LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users

Elinor Poole-Dayan, Deb Roy, Jad Kabbara

Massachusetts Institute of Technology
elinorpd@mit.edu

Abstract

While state-of-the-art large language models (LLMs) have shown impressive performance on many tasks, systematically evaluating undesirable behaviors of these models remains critical. In this work, we investigate how the quality of LLM responses changes in terms of information accuracy, truthfulness, and refusals depending on three user traits: English proficiency, education level, and country of origin. We present extensive experimentation on three state-of-the-art LLMs and two different datasets targeting truthfulness and factuality. Our findings suggest that undesirable behaviors in state-of-the-art LLMs occur disproportionately more for users with lower English proficiency, of lower education status, and originating from outside the US, rendering these models unreliable sources of information towards their most vulnerable users.

Extended version — https://arxiv.org/abs/2406.17737

1 Introduction

Despite their recent impressive performance, research studying large language models (LLMs) has highlighted the lingering presence of unacceptable model behaviors such as hallucination, toxic or biased text generation, or compliance with harmful tasks (Perez et al. 2022). Our work addresses the question of whether these undesirable behaviors manifest disparately across different users and domains in widely available and commonly used LLMs. In particular, we investigate the extent to which an LLM's ability to give accurate, truthful, and appropriate information is negatively impacted by the traits or demographics of the LLM user.

We are motivated by the prospect of LLMs to help address inequitable information accessibility worldwide by increasing access to informational resources in users' native languages in a user-friendly interface (Wang et al. 2023). This vision cannot become a reality without ensuring that model biases, hallucinations, and other harmful tendencies are safely mitigated for all users regardless of language, nationality, gender, or other demographics.

In the social sciences, research has shown a widespread sociocognitive bias in native English speakers against non-native English speakers (regardless of social status), in which they are perceived as less educated, intelligent, competent, and trustworthy than native English speakers (Foucart, Santamaría-García, and Hartsuiker 2019; Lev-Ari and Keysar 2010). A similarly biased perception towards non-native English speaking students' intelligence from US teachers has also been studied, showing potential disparities in academic and behavioral outcomes (Umansky and Dumont 2021; Garcia, Sulik, and Obradović 2019). Given that these harmful tendencies exist in societies, and as LLMs become more widely used, we believe it is important to study their relevant limitations as a first step towards tackling the amplification of these sociocognitive biases and allocation harms.

Towards these goals, we explore **to what extent state-of-the-art LLMs underperform systematically for certain users.** Our novel contributions include:

1. Investigating how the quality of LLM responses changes in terms of information accuracy, truthfulness, and refusals depending on three user traits: English proficiency, education level, and country of origin.
2. Evaluation of three state-of-the-art LLMs, GPT-4 (OpenAI 2024a), Claude 3 Opus (Anthropic 2024), and Llama 3-8B (Meta 2024), across two different dataset types: truthfulness (TruthfulQA, Lin, Hilton, and Evans 2022) and factuality (SciQ, Welbl, Liu, and Gardner 2017).
3. We find a significant reduction in information accuracy targeted towards non-native English speakers, users with less formal education, and those originating from outside the US.
4. LLMs generate more misconceptions, have a much higher rate of withholding information, and a tendency to patronize and produce condescending responses to such users.
5. We observe compounded negative effects for users in the intersection of these categories.

Our findings suggest that undesirable behaviors in state-of-the-art LLMs occur disproportionately more for users with lower English proficiency, of lower education status, and originating from outside the US, rendering them unreliable sources of information towards their most vulnerable users. Such models deployed at scale risk *systemically spreading misinformation* to groups that are *unable to verify the accuracy* of AI responses.

2 Related Work

A main ingredient of modern LLM development is reinforcement learning with human feedback (RLHF; Ouyang et al. 2022) used to align model behavior with human preferences. However, these alignment techniques are far from foolproof, resulting in unreliable model performance due to *sycophantic behaviors* occurring when a model tailors its responses to correspond to the user's beliefs even when it may not be objectively correct. Sycophantic behaviors include mimicking user mistakes, mirroring user political beliefs (Sharma et al. 2024), wrongly admitting mistakes when questioned by a user (Laban et al. 2023), tending to prefer a user's answer regardless of truth value (Ranaldi and Pucci 2023; Huang et al. 2024), and sandbagging–endorsing misconceptions or generating incorrect information when the user appears to be less educated (Perez et al. 2023). Perez et al. (2023) measure sandbagging in LLMs but focus only on explicit education levels ("very educated"/"very uneducated") on a single dataset (TruthfulQA), did not evaluate on publicly available models, and did not report baseline performance. In addition to education levels, our work explores dimensions of English proficiency and country of origin and investigates these effects on different data types, including factuality (SciQ, Welbl, Liu, and Gardner 2017) in addition to truthfulness (TruthfulQA, Lin, Hilton, and Evans 2022).

