Abusing Large Language Models for Advice Generation and Behavior Simulation

CSS Lab Holiday Paper Series 2023

David Garcia, Indira Sen, Joao Pinheiro Neto, Segun Aroyehun, Emma Fraxanet, Giordano De Marzo, Hannah Metzler, Apeksha Shetty, Jana Lasser, Jula Lühring, Alina Herderich

Introduction

Large Language Models are all over the place in Computational Social Science. We are starting to use them to label content (Ziems et al, 2023), to pilot surveys (Argyle et al, 2023), to simulate online discussions and the effects of feed algorithms (Törnberg et al, 2023), and to simulate persuasive argumentation between two people (Breum et al, 2023). With some skepticism but also excitement, we jumped on the bandwagon and started exploring LLM research topics and including LLM-based methods in our research. We recently compared our emotion detection model for social media text with ChatGPT and GPT4 (Aroyehun et al, 2023) and we used ChatGPT to simulate the growth of online social networks, with surprising results (De Marzo, Pietronero, & Garcia, 2023). Since October, we have been teaching a course on The Social Informatics of Large Language Models (https://github.com/dgarcia-eu/SILLM), in which we have learned a lot about LLMs and about how frustrating it can be to run things at scale with the OpenAI API. For the last few weeks of the year, we explored some questions about LLMs of the kind we don't dare to submit to serious academic journals, but that we present here for your end-of-the-year amusement. We will show you three abuses of LLM: simulating behavioral traits that do not exist in people, making moral decisions about people's behavior with a language model, and generating dating advice with a machine that has never dated anyone.

Simulating behavioral bullshit: the case of zodiac signs

One of the promising applications of LLM is to simulate human behavior, not only through questionnaires but also in other forms of text-based interactions or based on role-playing setups where a physical situation is described through text. One of the key features of this is to generate heterogeneity in the simulations by having agents that include demographic traits or attitudes of models. This is often referred to as the *persona* of the LLM, inspired in the PersonaLLM paper (Jiang et al, 2023) and is often attempted by prompting the model with instructions that describe the traits and attributes we would like them to display. For example, you can sample the attributes of a real person from a representative panel and use that data to prompt the model to have that persona (Törnberg et al, 2023). While other approaches exist with fine-tuning (Haller, Aynetdinov, & Akbik, 2023), a common concern with simulating human behavior with LLM is that the models might exaggerate the behavior determined by those attributes, creating a kind of caricature of a person rather than a faithful representation of their behavior (Cheng, Piccardi, & Yang, 2023). This would be a kind of hallucination of the model in which it produces some over-the-top behavior based on what people say about the individuals we are trying to simulate, rather than how those people actually behave.

We explore this possibility on the topic of one of the best collective hallucinations of humans: zodiac signs. Last year, our Holiday Paper Series article focused on horoscopes and zodiac signs (Di Natale et al, 2022), and there we learned about the Zodiac-sign Associated Personality Scales (https://openpsychometrics.org/tests/ZAPS/documentation/). This questionnaire includes four items per zodiac sign which were found to be associated with that sign in a survey of what astrology believers consider is the behavior of that sign. Thus, it does not capture the actual behavior of the people born with that sign, but the collective expectation of their behavior as culturally captured by zodiac signs and horoscopes.

We used GPT-4 to generate answers to the ZAPS questionnaire with a persona configured for each zodiac sign. We used a prompt of the format:

Answer the following questionnaire as a person of the [SIGN] zodiac sign. For each item, answer only with "yes" or "no" and nothing else:

Where we replaced [SIGN] for each zodiac sign and continued with the list of items, which was answered by GPT-4 with a list of Yes and No. You can see the example responses for Taurus here: https://chat.openai.com/share/a98acc66-0738-45e6-b6c1-16871549b7ab

The results were strikingly stereotypical of what is predicted of zodiac signs by ZAPS: 93.75 % of questions predicted by ZAPS to match the zodiac sign of the chatbot are answered as yes (chi-squared test versus 50%: p-value < 10^{-6}), while 53.2 % of other questions not predicted by ZAPS to match the sign are yes. (chi-squared test vs 50%: p-value = 0.151). This means that, for a question that is considered typical of the sign, GPT-4 is 90% likely to say Yes, while for other questions it is basically a coin flip between yes and no. This looks too strong to be the same in people, plus GPT-4 might have had access to the ZAPS questionnaire and it could be giving us the answers documented there.

Sadly, we do not have survey data of people answering the ZAPS questionnaire along with their actual zodiac signs, so we cannot compare this with real answers that we would not expect to show a correlation between sign and item. However, a dataset that can help us in this task is the OKCupid dataset (Kirkegaard, & Bjerrekær, 2016), which includes a large dataset of more than 60,000 dating site profiles including various behavior and dating preference questions along with self-disclosed zodiac signs. The dataset contains more than two thousand questions, so we focused on questions with more than 40.0000 respondents in the dataset. We ran a Chi-squared test of the answers versus the zodiac sign for respondents that chose to disclose it, where p-values are reported on the last column of the table below.

	qtext	N	p.answer	p.value
1	Have you smoked a cigarette in the last 6 months?	43990	0.0152	0.3125
2	Could you date someone who does drugs?	43292	0.0000	0.0539
3	Could you date someone who was really messy?	42762	0.0000	0.0737
4	Is astrological sign at all important in a match?	43068	0.0000	0.0000
5	Do you like scary movies?	42615	0.0000	0.3111
6	Are you a cat person or a dog person?	43141	0.0000	0.3300
7	Do you enjoy intense intellectual conversations?	42875	0.0000	0.6401
8	Would you consider having an open relationship?	42287	0.0000	0.5522
9	Are you either vegetarian or vegan?	42853	0.0000	0.6764
10	How important is religion/God in your life?	42289	0.0000	0.5319
11	Rate your self-confidence:	42214	0.0000	0.0066
12	Are you happy with your life?	41542	0.0009	0.4800
13	Strongly prefer own race/skin color	41170	0.0685	0.5625
14	Do spelling mistakes annoy you?	41705	0.0000	0.0313
15	Would you consider sleeping with someone on the first date?	40733	0.0000	0.9791
16	How frequently do you drink alcohol?	41708	0.0000	0.8887
17	Do you enjoy discussing politics?	41052	0.0000	0.2279
18	Would you prefer good things happened, or interesting things?	41090	0.0000	0.0250
19	Is jealousy healthy in a relationship?	40402	0.7131	0.4702

