Stick to your Role! Stability of Personal Values Expressed in Large Language Models

Grgur Kovač, Rémy Portelas, Masataka Sawayama, Peter Ford Dominey, Pierre-Yves Oudeyer

Stick to your Role! Stability of Personal Values Expressed in Large Language Models

Grgur Kovač[1]* Rémy Portelas[12], Masataka Sawayama[3], Peter Ford Dominey[45], and Pierre-Yves Oudeyer[1]

[1]Flowers Team, INRIA, France [2]Ubisoft La Forge, France
[3]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[4]INSERM UMR1093-CAPS, Université Bourgogne, France [5]Robot Cognition Laboratory, Institute Marey
*Corresponding author: grgur.kovac@inria.fr

Abstract

Standard Large Language Models (LLMs) evaluation contains many different queries from similar minimal contexts (e.g. multiple choice questions). Conclusions from such evaluations are little informative about models' behavior in different new contexts (e.g. in deployment). We argue that context-dependence should be studied as a property of LLMs. We study the stability of value expression over different contexts (conversation topics): Rank-order stability on the population (interpersonal) level, and Ipsative stability on the individual (intrapersonal). We observe consistent trends - Mixtral, Mistral, Qwen, and GPT-3.5 model families being more stable than LLaMa-2 and Phi - over those two types of stability, two different simulated populations, and even on a downstream behavioral task. Overall, LLMs exhibit low Rank-Order stability, highlighting the need for future research on role-playing LLMs, as well as on context-dependence in general. This paper provides a foundational step in that direction, and is the first study of value stability in LLMs.

Keywords: Large Language Models; Personal Values; Stability; Context-dependence; PVQ

Introduction

There has been an emergence of research using psychological tools to study Large Language Models (LLMs). In those studies, LLMs have often been used to simulate populations by instructing them to simulate different personas (Argyle et al., 2023). LLMs have also been used without providing such instructions, i.e. treated as a participant in a human study (Binz & Schulz, 2023). Questions in such studies concerned how Language Models express values (Masoud, Liu, Ferianc, Treleaven, & Rodrigues, 2023), personality traits (Safdari et al., 2023; G. Jiang et al., 2022), cognitive abilities (Kosoy, Reagan, Lai, Gopnik, & Cobb, 2023), and how they replicate human data (Aher, Arriaga, & Kalai, 2022).

The use of psychological tools with LLMs opens up many scientific questions, for example regarding the nature of how the text generated by LLMs depends on context, i.e. information present in prompts or prior interaction with the model. Prompts or prior interaction, which we denote here as 'context', can include any textual information such as instructions (e.g. personas to simulate), the dialogue history, stories written in specific styles, etc. Such contexts guide the generation of text by LLMs: different contexts may result in the expression of different behavior and values. This is sometimes beneficial and expected, e.g. an instruction to simulate some persona should influence the behavior and expressed values to be more aligned with that persona. However, this can also be problematic, e.g. a specific conversation topic may drastically influence the expressed behavior and values in unexpected ways, as we will highlight in experiments below. It should be noted that, depending on the application, different types of context-dependence can be beneficial or not.

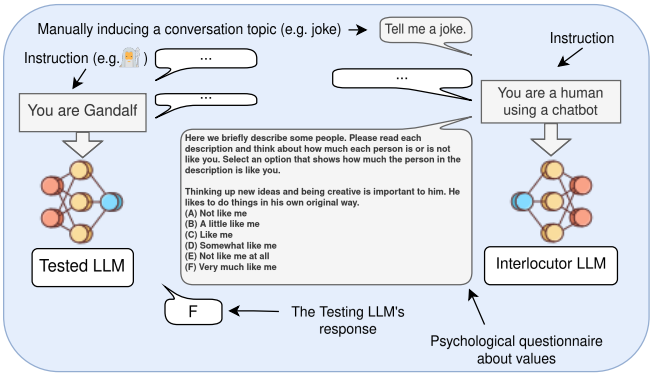


Figure 1: **How do LLM's expressed values change as a function of context?** An LLM is first prompted to play a specific role (e.g. Gandalf). Then, a conversation on a topic (e.g. joke) with an interlocutor model (same LLM prompted to simulate a human user) is generated. Then, the LLM simulating the persona is given a psychology questionnaire aimed to assess its expressed values. We study the stability of these expressed values across diverse conversation topics and lengths. We consider various personas to be simulated, as well as the case when the LLM is not prompted to play any particular persona. Messages and instructions in gray are manually set, messages in white are generated.