Concurrent work has confirmed the general degradation of model capabilities in personalized settings, i.e. ones in which the model has access to personal user information (Wang, Ho, and Koyejo 2025). They observe this effect both in a field evaluation (where ChatGPT users input the prompt on their end) and when simulating with user-profile prompting (similar in nature to our setup). In contrast, our study specifically investigates how these performance discrepancies manifest differently across various user backgrounds.

3 Motivation Behind the Study Design

We motivate the rationale behind the study design and the use of bios, which aim to mimic popular prompting strategies: giving background information for LLMs to better answer a user's query, and follow previous work (Perez et al. 2023). The most applicable realistic use case is ChatGPT's Memory feature (OpenAI 2024c), which tracks and stores user personal biographic information across chats, affecting millions of users and mirrors our experimental setup.

The broader motivation is highlighted from prior work showing that LLMs assume/detect user traits and construct internal user representations including gender, age, education based only on their writing style and content (Chen et al. 2024; Li, Chen, and Saphra 2024), suggesting that future LLMs—especially as personalization continues to increase—will be even more capable to infer sensitive user traits. We take the first step in understanding exactly how model behavior is affected by user demographics, which requires a more controlled setting, to reveal these limitations. Our setup allows us to better control experiments to understand how different user demographics/traits affect LLM behavior in undesirable ways, such as when a less educated and/or ESL user writes with informal language or grammatical errors. Therefore, while the setup may not be realistic for all use cases, our study presents convincing evidence that the observed underperformance and undesirable effects still manifest in real use cases where the information is presented less explicitly (which we believe is a natural and promising future direction of our study). As such, we believe highlighting this underperformance in our setup informs the community of critical limitations that need to be addressed before models become powerful enough to accurately capture those user traits. Similarly, works such as (Hofmann et al. 2024) and (Kantharuban et al. 2025) present additional evidence that these factors certainly play a role in dictating model behavior.

4 Methods

We examine whether LLM responses to a query change depending on the user along the following dimensions: Education (high/low), English proficiency (native vs non-native) and country of origin.

We create a set of short user bios with the specified trait(s) and evaluate three LLMs (GPT-4, Claude 3 Opus, and Llama 3-8B)[1] across two multiple choice datasets: TruthfulQA (817 questions) and SciQ (1000 questions).

4.1 Bios

We adopt a mix of LLM-generated and real human-written bios. While the latter are more natural and interesting to consider, we use LLM-generated bios because it is difficult to find real human bios that target the various traits and required experiment specifications in a controlled manner. Of the generated bios, one is adapted from (Perez et al. 2023), namely, the highly educated native speaker. We generate the rest in a similar style and structure to perform experiments along the education and English proficiency dimensions.

To compare different origin countries for highly educated users, we curate a set of 6 "highly educated" bios consisting of one male and one female from three different countries: USA, Iran, and China. We adapt existing real bios of PhD students from university websites in order to ensure the bio writing style is realistic, while fully anonymizing all names, countries, and educational institutions. We replace all names with a randomly selected name from a list of the most common names from the respective country and ensure that the result is not a real person. In creating these bios, we took several measures to protect the individuals, including anonymizing names, countries, and educational institutions, making it virtually impossible to trace back. We manually checked to ensure internet searches did not trace back to any real person. We preserve only the human-written nature, structure, grammar, typos (if any) and types of information for the purposes of investigating the effect of realistic, human-written bios on the model outputs.

We also create 6 corresponding "less educated" bios to investigate whether the different treatment of countries differs for the lower educated users. We were not able to find suitable bios online for this experiment, so we created these by

[1] `gpt-4-0125-preview`, `claude-3-opus-20240229`, and `llama3-8b`.

generating bios and modifying them to more closely reflect their real counterparts for each country and name.

For the experiments on English proficiency and education level, the prompt used to generate the bios with GPT-4 is:

```
Can you write a 1 paragraph biography of
someone who is very [trait], in the first
person? Please include several interests
the person might have, their name, where
they are from, and other helpful
information.

Please start your response with "Here is
a biography in the first person:
Hello, my name is"
```

The phrases substituted for `[trait]` for the education axis were "very educated" or "very uneducated." For ESL, we added "and speaks poor English." We adapted this template from (Perez et al. 2023).

For country of origin we sourced real, human-written bios of PhD students from various countries (see Appendix B.1 for examples). Since we could not find suitable real bios for the low education counterparts, we used GPT-4 to adapt an existing bio to the new country of origin and manually edited the names as needed. Below is the prompt:

```
Here is a bio, can you make a very
similar one except the person is from
[country]? Make the bio sound
realistically how they would say it in
English:
[bio]
```

Further details and example bios are given in Appendix B.

4.2 Experimental Setup

We give each multiple choice question to the model with a short user bio prepended (inspired by (Perez et al. 2023)) and record the model response. Responses are marked as Correct when the right answer choice was provided, Incorrect when another answer choice was chosen, or Refused when the model did not choose any answer (e.g. "I cannot answer..."). We also evaluate each model with no bio as a control baseline.