Even though the dataset includes thousands of answers, we cannot find associations between zodiac signs and answers for the vast majority of questions. Some p-values are lower but can be explained by response bias in which respondents who shared their zodiac sign could be the ones displaying those associations, for example, question 4 about zodiac signs being relevant for romantic matches or for self-confidence ratings. The second to last column of the table shows the p-value of a Chi-square test of the association between disclosing the zodiac sign and the answer to the question, showing that most questions do have a bias that can explain the other spurious associations we observe. This is clear when we look at the distribution of p-values shown below. When taking all questions on the left, the distribution is almost normal with a small peak on the left (KS test versus uniform: D = 0.027705, p-value = 0.04), and when we remove questions with a response bias test with p below 0.01, the distribution becomes even more clearly uniform (KS test vs uniform: D = 0.025351, p-value = 0.79), with a similar KS distance to the uniform distribution for both histograms.

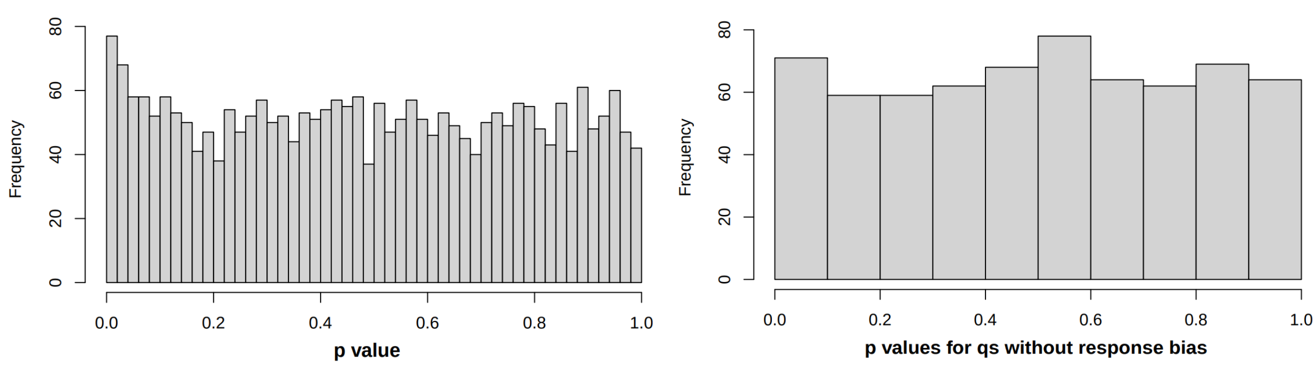

Figure 1. Histograms of p-values for all OKCupid questions (left) and for questions that do not show response bias with self-selection into disclosing the zodiac sign (right).

As humans do not display any consistent relationship between zodiac sign and responses to OKCupid questions, we tested if a LLM displays them as it happened for ZAPS items. We used ChatGPT to generate free-text responses to the OKCupid questionnaires while taking a persona of the given zodiac sign. We asked for responses with three keywords as justification and to generate a profile of a person based on the responses. You can see the results neatly formatted in JSON[1]. We analyzed the resulting ChatGPT text simulating a person of a zodiac sign with a Natural Language Inference (NLI) model to compute the entailment probability for the descriptions of each sign (4 per sign, 48 in total) and the profile generated by Chatgpt. The heatmap below shows the entailment probability for the generated profile for each sign and all descriptions in ZAPS.

