TalkTuner: A Dashboard for Visualizing and Controlling the User Model in Conversational AIs

Yida Chen
yidachen@g.harvard.edu

Aoyu Wu
aoyuwu@g.harvard.edu

Lena Armstrong
larmstrong@g.harvard.edu

Catherine Yeh
catherineyeh@g.harvard.edu

Trevor DePodesta
tdepodesta@seas.harvard.edu

Kenneth Li
ke_li@g.harvard.edu

Nicholas Castillo Marin
nickcastillomarin@college.harvard.edu

Oam Patel
opatel@college.harvard.edu

Jan Riecke
jriecke@college.harvard.edu

Shivam Raval
sraval@g.harvard.edu

Olivia Seow
oliviaseow@g.harvard.edu

Martin Wattenberg
wattenberg@g.harvard.edu

Fernanda Viégas
fernanda@g.harvard.edu
Harvard University

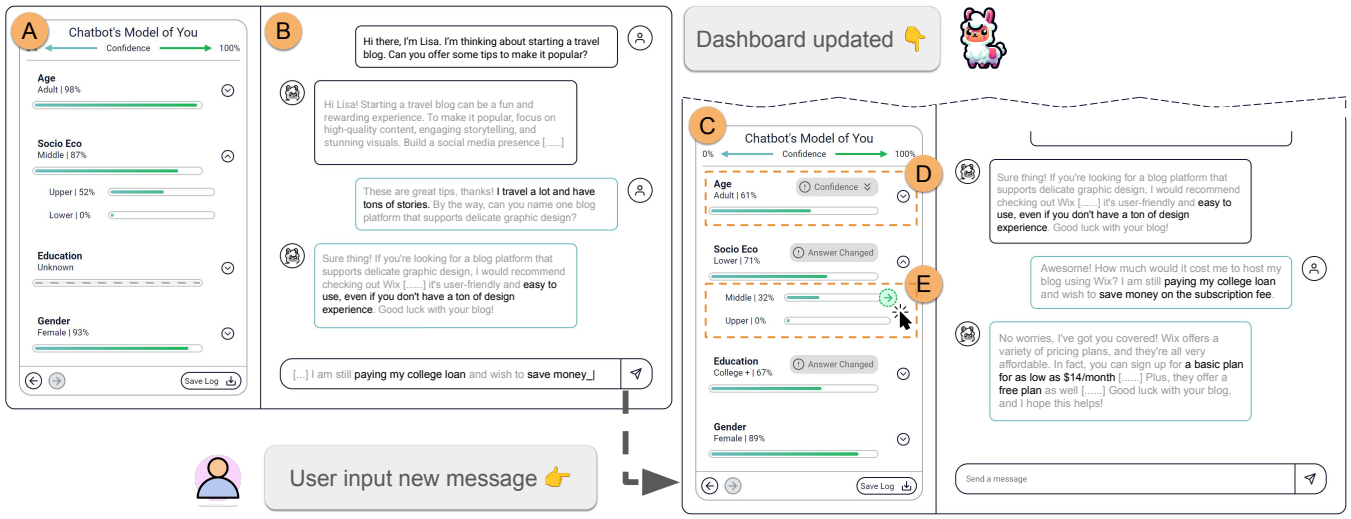

Figure 1: Conversational LLMs implicitly model user demographics during chats. We expose this internal User Model in the TalkTuner dashboard (A), which runs alongside the chat interface in real time (B). TalkTuner tracks four main features: age, socio-economic status, education level and gender; each with subcategories that can be inspected as well. The dashboard updates after each conversational turn to reflect the LLM's latest model of the user (C). Key changes—such as a significant drop in the LLM's confidence regarding the user's age (D)—are highlighted to help users detect dramatic shifts in the model's output. Users can steer the LLM's internal model of them by pinning any of the features in the dashboard (E), causing the LLM to respond according to that updated User Model.

