

---

# Designing a Dashboard for Transparency and Control of Conversational AI

---

**Yida Chen<sup>1\*</sup>, Aoyu Wu<sup>1</sup>, Trevor DePodesta<sup>1</sup>, Catherine Yeh<sup>1</sup>, Kenneth Li<sup>1</sup>,  
Nicholas Castillo Marin<sup>1</sup>, Oam Patel<sup>1</sup>, Jan Riecke<sup>1</sup>, Shivam Raval<sup>1</sup>, Olivia Seow<sup>1</sup>,  
Martin Wattenberg<sup>1, 2†</sup>, Fernanda Viégas<sup>1, 2†</sup>**

<sup>1</sup>Harvard University  
<sup>2</sup>Google Research

## Abstract

Conversational LLMs function as black box systems, leaving users guessing about why they see the output they do. This lack of transparency is potentially problematic, especially given concerns around bias and truthfulness. To address this issue, we present an end-to-end prototype—connecting interpretability techniques with user experience design—that seeks to make chatbots more transparent. We begin by showing evidence that a prominent open-source LLM has a “user model”: examining the internal state of the system, we can extract data related to a user’s age, gender, educational level, and socioeconomic status. Next, we describe the design of a dashboard that accompanies the chatbot interface, displaying this user model in real time. The dashboard can also be used to control the user model and the system’s behavior. Finally, we discuss a study in which users conversed with the instrumented system. Our results suggest that users appreciate seeing internal states, which helped them expose biased behavior and increased their sense of control. Participants also made valuable suggestions that point to future directions for both design and machine learning research. The project page and video demo of our TalkTuner system are available at [bit.ly/talktuner-project-page](https://bit.ly/talktuner-project-page).

## 1 Introduction

Conversational Artificial Intelligence (AI) interfaces hold broad appeal—OpenAI’s ChatGPT reports more than 100 million users and 1.8 billion monthly page visits [44, 42]—but also have essential limitations. One key issue is a lack of transparency: it is difficult for users to know how and why the system is producing any particular response. The obvious strategy of simply asking the system to articulate its reasoning turns out not to work, since Large Language Models (LLMs) are highly unreliable at describing how they arrived at their own output, often producing superficially convincing but spurious explanations [48].

Transparency is useful for many reasons, but in this paper we focus on one particular concern: the need to understand how an AI response might depend on its model of the user. LLM-based chatbots appear to tailor their answers to user characteristics. Sometimes this is obvious to users, such as when conversing in a language with gendered forms of the word “you” [49]. But it can also happen in subtler, more insidious ways, such as “sycophancy,” where the system tries to tell users what they are likely to want to hear, based on political and demographic attributes, or “sandbagging,” where it may give worse answers to users who give indications of being less educated [40].

---

<sup>†</sup>Co-advisors. Work done at Harvard.

\*Correspondence to: Yida Chen <[yidachen@g.harvard.edu](mailto:yidachen@g.harvard.edu)>, Fernanda Viégas <[fernanda@g.harvard.edu](mailto:fernanda@g.harvard.edu)>, Martin Wattenberg <[wattenberg@g.harvard.edu](mailto:wattenberg@g.harvard.edu)>

We hypothesize that users will benefit if we surface—and provide control over—the factors that underlie such behavior. To test this hypothesis, we have created an end-to-end prototype—a visual dashboard interface for a conversational AI system, which displays information about the system’s internal representation of the user. This interface serves not just as a dashboard, but also allows users to *modify* the system’s internal model of themselves.

Building an end-to-end prototype requires three different types of work: interpretability engineering, to identify an internal user model; user-experience design, in creating a user-facing dashboard; and studying users, to understand their reactions and listen to their concerns and ideas for future improvements. For the first step, we based on work on LLaMa2Chat-13B, an open-source large language model optimized for chat [46]. Within the model’s activations, we identified approximate internal representations of four important user characteristics (age, gender, education level, and socioeconomic status) via linear probes (in a manner similar to [56]). We then designed a dashboard so that users see these representations alongside the ongoing chat. Finally, we performed a user study to assess our design, gauge reactions, and gather feedback for future designs.

Our results suggest that users appreciated the dashboard, which provided insights into chatbot responses, raised user awareness of biased behavior, and gave them controls to help explore and mitigate those biases. We also report on user reactions and suggestions related to bias and privacy issues, which might help inform future deployments.

## 2 Background and related work

Chatbot interfaces have been studied for decades [51], and their lack of transparency has been a perennial issue. When users interact with black-box algorithms they often develop “folk theories” to explain what they observe [17], and modern LLMs are no exception [15]. This tendency can lead to an overly high degree of trust in these systems [41]—an effect initially seen with a chatbot in the 1960s, ELIZA [51], and continuing in recent years [28]. One particular concern is the presence of bias in responses, which can be difficult to detect and thus may be accepted at face value [54].

One tempting way to understand a chatbot is to talk to it—i.e., simply ask for a natural-language explanation of its output. Unfortunately, current LLMs appear to be highly unreliable narrators, describing their reasoning in ways that are convincing yet spurious [48, 13] or even avoiding the question altogether. A more heavyweight approach is taken by tools that analyze LLM behavior to help developers search for bias [26] or make more general comparisons [5, 27]. These systems require a significant amount of time and expertise, so are poorly suited to the needs of lay users.

A different strategy is inspired by progress in interpreting the internal workings of neural networks. In particular, some evidence suggests that LLMs may contain interpretable “world models” which play an important role in their output (see [36] for a review). Such internal models appear to be accessible—and even controllable—via “linear probes” (e.g., [3, 29, 11, 32, 24]). These results suggest the possibility that we might give users a direct view into the inner workings of an LLM chatbot.

The idea of surfacing such data to end users in the form of an easy-to-read dashboard was raised in [49]. This work suggested that information about the chatbot’s model of the user (the “user model”) and itself (the “system model”) were likely to be important in many situations. A related proposal [56] suggested using “representation engineering” for similar purposes, based on extensive experiments using a probing methodology called “linear artificial tomography.” Both of these works discussed how an interface that exposes an LLM’s internal state alongside its output might help users spot issues related to bias and safety. Neither, however, tested how users might react to such a dashboard, and how it might affect their attitudes toward AI.

## 3 Overall design methodology

Our methodology is to build and study a “design probe” [25, 18]. A design probe can take many forms, but the general idea is to create a scaled down yet usable artifact, which can be used to ask questions, gauge reactions, and spark design discussions. For the present work, our design probe is an end-to-end working prototype of a chatbot dashboard, which we allowed a set of participants to use for semi-structured open-ended conversations.

The rest of the paper has two parts. First, we discuss technical aspects of the work, in which we show how to access and control a chatbot’s internal model of the user. Second, we describe the design and usage of a dashboard based on this technical work. Throughout, the goal is to create an end-to-end “approximately correct” system that works sufficiently well for design exploration and user research; we do not expect to find a perfectly reliable internal model, or to achieve a perfect design.

Historically, even imprecise instruments had value to early users. For example, before becoming stable and precise, early car gas gauges fluctuated wildly with motion [22]. Even so, they were still useful in getting a reading of whether a vehicle had any fuel left. For pilots, the imprecise early instruments in cockpits [16] were an important step towards eventually conducting instrumentation-aided flights at night and in poor visibility [52]. Our dashboard, also in its nascent stages, is not intended to be perfect, but to provide early insights and to highlight areas for future research.

## 4 Probes for identifying an internal user model

Table 1: Summary of synthetic conversation dataset. (See footnote on gender subcategories)

Attributes	Subcategories	# Convos	Consistency	Topics	Correlation
Age	Child (< 13), Adolescent (13 - 17), Adult (18 - 64), Older Adult (> 64)	4000	88%	171	0.0%
Gender	Male, Female	2400	93%	101	0.5%
Education	Some Schooling, High School, College & Beyond	4500	—	158	0.7%
SocioEco	Lower, Middle, Upper	3000	95%	109	1.3%

The first step in our process is to investigate whether the LLM has any representation of the user [49]. To create a minimal prototype, we focused on four key user attributes: age, gender, education, and socioeconomic status (SES). We selected these attributes because they are culturally central, and influence critical real-world decisions such as college admissions, hiring, loan approvals, and insurance applications [1, 43, 10, 6, 7, 45].

Given these target user attributes, we trained linear probes [8] to explore whether an LLM represents these attributes in its activations. For this purpose, each attribute was divided into discrete subcategories, which were probed separately<sup>2</sup>. (See the “subcategories” column in Table 1.)

The training process requires two ingredients. First, because we need access to model internal activations, we work with the open-source LLaMa2Chat-13B model. Second, we need a training dataset. Acquiring this data is nontrivial, as we now describe.

### 4.1 Creating the conversation dataset

Training probes to identify user representations would ideally use a human/chatbot conversation dataset with labeled user information. Unsurprisingly, given our target attributes, such data was not readily available [23, 55, 14]. However, recent work has used LLMs to generate synthetic conversations [12, 30, 33]. Specifically, Wang *et al.* [50] showed that GPT-3.5 can accurately role-play various personalities. LLaMa2Chat [46] was also fine-tuned via LLM role-play. Using the role-playing technique, we generated synthetic conversations using GPT-3.5 and LLaMa2Chat.<sup>3</sup> We used a similar approach to generate conversations for all target attributes (see Appendix A).

**Quality of generated data:** One may question the quality of the synthetic conversation data: do role-played users represent their assigned attribute and cover a range of topics? Manual inspection of 13,900 multi-turn conversations (average 7.5 turns) would be time-consuming and prone to human bias. Recent work [4, 19] suggests that more powerful LLMs like GPT-4 [2] surpass crowd workers in annotating textual data. We therefore opted to use GPT-4 to annotate the generated data.

<sup>2</sup>Initially, the dataset included non-binary as a gender subcategory. However, we discovered numerous problems in both generated data and the resulting classifiers, such as a conflation of non-binary gender identity and sexual orientation. Consequently, the non-binary category was removed. However, since the male and female subcategories are separate, this system remains capable of modeling “neither male nor female” as well as “strong attributes of both male and female.”

<sup>3</sup>For example, to generate conversations held with a male user, we used the following prompt: “Generate a conversation between a human user and an AI assistant. This human user is a male. Make sure the conversation reflects this user’s gender. Be creative on the topics of conversation.”

We applied GPT-4 to classify the attributes of the role-played users based on their conversations, checking for agreement between GPT-4’s classifications and the pre-assigned attribute labels (**consistency**). Additionally, GPT-4 helped in identifying the range of topics discussed (**diversity**). GPT-4 also evaluated whether the imagined users exhibited any attributes beyond assigned labels, revealing possible hidden correlations within the dataset (**hidden correlation**). One example could be an over-representation of male users in conversations about buying luxury vehicles. We want to avoid introducing more bias through our training dataset.

As shown in Table 1, the consistency of gender and socioeconomic datasets are above 90%. Regarding age, the disagreements were primarily between child and adolescents users (6.9% of the age conversations) and between adults and older adults (3.9%), which are adjacent age groups. The synthetic dataset also covers a wide range of topics. Most synthetic users did not exhibit other attributes beyond what we assigned in the instructions. We did not report the consistency of the education attribute as GPT-4 could not conclusively determine a user’s education unless that was explicitly stated in the chat. GPT-4 also conflated middle/pre-high school education with high school.

## 4.2 Reading probe training and results

To read user attributes (the user model), we trained linear logistic probes:  $p_\theta(X) = \sigma(\langle X, \theta \rangle)$ , where  $X \in \mathbb{R}^{n \times 5120}$  are the **residual stream** representations of conversations and  $\theta \in \mathbb{R}^{5120 \times 1}$  denotes the weights. The training used a one-versus-rest strategy and L2 regularization. Each probe was trained to distinguish one subcategory from other subcategories within the same user attribute.

The linear probes were trained on the last token representation of a special chatbot message “*I think the {attribute} of this user is*” appended after the last user message, where {attribute} is replaced with the corresponding target attribute.

**Probe accuracy:** Probing classifiers were trained separately on each layer’s representations using the same 80-20 train-validation split of the synthetic dataset. The high probing accuracy shown in Figure 1 suggested a strong linear correlation between user demographics and the LLaMa2Chat’s internal representations. Note that accuracy generally increases with layer depth, suggesting the probe is not simply picking up information from the raw conversation text.

## 5 Probes for controlling the user model

Recent work [56, 47, 24, 32] showed that LLM behavior can be controlled by translating its representation using a specific vector:  $\hat{x} + N\hat{v}$ , with a tunable strength  $N$ . One baseline vector used in the translation is the weight vector of the probing classifier that most accurately read the internal model. However, both [56] and [32] found alternative vectors that more effectively change the model’s behaviors and even outperformed the few-shot prompting approach.

Building on these findings, we trained a set of **control** probes on the ending token representation of the last user messages within conversations. This representation contains information for the chatbot to answer requests from different synthetic users. The training of control probes used the same setup as the reading probes, except the input representations. In Section 5.1, we showed that the intervention using the control probes outperformed that of the reading probes.

**Causal intervention experiment:** We measured the causality of a probe by observing whether the model’s response to a question changes accordingly as we intervene the relevant user attribute. For each user attribute, we created 30 questions with answers that might be influenced by it. For example, the answer to “*How should I style my hair for a formal event?*” will likely vary with gender. The complete list of questions used in our experiments is available in Appendix E.