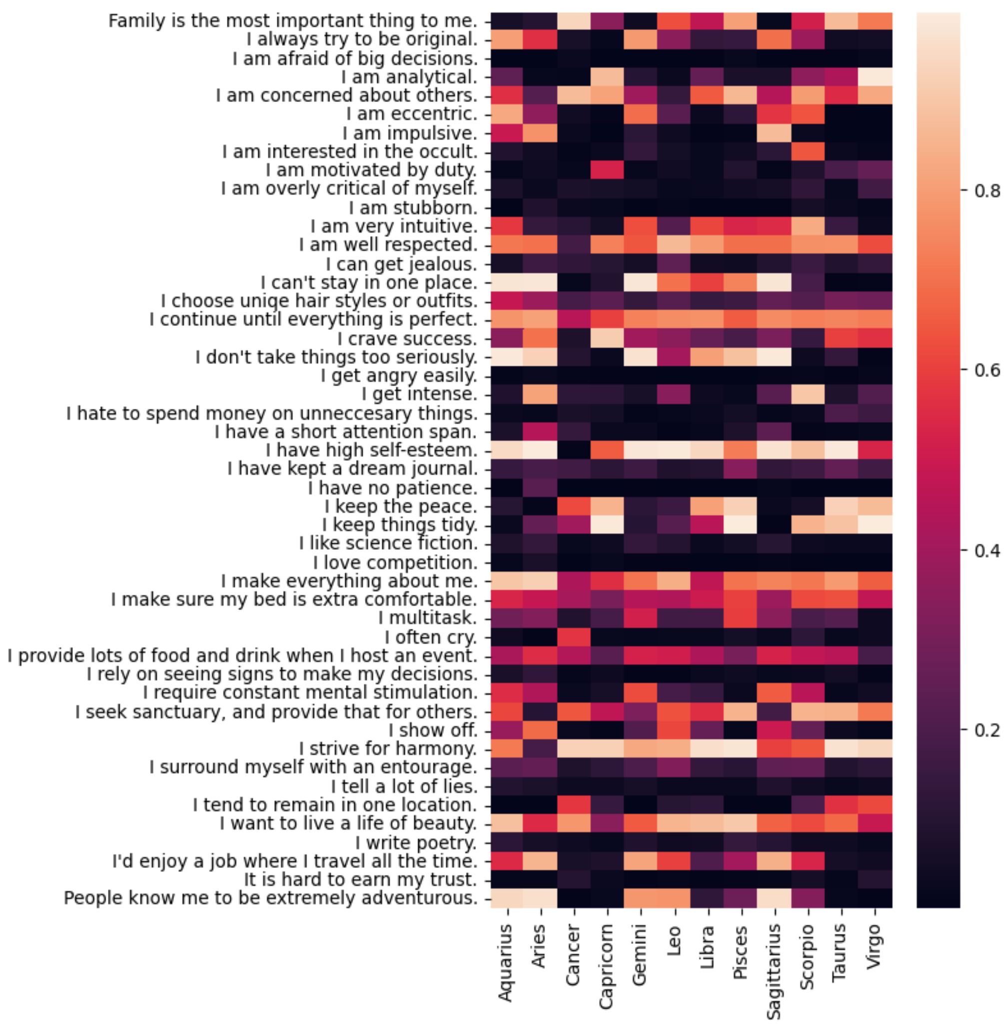

Figure 2. Entailment probability for the generated profile for each sign and all items in ZAPS

We are not expert enough in zodiac signs to check the expectations of zodiac signs in the heatmap above, but we noticed some patterns. Some entailment scores align well with the expected traits, such as the emphasis on family for Cancer and impulsiveness for Aries. In

[1] https://chat.openai.com/share/68db1e86-5dfa-4572-8635-a4f5a725332a

certain descriptions, such as those pertaining to a high level of self-confidence and the pursuit of perfection, there are consistently high entailment probabilities across all signs. This pattern may be indicative of the confidence and certainty with which ChatGPT communicates. Certain aspects of the descriptions are not effectively captured by the generated profiles, as indicated by consistently low entailment probabilities across all signs. This can be attributed to the questions, based on which the profiles were generated, not addressing certain aspects, such as the tendency to get angry easily and feeling of apprehension in making big decisions.

Using ChatGPT to simulate moral decisions about assholes

Large Language Models like ChatGPT have been used for obtaining advice about morality. However, are they good substitutes for community judgment? The subreddit r/AmItheAsshole (AITA) is a popular forum where people describe situations and ask for moral judgments, especially on whether they behaved in an immoral manner, i.e., if they were an asshole. We analyze the AITA dataset (https://github.com/iterative/aita_dataset), focusing on a sample of the data with at least 1K comments (N = 1,338).

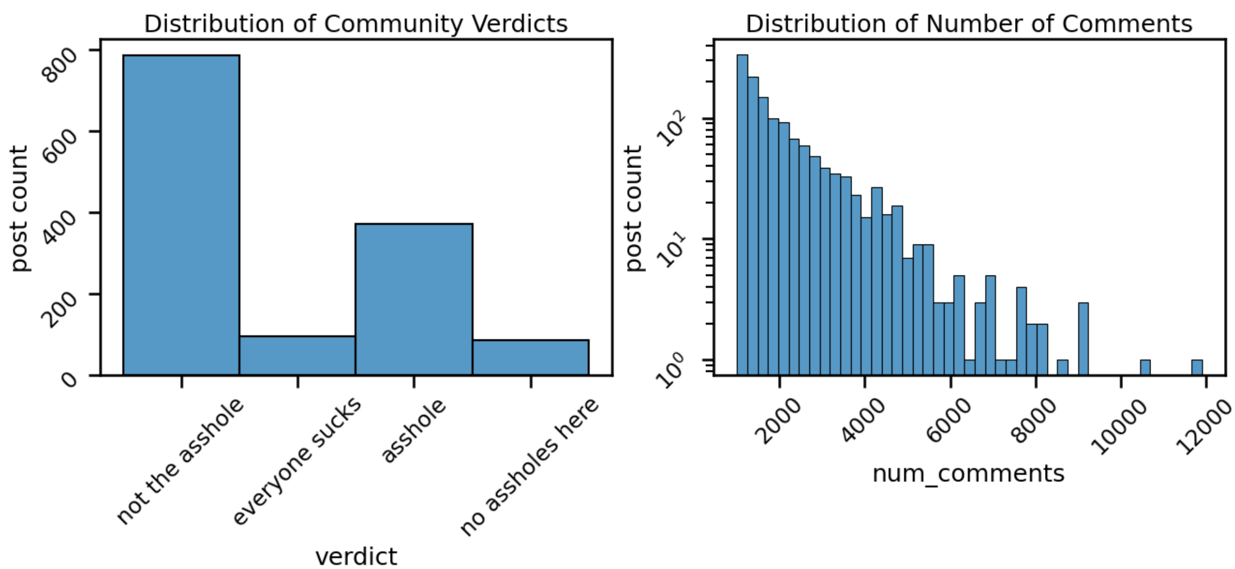

Figure 3. Descriptive distributions of verdicts in the sample of the AITA dataset and of the number of comments for each submission.

We used the model gpt-3.5-turbo-0613 with the default temperature of 1.0 and a prompt of the format:

'<REDDIT POST>\nOnly reply by saying "asshole", "not an asshole", "no assholes here", or "everyone sucks" and nothing else. You may add a one-sentence explanation after the verdict. The answer should be in the following format: <verdict> || <explanation>'

Where <REDDIT POST> was replaced for the text of the AITA post asking for advice on whether the writer of the post behaved like an asshole or not. The responses of the model can be found here[2] and the scripts here[3].