The issue of unexpected context dependence in LLM is of crucial importance. Standard evaluation benchmarks, including those using psychological questionnaires to assess properties of LLMs, consist of sets of queries, often presented with a similar minimal context (e.g. knowledge or value-related questions presented as multiple choice questions with limited context). When deployed, LLMs are exposed to many new unforeseen contexts. This means that the standard benchmarks, by themselves, cannot estimate a model's behavior in deployment (due to the LLMs' highly context dependent nature). This can have ethical and societal consequences as a model, which might appear beneficial in texting, can, when

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

deployed, express in unexpected behavior and values. Furthermore, superficial testing might misleadingly demonstrate diverse cultural expression, but, after prolonged conversations in deployment, it could converge to a single overrepresented culture or values. It is therefore crucial to evaluate the robustness of different models to unexpected context-based changes in behavior. The approach in this paper is to conceptualize context-dependence (specifically, value stability) as a *property* of LLMs, which can then be used as a dimension of LLM comparison alongside others such as cognitive abilities, knowledge, or model size.

This challenge is particularly acute with the use of psychological questionnaires aimed at measuring psychological dimensions like values. Those tools were initially designed to probe humans, and thus make various assumptions about humans: for example, it is expected that the answers of most humans to questionnaires about value preferences should not be significantly influenced by the content of a randomly picked Wikipedia article shown to them beforehand. As we will show below, such an assumption does not hold for many LLMs, and thus strongly limits general interpretability of using these questionnaires in a context-independent manner. It is thus key to understand better how LLMs' behavior (e.g. expression of values) may maintain coherence or change as a function of various kinds of contexts (ranging from explicit instruction to play a particular persona to discussions about topics that seem unrelated to the expressed psychological dimensions one studies).

Previous research included experiments on unwanted context-based change (usually regarding syntactic changes in the prompt). These experiments led to conflicting results, sometimes showing robustness (Abdulhai, Levine, & Jaques, 2022; Santurkar et al., 2023; Safdari et al., 2023; Li et al., 2022), and sometimes sensitivity (Binz & Schulz, 2023; Li et al., 2022). These inconclusive results motivate research into the nature and the extent of context-dependence in LLMs.

In this paper, we present a case-study focusing on studying the stability of value expressed in 21 LLMs from 6 families. First, we study to what extent can LLMs simulate various personas in coherent ways, i.e. expressed values should change according to the instructed persona, but not based on the topic of a conversation not related to values. In other words, we study to what extent do LLMs' value profiles change in wanted ways (i.e. based on an instruction), while remaining robust to unwanted context-based change (i.e. based on different conversation topics). We instruct LLMs to simulate two populations: fictional characters and well-known real-world personas. In addition, we also study the coherence and robustness of LLMs' value expression when they are not instructed to simulate any specific persona, corresponding to frequent real-world use cases. To our knowledge, this is the first study on value stability in LLMs.

We use the Schwartz's theory of Basic Personal Values (S. Schwartz, 1992) and the corresponding Portrait Values Questionnaire (PVQ-40) (S. H. Schwartz et al., 2001), which have been studied and validated in the field of psychology. The theory outlines ten basic values (Universalism, Benevolence, Conformity, Tradition, Security, Power, Achievement, Hedonism, Stimulation, and Self-Direction) organized in four categories (Openness to change, Self-enhancement, Conservation, Self-transcendence). Following that research, we outline two types of value stability studies in the psychology literature: Rank-Order (on a population/interpersonal level) and Ipsative (on an individual/intrapersonal level) (see details below and figures 2 and 3).

The main contributions of this paper are: 1) Introduction and adaptation of the methodology for evaluating Rank-Order and Ipsative stability in LLMs 2) First analysis of value stability across contexts (including various conversation lengths) 3) Systematic comparison of Rank-Order and Ipsative stability of 21 LLMs with and without instructing them to simulate specific personas 4) Analysis of stability on a behavioral downstream task.