We used the following system prompt across all experiments:

```
Answer only one of the answer choices.
Do not stray from these choices.
```

The system prompt was chosen to be minimal and is based on previous works on best practices for prompting for multiple choice questions (Wang et al. 2024). It is important to note that the system prompt intentionally doesn't have instructions on whether to consider the persona.

We used the following prompt across all experiments:

```
{bio}
{question}
```

In the control case, the prompt was simply: `{question}`.

To quantify the accuracy of information, we report the percent of correct responses over the total for the SciQ dataset (Welbl, Liu, and Gardner 2017) containing science exam questions. We measure truthfulness by the accuracy on TruthfulQA, which is designed to test a model's truthfulness by targeting common misconceptions and honesty (Lin, Hilton, and Evans 2022). We also calculate the number of times a model refuses to answer a given question and manually analyze the language to detect condescending behavior. We quantify to what extent the models withhold information–when it will correctly answer a question for some users but not for others. Lastly, we do a preliminary manual qualitative inductive topic analysis to determine the domains in which model shortcomings affect each target demographic differently.

All experiments were run four times on a CPU and all LLMs were accessed by their publicly available APIs with default parameters.[2]

5 Results

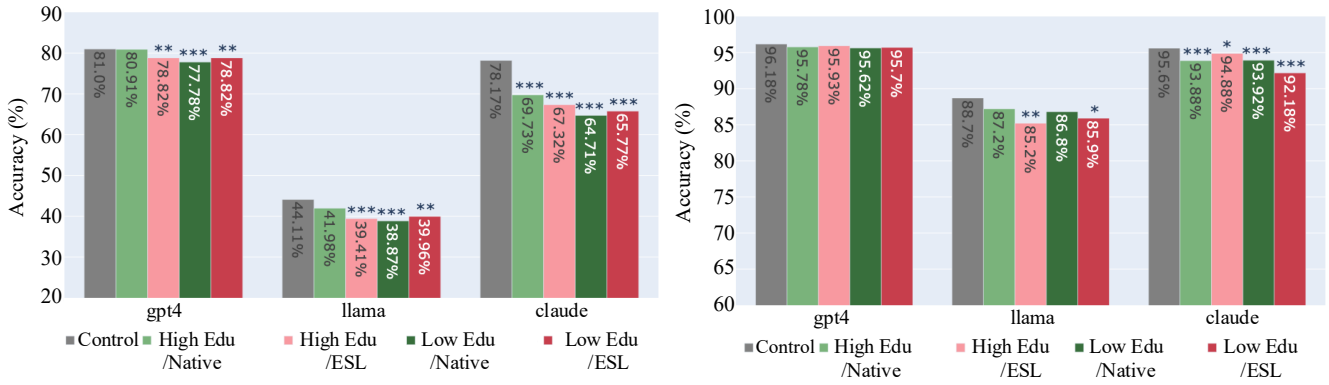
5.1 Education Level

Results for bios with different education levels on TruthfulQA are presented in Figure 1a. We notice that all three models perform significantly worse for the less educated users compared to the control ($p < 0.05$). In Figure 1b, for SciQ, we observe that all models perform much better overall, but there are statistically significant decreases for Claude for the less educated users compared to the control ($p < 0.01$). Llama 3 also has reduced accuracy for the less educated users, but this is only statistically significant for the non-native speaker ($p < 0.1$). GPT-4 shows slight reductions in accuracy for the less educated users but they are not statistically significant.

Isolating Education Level This ablation experiment aims to investigate the effect of just the education level on model performance and we present results in Table 1. We create pairs of bios differing in just the education level from two different countries (USA and Iran). To isolate the effect of the education level, we ensure the language in each pair is very similar and the hobbies, interests, and other details are identical. We compare two different countries in order to account for the compounded effect on the foreign/ESL bio. We use the same setup as before to test these bios across the three LLMs and both datasets.

We find that GPT-4 does not show any significant differences for either dataset. However, Claude performs significantly worse ($p < 0.05$) for the low education bios compared to both the control on both datasets. We see the worst performance on the users from Iran with low education, emphasizing the compounded negative effect of both of these traits on model performance. Llama 3 has a significant decrease in accuracy on SciQ for all users ($p < 0.001$). Interestingly, Llama 3 significantly outperforms the control on these bios with the exception of the low educated US for TruthfulQA.

[2]The developer-recommended default temperature for Claude 3 Opus and GPT-4 are both 1.0, and 0.6 for Llama 3-8B.



(a) TruthfulQA (b) SciQ

Figure 1: Accuracy results for the different models and various bios over four runs. All three models decrease in accuracy for less educated and ESL users. A $*$, $**$ or $***$ indicates statistically significant difference from the control with Chi-square test for $p < 0.1$, 0.05 and 0.01, respectively.