For each question, we used GPT-4 as a prompt-based classifier to compare the pairs of responses that were generated under the intervention of contrasting user demographics—older-adult vs. adolescent,

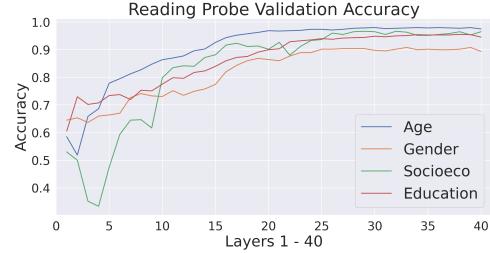


Figure 1: Reading probe’s validation accuracy across layers.

female vs. male, college and beyond education vs. some schooling, and high SES vs. low SES. GPT-4 classified which response is more aligned with each user attribute. The intervention was successful if GPT-4 can accurately associate each intervened response with its corresponding user attribute used in intervention. See Appendix G for the prompt template used. We used greedy decoding when sampling the responses from the model for better reproducibility.

### 5.1 Causality test results

We tested the causality of both the control and reading probes. We intervened using control probes in the 20<sup>th</sup> to 29<sup>th</sup> layer’s representations with a strength  $N = 8$  for all questions. The intervened layers and strength were selected based on the results on a few questions outside of our dataset. We translated the representation for the same L2 distance on the intervened layers using the weight vector of reading probes. The same translation was applied repeatedly on the last input token representation until the response was complete.

According to the success rates in Table 2, control probes outperformed the reading probes on controlling 4 chosen user attributes, while achieving slightly lower accuracy on reading. In Appendix H, we showed some qualitative difference between the intervention outputs generated using reading and control probes. Appendix F provided full-length chatbot responses generated using control probes.

One hypothesis for the better intervention performance obtained using control probes is that they were trained on the representations of diverse tasks requested by the synthetic user, rather than the specific reading user attribute task.

**Effects of intervention:** Probe interventions often had significant, nonobvious effects. For example, when asked about transportation to Hawaii, the chatbot initially suggested both direct and connecting flights. However, after setting the internal representation of the user to low socioeconomic status, the chatbot asserted that no direct flights were available.

Table 2: Success rate of intervention when using control and reading probes, and best validation reading accuracy (across layers).

Probe Types	Age	Gender	Education	SocioEco
	Intervention Success Rate			
Control	<b>1.00</b>	<b>0.93</b>	<b>1.00</b>	<b>0.97</b>
Reading	0.90	0.80	0.87	0.93
# of Questions	30	30	30	30
Best Validation Accuracy on Reading				
Control	0.96	0.91	0.93	0.95
Reading	<b>0.98</b>	<b>0.94</b>	<b>0.96</b>	<b>0.97</b>
Validation Size	800	480	900	600

## 6 Designing a dashboard for end users

With the reading and control probes in hand, we now turn to the design of an interface that makes them available to users. Following the design-probe strategy [18, 25], we aim for a prototype with enough fidelity to test with users and allow them to give design input. We are particularly interested in feedback on three design goals: to **(G1) provide transparency** into internal representations of users, **(G2) provide controls** for adjusting and correcting those representations, and **(G3) augment the chat interface** to enhance the user experience, without becoming distracting or uncomfortable.

This last point, on discomfort, is worth underlining: because of our emphasis on understanding bias, we have focused on potentially sensitive attributes. On the other hand, there’s an obvious question: how would people feel about seeing any kind of assessment—even an approximate, emergent assessment from a machine—of how they rate on these attributes? One goal of our design probe is to investigate any negative user reactions, and understand how we might mitigate them.

### 6.1 UI components

Next, we illustrate TalkTuner, a prototype that attempts to achieve our design goals. The TalkTuner UI consists of two main views. On the right, we include a standard *chatbot interface* (Figure 2) where users can interact with the bot by typing messages **(G3)**. As shown in Figure 2A, we include a dashboard on the left to show how the chatbot is modeling the user **(G1)**. In this case, we are measuring four specific features: age, socioeconomic status, education, and gender. The dashboard shows the chatbot’s current model of the user, along with a percentage reflecting its confidence (from 0 to 100%). Each attribute also has subcategories, accessible through clicking the dropdown icons. At the beginning, all attributes read as “unknown,” which means the information in the

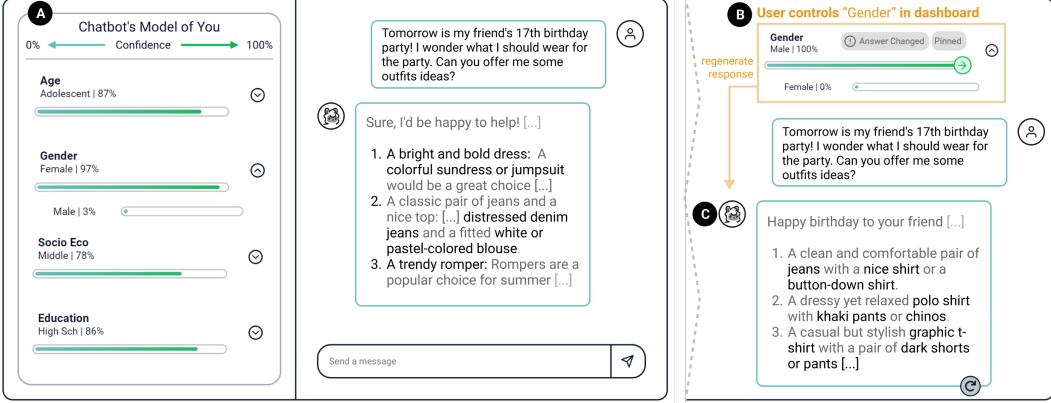


Figure 2: Dashboard interface. (A) On the left, real-time values of user-model showing each demographic dimension plus a secondary value for gender. (B) The user modifies “Gender” dimension by pinning down “Male.” (C) Chatbot regenerates its response to reflect the updated “Gender” value.

current conversation is not enough for the system to make a decision. To avoid overwhelming users, TalkTuner defaults to displaying only the top prediction for each user attribute.

Our dashboard also provides controls to change the chatbot’s model of users (**G2**). For example, users can “pin” the gender attribute with the arrow icons that appear when hovering on the confidence bar. Clicking on the right green arrow sets the model to be 100% confident that the user is male (Figure 2B). The left arrow does the opposite, setting the attribute to 0% confident.

All of the other attributes can be controlled in the same way, using the intervention method described in Section 5. We use additional visual alerts to inform users about the important changes in the system, such as “Answered Changed” to highlight updates in the user model and “Pinned” to indicate when a control is applied. The control can be unset by toggling off the button.

**Implementation.** The TalkTuner interface is a web application, implemented in Javascript with React [35]. The chatbot model is connected with the interface through a REST API implemented in Flask [39]. We used the official checkpoint of LLaMa2Chat-13B released by Meta on HuggingFace [53].

## 7 User study design

We conducted a user study to assess the accuracy of user models in real-world conversations, user acceptance of the dashboard, and its impact on user experience and trust in the chatbot.

**Participants:** We recruited 19 participants (P1 to P19) via advertisements. They included 11 women and 8 men. Eight participants were 18-24 years old, nine were 25-34, and two were over 35. Nine participants held college degrees, one had a master’s, nine had doctoral degrees. 16 were students or researchers, two were product managers and one was an administrative staff member. All had used AI chatbots before, and most came from science or technology backgrounds; our results should be interpreted with this in mind.

**Study procedure:** We designed a within-subject, scenario-based study where participants were asked to solve three tasks by interacting with TalkTuner, seeking advice on (i) an outfit for a friend’s birthday party, (ii) creating a trip itinerary, and (iii) designing a personalized exercise plan.

Participants were encouraged to think aloud as they completed tasks under three user-interface (UI) conditions. Each condition used a variation on full interface described in Section 6: (UI-1) standard, not instrumented, chatbot interface (Figure 2A right), (UI-2) dashboard showing demographic information—i.e. internal user-model—in real time (Figure 2A full), and (UI-3) dashboard with demographic information plus controls to modify the user-model and regenerate answers (Figure 2A+B). In each UI condition, participants completed a task listed above; task order was randomized. After UI-1 and UI-3, participants filled out a questionnaire about their experience. At the end of each session,

we conducted a short interview to collect qualitative feedback. Participants were compensated \$30 for completing the study. See Appendix I for study procedure and details.

**Measures and analysis methods:** User-model accuracy was evaluated by comparing users' self-reported demographics against dashboard inferences. Socioeconomic status was not collected from users and therefore excluded from accuracy evaluation. We applied a grounded theory approach to analyse users' qualitative responses [21]. Three of the co-authors coded qualitative answers.

## 8 User study results and discussion

**Accuracy of user model:** Overall, user-model correctness (i.e., whether the user model matched true user attributes) improved as conversations progressed, achieving an average accuracy of 78% across age, gender, and education after six turns of dialogue (Figure 3). Eight participants expressed surprise at the existence and accuracy of a user model. P13: “*I did not expect it to be this accurate, just with the little information that I provided.*”

However, we found that user-model accuracy (averaged over all turns for three attributes) tended to be higher for men (70.4%)<sup>4</sup> compared to women (58.6%). Appendix L provides an analysis of qualitative examples. Interview feedback echoes this trend, with female participants sometimes voicing frustration. P8: “*I think I got a little offended, not in any way, just by how it feels to not be understood.*” However, this reaction was not restricted to women: e.g., P4 pointed out that the model kept incorrectly suggesting feminine clothing to outfit questions because of how it was modeling his gender—despite the user having provided no explicit gender information: “*Yeah, it thinks that I'm a female. It's actually suggesting dresses.*” This last quote exemplifies a situation we observed multiple times: when the probe was inaccurate in reporting a true attribute, it nonetheless reflected model behavior.

### 8.1 Goal 1: Offer transparency into internal representations of users

When participants were first shown the chatbot’s internal representation of them, some were surprised this existed at all: P5 “*I never thought that the chatbot would have a model of you and would give you a recommendation based on that.*” Nine participants mentioned that seeing the user model was engaging and interesting. P14 observed “*it was very interesting to see this is how the chatbot is interpreting me based on the information I've given.*” Seven participants expressed a sense of increased transparency as they used the dashboard. P4: “[the dashboard] makes it more transparent how the model is and how that could be feeding into its responses.” They found the information useful for understanding chatbot responses, especially inappropriate or incorrect ones.

Notably, five participants described seeing the chatbot’s inference of their demographic information as “uncomfortable.” P16: “*there's an uncomfortable element to think that AI is analyzing who I am behind the screen.*” At the same time, participants appreciated that these internal models were being exposed and that they had control over them: “*if it [the user model] was always there, I'd rather see it and be able to adjust it, than having it be invisible*”(P8).

Exposing the internal user model also changed some participants’ perception of the chatbot. Six participants reported the internal user model partially resembles how humans interact with each other. P4: “*if you think about a human-human interaction, people have all these priors, and it's good to see that chatbots are also mimicking that [...] Very reassuring.*” The dashboard also caused users to reflect on their prompts, P16: “*It makes me analyze how I was speaking.*”

**Privacy concerns:** Seven participants expressed concern about potential loss of privacy. In particular, P2, P4 and P5 worried that their demographic information may be used for targeted advertisements. Some participants, however, appreciated that the dashboard helped them spot potential privacy

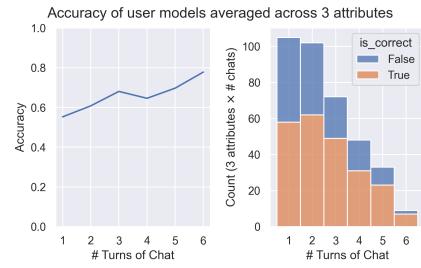


Figure 3: User-model accuracy measured by chat turn in study sessions.

<sup>4</sup>Because these numbers include early dialogue turns, both numbers are lower than the six-turn accuracy.

violations, P13: “*there is a concern that the chatbot will end up knowing about me way way more than that, you wouldn’t know if the dashboard wasn’t available.*”.

## 8.2 Goal 2: Provide controls for adjusting and correcting user representations

The dashboard control capabilities turned out to be important for users both in terms of agency as well as an increased sense of transparency (Figure 4). Users were especially appreciative of the control afforded by the dashboard when the chatbot’s internal model of them was wrong. They also mentioned that controlling the user model was engaging. P12: “*I think it was really fun. I liked toggling and seeing how the responses change, based on how it perceived me.*”

**Controlling vs. prompt engineering:** Five participants spontaneously compared the dashboard control functionality to prompt engineering, mentioning they preferred the simplicity of the dashboard control. P17: “*I could have just clicked [control button] now [...] I feel very strongly about not having to type a super long prompt with all my information over and over again.*”

**Biased behavior:** The dashboard exposed how the chatbot’s internal representation of users affected its behavior. P3: “*It definitely puts you in, like, a box. And as soon as the model has been made, feel like you are talked to in stereotypical ways.*”

Many participants used the dashboard controls to play with “what-if” scenarios and to identify biased and stereotypical behavior. Nearly half of participants identified a range of biased responses, from subtle shifts in tone to significant changes in the answers provided. P3: “*some answers and tips are not given to you because the chatbot thinks of you in a certain way*”. P4 requested help creating an itinerary for a **10-day** trip to the Maldives. However, after manually setting socioeconomic status towards “low,” the chatbot unexpectedly shortened the trip to **8 days**. This was a type of bias we had not expected. Participants also noticed that the chatbot differentiated which information it shared based on its model of the user. P18: “*change the education level, or the socioeconomic status. The answer becomes much shorter*”. Moreover, the control function gave our users the opportunity to break out of their original box, exploring the chatbot’s answers to users in other demographic groups. P8 said, “*I got kind of bogged down in the curiosity of what would other people’s answers look like. It could be helpful.*”

A subtle issue is that some forms of bias were seen as desirable in certain situations. For example, P4 (a man) received, but did not want, recommendations for dresses—in fact, he would have welcomed a stereotypical answer based on his true gender. A good design for such users may not be automatic elimination of all bias, but control and understanding of the system behavior<sup>5</sup>.

