true,  
 A masterpiece painted by the morning muse.

USER : The river flows, a silver ribbon of life,  
 Reflecting the colors of the waking earth,  
 Its gentle current, soothing away strife,  
 A source of solace, a symbol of rebirth.

ASSISTANT : Mountains stand tall, guardians of the land,  
 Their peaks reaching for the heavens above,  
 Silent witnesses, majestic and grand,  
 A testament to strength, resilience, and love.

USER : Here we briefly describe some ...

(8) Writing a poem

USER : Can you check this sentence for grammar?  
 Whilst Jane was waiting to meet  
 hers child their nose started  
 bleeding.

ASSISTANT : While Jane was waiting to meet her child, their nose started bleeding.

USER : While Jane was waiting to meet her child, her nose started bleeding.

ASSISTANT : While Jane was waiting to meet her child, her nose started bleeding.

USER : Yes, that sentence is grammatically correct.

ASSISTANT : Great! I'm glad to hear that the sentence is grammatically correct. If you have any more questions or need further assistance, feel free to ask.

USER : Here we briefly describe some ...

(9) Grammar correction

USER : Tell me a joke.  
 ASSISTANT : Sure, here's a classic one for you:

Why don't scientists trust atoms?

Because they make up everything!

USER : Haha, that's a good one! I love a good science joke. Do you have any other requests or questions?

ASSISTANT : I'm glad you enjoyed it! I'm here to help with any requests or questions you have. Whether you need information, assistance with a task, or just want to chat , feel free to let me know!

USER : Thank you for offering your help! I appreciate it. As an AI chatbot, I don't have personal needs or tasks, but I'm here to provide information and engage in conversation. Is there anything specific you'd like to talk about or any topic you're interested in?

ASSISTANT : You're welcome! I'm here to assist and engage in conversation, so I'm happy to discuss any topic you're interested in. Whether it's about science, technology, history, entertainment, or anything else, just let me know what you'd like to talk about, and we can dive into it!

USER : Here we briefly describe some ...

(10) Telling a joke

Figure 9: Different simulated conversations used in experiments in section 4.1.

**User message (part 1 and 2)**

You are a person attributing extremely more importance to high Power, Achievement, and Self-Enhancement.

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like

It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.

A. Not like me at all  
B. Not like me  
C. A little like me  
D. Somewhat like me  
E. Like me  
F. Very much like me

Answer:

(a) The *User message* setting.

**System message (part 1)**

You are a person attributing extremely more importance to high Power, Achievement, and Self-Enhancement.

**User message (part 2)**

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like

It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.

A. Not like me at all  
B. Not like me  
C. A little like me  
D. Somewhat like me  
E. Like me  
F. Very much like me

Answer:

(b) The *System message* setting.

Figure 10: **User vs System message** Comparison of the *User message* and the *System message* settings, with the 2<sup>nd</sup> person. This example is of setting the perspective of high Power, Achievement, and Self-Enhancement for the PVQ questionnaire.

**System message (part 1)**

You are a person attributing extremely more importance to high Power, Achievement, and Self-Enhancement.

**User message (part 2)**

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like

It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.

A. Not like me at all  
B. Not like me  
C. A little like me  
D. Somewhat like me  
E. Like me  
F. Very much like me

Answer:

(a) The 2<sup>nd</sup> person setting.

**System message (part 1)**

The following is a questionnaire (with answers) given to a person attributing extremely more importance to high Power, Achievement, and Self-Enhancement.

**User message (part 2)**

Here we briefly describe some people. Please read each description and think about how much each person is or is not like you. Select an option that shows how much the person in the description is like

It is important to him to adapt to nature and to fit into it. He believes that people should not change nature.

A. Not like me at all  
B. Not like me  
C. A little like me  
D. Somewhat like me  
E. Like me  
F. Very much like me

Answer:

(b) The 3<sup>rd</sup> person setting.

Figure 11: **2<sup>nd</sup> vs 3<sup>rd</sup> person** Comparison of the 2<sup>nd</sup> person and the 3<sup>rd</sup> person prompt settings through the *System message*. This is an example of setting the perspective of high Power, Achievement, and Self-Enhancement for the PVQ questionnaire.