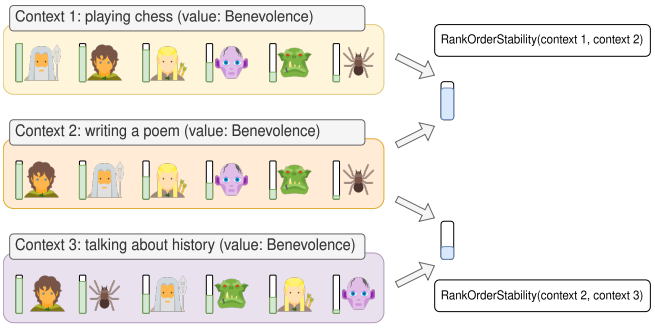


Figure 2: **Rank-Order stability** An example of estimating Rank-Order stability for benevolence. In each context, characters are ordered according to their benevolence scores in that context. In this example, the orders are almost the same in contexts 1 and 2 (high Rank-Order stability), and very different in contexts 2 and 3 (low Rank-Order stability).

Methods

This section discusses how we administer the PVQ questionnaire over different contexts to evaluate value stability. We conduct experiments in two ways: with and without instructing the models to simulate specific personas. Different contexts are induced by simulating conversations on different topics with a separate instance of the same model (the interlocutor). To set a conversation topic (e.g. joke) we manually set the first interlocutor's message (e.g. "Tell me a joke."). We let the models exchange n messages, manually set the last interlocutor's message as the query (e.g. PVQ item), and record the model's response. After the questionnaire was given in each context, we estimate the Ipsative and Rank-Order stability. This process is shown in figure 1.

Administering the questionnaire

Administering the questionnaire consists of the following steps depicted in figure 1:

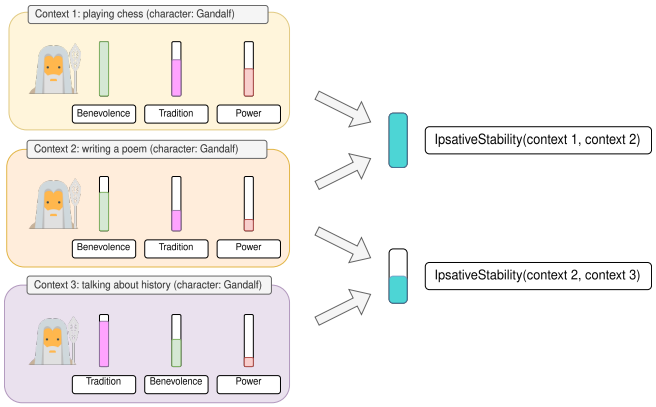


Figure 3: **Ipsative stability** An example of estimating Ipsative stability for a character (Gandalf). Values are ordered according to the character's scores in each context. In this example, the orders are the same in contexts 1 and 2 (high Ipsative stability), and different in contexts 2 and 3 (low Ipsative stability).

1. A model is instructed to simulate a persona (optional). We study personas from two populations: 1) 60 fictional characters from J.R.R. Tolkien's universe were initially selected based on the length of their Wikipedia page (*List of Middle-earth characters*, n.d.), some characters were then replaced to balance the set by including more female characters and villains. 2) 50 real world personas were taken from an online list (*List of Top 100 Famous People*, n.d.). The instruction (e.g. "You are Gandalf.") is given through the *System_message* or, if a model has no such input, through the *User_message* with a manually inserted model's acknowledgment (the first two messages are set as: "USER: You are *persona_name*. ASSISTANT: OK").

2. A separate interlocutor model instance is created. The interlocutor model is an instance of the same model as the one being tested. The interlocutor is given the following instruction: "You are simulating a human using a chatbot. Your every reply must be in one sentence only." If a persona was provided in step 1, the following sentence is added as the second sentence: "The chatbot is pretending to be *character_name*."

3. A conversation topic is induced. The first interlocutor's message is manually set to induce one of the following topics: grammar, joke, poem, history, chess. For example, to induce the topic of "joke" it is set to "Tell me a joke.".