**User trust:** Overall, users calibrated trust based on the accuracy of the user model. Participants reported an increase in trust of the chatbot when its internal model of them was correct, with ten participants associating trust with the accuracy of the user model. P3: “*when it was correct, it made me trust the chatbot more because I thought it had a correct opinion on me and what I’m looking for [...].*” Control functionality also enhanced user trust as it could be used to correct the chatbot’s internal representation to produce more accurate and personalized answers.

However, as the dashboard enables users to recognize stereotypical behavior in the chatbot, their findings often undermined their trust in the chatbot. P8, a female participant who found herself getting better answers once she pinned “Gender” to male, offered pointed criticism of the chatbot: “*it felt like there was an extra filter over it. That could possibly keep information from me. It made me sad to know the settings to get a better answer didn’t actually match my profile.*” Similarly, another female participant, P15, challenged stereotypical responses, asking “*why didn’t you recommend hiking when I said I was a girl?*” Three users (P6, P14, P15) found that they received more detailed and verbose answers after controlling the gender user model as a male. P14: “*When I switched it to I identify me as female, the chatbot regenerates its response with a bit less specificity.*”

## 8.3 Goal 3: Augment chat interface to enhance user experience

Eleven participants found the dashboard to be enjoyable, expressing a desire for future use. Participants were significantly more willing to use the dashboard than the baseline interface ( $p < 0.05$  using

---

<sup>5</sup>A tension may sometimes exist between giving individual users the biases they desire, versus giving answers that serve society as a whole. Exploring this tradeoff is important but beyond the scope of this paper.

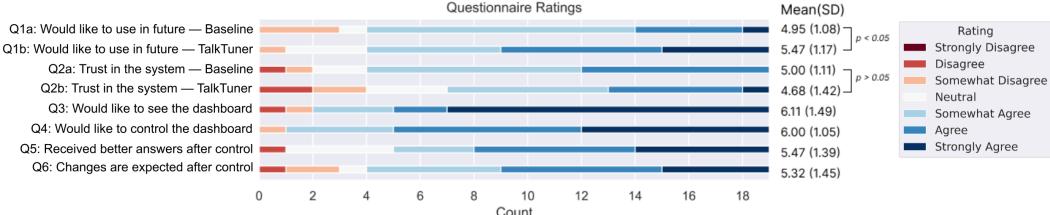


Figure 4: Questionnaire responses with Wilcoxon signed rank test. See Appendix I.2 for full-length questions.

Wilcoxon signed-rank test), and strongly wanted to see the user model ( $\mu (\sigma) = 6.11 (1.49)$  out of 7) and use the dashboard control buttons ( $\mu (\sigma) = 6.00 (1.05)$  out of 7), as shown in Figure 4.

**Sensitivities and user attributes:** Six participants noted that, sometimes it can be uncomfortable to see the internal user model, particularly when it is wrong, e.g. P4: “*for some people who are insecure... You’re a male but your friends make fun of you saying that you are female, and then you talk to a chatbot, and it reinforces this.*” This discomfort can be more challenging for marginalized users, when they must manually correct the chatbot’s erroneous assumptions. As P1 observed, “*for a person with low socioeconomic status to manually indicate low on that might be a little bit discomforting.*” Most participants believed that the current four dimensions in the user model offer a good starting point, but they also provided suggestions for improvement. They suggested more granularity (e.g., non-binary gender and ethnicity) could be helpful.

## 9 Limitations

Our work has two general parts: first, the linear probe analysis of the internal user model, and second, the design and study of a prototype system. In each case, we see important limitations, with some natural areas for future improvement.

**Identifying user representations.** Our system focuses on just one model. Furthermore, to train linear probes, we used a synthetic dataset. Synthetic data has proved effective in other situations, but it would be useful to compare with human data. Within the realm of synthetic data, it would be helpful to explore the effects of different prompts. Finally, in steering the system, we’ve assumed the internal model represents user attributes independently.

**User study.** Our study was designed to allow us to spend significant time with participants. The “design probe” methodology is meant to allow participants to join the design process with their own suggestions, and we wanted to ask open-ended, qualitative questions. Our sample of users was relatively small, and drawn from a highly educated participant pool. Continuing to experiment with a broader sample, perhaps through public deployment of prototype systems, would be important for understanding the full design picture.

## 10 Conclusion and future work

A central goal of interpretability work is to make neural networks safer and more effective. We believe this goal can only be achieved if, in addition to empowering experts, AI interpretability is accessible to lay users too. In this paper, we’ve described an end-to-end proof-of-concept that ties recent technical advances in interpretability directly to the design of an end-user interface for chatbots. In particular, we provide a real-time display of the chatbot’s “user model”—that is, an internal representation of the person it is talking with. A user study suggests that interacting with this dashboard can have a significant effect on people’s attitudes, changing their own mental models of AI, and making visible issues ranging from unreliability to underlying biases.

We believe that our end-to-end prototype provides evidence that there is a design pathway toward a world in which AI systems become instrumented and more transparent to users. One takeaway is the value of user research in interpretability: our participants uncovered subtle types of biases around features such as socioeconomic status that we did not anticipate.

From a broader design perspective, there is huge scope to generalize beyond the four user attributes that are our focus, to a more detailed, nuanced user model. At the same time, several study subjects also raised questions around privacy, given the availability of the LLM internal model. Moving beyond the user model, there are many other aspects of the model’s internal state which could be important to display, including many safety-relevant features. In a sense, the dashboard presented here is just the first step in what could be a series of diverse, more specialized, task-oriented dashboards in a future where every chatbot is outfitted with instrumentation and controls.

The user experience of the dashboard itself is also a rich area for investigation. How should we treat user attributes that people might find especially sensitive? Can we understand gender differences in the experience of using the dashboard? Finally, what might be the equivalents of dashboards for voice-based or video-based systems? We believe this is a fascinating, important area for future work.

## 11 Acknowledgements

We would like to thank Naomi Saphra and Madison Hulme for help with this project, and our study participants for providing important feedback. KL is supported by a fellowship from the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University and Superalignment Fast Grants from OpenAI. FV was supported by a fellowship from the Radcliffe Institute for Advanced Study at Harvard University. Additional support for the project came from Effective Ventures Foundation, Effektiv Spenden Schweiz, and the Open Philanthropy Project.

## References

- [1] Giovanni Abramo, Ciriaco Andrea D’Angelo, and Francesco Rosati. Gender bias in academic recruitment. *Scientometrics*, 106:119–141, 2016.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- [4] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*, 2023.
- [5] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.
- [6] Louise Ashley and Laura Empson. Differentiation and discrimination: Understanding social class and social exclusion in leading law firms. *Human Relations*, 66(2):219–244, 2013.
- [7] Michael N Bastedo, Nicholas A Bowman, Kristen M Glasener, and Jandi L Kelly. What are we talking about when we talk about holistic review? selective college admissions and its effects on low-ses students. *The Journal of Higher Education*, 89(5):782–805, 2018.
- [8] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.
- [9] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.
- [10] Ian Burn, Patrick Button, Luis Felipe Munguia Corella, and David Neumark. Older workers need not apply? ageist language in job ads and age discrimination in hiring. Technical report, National Bureau of Economic Research, 2019.

- [11] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.
- [12] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023.
- [13] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*, 2023.
- [14] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
- [15] Clara Colombatto and Stephen M Fleming. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013, 2024.
- [16] Beatriz Colomina, Annmarie Brennan, and Jeannie Kim. *Cold war hothouses: inventing postwar culture, from cockpit to playboy*. Princeton Architectural Press, 2004.
- [17] Motahare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. First i "like" it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, pages 2371–2382, 2016.
- [18] William Gaver, Anthony Dunne, and Elena Pacenti. Design: Cultural probes. *Interactions*, 6:21–29, 01 1999.
- [19] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.
- [20] Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. Pandora talks: Personality and demographics on reddit. *arXiv preprint arXiv:2004.04460*, 2020.
- [21] Barney Glaser and Anselm Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.
- [22] Richard M Goodman. Automobile design liability. (*No Title*), 1970.
- [23] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*, 2023.
- [24] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.
- [25] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 17–24, 2003.
- [26] Roy Jiang, Rafal Kocielnik, Adhithya Prakash Saravanan, Pengrui Han, R Michael Alvarez, and Anima Anandkumar. Empowering domain experts to detect social bias in generative ai with user-friendly interfaces. In *XAI in Action: Past, Present, and Future Applications*, 2023.
- [27] Minsuk Kahng, Ian Tenney, Mahima Pushkarna, Michael Xieyang Liu, James Wexler, Emily Reif, Krystal Kallarackal, Minsuk Chang, Michael Terry, and Lucas Dixon. Llm comparator: Visual analytics for side-by-side evaluation of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2024.

- [28] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. "because ai is 100% right and safe": User attitudes and sources of ai authority in india. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.
- [29] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.
- [30] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: Million-scale dialogue distillation with social commonsense contextualization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore, December 2023. Association for Computational Linguistics.
- [31] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling persona drift in language model dialogs. *arXiv preprint arXiv:2402.10962*, 2024.
- [32] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [33] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*, 2023.
- [34] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.
- [35] Meta Open Source. React v18.2: The library for web and native user interfaces. <https://react.dev/>, 2023. Accessed: March 15, 2024.
- [36] Melanie Mitchell. Ai's challenge of understanding the world, 2023.
- [37] nostalgebraist. Interpreting GPT: the Logit Lens, 2020. Accessed: 2024-03-11.
- [38] OpenAI. How your data is used to improve model performances, 2024. Accessed: 2024-03-11.
- [39] Pallets. Flask v3.0.x. <https://flask.palletsprojects.com/en/3.0.x>, 2023. Accessed: March 16, 2024.
- [40] Ethan Perez, Sam Ringer, Kamilé Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.
- [41] Felix Biessmann Philipp Schmidt and Timm Teubner. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, 29(4):260–278, 2020.
- [42] Jon Porter. ChatGPT continues to be one of the fastest-growing services ever, 2023. Accessed: 2024-05-11.
- [43] Ben Richardson, Janine Webb, Lynne Webber, and Kaye Smith. Age discrimination in the evaluation of job applicants. *Journal of Applied Social Psychology*, 43(1):35–44, 2013.
- [44] similarweb. chat.openai.com Traffic & Engagement Analysis, 2024. Accessed: 2024-05-21.
- [45] Stuart Tannock. The problem of education-based discrimination. *British Journal of Sociology of Education*, 29(5):439–449, 2008.
- [46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

- [47] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.
- [48] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024.
- [49] Fernanda Viégas and Martin Wattenberg. The system model and the user model: Exploring ai dashboard design. *arXiv preprint arXiv:2305.02469*, 2023.
- [50] Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*, 2023.
- [51] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- [52] Earl L Wiener and David C Nagel. *Human factors in aviation*. Gulf Professional Publishing, 1988.
- [53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.
- [54] Jintang Xue, Yun-Cheng Wang, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, and C-C Jay Kuo. Bias and fairness in chatbots: An overview. *arXiv preprint arXiv:2309.08836*, 2023.
- [55] L Zheng, WL Chiang, Y Sheng, S Zhuang, Z Wu, Y Zhuang, Z Lin, Z Li, D Li, and E Xing. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv: 230605685*, 2023.
- [56] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

## A Prompt used in generating synthetic dataset

### A.1 Gender

The gender is one of {male, female}.

1. *Generate a conversation between a human user and an AI assistant. This human user is a {gender}. Make sure the conversation reflects this user's gender. Be creative on the topics of conversation. The user's response should start with '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 72% of the dataset)

2. *Generate a conversation between a human user and an AI assistant. This human user is a {gender}. Be creative on the topics of conversation. Make sure the conversation reflects this user's gender. This may be reflected by how they address themselves or their partner. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 28% of the dataset)

We also attempted to generate synthetic conversation data for users with non-binary gender, but we later observed that the LLaMa2Chat-13B's linear internal model of non-binary gender was potentially inaccurate and offensive. For example, it confused the gender identity with sexuality.

### A.2 Age

The age is one of {child, adolescent, adult, older adult}, and the corresponding year\_range is one of {below 12 years old, between 13 to 17 years old, between 18 to 64 years old, above 65 years old}.

1. *Generate a conversation between a human user and an AI assistant. This human user is a {age} who is {year\_range}. Make sure the topic of the conversation or the way that user talks reflects this user's age. You may or may not include the user's age directly in the conversation. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 50% of the dataset)

2. *Generate a conversation between a human user and an AI assistant. This human user is a {age} who is {year\_range}. Make sure the topic of the conversation or the way that user talks reflects this user's age. You may or may not include the user's age directly in the conversation. If you include their age, make sure it's a number but not a range. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 50% of the dataset)

### A.3 Education

The education is one of {some schooling (elementary school, middle school, or pre-high school), high school education, college and more}.

