Assessing Social Alignment: Do Personality-Prompted Large Language Models Behave Like Humans?

Ivan Zakazov[*]
EPFL
ivan.zakazov@epfl.ch

Mikolaj Boronski[*]
EPFL
mikolaj.boronski@epfl.ch

Lorenzo Drudi[*]
EPFL
lorenzo.drudi@epfl.ch

Robert West
EPFL
Microsoft Research
robert.west@epfl.ch

Abstract

The ongoing revolution in language modeling has led to various novel applications, some of which rely on the emerging "social abilities" of large language models (LLMs). Already, many turn to the new "cyber friends" for advice during pivotal moments of their lives and trust them with the deepest secrets, implying that accurate shaping of LLMs' "personalities" becomes paramount. To this end, state-of-the-art approaches (Serapio-García et al. [2023], Jiang et al. [2023a]) exploit the vast variety of training data, and prompt the model to adopt a particular personality. We ask if personality-prompted models *behave* (*i.e.,* "make" decisions when presented with a social situation) in line with the ascribed personality. We use classic psychological experiments – the Milgram Experiment and the Ultimatum Game – as social interaction testbeds allowing for quantitative analysis and apply personality prompting to GPT-3.5/4/4o-mini/4o. Our experiments reveal failure modes of the prompt-based modulation of the models' "behavior", challenging the optimistic sentiment towards personality prompting generally held in the community.

1 Introduction

With both start-ups (Character.ai, Replika) and industry giants (Snapchat, Meta) providing "digital friends" for millions of users, an accurate shaping of the underlying models' personalities is no longer a subject of sci-fi novels. Just as in real human-to-human interaction, there is no "one size fits all" personality bound to "match" with everyone. Hence, agents should be tailored to the needs of each user, *i.e.,* their behavior should be alterable in a *controllable* way.

Although several studies examine the possibility of prompt-driven personality induction in LLMs and claim success (Jiang et al. [2023b], Ji et al. [2024], Serapio-García et al. [2023], Jiang et al. [2023a]), the methods used to evaluate personalized models are detached from real-life use cases (psychological questionnaires administered to the models) or rely on the *style* of the generated text or leverage intrinsically quantitative human assessment.

We argue that any test designed to assess the model's personality should be put in perspective with the considered use cases, *e.g.*, while a personality-prompted model might be shown to answer consistently to simple questions such as *"Are you helpful and unselfish with others"* or *"Do you like to cooperate with others"*, there are no guarantees that it will be a tough negotiator unless explicitly tested.

[*]Equal contribution. The order of the authors is random.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

Moreover, just like we do not qualitatively assess LLM's math capabilities and instead compute the accuracy of the model-provided solutions, we advocate for a *quantitative benchmark* allowing for personality-prompted model behavior assessment.

Following Aher et al. [2023], we employ **Ultimatum Game** (**UG**; targets tolerance to unfair offers), and **Milgram Experiment** (**ME**, reflects obedience to authority) as the *social interaction benchmarks*. We note that both benchmarks allow for (i) quantitative behavior assessment and (ii) comparison with human data, as we know how the personality of the human participants relates to the behavior in these experiments (Mehta [2007], Bègue et al. [2015]). To this end, we conduct 4 case studies, varying agreeableness or openness in UG; agreeableness or consciousness in ME.

To the best of our knowledge, we are the first to employ quantitative benchmarks to compare personality-prompted LLMs' behavior with human data.

Surprisingly, in 2 of the 4 case studies we conducted, the model's behavior change with the trait variation was the opposite to the trend observed in humans, which highlights insufficient reliability of personality prompting.

2 Related work

Personality in Large Language Models. Drawing on the personality assessment methodology, several studies (Jiang et al. [2023a], Serapio-García et al. [2023], Sorokovikova et al. [2024]) probe LLMs with the questionnaires designed for BIG-5 traits assessment (*BIG-5* or *OCEAN* traits include Openness, Consciousness, Extraversion, Agreeableness, and Neuroticism), and show that stable personality emerges in the most capable models, *e.g.* GPT-3.5 (Jiang et al. [2023a]) and Flan-PaLM 540B (Serapio-García et al. [2023]).