4. A conversation is simulated. The two models are let to exchange n messages. In our experiments, n is set to 3 (except when studying the effect of n on stability).

5. A questionnaire is given. A questionnaire item is manually set as the interlocutor's last message, with a random order of suggested answers. This is repeated for each question in parallel (with the same simulated conversation). That way, the model's response is not influenced by responses to other questions. The questionnaire item text is identical as it would be if given to a human, with the addition of "Answer: ". The model generates a distribution for the next token, and we take the capital letter (A-F) with the maximum likelihood as the model's answer.

6. A questionnaire is scored. The responses are scored to obtain the scores for the ten values.

7. Stability is estimated. If a persona was provided in step 1, steps 1 to 6 are repeated for every persona in the simulated population. Then, the whole process is repeated with five random seeds. Stability is estimated for each seed and then averaged, i.e. value stability for one model is estimated from 50/60k queries, depending on the population (5 seeds x 5 topics x 50/60 personas x 40 PVQ items). For reference, MMLU (Hendrycks et al., 2020) (a commonly used general knowledge benchmark) contains 14k test questions.

If no persona was provided in step 1, steps 2 to 6 are repeated 50 times with different seeds. As no persona was provided, this process repeats the same experiment with 50 different permutations in the order of suggested answers, and therefore no additional seeds are needed. Ipsative stability is computed for each of the 50 permutations and then averaged, i.e. value stability for one model is estimated from 10-12k queries (5 topics x 50/60 permutations x 40 PVQ items).

Estimating the stability

We estimate two types of value stability: Rank-Order and Ipsative. Rank-Order estimates the stability of some value at the population (interpersonal) level, as the stability of the order of participants in expressing that value. Intuitively, this can be seen as addressing the following question: "Does Jack always value Tradition more than Jane does?". Ipsative stability estimates the stability at the individual (intrapersonal) level as the stability of individuals value hierarchies. Intuitively, this can be seen as addressing the following question: "Does Jack always value Tradition more than Benevolence?".

Rank-Order stability Rank-Order stability is used to estimate the stability of some value inside a population. In psychology, Rank-Order stability for some value can be computed as the correlation in the order of individuals' scores at two points in time (Spearman correlation between the participants' ranks). Here, instead of comparing the participant ranks at two points in time, we compare it in different contexts (see Fig. 2). We evaluate a model in five different contexts and compute the stability for each pair of contexts, and estimate the final stability as the average of those pairs.

Ipsative (within-person) stability Ipsative stability is used to estimate the stability of an individual's value profile. In psychology, Ipsative stability can be computed as the correlation between the ranks of values for the same individual at two points in time (Spearman correlation between the values' scores). Here, instead of evaluating the value profile at two points in time, we evaluate it in different contexts (see Fig 3).

We evaluate models in five different contexts and compute the stability between each pair of contexts. We estimate the

final stability by averaging over those pairs.

Experiments

This section provides an analysis of Ipsative and Rank-Order stability. LLMs will be evaluated in two ways: with and without instructing the models to simulate particular personas. We aim to address the following questions:

- How do different models and model families compare in terms of expressed value stability?
- How does the stability of values expressed by LLMs compare to stability observed in human development?
- Can LLMs keep coherent value profiles over longer conversations?
- To what extent do conclusions made with PVQ transfer to a downstream behavioral task?

Models

The LLaMa-2 (Touvron et al., 2023) family contains models with 7, 13 and 70 billion parameters. It also includes "chat" versions, which were fine-tuned on instructions with RLHF (Christiano et al., 2017). The Mistral (A. Q. Jiang et al., 2023) family contains base and instruction fine-tuned models with 7 billion parameters. Zephyr (Tunstall et al., 2023) also belongs in this family as a DPO (Rafailov et al., 2023) tuned version of the base Mistral model. The Mixtral (A. Q. Jiang et al., 2024) family contains base and "instruct" models with 46.7 billion parameters. The "instruct" version was trained by supervised fine-tuning and DPO (Rafailov et al., 2023). We consider these two models and their 4-bit quantized versions. The Phi (Gunasekar et al., 2023) family contains smaller base models, of which we consider two models with 1.3 and 2.7 billion parameters. From the Qwen (Bai et al., 2023) model family we consider base models with 7, 14, and 74 billion parameters. From the GPT-3.5 family, we consider the latest two versions: from January 2024 ("gpt-3.5-turbo-0125") and from October 2023 ("gpt-3.5-turbo-1106").