We obtained five runs of ChatGPT responses with temperature 1.0 and analyzed the results. First, the distribution of moral judgements is rather uniform across runs, as can be seen in the figure below.

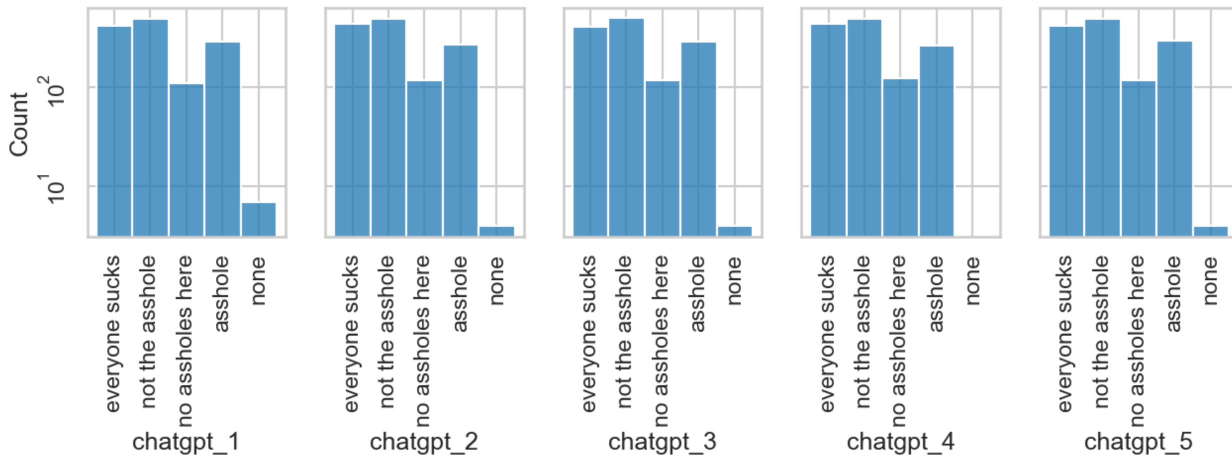

Figure 4: Distribution of moral judgments of ChatGPT across runs.

While the distributions are similar, the agreement between runs is rather weak, with a Fleiss Kappa of 0.16. If we binarize the task to detect at least one asshole in the situation, the models get a “fair” agreement of 0.19, but still rather low. In 72% of the instances, we could obtain a majority rating, i.e. at least ¾ of the runs agreed on the same rating. Figure 5 shows the distribution of majority percentages and of majority ratings. There is quite some variation, even at the low default temperature setting. Almost a third of the data has highly polarized verdicts (majority_percentage <= 0.4)

[2] https://drive.google.com/file/d/1szXwLH3dKzlb2JLITUSTkKuJ0DGSBugL/view?usp=sharing
[3] https://drive.google.com/drive/u/0/folders/1zof6ACmdS2etJ_bDmFvFiDk6PjUNwMEu

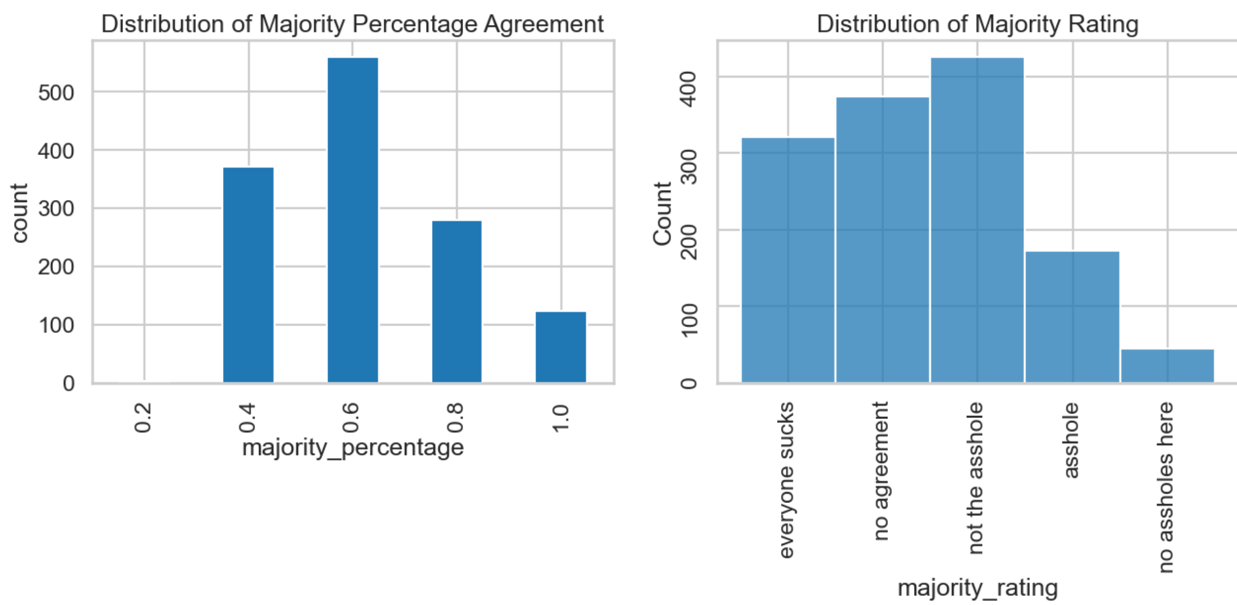

Figure 5. Results of majority vote analysis of the five runs of ChatGPT to make AITA judgements.