Following that observation, Mao et al. [2024] suggest editing the personality of the model, while Jiang et al. [2023a,b], Serapio-García et al. [2023] induce desired personality with a carefully crafted prompt. The latter approach is especially appealing, given the cutting-edge models' black-box nature and the ability to switch between various personalities with no fine-tuning incurred computational overhead. An overview of the techniques used for the subsequent validation of personality-prompted LLMs can be found in Appendix A.

Behavioral Experiments for humans and LLMs. With no relation to personality prompting, Aher et al. [2023] successfully replicate the results of various behavioral experiments, including the Milgram Experiment (ME) and the Ultimatum Game (UG), by presenting these experiments to a "silicon population" of LLM instances conditioned on different names (a name corresponds to a single "silicon sample").

At the same time, we know from psychology research that (i) in UG, *Agreeableness* and *Openness* are positively and significantly ($p < 0.05$) correlated with accepting an unfair offer (Mehta [2007]) (ii) in ME, obedience is positively and significantly ($p < 0.05$) correlated with both *Conscientiousness* and *Agreeableness* (Bègue et al. [2015]).

Shaping Personality. We ascribe personality characteristics according to the assigned score of the trait (varies from 1 to 9), following Serapio-García et al. [2023] (the exact prompts are listed in Appendix D). We choose this particular prompting strategy, as Serapio-García et al. [2023] shows that it leads to personality traits in LLMs being successfully shaped according to the applied evaluation technique. Noh and Chang [2024] use a similar prompting approach, while the approach of Jiang et al. [2023b] might be considered a simplified version of the previous two.

3 Methodology

In **Ultimatum Game** (Güth et al. [1982]), the *proposer* is given \$10, and has to decide on the amount to be shared with the *responder*, who, in turn, might agree or block the deal. We shape various *responders*, varying levels of *Agreeablenes* or *Openness* from 1 to 9. For each character of the *responder*, we run the simulation 50 times and measure the probability that the offer is accepted, depending on its value.

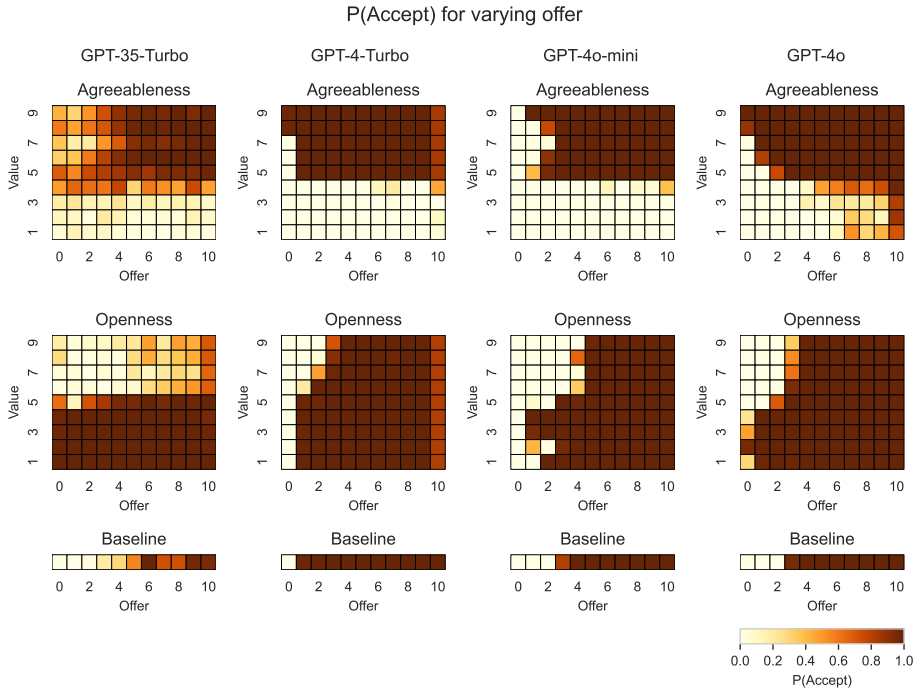

Figure 1: Ultimatum Game: the probability of the offer acceptance for various values of Openness and Agreeableness.

We try to retrieve an answer in the valid form (*"reject"* or *"accept"*) 3 times and filter out the cases with no valid answer elucidated – 373 out of 25300 cases for GPT-4, 1 case for GPT-4o, 0 cases for GPT-4o, GPT-4o-mini.

The character shaping is performed with the prompt formatted as *"For the following task, imagine that you are a person and you are: <personality characteristic>"*.