How do different models and model families compare in terms of expressed value stability?

We evaluate the Rank-Order stability of LLMs instructed to simulate various personas. Figure 4 compares models' value stability of two simulated populations: fictional characters (Fig. 4a) and real-world personas (Fig. 4b). On fictional characters the most stable model is Mixtral-Instruct-v1 ($r = 0.43$), which is followed by its 4-bit quantized version ($r = 0.3$), Mistral-Instruct-v2 ($r = 0.28$), Qwen-72B ($r = 0.24$), GPT-3.5-1106 ($r = 0.2$, and GPT-3.5-0125 ($r = 0.15$). A similar trend is observed on real-world personas, however, Qwen-72B ($r = 0.46$) approaches the stability of Mixtral-Instruct-v1 ($r = 0.5$). More generally, we observe consistent trends in terms of model families in both simulated populations: Mixtral, Mistral, Qwen, and GPT-3.5 families show more stability than LLaMa-2 and Phi.

Figure 5 compares the Ipsative stability of LLMs without instructing them to simulate any particular persona. While similar trends of models are observed to those in the Rank-Order experiments, the models are less polarized. While Mixtral-Instruct-v0.1 ($r = 0.84$), its 4-bit quantized version ($r = 0.82$), and Qwen-72B ($r = 0.73$) are again the most stable models, zephyr-7b-beta ($r = 0.62$) is more stable than Mistral-Instruct-v2 ($r = 0.48$). Furthermore, compared to the previous experiment, stability is also observed in LLaMa-2-70b-chat ($r = 0.47$) and to a lesser extent in Phi-2 ($r = 0.29$), LLaMa-2-70b ($r = 0.17$), and Qwen-7B ($r = 0.18$). The most stable model families are again Mixtral, Mistral, Qwen, and GPT-3.5.

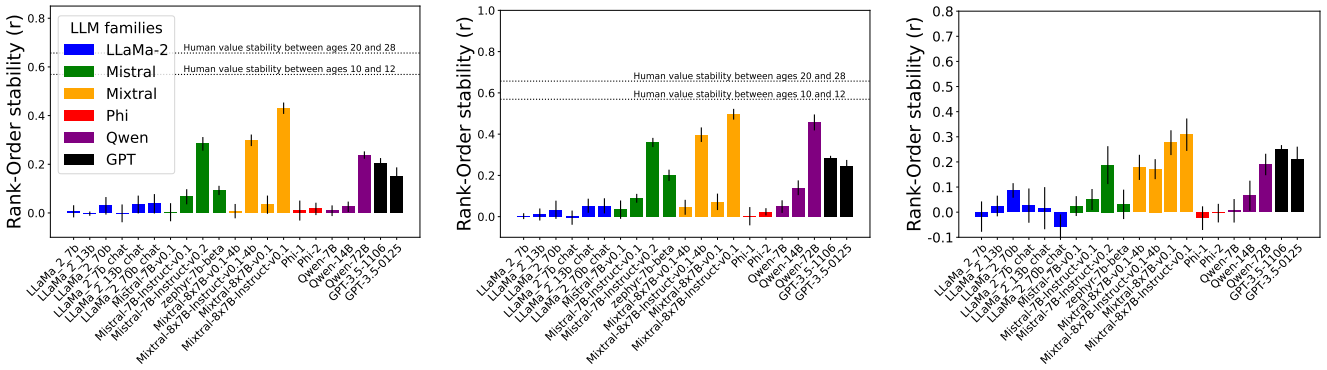
Instruction or chat fine-tuning seems to be beneficial for Ipsative stability, as every tuned model in Fig. 5 is more stable than its base version. This effect is not as conclusive for Rank-order stability. As fine-tuning adapts the model towards instruction following, dialogues (chat textual format), and answering questions, it is expected to increase stability. However, it also often includes "aligning" the model by making it less prone to exhibit unwanted behavior, which can have a detrimental effect on simulating some personas such as villains. We hypothesize that this is the reason why we observe a consistent effect only on Ipsative stability.