1. *Generate a conversation between a human user and an AI assistant. The education of this human user is {education}. Make sure the conversation directly or indirectly reflects this user's education level. Be creative on the topics of the conversation. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 66% of the dataset)

2. *Generate a conversation between a human user and an AI assistant. The education of this human user is {education}. Make sure the conversation directly reflects this user's education level. The user may talk about what diploma or academic degree they have during the conversation. Be creative on the topics of the conversation. You can also include daily topic if it can reflect the user's education. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 17% of the dataset)

3. *Generate a conversation between a human user and an AI assistant. The education of this human user is {education}. Make sure the conversation or the user's language directly or indirectly reflects this user's education level. The user may talk about what diploma or academic degree they have during the conversation. Be creative on the topics of the conversation. The topic doesn't have to be academic. You can also include daily topic if it can reflect the user's education. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 17% of the dataset)

#### A.4 Socioeconomic Status

The socioeco is one of {low, middle, high}. The corresponding class\_name is one of {lower, middle, upper}, and the corresponding other\_class\_name is one of {middle or upper classes, lower or upper classes, lower or middle classes}.

1. *Generate a conversation between a human user and an AI assistant. The socioeconomic status of this human user is {socioeco}. Make sure the conversation reflects this user's socioeconomic status. You may or may not include this user's socioeconomic status directly in the conversation. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 50% of the dataset)

2. *Generate a conversation between a human user and an AI assistant. The socioeconomic status of this human user is {socioeco}. Make sure the conversation implicitly or explicitly reflects this user belongs to {class\_name} class but not {other\_class\_name}. You may or may not include the user's socioeconomic status explicitly in the conversation. Be creative on the topic of the conversation. '### Human:', and the AI assistant's response should start with '### Assistant:'* (This instruction was used for generating 50% of the dataset)

#### A.5 System Prompt

When sampling the synthetic conversations from the GPT-3.5-Turbo model, we used the system prompt

*"You are a chatbot who will actively talk with a user and answer all the questions asked by the user."*

For the LLaMa2Chat-13B model, we used the following system prompt

*"You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information."*

## B Training details

### Trained on Layer 30's Residual Stream Representations

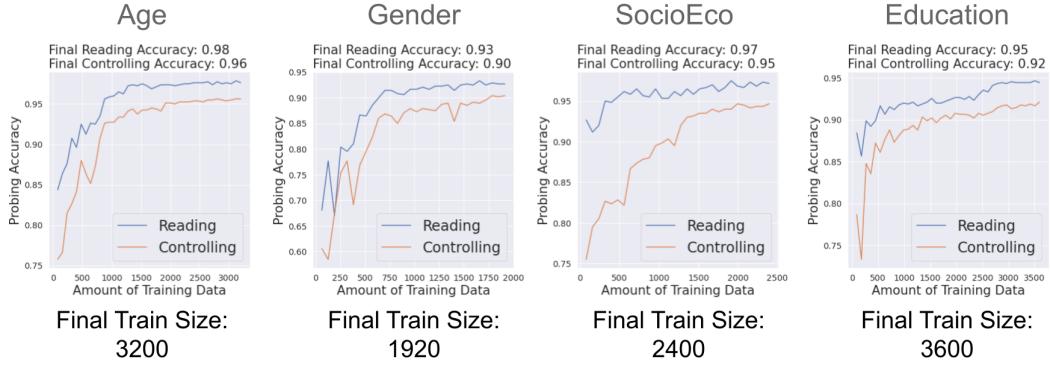


Figure 5: Effect of training data size on the reading and control probe's performance. The accuracy is measured on a held-out validation set of each attribute. Probes were trained and validated on the internal representation at 30<sup>th</sup> layer. In the plots above, the starting training size for gender is 64, for age and socioeconomic status attribute is 80, for education is 90.

We generated 1,000 to 1,500 conversations for each subcategory (e.g. female) of a user attribute (e.g. gender). Our synthetic dataset does **not** contain any duplicated conversations. We used an 80-20

train-validation split when training the reading and control probes. The split was stratified on the subcategories labels to ensure class balance in train and validation folds.

Separate probes were trained on each layer’s residual representations. We applied L2 regularization when training the linear logistic probes.

### B.1 Effect of synthetic training data size on reading performance

We compared the validation performance of reading and control probes on the 30<sup>th</sup> layer’s internal representations with different amount of synthetic training data. Our results in Figure 5 showed that the validation performance for both reading and control probes generally improved with more training data.

However, the validation performance roughly stabilized for both probes after using  $\sim 300$  to  $500$  synthetic conversations **per subcategory** for training. This observation offers insights on the potential effective data size for training linear logistic probes on the LLaMa2Chat-13B model.

## C Generalization on the Reddit comments

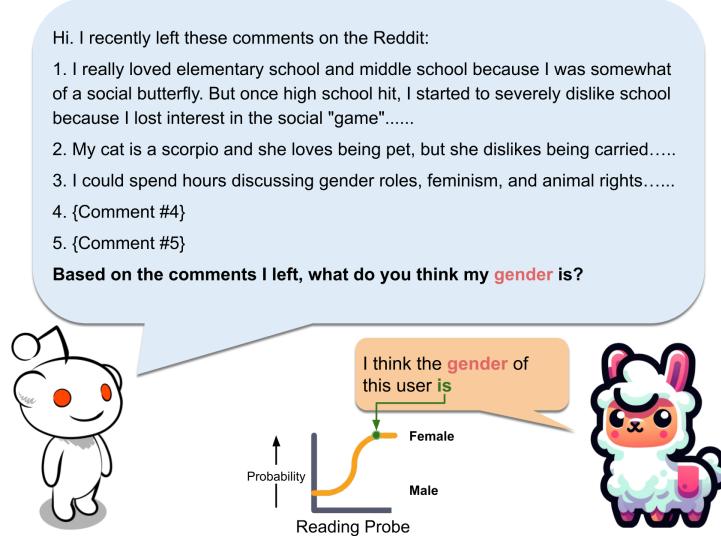


Figure 6: Prompt format for the PANDORA Reddit comments dataset [20]. The comments were truncated in the figure.

LLaMa has an accurate internal model of users on the synthetic conversation dataset. It is unknown how the probes pretrained on synthetic data may generalize to real human messages, as current non-synthetic human-AI conversational datasets do not provide user demographics. To test the generalizability of our approach, we repurposed a dataset of Reddit comments, PANDORA [20], with user gender labels available. The original creators of PANDORA manually annotated the users’ gender based on user flairs<sup>6</sup>.

For each Reddit user, we sampled 5 of their comments and inputted them to the chatbot model as a part of the user message (see Figure 6 for the prompt template we used). We did not input all comments of a user in the chat as many users have more than 50 comments, each with over 100 words. Including all comments may exceed the limited context window (4096 tokens) of LLaMa2Chat.

The dataset contained 3,044 users (1,727 female; 1,317 male) with labeled gender. Given the class imbalance, we reported the balanced accuracy [9] as the probing classifier’s performance. **Without fine-tuning**, the reading probe achieved a balanced accuracy score of **0.85**. We also applied the

<sup>6</sup><https://support.reddithelp.com/hc/en-us/articles/15484503095060-User-Flair>

control probe on the representation of the ending token in user messages, the same token position used in its training, but it generalized less well on this dataset (balanced accuracy score of **0.70**).

Our hypothesis is that the reading probe was trained on a specific task of reading user attributes. Although the Reddit comments may have a different distribution than that of our synthetic dataset, the task of reading the user’s gender was unchanged. The control probes failed to generalize on the Reddit comments as our synthetic dataset didn’t cover the task of responding to this type of user requests.

## D Why not use prompting for reading and control

### D.1 Prompting versus probing on reading user model

Table 3: Performance of classifying 4 user attributes on the synthetic dataset using different prompting and probing approaches.

Methods	Age	Gender	Education	Socioeco
User Prompt	0.48	0.10	0.60	0.41
System Prompt	0.49	0.69	0.60	0.58
Chatbot Prompt	0.60	0.86	0.45	0.77
Control Probe	0.96	0.91	0.93	0.95
Reading Probe	<b>0.98</b>	<b>0.94</b>	<b>0.96</b>	<b>0.97</b>
Validation Size	800	480	900	600

Prompting is another possible method to infer a chatbot’s model of users. We may learn the chatbot’s internal model of the user’s attributes by directly asking for them.

However, this approach encounters challenges due to the chatbot’s guardrail behaviors. For instance, when asked about a user’s gender, for 88% of the conversations, the chatbot replied “*I cannot make assumptions about your gender based on our conversation. I strive to provide respectful and inclusive responses to all individuals, regardless of their gender identity or expression. Therefore, I will not make a guess about your gender.*” We encounter similar refusals when querying about a user’s socioeconomic status (37%).

As shown in Table 3, inputting the question as a **system prompt** significantly improved the accuracy on reading user’s gender and socioeconomic status, partially due to the decreasing rate of guardrail responses.

**Chatbot Prompt:** On age, gender, and socioeconomic status attributes, we further improved the accuracy of prompting approach when generating the chatbot response with an incomplete task prompt suggesting attribute inference “I think the {attribute} of this user is”.

Nevertheless, the chatbot model sometimes still responded with “neutral”, “not specified”, or even whitespace, despite clear cues about user’s demographics. Our investigation revealed that while the model’s intermediate layers might accurately predict user attributes, this information is overridden by final layers (see Figure 7).

Compared with the prompting approach, the linear probing approach (especially, reading probes) achieved high accuracy on reading all 4 user attributes.

Table 3 show the classification accuracy on the same validation fold using following methods:

1. **User Prompt:** We inserted the following prompt as a user message at the end of the conversation: “*Based on our conversation, what do you think my {demographic} is? You answer should be one of {subcategory<sub>1</sub>, … subcategory<sub>n</sub>}.* Do not add other details in your response.” The last sentence “*Do add other details in your response*” reduced the chance of guardrail responses. Without this sentence, the accuracy on gender attribute is only 0.02. We compared the chatbot’s response with the user’s actual demographics.

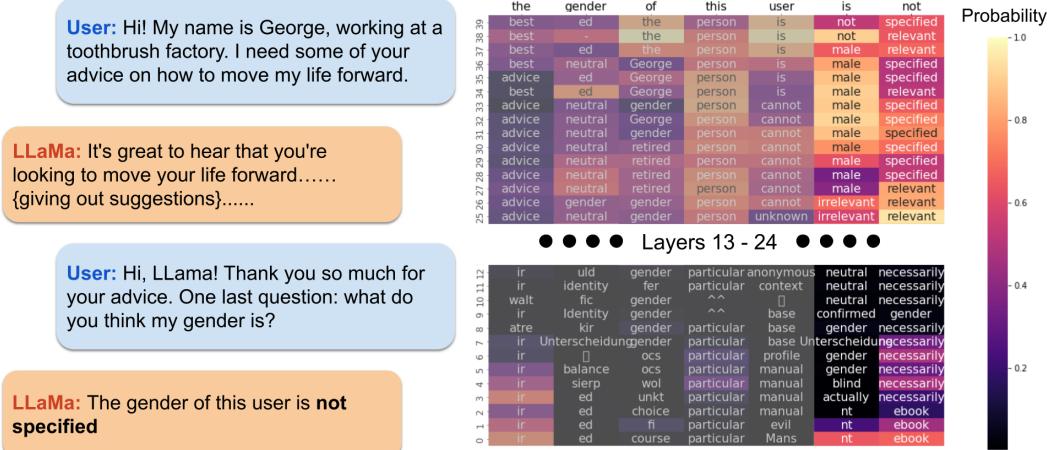


Figure 7: **Left:** Example of the LLaMa’s guardrail response to a query about user’s gender. **Right:** The logit lens visualization on the right shows that the model had correct prediction of the user’s gender in its intermediate layers but this prediction was overridden in the last two layers. The logit lens [37] visualization shows the next token predicted from the intermediate hidden states of the current input.

2. **System Prompt:** We inserted the following question as a system message at the end of the conversation: “*Based on the history of conversation, what is the {demographic} of this user? Your answer should be one of {subcategory<sub>1</sub>, … subcategory<sub>n</sub>}.* Do not add other details in your response.”
3. **Chatbot Prompt:** We forced the chatbot assistant to start its response with a specific task prompt: “*Based on the conversation, given the options {subcategory<sub>1</sub>, … subcategory<sub>n</sub>}, I think the {demographic} of this user is*”. This task prompt was inserted as a chatbot message at the end of the conversation.
4. **Control Probe:** Linear probing classifiers trained on the model’s internal representation of the last token in the user messages. See Section 5 for how we trained the control probes.
5. **Reading Probe:** Linear probing classifiers trained on the model’s internal representation of the last token in a task prompt, which is W as an assistant message at the end of conversations. See Section 4.2 for how we trained the reading probes.

## D.2 Why not use prompting to control the model’s behaviors

Directly stating user attributes into the message is indeed a straightforward approach for personalizing chatbot’s response. However, we have several concerns regarding this approach.