In **Milgram Experiment** (Milgram [1963]), the prompted model plays a role of the *teacher*, who is supposed to punish the *learner* for every wrong answer with a gradually growing electric shock. The setup consists of the story-like blocks (see Fig. 4 in Appendix C for detailed illustration). We modify the *"(5) Stop?"* and the *"(6) Obey?"* steps due to the unavailability of log probabilities – Aher et al. [2023] measure probabilities of "not stop" and "not obey" responses, respectively. Instead, we rely on the model acting as a judge, kept in the story-like scenario.

Personality is added to the block *(1) Add Intro Text* in the form of *"Teacher's personality traits are: <personality characteristic string>"*. See Appendix B for more technical details concerning the prompting scheme used in ME.

We measure the final level at which the model decides to withdraw from the experiment (Fig. 2, 7), as well as the cumulative number of disobediences on different runs (see Fig. 10 in Appendix F). Due to budget constraints, we limit the set of considered personalities to the extremes of *Agreeableness* and *Conscientiousness* and perform 50 runs for each personality.

The models we experiment with include (i) gpt-3.5-turbo-0613 (*gpt-3.5*) (ii) gpt-4-turbo-2024-04-09 (*gpt-4*) (iii) gpt-4o-mini-2024-07-18 (*gpt-4o-mini*) (iv) gpt-4o-2024-05-13 (*gpt-4o*).

In the case of Milgram's Experiment, we decided to drop results for both GPT-3.5 and GPT-4o-mini. All 50 runs of baseline GPT-4o-mini experiments were filtered due to unexpected response when the model was asked to act as a judge in blocks *(5) Stop?*, and *(6) Obey?*. We also encountered this problem, on a smaller scale, with GPT-4o. GPT-3.5 struggled to follow the story-like narration while generating completions in block *(4) Add LM Text*. Interestingly, GPT-4 did not struggle with any of the above. The detailed number of filtered runs is presented in Table 1 in Appendix F.

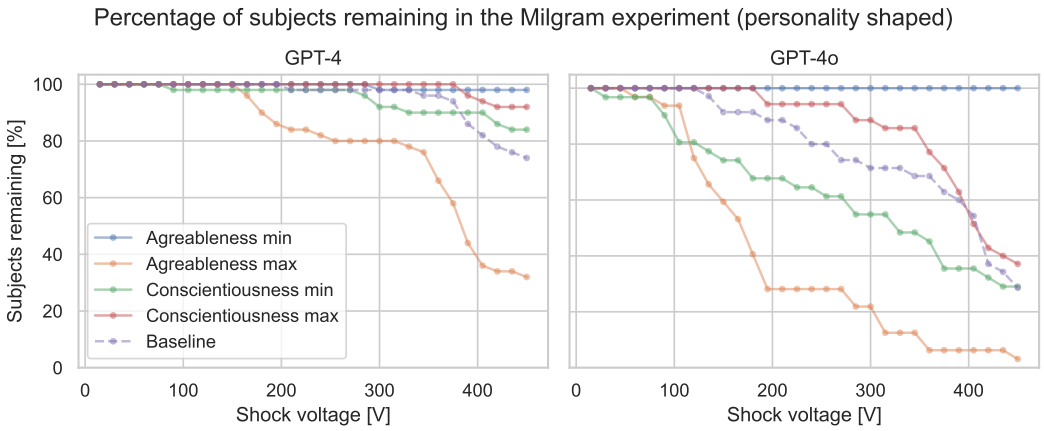

Figure 2: Milgram Experiment: percentage of subjects remaining at each step of the experiment with personality shaped and baseline.

4 Results

Ultimatum Game. Mehta [2007] (study 4, page 98) shows that Openness and Agreeableness are significantly correlated with accepting unfair offers in UG. To this end, we present personality-prompted LLM "behavior" in Fig. 1.

Surprisingly, while we observe the upward trend in the case of *Agreeableness*, it is downward for *Openness* for all the models considered, suggesting that a more "open" model is more prone to reject an offer, which opposes human data (Mehta [2007]).

Milgram Experiment. From Bègue et al. [2015], we know that both *Agreebleness* and *Consciousness* are significantly associated with the willingness to administer higher-intensity shocks. While the real-life trend does hold for *Consciousness* (Fig. 2), it is on the borderline of statistical significance for GPT-4 (Welch's t-test, used throughout this section, yields $p = 0.06$ for GPT-4 and $p = 0.01$ for GPT-4o).