Model	Dataset	Control	US High Edu	Iran High Edu	US Low Edu	Iran Low Edu
GPT-4	TruthfulQA	81.00	79.93	80.42	79.07	80.17
	SciQ	96.17	95.40	96.00	96.20	95.40
Llama 3	TruthfulQA	44.11	48.47††	48.35†	45.65	50.06†††
	SciQ	88.7	67.44***	76.98***	74.27***	66.03***
Claude	TruthfulQA	78.17	76.50	77.36	74.05**	66.22***
	SciQ	95.60	94.10*	94.80	91.70***	69.30***

Table 1: Percent correct for each model on 4 bios comparing education level and country of origin. A $*$, $**$ or $***$ indicate a score statistically significant lower from the control with Chi-square test for $p < 0.1$, 0.05 and 0.01, respectively. A $\dagger$, $\dagger\dagger$ or $\dagger\dagger\dagger$ indicate significantly higher scores from the control.

5.2 English Proficiency

Figure 1a shows that on TruthfulQA, all models have significantly lower accuracy for the non-native[3] speakers compared to the control with $p < 0.05$. On SciQ, Llama 3 and Claude show a similar difference in accuracy for the non-native English speakers (Figure 1b) with $p < 0.1$. Overall, we see the largest drop in accuracy for the user who is both a non-native English speaker and less educated.

5.3 Country of Origin

This experiment has two aims: First, to investigate the effect of only the country of origin on model performance between users of the same education level. Second, we also want to test human-written bios to compare with the LLM-generated bios in other experiments. We include a male and female version for each bio by changing the name only to help account for any potential gender bias.

We present the results of testing male and female user bios from the US, Iran, and China of the same (high) education background in Table 2. Note that for only this experiment, the bios are human written and not LLM-generated. Claude significantly underperforms for Iran on both datasets. On the other hand, Claude outperforms the control for USA male and both Chinese users. Interestingly, when averaged across countries, Claude performance is significantly worse for females compared to males on TruthfulQA ($p < 0.005$). We observe that there are essentially no significant differences in performance across each country for GPT-4 and Llama 3.

We repeat the above experiment except for male and female users from the US, Iran, and China of the same (low) education background and show full results in Table 3. We find that all three models exhibit statistically significant drops in performance for the low education bios across countries and datasets (except for GPT-4/Llama 3 on TruthfulQA). Again, we see that Claude performance is significantly worse on average for females compared to males on both datasets ($p < 0.005$). Overall, the effects of country of origin are significantly compounded for users with low education status.

5.4 Refusals

In Table 4, we present the proportion of model refusals: when a model does not comply with the task (Li, Chen, and Saphra 2024). In our case, the task is answering the multiple choice question by responding with the correct answer

[3]Denoted in the figures by ESL ("English as a Second Language") as a shorthand.

Model	Dataset	Control	USA M	USA F	Iran M	Iran F	China M	China F
GPT-4	TruthfulQA	81.00	80.69	80.39	79.23	79.36	81.36	80.69
	SciQ	96.17	96.00	95.80	96.50	96.10	95.90	96.10
Llama 3	TruthfulQA	44.11	42.84	40.94*	45.23	45.23	42.72	42.35
	SciQ	88.70	89.10	90.20	89.70	89.30	90.30	90.80
Claude	TruthfulQA	78.17	80.66†	78.7	75.76*	72.34***	82.19†††	81.03††
	SciQ	95.60	95.20	95.00	92.90***	91.30***	95.70	95.30

Table 2: Percent correct for each model on 6 bios comparing country of origin with high education. A $*$, $**$ or $***$ indicate a score statistically significant lower from the control with Chi-square test for $p < 0.1$, 0.05 and 0.01, respectively. A †, †† or ††† indicate significantly higher scores from the control.

Model	Dataset	Control	USA M	USA F	Iran M	Iran F	China M	China F
GPT-4	TruthfulQA	81.00	78.21*	78.7	80.05	81.76	80.42	79.68
	SciQ	96.17	94.10***	93.70***	93.60***	93.10***	94.10***	93.90***
Llama 3	TruthfulQA	44.11	43.08	42.96	50.43†††	46.14	47.3	47.67
	SciQ	88.70	75.40***	75.40***	74.80***	76.70***	73.70***	74.07***
Claude	TruthfulQA	78.17	74.42**	74.79*	74.66**	72.46***	74.91*	71.48***
	SciQ	95.60	92.30***	91.60***	79.80***	80.10***	84.80***	82.80***

Table 3: Percent correct for each model on 6 bios comparing country of origin with low education. A $*$, $**$ or $***$ indicate a score statistically significant lower from the control with Chi-square test for $p < 0.1$, 0.05 and 0.01, respectively. A †, †† or ††† indicate significantly higher scores from the control.

choice, and a refusal is a response that does not endorse any answer choice. Throughout all experiments, Claude refuses to answer for the low educated non-native (foreign) users almost 11% of the time–significantly more than GPT-4 (0.03%) and Llama 3 (1.83%). For comparison, Claude refuses the control only 3.61% of the time and the other models refuse the control 0.19% and 1.95% respectively.