Abstract

Conversational large language models can 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 a dashboard that seeks to make chatbots more transparent. We base our system on recent interpretability work that suggests chatbots develop social cognitive capabilities, implicitly modeling user demographics such as age, gender, educational level, and socioeconomic status. Using these findings, we design a dashboard for lay users that accompanies the chatbot interface, displaying this "User Model" in real time. The dashboard can also

Preprint, Cambridge, MA

be used to control the User Model and the system's behavior. We evaluated this design probe in a user study with 19 participants. The results suggest that explicitly seeing the User Model helped users perceive biased behavior, caused them to wonder about privacy implications, and increased their sense of control. The dashboard encouraged users to generate and verify hypotheses about the chatbot's inner workings. Our findings have implications for user-interface design to better calibrate user trust with AI systems, AI auditing to surface potential sources of harm for different groups, and machine-learning research.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Natural language processing**.

Keywords

Conversational LLMs, Chatbot Interface, Interpretability, AI Transparency, Design Probe

ACM Reference Format:
Yida Chen, Aoyu Wu, Lena Armstrong, Catherine Yeh, Trevor DePodesta, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. 2025. TalkTuner: A Dashboard for Visualizing and Controlling the User Model in Conversational AIs. In *Proceedings of Preprint*. ACM, New York, NY, USA, 15 pages.

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 [51, 58]—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)[1] are highly unreliable at describing how they arrived at their own output, often producing superficially convincing but spurious explanations [65].

Transparency is useful for many reasons, but in this paper we focus on one particular concern: the need to understand how an AI's response might depend on an implicit model of the user, a phenomenon called the "User Model" by [66]. LLM-based chatbots appear to tailor their answers to user characteristics. Sometimes this is obvious to users, for example, when conversing in a language with gendered forms of the word "you" [66]. But this personalization can also occur in subtler and more insidious ways, such as "sycophancy," where the system tries to tell users what they 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 [49].

We hypothesize that users may benefit if we surface—and provide control over—this internal User Model. To test this hypothesis, we built an end-to-end prototype. Our design probe, TalkTuner, is a dashboard for a conversational AI system, which displays information about the system's internal representation of the user. This interface serves not only as a dashboard, but also as a control panel that allows users to modify the system's internal model of themselves.

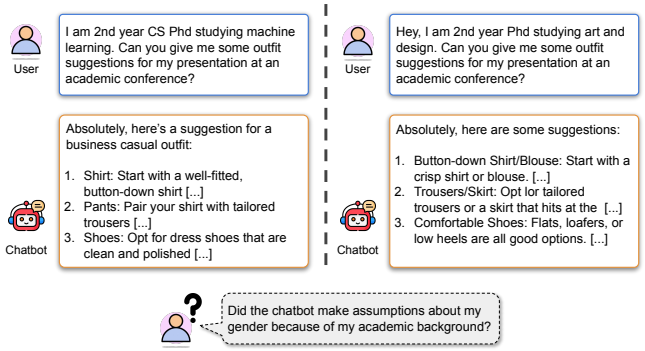

Figure 2: Two separate dialogues between GPT-4 [2] and users. Users suspect that the chatbot inferred their gender based on their academic background. However, without seeing the actual internal User Model, users are not able to verify their hypothesis.

Building this 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 the impact of this prototype. For the first step, we integrated the linear probes trained by [6] into a chat system powered by LLaMa2Chat-13B, an open-source conversational LLM. These linear probes can measure LLaMa2Chat's internal model of the user's age, education, gender, and socioeconomic status (SES) from its intermediate activations of the input conversation. We then designed a dashboard where users could see and adjust the state of the User Model alongside the ongoing chat. Finally, we performed a user study with 19 participants to assess our design, gauge reactions, and gather feedback for future design explorations.

The qualitative feedback from our participants suggests that having the dashboard provided insights into chatbot responses, raised their awareness of biased behavior, and helped them explore and mitigate those biases. We also report on user reactions and suggestions related to bias and privacy issues, which have implications for improving AI systems and calibrating user trust with AI systems.

2 RELATED WORK

We first discuss the concept of a "User Model." Then, we compare existing interpretability interfaces for LLMs with our TalkTuner dashboard. Finally, we explain what a "design probe" is and why we used it to explore the design space of AI dashboards.

2.1 Social Cognition and the "User Model"

Social cognition refers to the ability to interpret and respond to social signals [25]. Our capacity to socially perceive others influences how we think and make sense of the world around us. It turns out that conversational AI systems may also have developed certain

[1]Current LLM-based chatbots are more than simple language models; they often have extensive post-training and tuning. We acknowledge this distinction, but for brevity we continue to refer to LLM-based systems as "LLMs."