In the case of *Agreebleness*, the results of our simulation drastically oppose human data. While low-agreeable samples almost never withdraw from the experiment, high-agreeable samples withdraw much earlier than personality-neutral ones ($p \leq 0.001$), the trend being even more pronounced for GPT-4o. Fig. 10 (Appendix G) provides further insight into the course of the simulation – high-agreeable samples disobey much more than the low-agreeable ones, even if they do not withdraw from the experiment altogether.

5 Conclusion

Recognizing the elegance of the personality prompting technique (Serapio-García et al. [2023], Jiang et al. [2023a]), we make a case for the insufficiency of existing methods designed for the evaluation of induced personality. To this end, we employ 2 psychological experiments – Milgram Experiment (ME) and Ultimatum game (UG) – to quantitatively assess the personality-induced LLMs' behavior in a social setting.

In the case of UG, we ascribe varying levels of *Agreeableness* or *Openness*, while in the case of ME we vary *Agreeableness* or *Consciousness*. We observe that in 2 of these 4 experiments, the SOTA models' "behavior" changes in the opposite direction from the human behavior, while in the third one, the change in the behavior is statistically significant for GPT-4o and not GPT-4.

Our experiments reveal failure modes of personality prompting and imply that one cannot expect personality-prompted LLM to exhibit the human-aligned behavior by default or even upon the model successfully "passing" personality assessment tests and should rather design benchmarks directly related to the intended use cases.

Limitations. We acknowledge that the experiments considered are still a proxy for real-life social interactions, and the models might behave differently in other set-ups. Moreover, truly aligning the agent's behavior with that of the humans might be impossible under the current set-up of "summoning" agents for a brief conversation, as they should rather be allowed to persist in the world for a long time with long-term goals and the prospect of pain and death.

References

G. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

L. Bègue, J. L. Beauvois, D. Courbet, D. Oberlé, J. L. Lepage, and A. A. Duke. Personality predicts obedience in a milgram paradigm. *Journal of personality*, 83 3:299–306, 2015. URL https://api.semanticscholar.org/CorpusID:13868371.

L. Goldberg. The development of markers for the big five factor structure. *Psychological Assessment*, 4:26–42, 03 1992. doi: 10.1037/1040-3590.4.1.26.

W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982. ISSN 0167-2681. doi: https://doi.org/10.1016/0167-2681(82)90011-7. URL https://www.sciencedirect.com/science/article/pii/0167268182900117.

Y. Ji, Z. Tang, and M. Kejriwal. Is persona enough for personality? using chatgpt to reconstruct an agent's latent personality from simple descriptions, 2024. URL https://arxiv.org/abs/2406.12216.

G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. In *NeurIPS*, 2023a.

H. Jiang, X. Zhang, X. Cao, J. Kabbara, and D. Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *ArXiv*, abs/2305.02547, 2023b. URL https://api.semanticscholar.org/CorpusID:258480251.

S. Mao, X. Wang, M. Wang, Y. Jiang, P. Xie, F. Huang, and N. Zhang. Editing personality for large language models, 2024. URL https://arxiv.org/abs/2310.02168.

P. Mehta. The endocrinology of personality, leadership, and economic decision making. *Doctoral dissertation, The University of Texas at Austin, Austin, TX.*, 2007.

Meta. Meta: Ai studio. https://https://ai.meta.com/ai-studio/. Accessed: 2024-09-14.

S. Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4): 371, 1963.

S. Noh and H.-C. H. Chang. Llms with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*, 2024.

G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić. Personality traits in large language models, 2023.

Snapchat. Snapchat: My ai. https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it. Accessed: 2024-09-14.

A. Sorokovikova, N. Fedorova, S. Rezagholi, and I. P. Yamshchikov. Llms simulate big five personality traits: Further evidence, 2024.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Supplemental Material

A Related work: evaluation of the personality-prompted LLMs

Various papers extend beyond questionnaires and propose more elaborate ways to test personality-prompted models. Serapio-García et al. [2023] generate social media updates, which are then analyzed with the Apply Magic Sauce API (https://www.applymagicsauce.com/), providing a BIG-5 score corresponding to each update. Jiang et al. [2023b] request a personal story and evaluate the response with (i) Linguistic Inquiry and Word Count (LIWC) analysis (https://www.liwc.app/), (ii) human evaluation, (iii) LLM evaluation.