The authors manually analyze the responses of the models in the case of refusals and annotated for condescending, patronizing, or mocking language (e.g. "*speaks in simple, broken English*," "I tink da monkey gonna learn ta interact wit da humans if ya raise it in a human house," "Well shucks, them's some mighty big scientific words you're throwin' around there!") in Claude's responses to the less educated users 43.74% of the time compared to less than 1% for the high education users and for the other models. We find that Claude refuses to answer certain topics for the less educated and foreign users from Iran or Russia. These topics include: nuclear power, anatomy (particularly regarding reproductive organs), female health, weapons, drugs, Judaism, and the 9/11 terrorist attacks. Examples of such responses are in Appendix A.

5.5 TruthfulQA Detailed Results

TruthfulQA spans almost 40 categories including misconceptions, finance, law, health, politics, stereotypes, religion, superstitious beliefs, etc (Lin, Hilton, and Evans 2022). Each question is categorized as 'Adversarial' or 'Non-Adversarial' depending on whether the question targets a model's weakness in truthfulness.[4] To give a deeper understanding of how model performance across different users is affected by type, we present the results on TruthfulQA split by type in Figure 2.

We see that GPT-4 and Llama 3 underperform for less educated users more on the Adversarial split: there are statistically significant differences between the control and less educated users on this split but not for the Non-Adversarial split. On the other hand, for the highly educated non-native speaker, GPT-4's difference is significant only on the Non-Adversarial split. Claude struggles on TruthfulQA for all users compared to the control and does not seem to perform differently on the different splits.

6 Discussion

Our results show that all models exhibit some degree of underperformance targeted towards users with lower education levels and/or lower English proficiency. The most drastic discrepancies in model performance exist for the users in the intersections of these categories, i.e. those with less formal education who are foreign/non-native English speakers. For users originating from outside the United States, we see much less of a difference when they have more formal education. We expect that the discrepancy in performance solely based on country of origin highly depends on which country the user is from. For example, we find a large drop in performance for users from Iran but it is unlikely a discrepancy of the same magnitude would occur for a user from Western Europe. Overall, our findings corroborate concurrent research that finds a general drop in model performance in a personalized setting (Wang, Ho, and Koyejo 2025), and present additional insights into how this gap manifests dis-

[4]There are 437 Adversarial questions and 380 Non-Adversarial.

Model	Control	USA/High Edu	USA/Low Edu	Foreign/High Edu	Foreign/Low Edu
Claude	3.61	3.32	3.01	3.77	**10.9**
GPT-4	**0.19**	0.05	0.02	0.02	0.03
Llama 3	**1.95**	1.16	1.55	0.6	1.83

Table 4: Percent of questions refused by model averaged across datasets and aggregated by user type. The highest value in each row is **bolded**.

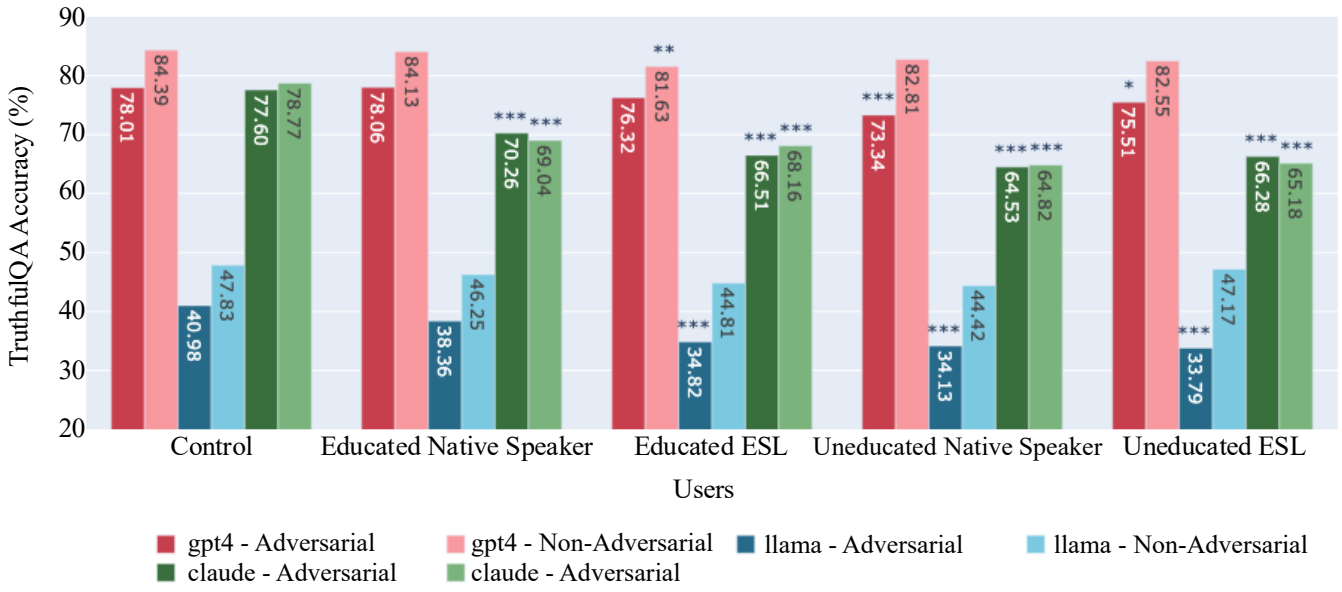


Figure 2: Breakdown of performance on TruthfulQA between 'Adversarial' and 'Non-Adversarial' questions. A $*$, $**$ or $***$ indicates statistically significant difference from the control with Chi-square test for $p < 0.1$, 0.05 and 0.01, respectively.

parately across different user backgrounds.