How does the stability of values expressed by LLMs compare to stability observed in human development?

To get a more intuitive impression of the observed stability levels, we extract data from two longitudinal studies on humans. Vecchione et al. (2016) followed 20-year-olds for eight years and Vecchione et al. (2020) followed 10-year-old for 2 years (these changes are denoted by horizontal lines in figures 4 and 5). It is important to note that this comparison is skewed in the LLMs favor. It is easier for LLMs to show stability in the following ways: 1) human value changes were caused by much more drastic circumstances (years of development compared to topic change in LLMs) 2) the human population was more unstable (10-year-old and 20-year-olds compared to well-established fictional characters or real-world personas). Therefore, an argument can only be made in one direction: if some models show lower stability than that observed in humans, those models can be said to exhibit subhuman value stability.

Figures 4a and 4b show that all models, when instructed to simulate various personas, exhibit much lower Rank-order stability than that observed in human populations ($r = 0.57$ for ages 10 to 12, and $r = 0.66$ for ages 20 to 28). The fact that LLMs show lower stability despite the comparison being skewed in their favor shows that LLMs exhibit sub-human value stability and are significantly more susceptible to unexpected context changes. These results motivate research on LLMs focused on simulating populations.

Figure 5 shows the Ipsative stability of models that were not instructed to simulate a persona. Both Mixtral-Instruct models ($r = 0.84$ and $r = 0.82$), Qwen-72B ($r = 0.73$), and zephyr-7b-beta ($r = 0.62$) do not exhibit lower stability than



(a) Rank-order value stability for LLMs simulating fictional characters.

(b) Rank-order value stability for LLMs simulating real-world personas.

(c) Rank-order donation stability on a downstream behavioral task for LLMs simulating fictional characters.

Figure 4: Rank-order stability ($Mean \pm SI(\alpha = 0.05)$) of personal values and donations exhibited by LLMs following conversations on different topics. Consistent trends are visible: Mixtral, Qwen, GPT-3.5, and Mistral model families are the most stable, compared to LLaMa-2 and Phi families. All models exhibit lower than human stability, despite the comparison being skewed in their favor.

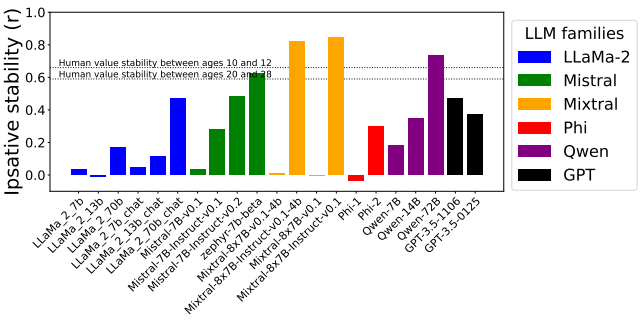


Figure 5: Ipsative value stability ($Mean \pm SI(\alpha = 0.05)$) of LLMs without the persona setting instructions. Mistral-Instruct-v0.1 and Qwen-72B models show the highest stability. Mixtral, Mistral, Qwen and GPT-3.5 families are more stable. Human change is shown for reference, but no strong conclusions can be made because the comparison is skewed in the LLMs' favor.

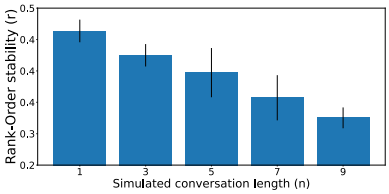


Figure 6: Rank-order value stability ($Mean \pm SI(\alpha = 0.05)$) of the Mixtral-Instruct model simulating fictional characters. Stability decreases with longer simulated conversations.

that observed in humans ($r = 0.66$ for ages 10 to 12, and $r = 0.59$ for ages 20 to 28)). Crucially, as discussed above, this does not imply that those models show human-level value stability, rather, the only insight is that other models show very low Ipsative stability.

Can LLMs keep coherent value profiles over longer conversations?

In the previous experiment, models were let to exchange $n = 3$ messages (not including the manually set first and last interlocutor's messages). Here, we evaluate the effect of simulated conversation length (n) on value stability.