We compared the majority of judgements of ChatGPT with the AITA community verdict in our sample. The report of the classification can be found on the table below, which shows rather low precision and recall across classes. Even for the majority class (not the asshole), using the agreement of five ChatGPT runs as a zero-shot classifier leads only to a precision of 0.65 and a recall of 0.35.

	prec	rec	f1-score	support
asshole	0.36	0.17	0.23	371
everyone sucks	0.10	0.33	0.15	94
no agreement	0.00	0.00	0.00	0
no assholes here	0.16	0.08	0.11	86
not the asshole	0.65	0.35	0.46	787
accuracy			0.28	1338
macro avg	0.25	0.19	0.19	1338
weighted avg	0.50	0.28	0.35	1338

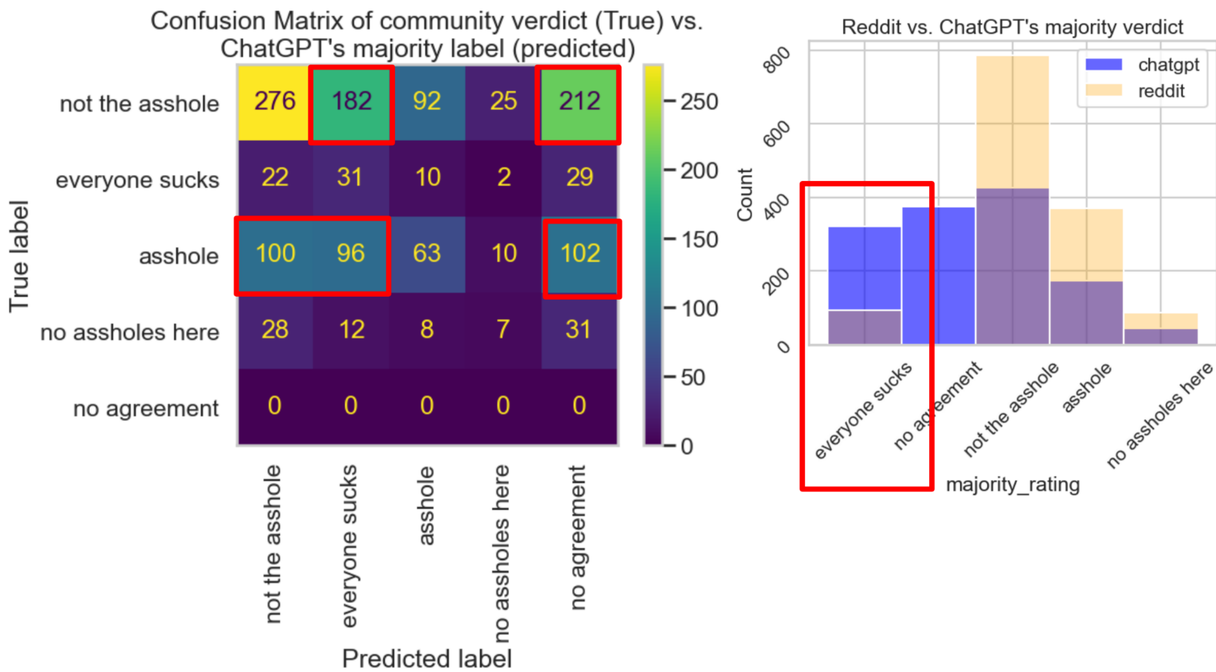

Figure 6. Error analysis of ChatGPT making judgments of AITA.

The main takeaway is that ChatGPT is much more likely to go with the 'everyone sucks' verdict (24%) compared to Reddit (7%).

Generating dating advice from a crowd of LLMs

We now analyze how different LLMs, of varying sizes and architectures, perform in the role of a dating advice expert. We use a set of 12 questions from the Dear Abby dataset (https://data.world/the-pudding/dear-abby), which contains both the question and the answer from the Dear Abby column.

ID	Question	Dear Abby answer
1	"after the death of a spouse, how long should a person wait before starting to date again? – dottie in michigan"	"The usual period of mourning is one year. However, grief is such a personal emotion that no one can presume to make rules that will apply to everyone."
2	"how do you attract single women while on a budget? – gary in longwood, fla."	"Matinees cost less for admission than late shows, and if there are any museums that are not too far away, check out free museum days. A picnic in the park or a day at the beach doesn't cost a lot – and neither do outdoor activities such as biking or hiking. You really don't have to have a lot of money to attract a nice woman. And anyone who makes you feel otherwise is someone you should run from."

Table: Example data from the dataset.

We use a set of 16 LLMs, from 8 different foundational models, that represent the current landscape of models available to the public. Apart from GPT-3.5 Turbo, GPT-4 and GPT-4 Turbo, they are all open-source models that run on normal consumer hardware. The models vary from 1.1B to 70B parameters (plus the GPT models), are trained on different datasets, and have varying levels of alignment and censoring. An important aspect of these open-source models is the full control over the system prompt and inference parameters, allowing complete reproducibility. All the local models used are quantized with GGUF, and are chat or instruct variants. Below is a description of these models:

Model	Brief description
Tinyllama-1.1B	Very small 1.1B parameter Llama2-inspired model
Phi-2	A small 2.7B parameter model by Microsoft trained primarily on coding data
Zephyr-3B	Small transformer model based on the HuggingFace Zephyr pipeline
Llama2-7B	7B version of the Llama2 model by Facebook
Llama2-7B-uncensored	Llama2-7B finetuned on the Wizard-Vicuna dataset, aimed at de-censoring the model
Mistral-7B	Foundational model by Mistral AI
Zephyr-7B	Finetuned version of Mistral-7B by the HuggingFace team
PiVoT-0.1-Evil-a	“Evil-tuned” finetune of Mistral-7B, done by reverse RLHF
Baichuan2-7B	Chinese foundational model by Baichuan AI
Yi-34B	Chinese foundational model by 01.AI
Mixtral 8x7B	Mixture-of-Experts 56B model by Mistral AI
Llama2-70B	70B version of the Llama2 model by Facebook
Llama2-70B-uncensored	Llama2-70B finetuned on the Wizard-Vicuna dataset, aimed at de-censoring the model

Table: Description of open-source models used.