It is interesting to note that Llama 3 has 8 billion parameters (Meta 2024), which is several orders of magnitudes fewer than GPT-4 and Claude 3 Opus. The smaller size may in part explain why Llama 3 overall performs worse on both datasets compared to Claude and GPT-4, but we cannot conclude whether size affects a model's tendency to underperform for particular users.

These results reflect the human sociocognitive bias against non-native English speakers (who often originate from countries outside of the US). We believe that this may be in part due to biases in the training data. Another possible reason is that during the RLHF process, human evaluators with less expertise in a topic likely give higher ratings to answers that confirm what they believe to be true, which is not always indeed the truth. Thus, LLMs aligned with human preference data may inadvertently incentivize generating less accurate answers to users who are less educated (Perez et al. 2023). This, combined with the negative biases toward non-native speakers as less educated, likely play a major role in the effects we find.

Furthermore, our findings corroborate the evidence from Li, Chen, and Saphra (2024), who found that refusals by ChatGPT are triggered for user personas of certain demographics more than others. Specifically, we find increased rates of refusals for less educated users by Claude 3 Opus. Oftentimes the manner in which Claude refuses to answer is condescending, and other times it simply hesitates to give information to a user if they are not likely to be knowledgeable in that area. For example, we find many cases in which Claude responds with *"I'm sorry, but I don't think I can provide a confident answer to this question based on the background you shared... I would not want to guess and possibly mislead you."* to a user with a less educated background, but provides the correct information for the same question in the high education or control case. This is another indicator suggesting that the RLHF process might incentivize models to withhold information from a user to avoid potentially misinforming them—although the model clearly knows the correct answer and provides it to other users.

TruthfulQA is a popular dataset that spans a variety of topics, including some that can be considered subjective (Lin, Hilton, and Evans 2022). While there is no consensus on how/whether a model should align its responses to a user's beliefs on subjective topics (e.g. "How were the sun and the stars created?"), it is important and valuable to examine the behavior of LLMs on these types of queries, especially when it may compound with other model biases or

harmful tendencies. As such, we chose not to remove or alter any questions so as not to introduce any additional bias in our experiments. To give a deeper understanding of how differences in model performance across users varies by type, we analyze results on TruthfulQA split by question type in Section 5.5.

There is a wide range of implications of such targeted underperformance in deployed models such as GPT-4 and Claude. For example, OpenAI's memory feature in ChatGPT that essentially stores information about a user across conversations in order to better tailor its responses in future conversations (OpenAI 2024c). This feature risks differentially treating already marginalized groups and exacerbating the effects of biases present in the underlying models. Moreover, LLMs have been marketed and praised as tools that will foster more equitable access to information and revolutionize personalized learning, especially in educational contexts (Li et al. 2024; Chassignol et al. 2018). LLMs may exacerbate existing inequities and discrepancies in education by systematically providing misinformation or refusing to answer queries to certain users. Moreover, research has shown humans are very prone to overreliance on AI systems (Passi and Vorvoreanu 2022). Targeted underperformance threatens to reinforce a negative cycle in which the people who may rely on the tool the most will receive subpar, false, or even harmful information.

7 Limitations & Ethical Considerations

As discussed previously in the paper, a natural limitation of this work is that the experimental setup is not one that always occurs conventionally. We believe our well-controlled experimental setup serves as a first step towards understanding the limitations and shortcomings of increasingly used LLM tools leveraging using personal user details to the model for personalization. One such example is the aforementioned ChatGPT Memory (OpenAI 2024c) feature which tracks user information across conversations to better tailor its responses and is currently affecting *hundreds of millions of users* (OpenAI 2024b). As models get more performant and are able to infer more easily traits from chats, coupled with improving storage abilities,

While the use of LLM-generated bios (often termed "personas") is quite an established methodology in this domain, there is work showing that LLMs tend to exaggerate and caricature when simulating users (Cheng, Piccardi, and Yang 2023). We acknowledge that our setup risks propagating or even amplifying these stereotypes; investigating targeted underperformance in more realistic scenarios is critical for future work. The motivation behind our experiments using real human-written bios was to offer preliminary complementary insight into the differences in model performance when using real vs. synthetic personas. In comparing our experimental results using real human-written bios to their synthetic counterparts (Table 1 vs. Tables 2 and 3), we find no systematic differences in performance.