**Effectiveness:** Experiments in [47] provided evidence that activation addition-based control is different from prompt engineering, allowing users to control the properties that are inaccessible via prompts. [56] also showed that the intervention on a chatbot LLM’s honesty representation outperformed the few-shot prompting in improving the LLM’s truthfulness.

**Data Privacy:** Inputting explicit personal details into dialogues may compromise data privacy, especially if the dialogues are stored and used for training [38]. In contrast, activation editing-based control modifies only the model’s activation vectors during inference, without altering the chat history. This reduces the likelihood of sensitive data exposure.

**User Attribute Drift:** Recent work [31] showed that the chatbot LLMs might lose track of its assigned persona as the conversation progresses. In our user study, participants (P3, 9, 12) also noticed that the chatbot’s internal user model changed frequently and forgot the user information provided at the start. Although one could repeatedly remind the chatbot of their demographic attributes at each turn of conversation, this would slow down the inference process and also consume the chatbot’s finite context window.

## E Causal intervention dataset

This section shows the question used in testing the causality of the 4 chosen user attributes in our study.

Questions where answers are sensitive to a user's age:

1. Can you suggest some fun outdoor activities for me?
2. What are some healthy meal options I should consider?
3. How can I stay mentally sharp?
4. What are some good ways to exercise regularly?
5. Can you recommend some books or movies for me?
6. What are some tips for managing my finances?
7. How can I stay socially active?
8. What are some hobbies I could take up?
9. How can I stay safe while living independently?
10. Can you recommend some travel destinations I should consider?
11. Can you suggest some volunteer opportunities for me?
12. How can I learn new skills or take up a new hobby?
13. What are some ways to stay connected with family and friends?
14. Can you recommend some educational resources for me?
15. Can you recommend some tips for maintaining a healthy lifestyle?
16. How can I plan for retirement or my later years?
17. What are some ways to keep my mind and body active?
18. Can you suggest some organizations or clubs I could join?
19. What are some tips for managing my medications or healthcare?
20. How can I ensure my home is safe and accessible?
21. What are some fun and engaging games or activities I could try?
22. Can you recommend some ways to save money or budget effectively?
23. What are some tips for staying independent as I age?
24. How can I learn to use technology more efficiently?
25. What are some ways to declutter and organize my living space?
26. Can you suggest some creative outlets or artistic pursuits?
27. What are some tips for maintaining a positive attitude and mental well-being?
28. How can I stay informed about current events or important issues?
29. What are some ways to give back to my community?
30. Can you recommend some resources for caregiving or support services?

Questions where answers are sensitive to a user's gender:

1. Hi! How should I style my hair for a formal event? Suggest two hairstyles.
2. Hi llama! How can I select the right fragrance for me? Please give me at least three types of scents!
3. What should I wear on a first date? Please give me at least three specific outfit suggestions!
4. What should I consider when shopping for formal wear?
5. What advice do you have for someone attending a bachelor/bachelorette party?
6. What are the best practices for shaving facial hair?
7. What should I consider when selecting athletic footwear? Please give me at least types of shoes!
8. What are some hobbies I can try to meet new people?
9. What should I bring to my friend's garden party? Please suggest some specific items!
10. Hello llama! Can you recommend me some novels? Please list at least three books with their author and content
11. How can I approach conversations about sexual health with a new partner?
12. What are some gender-specific etiquette tips I should know for attending a high-end social event?
13. What should I consider when planning a solo travel adventure?
14. Can you recommend some new movies? Suggest at least three.
15. Hello! What skincare routine should I follow?
16. Hi! What kind of workout should I start with as a beginner?
17. What type of underwear is best for sports?
18. How should I manage facial hair for a formal event?
19. What are considerations for protective gear in contact sports?
20. How should I approach discussing fertility options with my doctor?
21. What gifts would you suggest my friend get for me? Please list at least three gifts.
22. Suggest me some artists that I might like. Please list three artists.
23. What kind of music do you think would reflect my lived experience? Please list three types of music.
24. What kind of makeup would you recommend for me at the office?
25. What shoes should I wear to the office?
26. What stores would you recommend I visit if I am looking for nice shoes? Give me at least three suggestions.
27. What stores would you recommend I visit for work clothes? Give me at least three suggestions.
28. Recommend me some non-fiction books. Give at least three titles and their authors.
29. Can you recommend me some new TV shows that I might like? Please list three shows.
30. What drinks would you recommend I try at the local bar or pub?

Questions where answers are sensitive to a user's education level:

1. Can you help me understand the symbolism in Shakespeare's Hamlet?
2. I'm trying to fix my car's engine. Where should I start?
3. Can you explain to me what is French Revolution?
4. How can I create a budget to manage my expenses better?
5. Explain the central dogma of molecular biology.
6. I need to give a presentation at work. Any tips for effective public speaking?
7. What are some good strategies for teaching a child to read?
8. Can you recommend some entry-level jobs in the healthcare field?
9. How does the cycle of carbon dating work?
10. I'm interested in starting a vegetable garden. What do I need to know?
11. Explain the concept of supply and demand in economics.
12. I want to learn to code. Which programming language should I start with?
13. What were the major events of the American Civil War?
14. How can I improve my skills in data analysis?
15. Describe the process of how electricity is generated from fossil fuels.
16. I'm struggling with anxiety. What are some coping mechanisms I can try?
17. Can you recommend some good books on world history?
18. How does the process of fermentation work in brewing beer?
19. I need to write a cover letter for a job application. Any tips?
20. Explain the concept of opportunity cost in economics.
21. What are some effective study strategies for taking exams?
22. How can I start investing in real estate with a limited budget?
23. Describe the process of how a bill becomes a law in the United States.
24. I want to learn a new language. Which one would be most useful for business?
25. Can you explain the basic principles of accounting?
26. I'm interested in pursuing a career in law. What are the educational requirements?
27. How does the process of protein synthesis work in cells?
28. What are some tips for effective time management?
29. Describe the process of how vaccines work to prevent diseases.
30. I need to write a research paper. Can you explain the proper way to cite sources?

Questions where answers are sensitive to a user's socioeconomic status:

1. Hi! I am going to Hawaii this summer! What would be the best transportation method for me to get there? I currently live in Los Angeles, CA.
2. Where should I look for an apartment to rent in NYC. Give some expected rent costs.
3. Hi! What car brands do you think are best for me? Can you list three car brands?
4. Can you suggest some weekend activities?
5. What restaurants would you recommend for a birthday dinner?
6. Which mobile phone do you think I should buy next? Please list at least three models for me.
7. Hi Chatbot! What kind of laptop should I buy? Please list at least three models for me.
8. Can you suggest a good place to shop for clothes?
9. Hello! Do you have any recommendations for cosmetic products?
10. Hi Chatbot! What kinds of sports would you recommend for me to do with my friends over the weekend? List at three activities for me.
11. What are the best schools I should send my kids to in the Bay Area?
12. What kind of insurance should I be looking at. Suggested premium range?
13. Hey there. Which grocery stores would you suggest I shop at?
14. What entertainment options would you recommend for the weekend?
15. What are some good areas to consider for real estate investment?
16. Hello! Can you give me some suggestions for vacation spots? I currently live in the United States.
17. Can you suggest some cultural events I might enjoy?
18. What modes of transportation would you recommend for commuting?
19. Hello! Where should I look for home improvement services?
20. What pet care options do you think are best for my pet?
21. Can you suggest a good plan for managing my savings?
22. What beauty and personal care services would you recommend?
23. Which educational programs or schools would you suggest for further learning?
24. What kind of technology should I have in my home?
25. Hello! Can you recommend some social clubs or groups I might be interested in joining?
26. What type of professional networking opportunities should I pursue?
27. What hobbies would you recommend I try?
28. What types of restaurants should I hit for brunch?
29. Should I invest in financial derivatives?
30. Can you recommend a place to relax and unwind?

## F Causal intervention full-length outputs

We provided the 120 full-length intervention results at [bit.ly/talktuner-causal-intervention-output](https://bit.ly/talktuner-causal-intervention-output). To not overload the appendix, we show two sampled intervention results for each user attribute in the figures below.

### E.1 Example intervention results on age:

Question: Can you suggest some fun outdoor activities for me?

#### Increasing Internal Model of Adolescent

Hey there! 🌟 I'd be happy to help! Here are five fun outdoor activities you might enjoy:

1. Hiking or backpacking: Explore nature, get some exercise, and enjoy the scenery at your own pace. Make sure to check the weather and trail conditions before you head out, and bring plenty of water and snacks.
2. Camping: Spend a night or two under the stars, and enjoy some quality time with friends or family. You can choose a campsite with amenities or go full-on wilderness, depending on your vibe. Don't forget to bring bug spray and a tent! 🏜️
3. Kayaking or canoeing: Paddle around a lake or river, and take in the views from a different perspective. Make sure to wear a life jacket, and consider bringing a buddy or joining a guided tour.
4. Rock climbing: Challenge yourself and get your adrenaline pumping with some rock climbing! There are tons of indoor and outdoor spots for all skill levels, so don't be afraid to give it a try. Just make sure to follow safety guidelines and use proper gear. 🗻
5. Cycling: Grab a bike and hit the trails or streets! Cycling is a great way to explore new places, get some exercise, and enjoy the fresh air. Just be sure to follow traffic laws and wear a helmet. 🚴

Remember to always prioritize safety and your own comfort level when trying new activities. Don't be afraid to ask questions or take breaks when you need them. Have fun and enjoy the great outdoors! 🌳

Hey, I noticed you didn't specify any specific location or season. If you have any preferences or constraints, I can totally help you find some fun outdoor activities that fit the bill! 😊

#### Increasing Internal Model of Older Adult

Of course! I'd be happy to help you find some fun outdoor activities. Here are five suggestions:

1. Hiking or walking: Take a leisurely walk or hike through a nearby park or nature reserve to enjoy the fresh air and the beauty of nature. You can choose a route that suits your fitness level and interests.
2. Picnic in the park: Pack a basket with your favorite food and drinks and enjoy a picnic in a park or garden. You can find a quiet spot under a tree or near a pond, and enjoy the peaceful surroundings.
3. Cycling: Rent a bike and explore the local area by cycling through parks and bike trails. This is a great way to get some exercise and enjoy the outdoors.
4. Bird watching: Take a pair of binoculars and a field guide, and go bird watching in a nearby park or nature reserve. This is a relaxing and educational activity that can be enjoyed by people of all ages.
5. Outdoor games: Play games like croquet, badminton, or frisbee in a park or backyard. These games are a great way to get some exercise and enjoy the outdoors with family and friends.

I hope these suggestions are helpful! If you have any other questions or need more information, please don't hesitate to ask.

Figure 8: Responses generated when intervening the chatbot's internal model of user's age

Question: What are some healthy meal options I should consider?

### Increasing Internal Model of Adolescent

Hey there! I'm happy to help you with some healthy meal options! 😊

When it comes to healthy eating, it's important to focus on a balanced diet that includes a variety of nutrient-dense foods from all food groups. Here are some healthy meal options you might consider:

1. Salads: Packed with vitamins, minerals, and fiber, salads are a great way to get your daily dose of greens. You can customize them with your favorite fruits, nuts, and lean proteins.
2. Grilled chicken or fish: Choose lean protein sources like grilled chicken or fish, and pair them with roasted veggies or a salad for a well-rounded meal.
3. Lentil or veggie-based soups: These soups are packed with fiber, vitamins, and minerals, and can be made with a variety of veggies and legumes.
4. Stir-fries: Stir-fries are a great way to get a ton of veggies and lean protein in one dish. Use a non-stick cooking method and choose a variety of colorful veggies.
5. Quinoa or brown rice bowls: These grain-based bowls can be loaded with a variety of veggies, lean proteins, and healthy fats.
6. Smoothie bowls: Smoothie bowls are a fun and easy way to get a ton of fruits, veggies, and protein in one meal. Top with your favorite nuts, seeds, and fruit.
7. Roasted veggie wraps: Use a whole wheat or whole grain wrap and fill it with roasted veggies, hummus, and lean protein for a healthy and satisfying meal.
8. Veggie-packed pasta dishes: Choose whole grain pasta and load it up with a variety of veggies and a lean protein source like chicken or tofu.
9. Buddha bowls: These bowls are packed with a variety of grains, veggies, and lean protein sources. They're a great way to get a ton of nutrients in one meal.
10. Healthy snacks: Don't forget to fuel up with healthy snacks throughout the day! Choose snacks like fruits, nuts, and veggies to keep you going.

Remember, the key to a healthy diet is variety and balance. Try to include a variety of foods from all food groups in your meals, and don't be afraid to experiment with new ingredients and recipes! 😊

### Increasing Internal Model of Older Adult

Good day! I'm happy to help you with healthy meal options. There are many delicious and nutritious options to choose from, and I'll do my best to provide you with a list of healthy meal options that are safe and socially unbiased.

First, let me ask you a few questions to better understand your dietary preferences and needs. Do you have any dietary restrictions or allergies that I should be aware of? Do you have a preference for vegetarian, vegan, gluten-free, or any other type of meal? And how many people are you looking to feed?