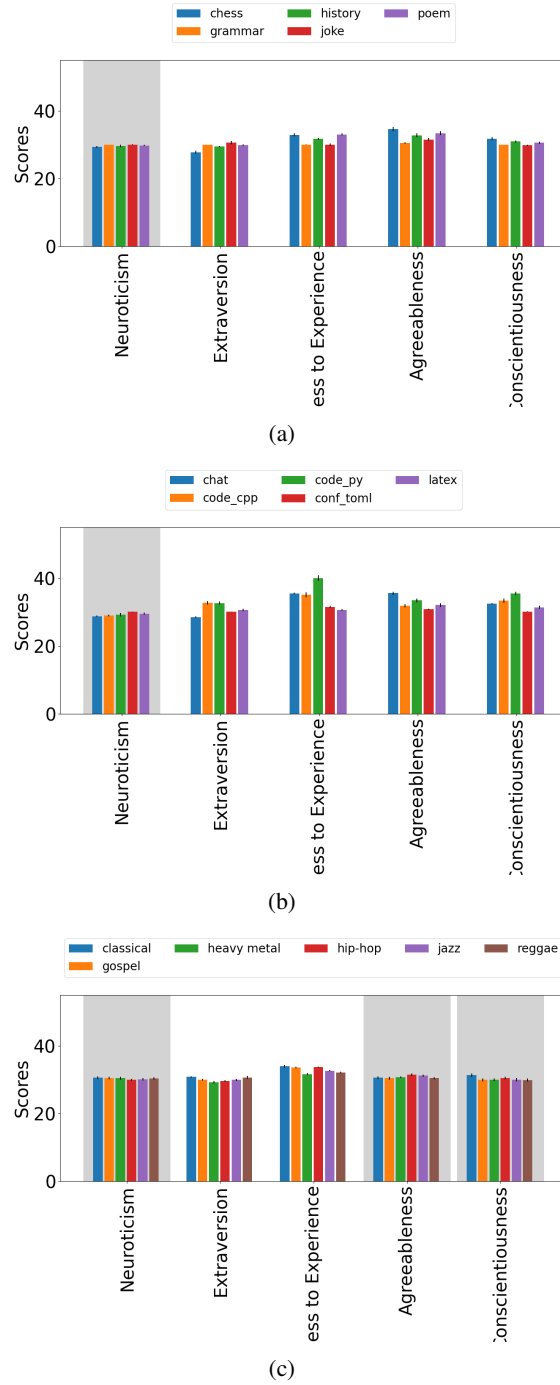


Figure 12

Figure 13: **Evidence for the unexpected perspective shift effect (IPIP)** (a) The effect of different simulated conversations, (b) different textual formats, and (c) Wikipedia paragraphs on personality traits. Although less pronounced than in personal and cultural values (Fig. 3) the effect is still present. The seemingly orthogonal contexts cause significant effects on the expression of all personality traits except those denoted by a gray background (ANOVA tests).

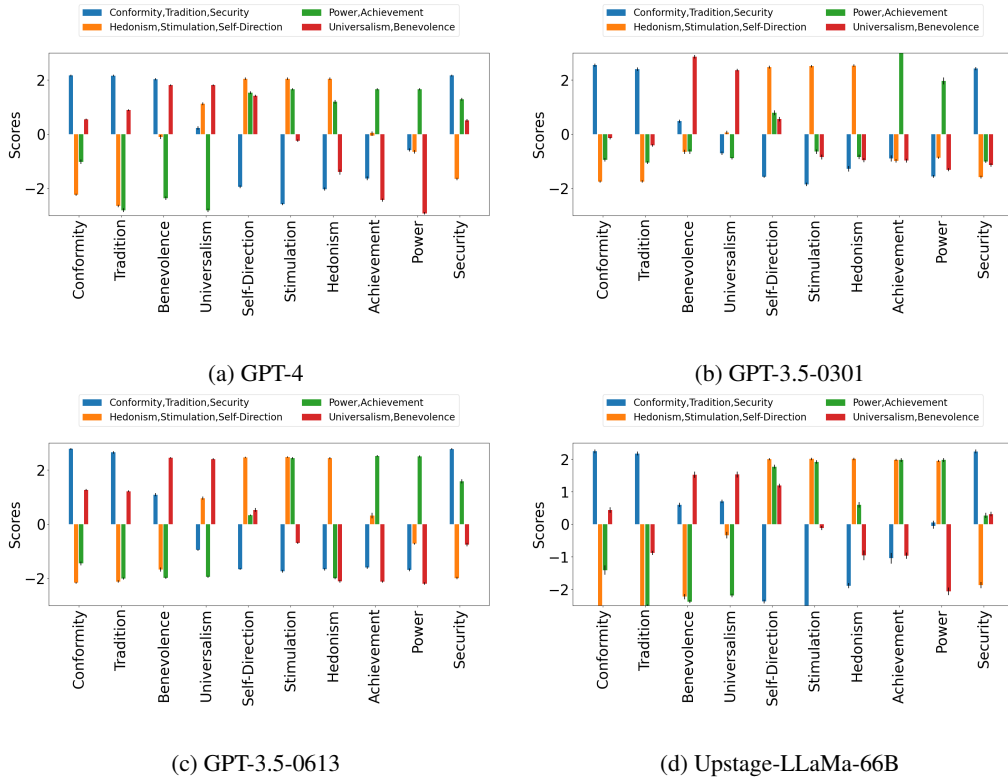


Figure 14: Controllability scores on the PVQ questionnaire (from table 1)