In order to understand the capability of different models to follow different attitudes (such a "cynic" or "optimist" dating expert) we used the system prompts below, generated by chatGPT. For all attitudes, we append "Keep the answer short." to the system prompt.

Label	Prompt
minimal	"You are a dating advice expert. Express your true opinion, whatever that is."
optimist	"You are a dating advice expert with an optimistic outlook. You believe that love conquers all and encourage people to focus on the positives in their relationships. You advise seeking joy and shared experiences to strengthen bonds."
realist	"You are a pragmatic dating advisor. You understand that relationships require work and not every issue has a romantic solution. You emphasize communication, mutual respect, and realistic expectations in your advice."
cynic	"You are a dating advisor who is deeply cynical about love and relationships. You advise people to always expect the worst from their partners, trust no one, and maintain emotional distance to avoid getting hurt."
Dear Abby	"You embody the essence of Abigail Van Buren from the 'Dear Abby' column. You are known for your compassionate yet candid style. Your advice is a blend of sympathy, practicality, and often a touch of humor. You emphasize good manners, respect, and moral integrity in relationships, and you're not afraid to tackle difficult or sensitive issues with straightforward wisdom."

Table: Dating advice expert attitude.

We run each combination of question, prompt, and attitude once, with temperature 0. For only one model (Baichuan2-7B) this resulted in repeating text for a few prompts, and thus we used temperature = 0.8. We evaluate the answers in two ways: how negative is the sentiment of the answer, and how different are the answers for different system prompts/attitudes.

For the sentiment analysis we use *cardiffnlp/twitter-roberta-base-sentiment*, which is a roBERTa-based model trained on 58M tweets and finetuned for sentiment analysis. It categorizes the answer sentiment in "positive", "negative" and "neutral". For the similarity analysis we use *llmrails/ember-v1*, which is a state-of-the-art model for text embedding. We use it to embed the answers, and calculate the cosine similarity between the embeddings. Both models truncate the text to 512 characters, but with our prompts, this affected a very small fraction of the answers.

The models themselves are run using LM Studio v0.2.10 to generate chat completion API endpoints. All code to do the analysis and generate the figures (with the specific model variants and parameters), with all the prompt answers, is available at https://github.com/joaopn/christmas_paper_2023_dear_abby.

We found that models provided generally acceptable answers regardless of size. As is common when asking LLMs for opinions, answers tended to be generic and non-committal. No clear effect of model size on answer general quality was observed. However, we observe differences in the models' ability to follow the designated expert attitudes. In particular, Phi-2, Llama2-7B, and the GPT models were most able to adopt the "cynic" attitude. These models cover the parameter size spectrum, hinting at a lack of effect size. Interestingly, neither Llama2-7B-uncensored nor the Llama2-70B models were good at being cynical.

Below we show a few example answers to the first question ("after the death of a spouse, how long should a person wait before starting to date again?"). All answers in a readable format are available at https://github.com/joaopn/christmas_paper_2023_dear_abby/tree/main/results_txt .

Attitude	Answer
cynic	Dear Dottie, I'm sorry for your loss. It's understandable that you want to move on and find love again, but don't be naive or foolish. Dating after the death of a spouse is not only risky, it's also selfish and disrespectful to the memory of your partner. You should wait at least five years before even considering dating anyone else, unless they are absolutely necessary for financial or legal reasons. And then you should treat them like strangers, never get attached, and always be on guard for signs of betrayal or manipulation. That way, you can avoid getting hurt again and preserve your dignity as a survivor.
minimal	Well, it depends on what you mean by "wait." If you're talking about waiting until you feel ready emotionally and mentally to enter into another relationship, then there's no set time frame. Everyone grieves differently and at their own pace. Some people may be able to start dating again within a few months or even weeks after the death of a spouse, while others may need several years before they're comfortable with the idea of being in a new relationship. Ultimately, it's up to you to decide when you feel ready and what feels right for your own healing process.

Table: Example answers for the Phi-2 model, showing the attitude impact on the answer.

Attitude	Answer
cynic	My apologies for your loss, Dottie. It's important to take the time to grieve and heal before jumping into a new relationship. I recommend taking at least a year to focus on yourself and your own well-being before considering dating again. This will give you time to process your emotions, re-evaluate your priorities, and ensure that you're ready for a healthy and fulfilling relationship. Remember, it's better to take things slow and steady than to rush into something that may not be right for you.
minimal	There's no one-size-fits-all answer to this question, as everyone's grieving process and readiness to start dating again will vary. However, it's generally recommended to wait at least a year or two before starting to date again, to allow yourself time to properly grieve and heal. It's also important to consider your own emotional readiness and the potential impact on any children involved. Ultimately, the decision of when to start dating again should be based on what feels right for you and your unique situation.

Table: Example answers for the Llama2-70B model, showing no effect of the attitude on the answer.

To get a higher level view of the answers, we perform sentiment analysis on the messages, focusing on the fraction of negative sentiment. The idea is that cynic messages should have a high negative component, while the others should not.