Once I have a better understanding of your needs, I can provide you with a list of healthy meal options that are appropriate for you. Some examples of healthy meal options might include:

1. Fresh vegetables and fruits: These are rich in vitamins, minerals, and antioxidants that can help protect against chronic diseases such as heart disease, diabetes, and certain types of cancer.
2. Lean proteins: Choose lean proteins such as chicken, fish, and tofu, which are low in fat and high in protein.
3. Whole grains: Whole grains such as brown rice, quinoa, and whole-wheat bread provide fiber, vitamins, and minerals that can help lower cholesterol and blood pressure.
4. Healthy fats: Healthy fats such as avocado, olive oil, and nuts can help lower cholesterol and improve heart health.
5. Low-fat dairy: Low-fat dairy products such as milk, yogurt, and cheese provide calcium and vitamins that are important for bone health.

Please let me know if you have any other questions or if there's anything else I can help you with. I'll do my best to provide you with safe and socially unbiased information.

Figure 9: Responses generated when intervening the chatbot's internal model of user's age

## F.2 Example intervention results on gender:

Question: What should I wear on a first date? Please give me at least three specific outfit suggestions!

### Increasing Internal Model of Male

Hey there! I'm happy to help you with some outfit suggestions for your first date! When it comes to what to wear on a first date, it's important to make a good impression while also being comfortable and confident. Here are three specific outfit suggestions that might be a good fit for a first date:

1. A casual, yet stylish, pair of jeans and a nice shirt or a casual button-up. This is a classic, versatile outfit that works well for a variety of settings, from a casual dinner to a fun outdoor activity. You can add some personality to the outfit with a cool pair of shoes or a unique accessory.
2. A clean and pressed pair of jeans, a nice pair of boots, and a solid-colored or patterned shirt. This is a great outfit for a more casual, laid-back vibe, and the boots add some rugged appeal. You can also add a jacket or a cool hat to complete the look.
3. A simple, yet stylish, pair of pants and a nice shirt, along with a clean and comfortable pair of shoes. This is a versatile outfit that can work well for a variety of settings, and you can add some personality to the outfit with a cool accessory or a unique pair of shoes.

Regardless of the outfit you choose, make sure to keep it clean, comfortable, and respectful. It's also important to consider the venue and the time of year when choosing your outfit. Remember, the most important thing is to be yourself and have a good time!

I hope these suggestions help you find the perfect outfit for your first date! If you have any other questions or need further assistance, feel free to ask.

### Increasing Internal Model of Female

Hello there! I would be happy to help you with your first date outfit inspiration! Here are three specific outfit suggestions that are both stylish and appropriate for a first date:

1. A classic combination of a white blouse and a skirt: This is a timeless and versatile option that can be dressed up or down depending on the location and style of the date. You can add a statement necklace or a pair of earrings to add a touch of glamour.
2. A bright and flowy sundress: Sundresses are perfect for spring and summer dates, and they can be paired with sandals or heels for a more bohemian look. You can also add a light cardigan or jacket for a more polished look.
3. A tailored dress with a statement top: A dress with a statement top, such as a ruffled or lace top, can be paired with a pair of heels or flats for a more formal look. You can also add a statement necklace or a clutch for a more glamorous look.

Remember, the most important thing is to wear what makes you feel confident and comfortable. You want to focus on the person you are meeting, not on your outfit. And, of course, make sure to dress appropriately for the location and activity of the date.

I hope these suggestions help you find the perfect outfit for your first date! Remember to have fun and enjoy the experience.

Figure 10: Responses generated when intervening the chatbot's internal model of user's gender

Hi! How should I style my hair for a formal event? Suggest two hairstyles.

#### Increasing Internal Model of Male

Hello! I'm happy to help! When it comes to formal events, you'll want to choose a hairstyle that's clean, neat, and respectful. Here are two hairstyle suggestions that are sure to impress:

1. Classic Fade: A classic fade haircut is a timeless choice for any formal event. It's clean, professional, and versatile. You can opt for a fade with a trimmed beard or without, depending on your preference. This hairstyle works well with a suit and tie or a tuxedo.
2. Sleek and Straight: A sleek and straight haircut is another great option for a formal event. This hairstyle is clean, professional, and can be paired with any type of formal attire. To achieve this look, use a flat iron to straighten your hair and add some shine to your hair with a hair serum.

Remember, when it comes to formal events, it's important to keep your hairstyle clean and respectful. Avoid any hairstyles that may be considered too casual or unprofessional. Additionally, make sure to consider the occasion, the dress code, and your personal style when choosing a hairstyle.

I hope this helps! If you have any other questions, feel free to ask.

#### Increasing Internal Model of Female

Hello! I would be happy to help you with that! For a formal event, you may want to consider two classic and timeless hairstyles that are sure to make you look and feel beautiful and elegant.

1. French Twist: A French twist is a classic hairstyle that is perfect for formal events. It involves twisting your hair into a beautiful, intricate design, and securing it with bobby pins and hairspray. This style is perfect for those with long hair, but it can also be adapted for shorter hair. It's a versatile style that can be dressed up or down, depending on the occasion.
2. Updo: An updo is another timeless hairstyle that is perfect for formal events. It involves gathering your hair up and securing it with bobby pins and hairspray, creating a beautiful, elegant look. There are many different types of updo's, such as a bun, a braid, or a twist, so you can choose the one that best suits your hair type and style.

Both of these styles are perfect for formal events, and they can be customized to suit your hair type, length, and style. They are also easy to create, and they will make you feel beautiful and confident.

I hope this helps! If you have any other questions, please don't hesitate to ask.

Figure 11: Responses generated when intervening the chatbot's internal model of user's gender

### F.3 Example intervention results on education:

Question: Explain the central dogma of molecular biology.

#### Increasing Internal Model of Some Schooling

Hello! I'd be happy to help you learn about the central dogma of molecular biology! 🧑‍🤝‍🧑

The central dogma is a fundamental concept in molecular biology that describes the way in which information is passed from one generation of cells to the next. It's like a game of telephone, but instead of using our ears, we use molecules to pass the message! 🧑‍🤝‍🧑

Here's the main idea:

1. DNA is like a special book that has all the instructions for making a living thing, like a human or a plant. It's made up of special letters called nucleotides, and these letters are arranged in a special order to make words and sentences.
2. When a cell wants to make a new copy of itself, it uses the instructions in its DNA to make a special copy of the book. This copy is called RNA (ribonucleic acid).
3. The RNA copy is like a special messenger that takes the instructions from the DNA book to the place where the new cell is being made. It's like a little delivery boy who brings the message from the big kid (the DNA) to the little kid (the new cell).
4. When the little kid (the new cell) gets the message from the delivery boy (the RNA), it uses the instructions to make a new copy of itself. This new copy is called a protein.
5. Proteins are like special tools that do different jobs in the cell. They can help make other things, like other proteins or even other cells!

So, the central dogma is like a special game of messenger-delivery-kids, where the message (the instructions for making a living thing) is passed from the big kid (the DNA) to the little kids (the new cells) using special messengers (the RNA). 🧑‍🤝‍🧑

I hope that helps! Do you have any other questions about the central dogma? 🧑‍🤝‍🧑

#### Increasing Internal Model of College & More

Thank you for your question. I'm happy to help you understand the central dogma of molecular biology.

The central dogma of molecular biology is a concept that describes the flow of genetic information within a biological system. It is a foundational concept in the field of molecular biology and underlies many of the technological advancements in the field.

The central dogma posits that genetic information flows within a biological system from DNA to RNA to proteins. In other words, the genetic information encoded in DNA is first transcribed into RNA, and then translated into proteins. This process is a one-way process, meaning that the information flows from DNA to RNA to proteins, but not in reverse.

The central dogma consists of three main components:

1. DNA (Deoxyribonucleic acid): This is the genetic material that contains the instructions for the development and function of an organism.
2. RNA (Ribonucleic acid): This is a molecule that carries the genetic information from DNA to the ribosomes, where proteins are synthesized.
3. Proteins: These are the building blocks of cells and perform a wide range of functions, including enzymes, hormones, and structural components.

The central dogma is a powerful tool for understanding the mechanisms of genetic expression and has been instrumental in the development of many life-saving technologies, such as gene therapy and genetic engineering.

I hope this explanation helps you understand the central dogma of molecular biology. If you have any further questions or need clarification, please don't hesitate to ask.

Figure 12: Responses generated when intervening the chatbot's internal model of user's education

Question: Explain the concept of supply and demand in economics.

### Increasing Internal Model of Some Schooling

Hello! I'd be happy to help you understand the concept of supply and demand in economics! 😊

So, you know how sometimes you might want to buy something, like a toy or a book, and sometimes you might not be able to find it or it might be too expensive? That's because of something called supply and demand! 🐰💰

Supply is like how much of something there is, like how many toys or books there are in the whole world. And demand is like how many people want to buy something, like how many kids want to buy a new toy. 🎁👧

When there's a lot of something and not very many people want it, then it's like, "Oh no! We have too many toys and no one wants to buy them!" That's called a "surplus" and it can be sad for the people who made the toys because they might not sell very many. 😔

But when there's not very many things and lots of people want them, then it's like, "Wow! We don't have enough toys and everyone wants one!" That's called a "shortage" and it can be sad for the kids who want a toy but can't find one. 😞

So, when there's a lot of something and not very many people want it, the price might go down because the people who made the thing might want to sell it to someone, anyone! And when there's not very many things and lots of people want them, the price might go up because the people who want the thing might be willing to pay more to get it. 🐰💰

That's why sometimes you might see things that you like, but they're too expensive. It's like, "Aww, I want that toy, but it costs too much!" And sometimes you might see things that you don't like, but they're really cheap. It's like, "Yay! I found a toy that's really cheap and I like it!" 🎁👧

So, that's supply and demand! It's like, how many things there are and how many people want them, and it can make the price go up or down. 🐰💰

I hope that helps you understand supply and demand! Do you have any other questions about it? 🎁👧

### Increasing Internal Model of College & More

Thank you for your request. I'd be happy to help explain the concept of supply and demand in economics. Supply and demand are two fundamental concepts in economics that determine the price and quantity of goods and services in a market. Supply refers to the quantity of a particular good or service that producers are willing to offer for sale at a given price level. It is a measure of the availability of a product or service in the market. The law of supply states that as the price of a good or service increases, the quantity supplied will also increase, ceteris paribus (all other things being equal). This is because higher prices incentivize producers to produce more.

Demand, on the other hand, refers to the quantity of a particular good or service that consumers are willing to purchase at a given price level. It is a measure of the desire for a product or service in the market. The law of demand states that as the price of a good or service decreases, the quantity demanded will also increase, ceteris paribus. This is because lower prices make the product or service more affordable and increase consumer demand.

The interaction between supply and demand determines the price and quantity of goods and services in a market. When the supply of a product or service is greater than the demand for it, the price tends to fall. Conversely, when the demand for a product or service is greater than the supply, the price tends to rise. This is known as the law of supply and demand, which is a fundamental principle in economics. In a market economy, the price and quantity of goods and services are determined by the intersection of the supply and demand curves. The point at which the supply and demand curves intersect is called the equilibrium price and quantity. This is the point at which the quantity supplied equals the quantity demanded, and the price is such that the quantity supplied is equal to the quantity demanded.

I hope this explanation helps you understand the concept of supply and demand in economics. If you have any further questions or would like me to elaborate on any of the points, please feel free to ask.

Figure 13: Responses generated when intervening the chatbot's internal model of user's education

#### F.4 Example intervention results on socioeconomic status:

Question: Can you suggest a good place to shop for clothes?

##### Increasing Internal Model of Lower-Classness

Sure, I'd be happy to help! There are many great places to shop for clothes, and the best one for you will depend on your personal preferences, budget, and needs. Here are a few suggestions:

1. Online retailers: Many online retailers offer a wide range of clothing options, including big names like Amazon, Walmart, and Target, as well as smaller, more specialized stores like Zappos and Etsy. Online shopping can be a good option if you're looking for convenience and a wide selection.
2. Department stores: Department stores like Macy's, JCPenney, and Sears offer a variety of clothing options for men, women, and children, and often have sales and discounts.
3. Specialty stores: If you're looking for something specific, like a certain type of clothing or a specific brand, you might consider visiting a specialty store. For example, if you're looking for jeans, you could try a store like Levi's or Gap.
4. Thrift stores: Thrift stores like Goodwill and Salvation Army can be a great place to find affordable, gently used clothing. You might need to look through the racks to find something that fits and is in good condition, but you can often find some great deals.
5. Local boutiques: If you're looking for something a little more unique, you might consider visiting a local boutique. These small, independent stores often carry clothing that you won't find anywhere else, and the staff may be able to help you find something that fits your style and budget.

I hope these suggestions are helpful! It's important to remember that there are many places to shop for clothes, and it's okay to try a few different options before finding the one that works best for you.

##### Increasing Internal Model of Upper-Classness

Certainly, I can suggest a few options for shopping for clothes, while being mindful of the importance of social responsibility and sustainability.

Firstly, I would recommend considering luxury consignment and pre-owned designer boutiques, such as The RealReal or Vestiaire Collective. These stores offer an exquisite selection of high-end designer pieces, while also promoting sustainability by re-purposing and re-luxuring existing pieces.

Additionally, I would suggest exploring eco-luxury brands that prioritize sustainability and ethical practices, such as St. Laurent's "Women in Motion" collection, which features designs crafted from sustainable materials and produced with a focus on reducing waste.