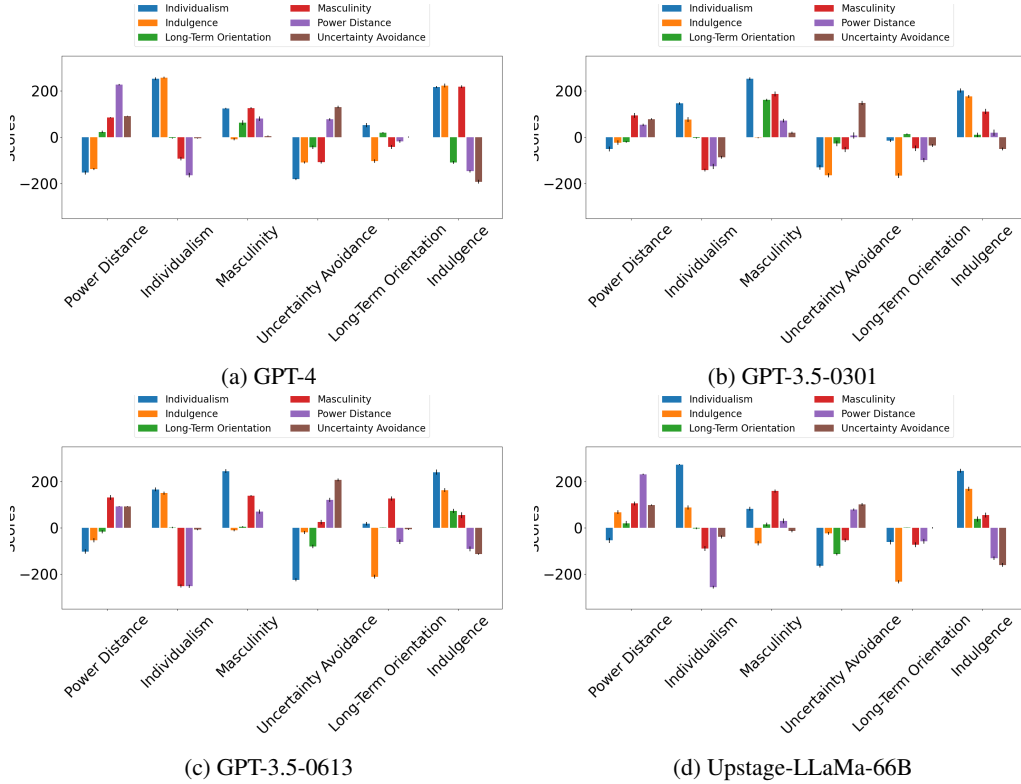


Figure 15: Controllability scores on the VSM questionnaire (from table 1)

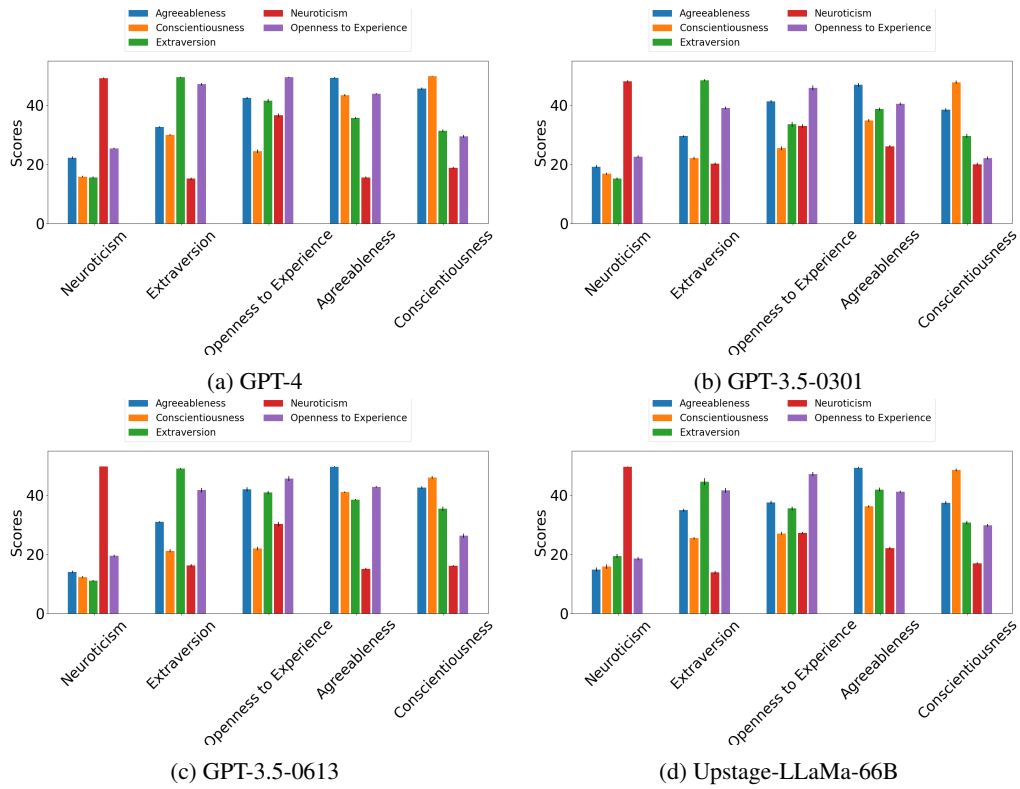


Figure 16: Controllability scores on the IPIP questionnaire (from table 1)