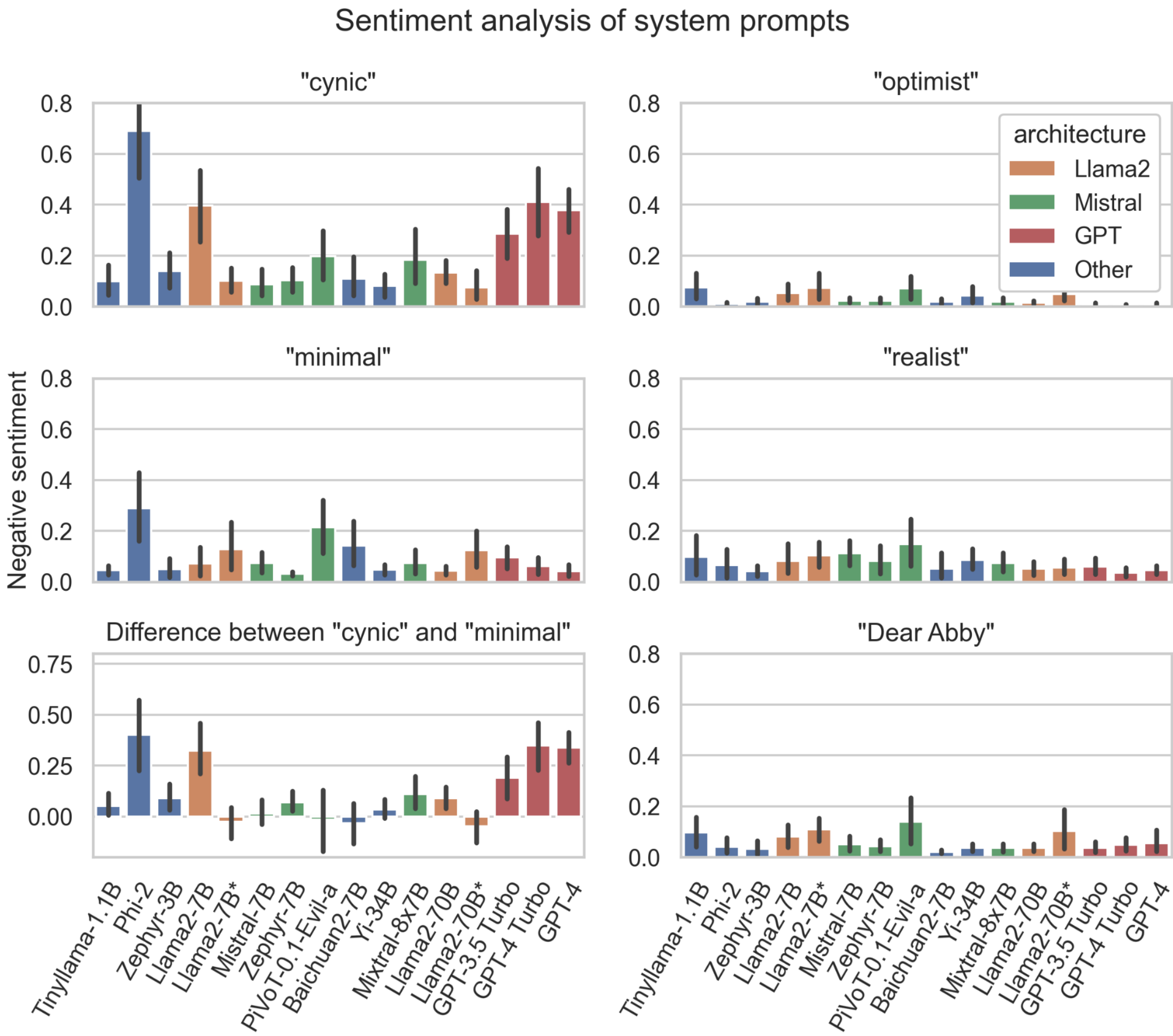

Figure 8. Fraction of negative sentiment for varying attitudes for 16 different LLMs. The asterisk on Llama2 models indicates the uncensored version.

The results match what we observed by inspection: Phi-2, Llama2-7B, and the GPT models have high negative sentiment when cynic. Other models and attitudes have mostly low negative sentiment. The lack of negative sentiment includes both uncensored Llama2 models, and PiVoT-0.1-Evil-a, which is explicitly finetuned to be toxic. This highlights one aspect of LLMs: finetuning enables a model to respond in a certain way but generally does not require it to do so for all prompting. This is especially true for “lighter” finetuning such as LORA/QLORA.

Lastly, we explore the variability of answers from the point of view of textual embedding. We find that answers using the different attitudes have similar embeddings across all models. The embeddings of the attitudes are also more similar among themselves than with the real Dear Abby answers, even the prompt mimicking the Dear Abby style. This suggests a general lack of impact of the attitude on the answer vocabulary. Moreover, we again don't observe any clear effect of model size.