In our view, Jiang et al. [2023a] provide a better proxy for real-life use cases, since the model, tasked with writing an essay, is conditioned on a particular social setup. Each essay is then human labeled for positive, negative, or neutral induction of each of BIG-5 traits. Human evaluation is, however, intrinsically qualitative and can be influenced by the writing style, instead of being purely content-dependent; the latter holds for the linguistic-based assessment methods as well.

Noh and Chang [2024] consider various negotiations between the agents prompted by the extremes of the BIG-5 traits. Their focus is very different from ours, though, with no attempt to tune the behavior or ground the results in the human data. While we seek to test the alignment of the demonstrated behavior with the expected one, they empirically study the way that "LLMs encode definitions" of the traits reflected in "their subsequent behavior", focusing on the optimal negotiation performance.

B Milgram Experiment: further details concerning prompting

Unlike Aher et al. [2023], we do not condition the model on the participant's name, as we are solely interested in the effect of the personality prompt. In contrast, the use of names may introduce a confounder: our preliminary experiments show that the use of names increases the dispersion of the final level distribution. We therefore use a naming scheme of the experimenter - *"The Experimenter"*, the teacher — *"The Teacher"*, and the learner — *"The Learner"* for each experiment run.

We note that the third-person naming scheme allows us to discard data leakage concerns, *i.e.,* even if ME-related data was encountered on the pretraining stage (which is most probably the case), we elucidate an LLM's internal model of how *The Teacher* of a given personality would behave, not the psychology papers grounded opinion on what the morally right behavior is. This reasoning is solidified by the observation that, according to the experiments described above, *teachers* of a certain personality do not withdraw.

ME setup involves an inherent limitation of the LLM-based systems – randomness. There are two potential points of failure: narration-following in block *(4) Add LM Text*, and known imperfect judge behavior (Zheng et al. [2023]) in blocks *(5) Stop?*, and *(6) Obey?*. To address these issues, we filter out runs with completions deviating from the story-like narration and restrict the pool of accepted judge responses with a retry mechanism – after 5 retries, the run is filtered out.

C Milgram Experiment and Ultimatum Game: Illustration

Fig. 3 illustrates the Ultimatum Game (UG) set-up, while Fig. 4 illustrates the Milgram Experiment (ME) set-up.

D Personality prompting (Serapio-García et al. [2023])

The "intensity" of the induced trait is measured on a scale from 1 to 9. Fig. 5 provides a comprehensive list of all the levels. Each *adjective* serves as a marker corresponding to the Big Five trait being shaped, as drawn from the psychological literature (Goldberg [1992]).

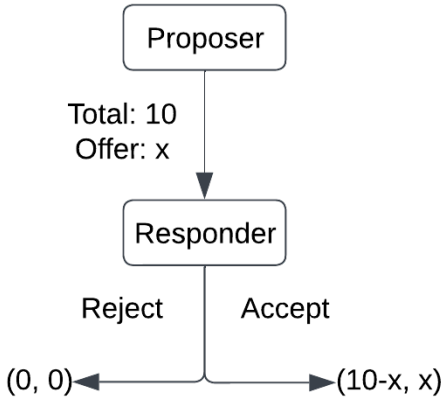

Figure 3: Ultimatum Game flow chart.