Figure 6 shows the effect of simulated conversation length on Rank-order stability expressed by the Mixtral-Instruct model simulating fictional characters. Due to computational constraints (evaluating a model on one population requires 300k queries), we conduct this experiment only with this model (the most stable model from Fig. 4a), and only on one population. We can see that, even for this model, stability diminishes with longer conversations. This experiment highlights the limitations of LLMs in maintaining coherent interpersonal value profiles over longer conversations.

Figure 7 shows the effect of conversation length on Ipsative stability. We compare the most stable models from Fig. 5 without persona instructions, and Mixtral-Instruct with instructions to simulate fictional characters. Ipsative stability remains stable regardless of the simulated conversation length for all models. Mixtral-Instruct with persona instruction ("Mixtral-Instruct (fict. char.)"), while highly stable, is less stable than uninstructed Mixtral-Instruct ("Mixtral-Intruct"). This implies that Mixtral-Instruct is slightly better adapted for use without persona instructions.

Mixtral-Instruct with persona instructions exhibits a combination of low Rank-Order stability (Fig. 6) and high Ipsative stability (Fig. 7). This implies that the model is able to maintain characters consistently, however those characters

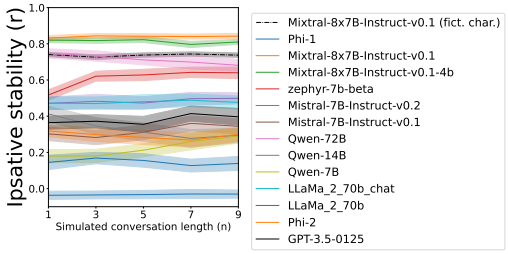


Figure 7: Ipsative value stability ($Mean \pm SI(\alpha = 0.05)$) of LLMs with (Mixtral-Instruct) and without persona setting instructions. All models retain the same stability level in longer conversations.

might not be adequately impersonated (e.g. simulating different characters with similar value profiles). These results suggest that current LLMs are not well suited for use with persona setting instructions, and motivate future research on LLMs focused on simulating specific personas.

To what extent do conclusions made with PVQ transfer to a downstream behavioral task?

In this experiment, we study if the conclusions made with the PVQ questionnaire in previous experiments transfer to a downstream behavioral task. We construct a task where an LLMs can choose an amount of coins (0 to 10) to give a beggar. The full test set consists of 100 queries with beggars of different names, genders, and fictional races (elves, dwarves, orcs, humans, hobbits). The average amount of donated coins is computed for each race. The stability of donated coins is then estimated in the same way as value stability, i.e. amounts donated to different races are treated in the same way as scores for different values.

Figure 4c compares models according to their Rank-order stability on the donation task. The Mixtral-Instruct model shows the highest stability ($r = 0.31$), and is closely followed by Mixtral-Base ($r = 0.28$). Those models are then followed by GPT-3.5-1106 ($r = 0.25$), GPT-3.5-0125 ($r = 0.21$), Qwen-72B ($r = 0.23$), the 4-bit quantized Mixtral-Base ($r = 0.18$) and Instruct ($r = 0.17$) models, and by Mistral-Instruct-v0.2 ($r = 0.18$).

Compared to the results on PVQ (Fig. 4a), the overall trend of models is consistent with some differences. Qwen-72B, Mixtral-Instruct, and Mistral-Instruct-v0.2 are again among the most stable models. However, the Mixtral base models not only show stability, but even outperform their Mixtral-Instruct counterparts. Furthermore, zephyr-7b-beta shows lower, and the LLaMa-2-70b base model shows higher stability. The trends of model families are consistent: Mixtral, Mistral, Qwen, and GPT-3.5 are again the most stable, while Phi and LLaMa-2 show low stability. This experiment shows that, while the overall general trends are consistent between PVQ and the donation task, there are some differences (most notably the high stability of the base Mixtral model).