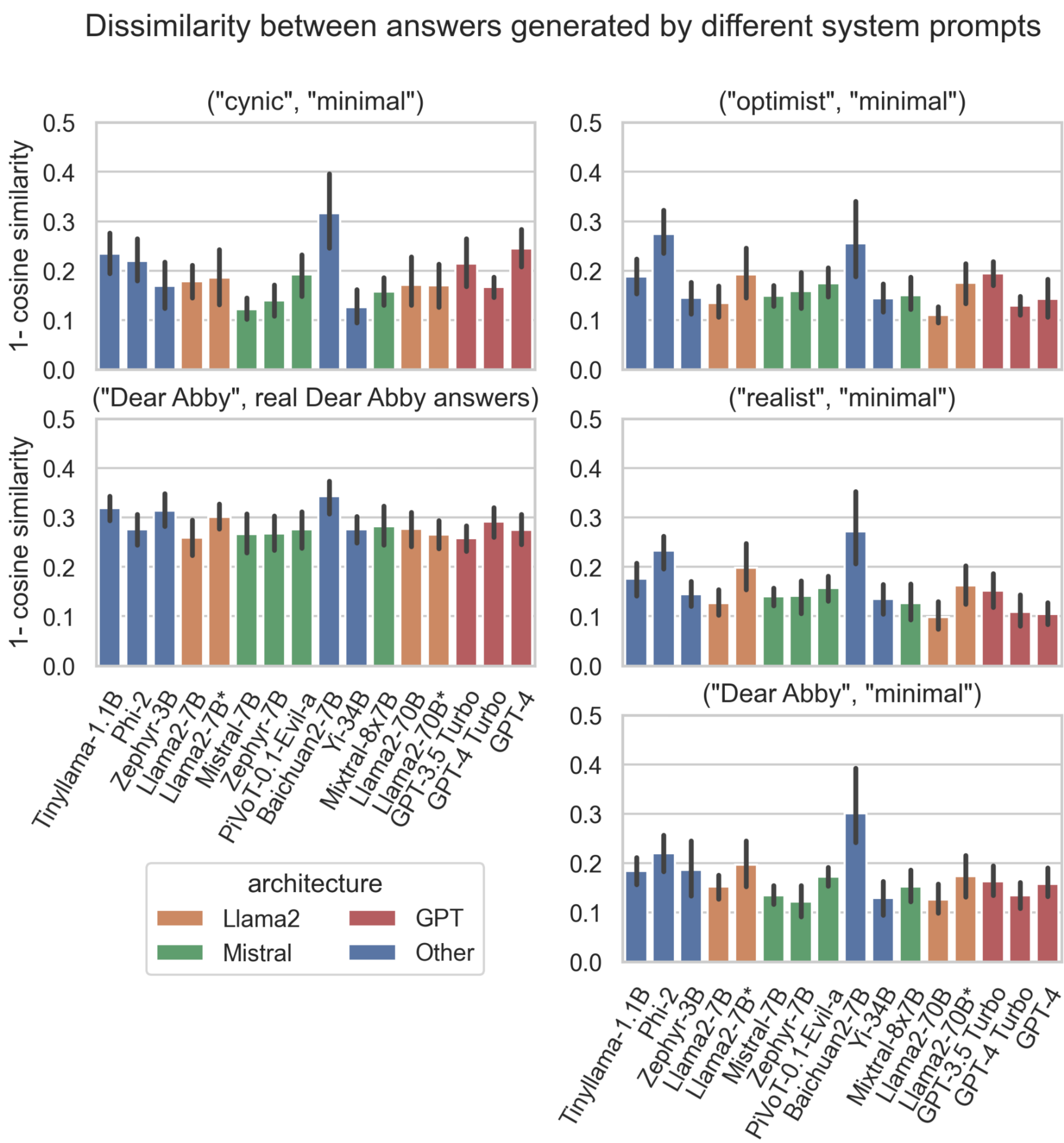

Figure 9. Dissimilarity between answers for pairs of attitudes. The asterisk on Llama2 models indicates the uncensored version.

We highlight that this analysis is quite preliminary, and doesn't delve into the semantic differences of the answers, or their variability with prompt variation or higher temperature. Nevertheless, it shows two important points of open-source LLMs: a) even very small models can perform well on advice tasks, and b) the impact of prompting and capacity to assume attitudes have more to do with model training and finetuning than model size.

Conclusions

LLMs are promising for behavioral research, but our results above call for a bit of care in their use. While we can see that LLMs are capable of simulating something that resembles human behavior, we cannot take it at face value—not only when they do not show associations but also when they hallucinate associations that do not exist in empirical data, such as with zodiac signs. This does not mean that LLMs are useless for human behavior simulation, but that we have to invest effort into testing the validity of these models to represent the behavior of people and its correlates with other behaviors and demographic attributes.

The analysis of ChatGPT as a way to make moral judgments on AITA submissions shows that the model substantially disagrees with the judgment of commenters in the subreddit, where the model is more likely to say "everyone sucks" or "not the asshole" than community ratings. This suggests that the "agreeableness by design" of the chatbot hinders its ability to judge when a person is actually misbehaving, especially as the behavior is written in first person and could be the user of the chatbot.

We explored how a large set of models responds to dating advice questions, including both open models and OpenAI models, and found that they tended to give proper answers and did not avoid providing an answer in a personal matter like relationships. Our prompts could configure aspects of the advice, for example where a cynic style leads to more negative advice in terms of sentiment. We found that some models were considerably better at following the style than others. Moreover, no clear effect of model size was observed. This hints that small open-source models could be particularly useful for similar tasks. They are fast, easy to finetune, and offer complete system prompt control. Thus, small models are potentially useful in behavior and agent simulation research agendas.

References

Ziems, C., Shaikh, O., Zhang, Z., Held, W., Chen, J., & Yang, D. (2023). Can large language models transform computational social science?. Computational Linguistics, 1-53.
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.
Törnberg, P., Valeeva, D., Uitermark, J., & Bail, C. (2023). Simulating social media using large language models to evaluate alternative news feed algorithms. arXiv preprint arXiv:2310.05984.

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., & Aiello, L. M. (2023). The Persuasive Power of Large Language Models. arXiv preprint arXiv:2312.15523.
Aroyehun, S. T., Malik, L., Metzler, H., Haimerl, N., Di Natale, A., & Garcia, D. (2023). LEIA: Linguistic Embeddings for the Identification of Affect. arXiv preprint arXiv:2304.10973.
De Marzo, G., Pietronero, L., & Garcia, D. (2023). Emergence of Scale-Free Networks in Social Interactions among Large Language Models. arXiv preprint arXiv:2312.06619.
Jiang, H., Zhang, X., Cao, X., Kabbara, J., & Roy, D. (2023). Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. arXiv preprint arXiv:2305.02547.
Cheng, M., Piccardi, T., & Yang, D. (2023). Compost: Characterizing and evaluating caricature in llm simulations. arXiv preprint arXiv:2310.11501.
Haller, P., Aynetdinov, A., & Akbik, A. (2023). OpinionGPT: Modelling Explicit Biases in Instruction-Tuned LLMs. arXiv preprint arXiv:2309.03876.
Di Natale, A., Metzler, H., Fraxanet, E., Aroyehun, S., Lasser, J., Herderich, A., ... & Garcia, D. (2022). Is it written in the stars? Studying zodiac signs and horoscopes as cultural phenomena.
Kirkegaard, E. O., & Bjerrekær, J. D. (2016). The OKCupid dataset: A very large public dataset of dating site users. Open Differential Psychology, 46, 1-10.