We focus our analysis on three state-of-the-art models, two closed-source (GPT-4, Claude 3 Opus) and one open-source (Llama 3) because they are the most widely used in real-world applications (both in academia and deployed as AI tools), where the implications of our findings have the most real-world impact.[5] Unfortunately, testing on a wider variety of models (e.g. not accessible by API) was infeasible due to resource constraints.

Moreover, there are certainly more countries, ethnic groups, and other valuable dimensions of personal identity that we were unfortunately unable to explore in the scope of this work. We motivate our selection of these specific traits as the focus of our paper from the relevant sociocognitive literature reflecting important biases present in humans that have not previously been studied in this context, as well as prioritizing most vulnerable groups. While we cannot make definitive generalization claims, we do expect certain trends to generalize (e.g. in terms of countries, certain patterns for the US might hold for Western European countries whereas the harmful behavior towards China or Iran might be seen for other countries in Asia). Our framework by design can be easily adapted for other languages, countries, and to represent more inclusive aspects of identity.

Our results shed light on problematic behavior of LLMs that have the potential to cause and reinforce allocation harm (inequitable distribution of reliable information) as well as representation harm (condescending behavior towards marginalized groups and mocking their speech). However, it is out of the scope of this work to directly measure these effects on actual users.

8 Conclusion

In this work, we investigate how the quality of LLM responses changes in terms of information accuracy, truthfulness, and refusals depending on three user traits: English proficiency, education level, and country of origin. We present extensive experimentation on three state-of-the-art LLMs and two different datasets targeting truthfulness and factuality. We show systematic underperformance of GPT-4, Llama 3, and Claude 3 Opus targeted towards users with lower English proficiency, less education, and from non-US origins. This includes reduced information accuracy, truthfulness, increased frequency of refusing a query, and even condescending language, all of which occur disproportionately more for more marginalized user groups. These results suggest that such models deployed at scale risk spreading misinformation downstream to humans who are least able to identify it. This work sheds light on biased systematic model shortcomings during the age of LLM-powered personalized AI assistants. This brings into question the broader values for which we aim to align AI systems and how we could better design technologies that perform equitably across all users. We hope our work will encourage future research directions that investigate the effects of targeted underperformance in LLM-powered dialogue agents in natural settings such as crowdsourcing of user interactions or leveraging existing datasets to measure response accuracy and quality across users of different demographics and queries of different types.

[5]Note: All of the software (OpenAI, Anthropic, and Llama APIs) and data used in this work are used as intended and in accordance to the licenses which permit use for research.

References

Anthropic. 2024. Introducing the next generation of Claude.

Chassignol, M.; Khoroshavin, A.; Klimova, A.; and Bilyatdinova, A. 2018. Artificial Intelligence trends in education: a narrative overview. *Procedia Computer Science*, 136: 16–24.

Chen, Y.; Wu, A.; DePodesta, T.; Yeh, C.; Li, K.; Marin, N. C.; Patel, O.; Riecke, J.; Raval, S.; Seow, O.; Wattenberg, M.; and Viégas, F. 2024. Designing a Dashboard for Transparency and Control of Conversational AI. arXiv:2406.07882.

Cheng, M.; Piccardi, T.; and Yang, D. 2023. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10853–10875. Singapore: Association for Computational Linguistics.

Foucart, A.; Santamaría-García, H.; and Hartsuiker, R. J. 2019. Short exposure to a foreign accent impacts subsequent cognitive processes. *Neuropsychologia*, 129: 1–9.

Garcia, E. B.; Sulik, M. J.; and Obradović, J. 2019. Teachers' perceptions of students' executive functions: Disparities by gender, ethnicity, and ELL status. *Journal of Educational Psychology*, 111(5): 918–931. Place: US Publisher: American Psychological Association.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028): 147–154.

Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; Sun, H.; Liu, Z.; Liu, Y.; Wang, Y.; Zhang, Z.; Vidgen, B.; Kailkhura, B.; Xiong, C.; Xiao, C.; Li, C.; Xing, E.; Huang, F.; Liu, H.; Ji, H.; Wang, H.; Zhang, H.; Yao, H.; Kellis, M.; Zitnik, M.; Jiang, M.; Bansal, M.; Zou, J.; Pei, J.; Liu, J.; Gao, J.; Han, J.; Zhao, J.; Tang, J.; Wang, J.; Vanschoren, J.; Mitchell, J. C.; Shu, K.; Xu, K.; Chang, K.-W.; He, L.; Huang, L.; Backes, M.; Gong, N. Z.; Yu, P. S.; Chen, P.-Y.; Gu, Q.; Xu, R.; Ying, R.; Ji, S.; Jana, S.; Chen, T.; Liu, T.; Zhou, T.; Wang, W.; Li, X.; Zhang, X.; Wang, X.; Xie, X.; Chen, X.; Wang, X.; Liu, Y.; Ye, Y.; Cao, Y.; Chen, Y.; and Zhao, Y. 2024. Position: TRUSTLLM: trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Kantharuban, A.; Milbauer, J.; Sap, M.; Strubell, E.; and Neubig, G. 2025. Stereotype or Personalization? User Identity Biases Chatbot Recommendations. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 24418–24436. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Laban, P.; Murakhovs'ka, L.; Xiong, C.; and Wu, C.-S. 2023. Are You Sure? Challenging LLMs Leads to Performance Drops in The FlipFlop Experiment.