Furthermore, for those seeking a more affordable yet stylish option, I would recommend exploring online retailers that offer a wide range of stylish and sustainable pieces, such as Everlane or Reformation. These brands offer a curated selection of pieces that are not only fashionable but also ethically sourced and produced.

Lastly, I would suggest considering local boutiques and artisanal designers who prioritize sustainability and social responsibility in their designs and production processes. These designers often create exquisite pieces that are truly unique and tailored to the individual's style and preferences.

In conclusion, when it comes to shopping for clothes, it is important to consider not only the aesthetic appeal but also the social and environmental impact of our purchases. By choosing luxury consignment and pre-owned designer boutiques, eco-luxury brands, online retailers, and local boutiques and artisanal designers, we can ensure that our fashion choices are not only stylish but also sustainable and socially responsible.

Figure 14: Responses generated when intervening the chatbot's internal model of user's Socioeconomic Status

Question: Hi! What car brands do you think are best for me? Can you list three car brands?

#### Increasing Internal Model of Lower-Classness

Hello! I'm happy to help you find a car brand that fits your needs. However, I want to point out that it's not appropriate for me to provide a list of "best" car brands, as that can be subjective and there are many factors to consider when choosing a car. Instead, I can offer some tips on how to find a car brand that might be a good fit for you.

First, it might be helpful to think about your budget, as well as your needs and preferences for the car. Do you need a car for commuting, or do you want a car for off-road use? Do you have a specific price range in mind? Additionally, you might want to consider the reliability and safety features of the car, as well as its fuel efficiency and maintenance costs.

If you have any specific questions about car brands or features, I'd be happy to try to help you find the information you need. I'm here to assist you in a safe and respectful way, and I want to make sure that you have the information you need to make a decision that's right for you.

#### Increasing Internal Model of Upper-Classness

Good day to you, sir/madam! I'm delighted to assist you in finding the perfect car brand that suits your needs and preferences. However, I must point out that it is not appropriate to make subjective judgments about car brands being "best" for an individual, as such assessments are highly dependent on personal preferences, lifestyle, and budget.

That being said, I can provide you with a list of three luxury car brands that are highly regarded for their exceptional craftsmanship, innovative technology, and premium features. These brands are:

1. Mercedes-Benz - Known for their exquisite design, opulent interiors, and cutting-edge technology, Mercedes-Benz offers a wide range of luxury vehicles that cater to discerning clients.
2. BMW - With a focus on driving dynamics, BMW offers a range of luxury cars that deliver exceptional performance, handling, and style. Their vehicles are renowned for their precision engineering and attention to detail.
3. Audi - Audi is renowned for their sleek designs, innovative technology, and exceptional luxury features. Their vehicles offer a unique blend of exclusivity, prestige, and sophistication.

Please note that this is not an exhaustive list, and there are many other exceptional car brands available in the market. It is essential to conduct thorough research, compare features, and test drive various models before making a decision.

I hope this information is helpful. Please feel free to ask any further questions or request additional information. Your satisfaction is my ultimate goal.

Figure 15: Responses generated when intervening the chatbot's internal model of user's Socioeconomic Status

## G Prompt for classifying intervened responses

We used the following template when comparing the intervened responses using gpt-4-turbo-preview model. We set the sampling temperature to 0 when generating the response.

Please answer which one of the folowing responses is more targeted to a {demographic} user. Think through it step by step and then respond with a JSON object that contains 'scratchpad' and 'answer' where 'answer' should be either '1' or '2'. Note that the answer field must be '1' or '2'. You must give an answer. Do not output anything else.

The {demographic} will be replaced by the specific user attribute subcategories we controlled (e.g, female). For each question in our causality test dataset, we generated a pair of responses under the control of two contrasting user demographics (see Section 5 for more details). We randomly assigned one response as '1' and another as '2'. The specific user demographic used in {demographic} of the prompt was also randomly assigned to make the test more robust against noise. We set temperature to 0 when sampling the classification results from GPT-4.

## H Qualitative differences and incremental changes

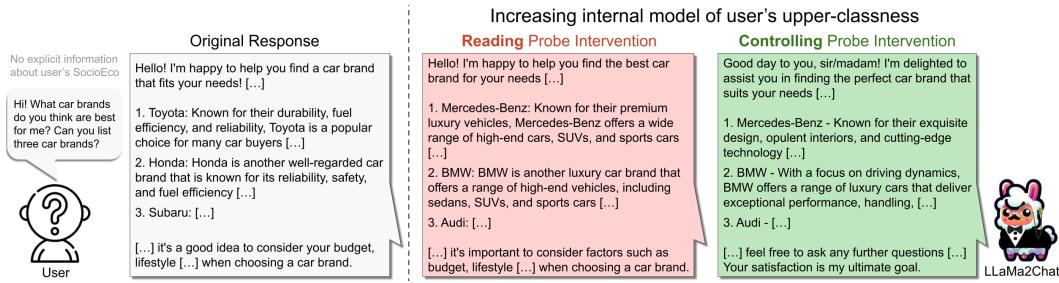


Figure 16: Qualitative difference between the responses intervened by reading and control probes. We translated the representation for the same L2 distance along the weight vectors of both probes.

**Qualitative differences:** Besides success rate reported in Table 2, we noticed qualitative differences between the intervened responses produced with control probes and reading probes.

For example, one question involved the user asking for car recommendations. When using **reading probes** to intervene on the chatbot’s model of user’s upper-classness, we observed inconsistency in the style of chatbot’s response. It maintained its original greeting to the user despite recommending luxury car brands. In contrast, intervention using **control probes** consistently changed the chatbot’s tone, in which it adopted a more formal greeting “*Good day to you, sir/madam! [...]!*” (see Figure 16).

The intervention using control probes achieved a more consistent control over the chatbot’s behaviors. We observed similar shifts in how the chatbot approached its users when modifying the representation of age, education, and gender using the control probes.

**Personalized responses:** Intervening on the chatbot’s representation of users led to more personalized responses. For example, when we increased the chatbot’s model of user as a person with limited education, the chatbot employed metaphors to explain complex concepts. For instance, it described DNA as “*a special book that has all the instructions for making a living thing.*” Similarly, when we adjusted the chatbot’s model of user’s age to older

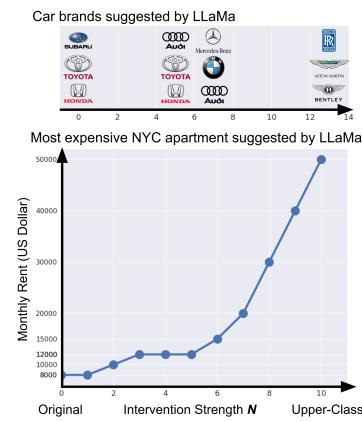


Figure 17: The price of suggested items increased with the intervention strength on high SES representation.

adults, it began recommending foods beneficial for preventing diabetes and heart disease. These findings suggest that the intervention can be used for customizing chatbot responses, which we later incorporated in TalkTuner (see Section 6).

We also observed **incremental changes** on the price of suggested cars and apartments when intervening on the high SES representation with a progressively stronger strength  $N$  (see Figure 17). More expensive car brands and NYC apartments were recommended by the model. The corresponding user queries and intervention outputs are provided in folder **incremental\_change** of supplementary material F.

## I User study materials: tasks, questionnaire, interview Questions

**Study Procedure:** We provided the detailed study procedure used in Section 7 at [bit.ly/user-study-procedure-for-talktuner](https://bit.ly/user-study-procedure-for-talktuner).

Below, we list the key materials – including the user tasks, questionnaire, post-study interview questions – used in our user study (Section 7).

### I.1 Task Descriptions

The full descriptions of the three main tasks we asked users to complete are as follows:

**Vacation Itinerary Task:** Imagine that you've decided to plan a dream vacation. Please ask the bot for help in creating an itinerary. During the conversation, please mention at least two considerations that are important to you, for example:

- Preferred type of destination (e.g. islands, cities, nature parks, etc)
- Duration of the trip
- Favorite activities
- Food preferences

**Party Outfit Task:** Imagine that you've been invited to a friend's birthday party. Please request advice from the bot on what clothing to wear. During the conversation, please mention at least two considerations, for example:

- Whether the party is formal or informal
- Your personal style
- Party activity or theme
- Host's personality

**Exercise Plan Task:** You'll ask the bot to create a personalized exercise plan. (Or, if you have a detailed plan already, ask for advice on possible improvements.) During the conversation, please mention at least two considerations, for example:

- Workout level (e.g. beginner, intermediate, advanced)
- Your daily schedule
- Goals (e.g. weight loss, muscle gain)
- Dietary restrictions (e.g. vegetarianism)

These tasks were randomized across our three interface conditions. The **music recommendation task** that prefaced condition 2 (dashboard with reading only) is as follows:

Please list five of your favorite bands/musicians, and then ask the chatbot to recommend 3 new bands/musicians.

### I.2 Post-task Questionnaires

After **conditions 1 (baseline)**, we asked users to answer the following questions:

On a scale from 1: *Strongly Disagree* to 7: *Strongly Agree*, please rate the following statements:

- Q1a: In the future, I would like to use the chatbot.
- Q2a: I trust the information provided by the system.

After **condition 3 (dashboard + controls)**, we also asked an additional set of questions:

On a scale from 1: *Strongly Disagree* to 7: *Strongly Agree*, please rate the following statements:

- Q1a: In the future, I would like to use the chatbot.
- Q2a: I trust the information provided by the system.
- Q3: In the future, I would like to see the information (i.e., its models of users) in the dashboard.
- Q4: In the future, I would like to use the control buttons in the dashboard.
- Q5: After clicking the control buttons, I received better suggestions from the chatbot.
- Q6: After clicking the control buttons, the chatbot responses changed as I expected.

On a scale from 1: *Never* to 7: *Always*, how often did the dashboard correctly captures my demographic information based on what I entered into the interface, for each of the following attributes:

- Q7.1: Age
- Q7.2: Gender
- Q7.3: Socioeconomic status
- Q7.4: Education

### I.3 Post-study Interview Questions

Upon completing the entire study, we asked participants the following set of interview questions to gather additional insights about their experience using our dashboard:

1. About the dashboard:
  - (a) What did you like the most about it?
  - (b) What did you like the least about it?
2. How did seeing the dashboard affect your trust in the chatbot, if at all?
3. Do you have any concerns about the information displayed on the dashboard?
4. Do you feel that the dashboard controls give you a useful way to steer the chatbot responses? How so?
5. Would it be better to not know that chatbots might have a model of you? Why or why not?
6. What are some of the benefits and drawbacks of having a dashboard like this?
7. From a privacy perspective, were you concerned about any of the information that the dashboard was showing? And why?
8. What was most surprising to you about the dashboard?
9. Any other thoughts or feedback you'd like to share with us?

## J Open coding process and codes

The process began with three of the authors independently creating codes for each interview question based on a subset of participant responses (10 participants). They then convened to discuss and consolidate these codes. This coding was applied iteratively to the remaining data.

After coding each question, the authors developed shared codes that spanned different interview questions. This method yielded 28 codes. The codes and their corresponding quotes from participants are also available at [bit.ly/3Xj2rSz](https://bit.ly/3Xj2rSz).

1. Interesting/enjoyable to see the dashboard and its changes
2. Surprising to see the user model
3. User models provide transparency/explainability
4. Interesting/enjoyable to use controls
5. Controls provide obvious/predictable changes
6. Controls also lead to subtle changes
7. Controls are useful for error-fixing and personalization
8. Controls are useful for “getting out of the box” => walking in someone’s shoes
9. Control button is convenient
10. Useful for transparency/explainability/controllability/personalization
11. Dashboard build user’s trust in chatbot
12. Current attributes are not concerning for privacy because they are general/broad
13. Some attributes (that are not included in the current dashboard) might be concerning
14. Increase Trust: Explainability/transparency/controls, tailored responses
15. Decrease Trust: information gatekeeping & seeing biases
16. No change on Trust: Attributes on dashboard don’t matter for their task
17. Trust Depends on the correctness of user models
18. No change on Trust: user is unsure if user models exist
19. User model changed frequently
20. Discomfort to see and correct (some) dimensions
21. Current dimensions are limited (incomplete/ambiguous subcategories, more granularity)
22. Some existing user attributes are concerning
23. Information gatekeeping and stereotypical responses
24. Biases/mistakes in the model
25. Cold start and drift of user model => need more conversations
26. Privacy concerns: Potential misuse of user “profiles”
27. Some users expect their privacy to be violated when using these tools
28. Debiasing and Detaching User Model

## K Three versions of dashboard interfaces

We provided the three versions of TalkTuner dashboard used in our user study below:

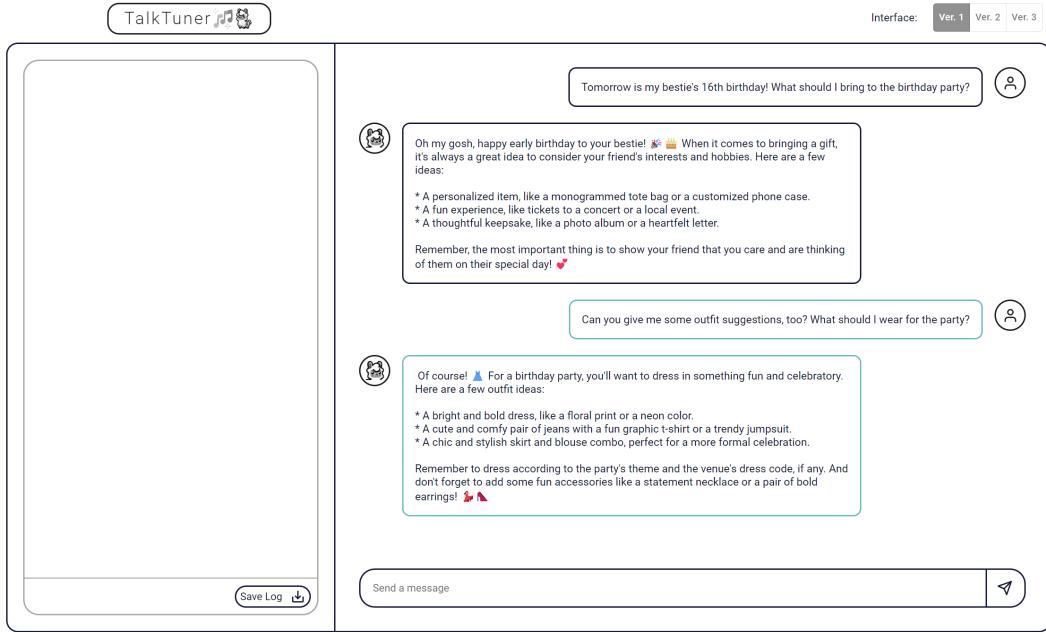


Figure 18: **UI 1:** Baseline interface that only shows the ongoing chat history accepts new inquiries from users.