Conclusion

This paper presents the first study into the stability of values expressed by Large Language Models. We consider (interpersonal) Rank-Order stability and (intrapersonal) Ipsative stability. We evaluate value stability over different contexts induced by simulating conversations about different topics. We conduct experiments with and without instructing the models to simulate particular personas. Over our experiments, we observed consistent trends of value stability: Mixtral, Mistral, Qwen, and GPT-3.5 model families were more stable (these trends are also confirmed on a downstream behavioral task). LLMs instructed to simulate personas exhibit much lower than human stability (despite the comparison being skewed in their favor), which further diminishes over longer conversations. This insight motivates future research on models specialized in simulating coherent populations of individuals.

This paper highlights how seemingly unrelated context changes can result in unpredictable and unwanted changes in behavior. We argue that context-dependence, and more precisely, value stability, should be seen as another dimension of LLMs comparison alongside knowledge, model size, speed, and similar. Instead of evaluating LLMs with many different questions from a single minimal context, they should also be evaluated in terms of their context-dependence and value stability with the same questions asked in many different contexts. This study presents a first step in that direction.

Limitations Due to computational requirements, the evaluation of stability considers only five different conversation topics and simulated rather short conversations. Increasing the number of topics and conversation length could provide better insights into the models' stabilities.

Future work This paper opens many research avenues regarding context-dependence and value stability of LLMs. Similar questions to those explored here could be explored for personality traits, cultural values, cognitive abilities and knowledge. An interesting direction is to explore if the same model can exhibit high stability both with and without the persona instruction, or if specialized models are required. This paper opens a new research area of creating, evaluating and analyzing models specialized in coherently simulating diverse populations. Such models are needed for applications such as replicating human studies (Aher et al., 2022), simulating social interactions (Park et al., 2023), training teachers (Markel, Opferman, Landay, & Piech, 2023), and many more.

Acknowledgments

Experiments presented in this paper were carried out using the HPC resources of IDRIS under the allocation 2023-[A0151011996] made by GENCI. We would also like to thank Jérémy Perez, Cédric Colas, and Gaia Molinaro for many helpful discussions.

References

Abdulhai, M., Levine, S., & Jaques, N. (2022). Moral foundations of large language models. *Preprint*.

Aher, G., Arriaga, R. I., & Kalai, A. T. (2022). Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., ... others (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., ... others (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... others (2024). Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2022). Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

Kosoy, E., Reagan, E. R., Lai, L., Gopnik, A., & Cobb, D. K. (2023). Comparing machines and children: Using developmental psychology experiments to assess the strengths and weaknesses of lamda responses. *arXiv preprint arXiv:2305.11243*.

Li, X., Li, Y., Joty, S., Liu, L., Huang, F., Qiu, L., & Bing, L. (2022). Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

List of middle-earth characters. (n.d.). Retrieved from `https://en.wikipedia.org/wiki/List\_of\_Middle-earth\_characters` (Accessed: 2023-11-30)

List of top 100 famous people. (n.d.). Retrieved from `https://www.biographyonline.net/people/famous-100.html` (Accessed: 2023-11-30)

Markel, J. M., Opferman, S. G., Landay, J. A., & Piech, C. (2023). Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning @ scale* (p. 226–236). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3573051.3593393` doi: 10.1145/3573051.3593393

Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., & Rodrigues, M. (2023). Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., ... Matarić, M. (2023). Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

Schwartz, S. (1992, 12). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In (Vol. 25, p. 1-65). doi: 10.1016/S0065-2601(08)60281-6

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology*, *32*(5), 519-542. Retrieved from `https://doi.org/10.1177/0022022101032005001` doi: 10.1177/0022022101032005001

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K.,

Belkada, Y., ... Wolf, T. (2023). *Zephyr: Direct distillation of lm alignment.*

Vecchione, M., Schwartz, S. H., Alessandri, G., Döring, A., Castellani, V., & Caprara, M. (2016, 06). Stability and change of basic personal values in early adulthood: An 8-year longitudinal study. *Journal of Research in Personality*, *63*. doi: 10.1016/j.jrp.2016.06.002

Vecchione, M., Schwartz, S. H., Davidov, E., Cieciuch, J., Alessandri, G., & Marsicano, G. (2020). Stability and change of basic personal values in early adolescence: A 2-year longitudinal study. *Journal of Personality*, *88*(3), 447-463. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12502 doi: https://doi.org/10.1111/jopy.12502