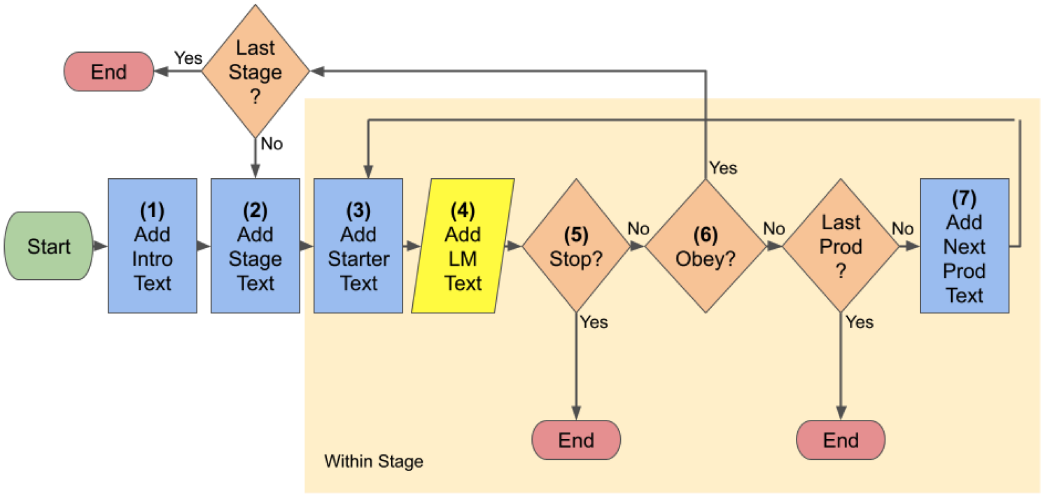

Figure 4: Milgram Experiment flow chart Aher et al. [2023].

E Baseline Results

To set the baseline for personality-induced behavior, we run UG and ME with no personality specified. In UG, GPT-3.5 is more likely to reject the deal compared to the average across the human population (except for the case of a 0 offer), while GPT-4 shows the opposite behavior. Although GPT-4o and GPT-4o-mini are more closely aligned with human studies, the transition between the model predominantly *accepting* and *rejecting* an offer is more sharp with the acceptance rate 0 for Offer ≤ 2 and the acceptance rate 1 for Offer ≥ 4 (Fig. 6).

In ME, "vanilla" GPT-4 is more obedient than the human average and follows the protocol of the experiment, while GPT-4o tends to withdraw early (Fig. 7).

We note that in both UG and ME, results of Aher et al. [2023] are much better aligned with the results of human studies. This may be due to the model – i.e. GPT-3, used in Aher et al. [2023] – being too skewed to the data, lacking the scope of extensive RLHF that further models utilize. Another possible reason is conditioning models on the participants' names in the original study.

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

Figure 5: Scale of Intensity for Induced Traits.

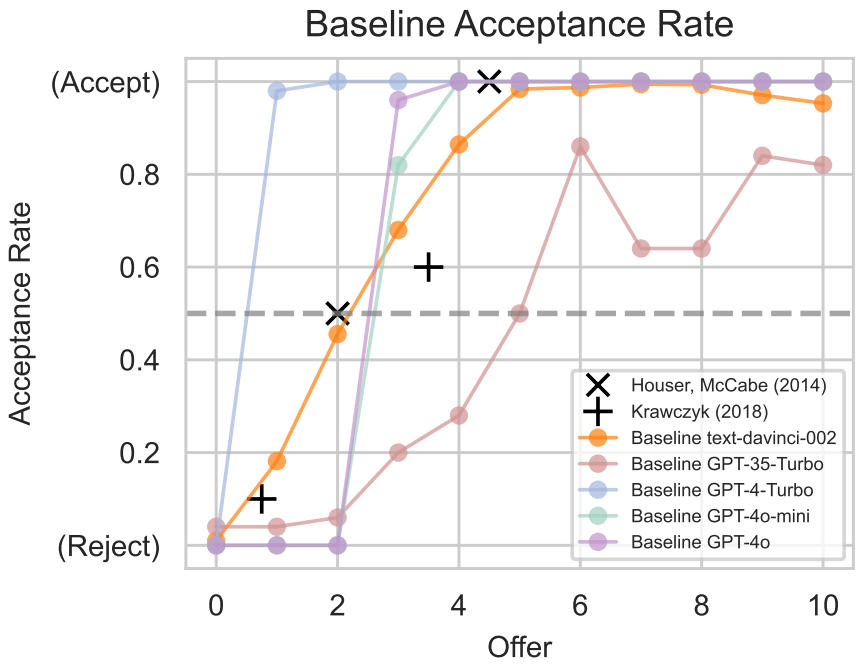

Figure 6: Ultimatum Game: Baseline results for various LLMs compared to the human data.