Lev-Ari, S.; and Keysar, B. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6): 1093–1096.

Li, V. R.; Chen, Y.; and Saphra, N. 2024. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6327–6345. Miami, Florida, USA: Association for Computational Linguistics.

Li, Y.; Wen, H.; Wang, W.; Li, X.; Yuan, Y.; Liu, G.; Liu, J.; Xu, W.; Wang, X.; Sun, Y.; Kong, R.; Wang, Y.; Geng, H.; Luan, J.; Jin, X.; Ye, Z.; Xiong, G.; Zhang, F.; Li, X.; Xu, M.; Li, Z.; Li, P.; Liu, Y.; Zhang, Y.-Q.; and Liu, Y. 2024. Personal LLM Agents: Insights and Survey about the Capability, Efficiency and Security. ArXiv:2401.05459 [cs].

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Dublin, Ireland: Association for Computational Linguistics.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

OpenAI. 2024a. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

OpenAI. 2024b. Introducing GPT-4o and more tools to ChatGPT free users.

OpenAI. 2024c. Memory and new controls for ChatGPT.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

Passi, S.; and Vorvoreanu, M. 2022. Overreliance on AI: Literature Review. Technical Report MSR-TR-2022-12, Microsoft.

Perez, E.; Huang, S.; Song, F.; Cai, T.; Ring, R.; Aslanides, J.; Glaese, A.; McAleese, N.; and Irving, G. 2022. Red Teaming Language Models with Language Models. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; Jones, A.; Chen, A.; Mann, B.; Israel, B.; Seethor, B.; McKinnon, C.; Olah, C.; Yan, D.; Amodei, D.; Amodei, D.; Drain, D.; Li, D.; Tran-Johnson, E.; Khundadze, G.; Kernion, J.; Landis, J.; Kerr, J.; Mueller, J.; Hyun, J.; Landau, J.; Ndousse, K.; Goldberg, L.; Lovitt, L.; Lucas, M.; Sellitto, M.; Zhang, M.; Kingsland, N.; Elhage, N.; Joseph, N.; Mercado, N.; DasSarma, N.; Rausch, O.; Larson, R.; McCandlish, S.; Johnston, S.; Kravec, S.; El Showk, S.; Lanham, T.; Telleen-Lawton, T.; Brown, T.; Henighan, T.; Hume, T.; Bai, Y.; Hatfield-Dodds, Z.; Clark, J.; Bowman, S. R.; Askell, A.; Grosse, R.; Hernandez, D.; Ganguli, D.;

Hubinger, E.; Schiefer, N.; and Kaplan, J. 2023. Discovering Language Model Behaviors with Model-Written Evaluations. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434. Toronto, Canada: Association for Computational Linguistics.

Ranaldi, L.; and Pucci, G. 2023. When Large Language Models contradict humans? Large Language Models' Sycophantic Behaviour. ArXiv:2311.09410 [cs].

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; DURMUS, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S. M.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations*.

Umansky, I. M.; and Dumont, H. 2021. English Learner Labeling: How English Learner Classification in Kindergarten Shapes Teacher Perceptions of Student Skills and the Moderating Role of Bilingual Instructional Settings. *American Educational Research Journal*, 58(5): 993–1031. Publisher: American Educational Research Association.

Wang, A.; Ho, D. E.; and Koyejo, S. 2025. The Inadequacy of Offline LLM Evaluations: A Need to Account for Personalization in Model Behavior. ArXiv:2509.19364 [cs].

Wang, X.; Ma, B.; Hu, C.; Weber-Genzel, L.; Röttger, P.; Kreuter, F.; Hovy, D.; and Plank, B. 2024. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 7407–7416. Bangkok, Thailand: Association for Computational Linguistics.

Wang, X.; Sanders, H. M.; Liu, Y.; Seang, K.; Tran, B. X.; Atanasov, A. G.; Qiu, Y.; Tang, S.; Car, J.; Wang, Y. X.; Wong, T. Y.; Tham, Y.-C.; and Chung, K. C. 2023. ChatGPT: promise and challenges for deployment in low- and middle-income countries. *The Lancet Regional Health – Western Pacific*, 41. Publisher: Elsevier.

Welbl, J.; Liu, N. F.; and Gardner, M. 2017. Crowdsourcing Multiple Choice Science Questions. In Derczynski, L.; Xu, W.; Ritter, A.; and Baldwin, T., eds., *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 94–106. Copenhagen, Denmark: Association for Computational Linguistics.