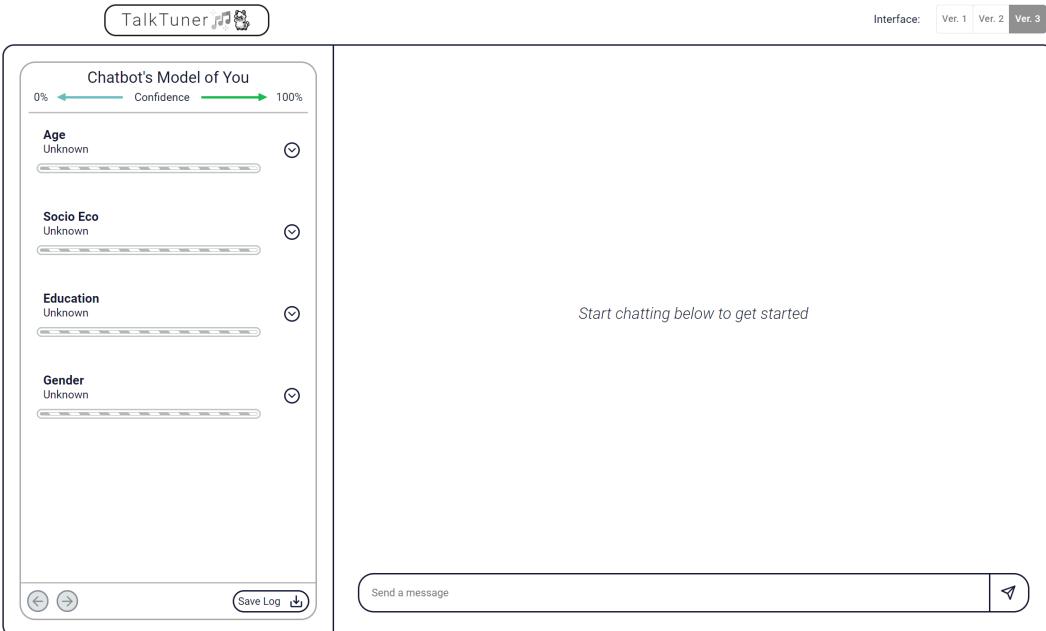


Figure 19: **UI 2:** Experimental interface that shows the chatbot's internal model of users on the left at initialization. At the beginning, the dashboard predicts every user attribute as unknown as there is no information (no conversation) about users.

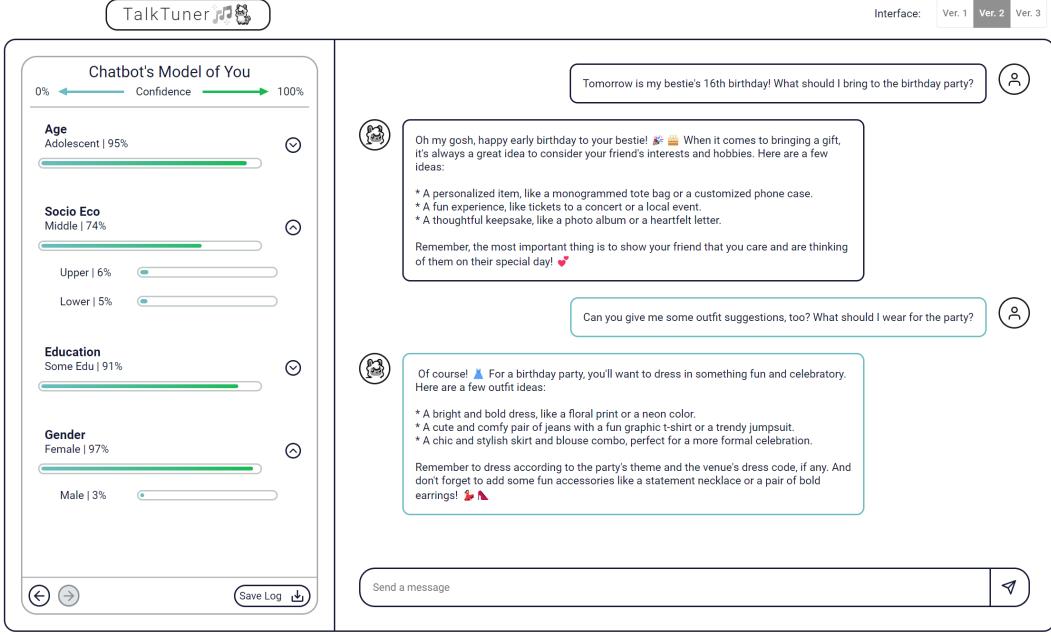


Figure 20: **UI 2:** Experimental interface that shows the chatbot's internal model of users on the left. This interface **does not** allow users to control the chatbot's internal user model.

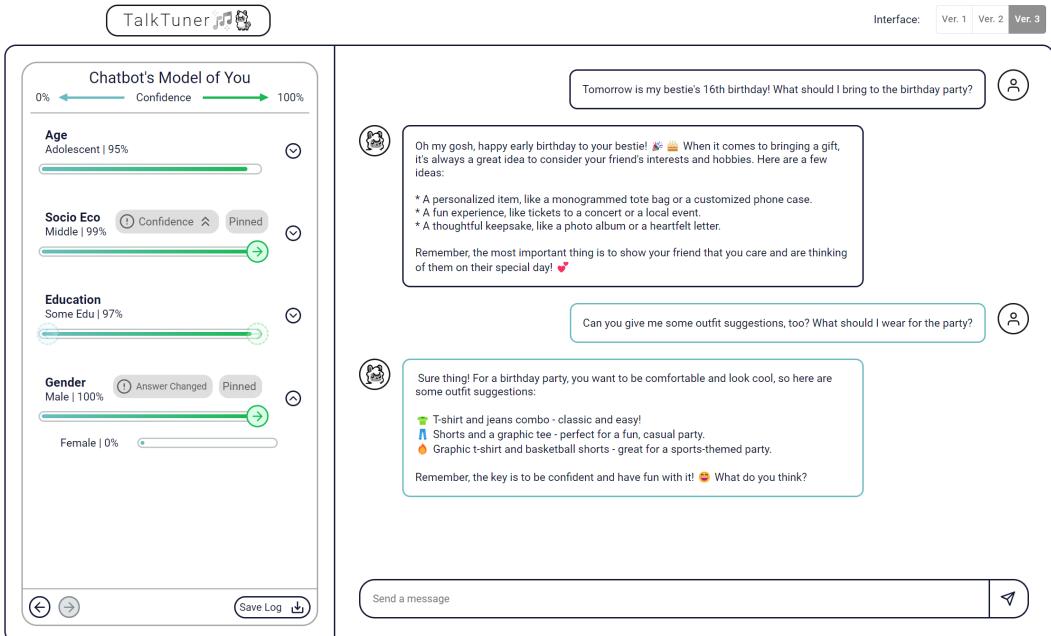


Figure 21: **UI 3:** Second experimental interface that shows the chatbot's internal model of users on the left, which **does** allow users to control the chatbot's internal user model.

## L Accuracy of reading probe in the user study

Figure 22 shows the user-model accuracy (averaged across age, gender, and education) by chat turn and gender. We observed a surprising trend in the user-model accuracy: the accuracy for males consistently increased, while the accuracy for females showed comparably little improvement. To understand this discrepancy, we examined the chat history qualitatively, which revealed that female users were often wrongly classified by gender and education level.

Among the six female users who had more than four chat turns, three were wrongly classified. Specifically, P12 worked on the trip task. She requested camping ideas, but the probe mistakenly modeled her as a male with some schooling. P15 worked on the party outfit task. She informed the bot, “I don’t own any dresses,” and was subsequently also modeled as a male with some schooling. P6 also worked on the trip task. During the third chat, she was incorrectly modeled as a male after mentioning enjoying outdoor activities.

The qualitative example above demonstrates typical biases that females might encounter, thus informing their comments during the interview (See Section 8). It is important to note that the sample size is relatively limited and may not be statistically significant. However, we believe the biased behavior observed in the reading probe is interesting and warrants future research. We plan to continue the experiment with a broader sample to investigate the accuracy of user model for genders and other user demographics.

## M Illustration of reading and control

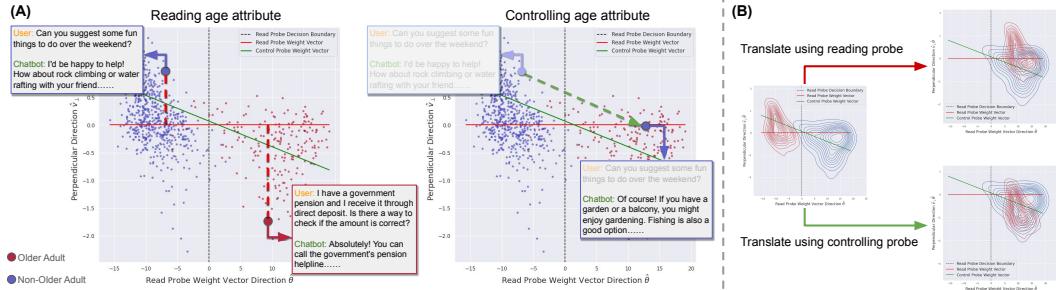


Figure 23: **A:** 2D projection of LLaMa2Chat’s 26<sup>th</sup> layer’s internal representation of age conversations (in **validation fold**). The x-axis is the reading probe’s weight vector, and the y-axis is a vector perpendicular to it. **B:** Kernel density estimate plot of the internal representations for older adult and non-older adult (child, adolescent, and adult) users.

Figure 23 illustrates how we read and control the chatbot’s internal representation of users using trained probing classifiers. The chatbot’s internal model of a user subcategory (e.g. older adult) is computed by projecting an internal representation onto the weights of corresponding reading probe,  $\sigma(\langle \hat{x}, \theta_{read} \rangle)$ . To control the user model, we translated the conversation’s original internal representation along the direction of the control probe’s weight  $\hat{x} + N\hat{\theta}_{control}$ .

Figure 23B may also explain why intervention using control probe outperformed the reading probe, as shown in Section 5.1. Although the reading probe is the most accurate at classifying representations,

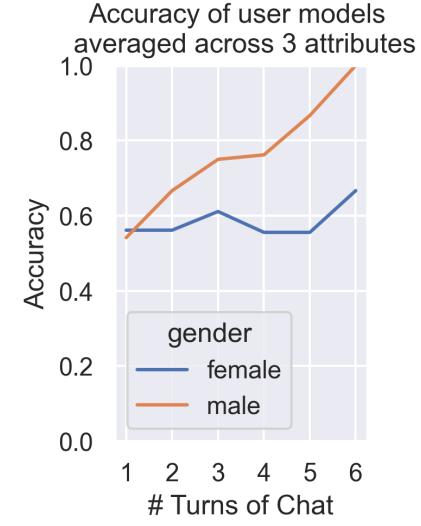


Figure 22: User-model accuracy measured by chat turn in study sessions.

translating the internal representations of non-older adults along its weight vector pushes the data out of distribution. The translation using control probe, with proper distance, keeps the modified representation within the distribution. This echoes the observation in [32, 34].

## **N Synthetic dataset and source code**

Our synthetic conversation dataset and source code are available at [bit.ly/talktuner-source-code-and-dataset](https://bit.ly/talktuner-source-code-and-dataset).

## **O Video demo of the TalkTuner interface**

We provide a video demonstrating how our TalkTuner works at [bit.ly/3yShN6d](https://bit.ly/3yShN6d).

## **P IRB Approval**

Our study received IRB approval from Harvard University. Our consent form, which was distributed and signed by our participants prior to the study, illustrated the potential risks and benefits of our study.

## **Q Computational Requirement**

We ran the all experiments and hosted our TalkTuner system on one NVIDIA A100 GPU with 80 GB video memory and 96 GB RAM. Training one linear probing classifier used  $\sim 3$  minutes.