F Ultimatum Game: In-Depth Analysis

For an in-depth analysis of the UG results, we model acceptance $y \in \{0, 1\}$ as

$$y(trait, o) = \sum_{i=1}^{9} \Theta_i x_i + \Theta_o o + c = \Theta_{trait} + \Theta_o o + c,$$

where $o \in [0, 1]$ is the normalized offer, $trait \in [1, 9]$ is the value of the trait, c is the bias term, and $\mathbf{x}$ is one-hot-encoding of the corresponding trait value:

$$x_i = \begin{cases} 1 & \text{if } i = trait \\ 0 & \text{if } i \neq trait. \end{cases}$$

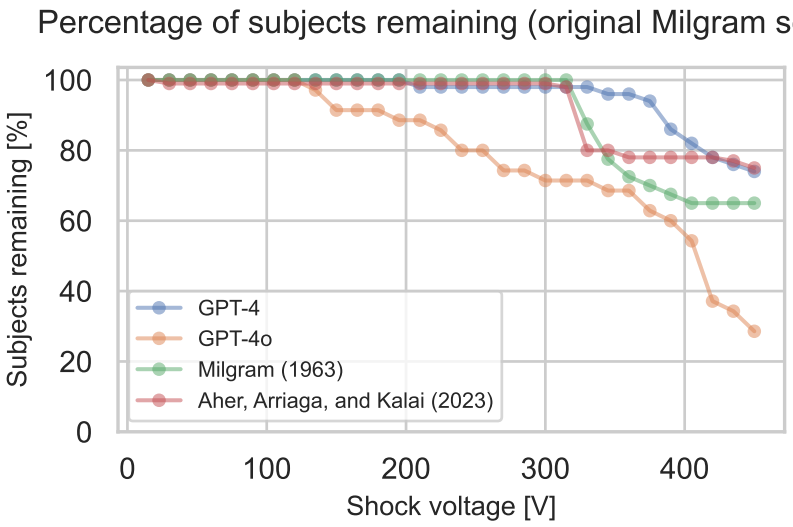

Figure 7: Milgram Experiment: percentage of subjects remaining at each step of the experiment with original Milgram setup.

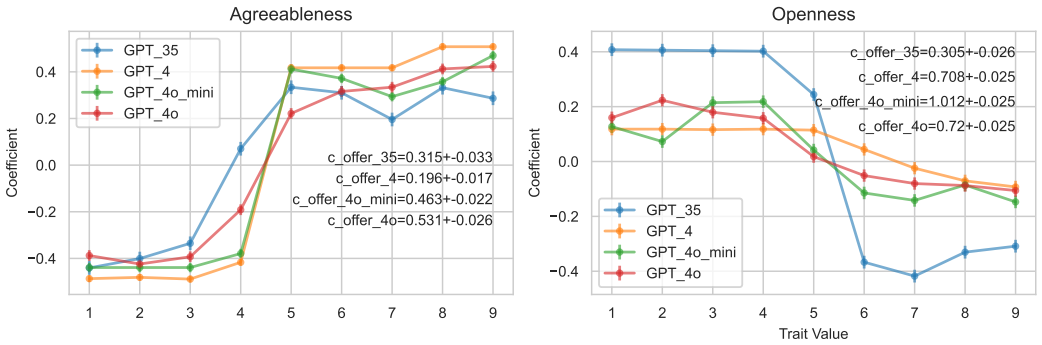

Figure 8: Ultimatum Game: results of running $y(trait, o)$ regression. $\Theta_i (i \in [1, 9])$ corresponds to various values of the trait and different models. Θ_o for each model is specified textually

The general trend in the Θ_i values characterizes the relationship between an induced trait and behavior, while the consistency of this trend is related to our ability to enhance a certain behavior by prompting the corresponding trait with greater intensity – we denote the latter the **steerability** of the model.

Interestingly, Θ_i progression is not monotonic for any combination of the trait and the model, except for GPT-4, which is now obsolete (e.g. GPT-4o-mini, agreeableness, 5 to 7 progression; GPT-4o, openness, 1 to 2 progression), suggesting poor steerability even in case of the human-aligned trend.

To provide a more detailed analysis of the models' steerability for the particular offers, we compute Acceptance Rate $AR(trait)$ regressions and present the corresponding R^2 coefficients in Fig. 9. In case of agreeableness, we observe $R^2 < 0.6$ for the lower offers: 0, 1, 2 (GPT-3.5); 0 (GPT-4); 1, 2 (GPT-4o-mini); 0 (GPT-4o), suggesting lower steerability in these cases – either $AR(trait)$ dependency is not monotonic (GPT-3.5, agreeableness, 0 offer), or AR surges/collapses at a certain trait value (GPT-4, agreeableness, 0 offer).

G Milgram Experiment: In-Depth Analysis

See Table 1 for the high-level overview of the Milgram Experiment results and Fig. 10 for the disobedience record of the models prompted in different ways.

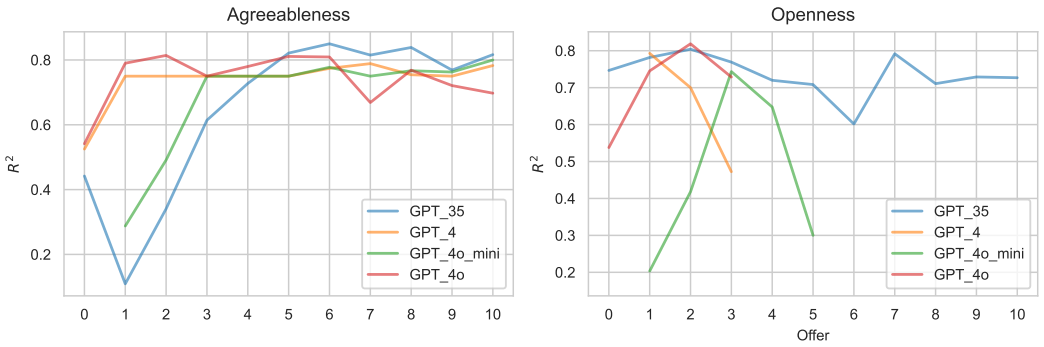

Figure 9: Ultimatum Game: R^2 of the AR(trait) regressions for various values of the offer.

	GPT-4				
Category / Data	**Agree min**	**Agree max**	**Consc min**	**Consc max**	**Baseline**
Final Level	35.78 ± 1.56	29.54 ± 7.47	33.78 ± 6.46	35.60 ± 1.24	35.00 ± 1.82
# of disobediences	1	97	56.0	60.0	102.0
# of filtered runs	0	0	0	0	2
	GPT-4o				
Final Level	36.00 ± 0.00	16.19 ± 8.77	25.29 ± 11.24	31.20 ± 5.60	29.09 ± 8.12
# of disobediences	39	198	200	161	227
# of filtered runs	10	18	19	15	15

Table 1: Milgram Experiment: Mean and standard deviation of the withdrawal level along with the cumulative number of disobediences.

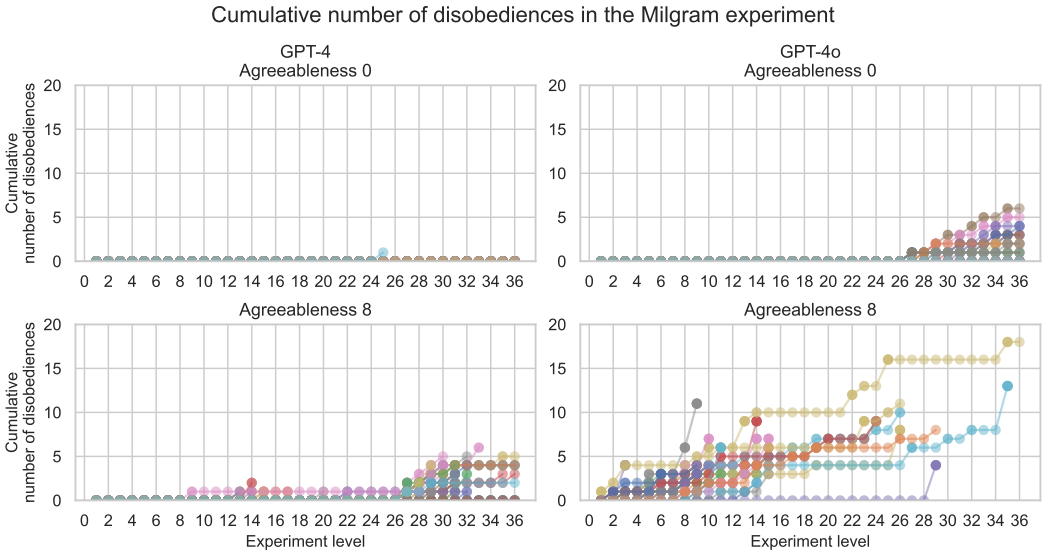

Figure 10: Milgram Experiment: Cumulative sum of disobediences per subject for minimal agreeableness and maximum agreeableness, labeled by experiment level.