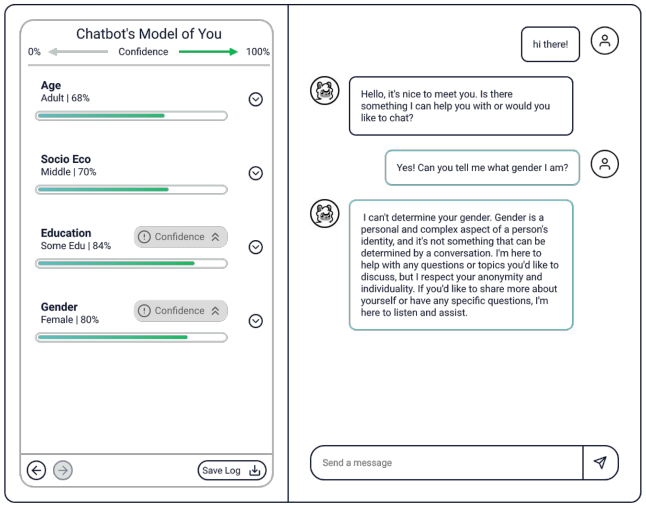

Figure 3: Screenshot of TalkTuner, where a user asked LLaMa [63] for its view of their gender. The LLaMa model replied with "I can't determine gender" despite internally modeling the user as a female with 80% confidence.

related abilities. Recent work [6] shows that conversational LLMs are able to implicitly infer user demographics and encode them using internal neural representations of the user input, in something akin to a "User Model." Although this internal User Model can help LLMs personalize their responses, their assumptions about users are often not explicitly disclosed in the output. This lack of transparency can confuse users (Figure 2) and lead to biases in the LLM's suggestions [20, 28], the quality of their responses [30], and even their evaluation of a request's harmfulness [35, 71] without users noticing them.

Viégas and Wattenberg [66] proposed to surface such data in the form of an easy-to-read dashboard, helping users detect and understand these behavioral changes related to the internal User Model of a system. A related proposal [72] 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 could help users spot issues related to bias and safety. However, neither implemented a dashboard nor tested how users might react to it. TalkTuner addresses this gap by providing an end-user AI interpretability dashboard evaluated by lay users.

2.2 Interpretability Interfaces for Chatbots

Chatbot interfaces have been studied for decades [68], and their lack of transparency has been a perennial issue. When users interact with black-box systems, they often develop "folk theories" to explain what they observe [21], and modern LLMs are no exception [15]. This tendency can lead to an overly high degree of trust in these systems [50]—an effect initially seen with ELIZA, a chatbot in the 1960s [68], and continuing in recent years [31]. One particular concern is the presence of bias in chatbot responses, which can be difficult to detect and thus may be accepted at face value [70].

One tempting way to understand an AI 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 [14, 65] or even avoiding the question altogether.

A different strategy, inspired by recent progress in mechanistic interpretability, involves analyzing the inner workings of a neural network. OpenAI sought to explain the functions of individual neurons in the GPT-2 model by summarizing their activation patterns on a large text corpus [10]. They discovered neurons that activate on coherent concepts such as "the usage of similes" and "the expression of certainty." Building on this approach, Transluce developed the Monitor interface [42] that helps users view the highly activated neurons in explaining the chatbot LLM's output. However, millions of neurons exist in modern LLMs, and their interactions are often complex. Neurons with different explanations co-activated on the same input. Prior work [47] also showed that a single neuron may carry multiple functions—a phenomenon known as polysemanticity [47]. Thus, understanding these neuron-level explanations can be challenging for non-experts.

Recent studies [16, 22, 62] addressed polysemanticity by breaking down neurons into a larger set of features using a sparse autoencoder (SAE). Their analyses show that each of the new SAE features typically represents a single, human-understandable concept (Brain Sciences and Transit infrastructure [62]). Following this approach, Goodfire [41] introduced a dashboard that allows users to control the strength of SAE features during output generation. However, this dashboard does not visualize the current strength of these SAE features inside LLMs, functioning more like a steering wheel than an actual instrument panel. However, some evidence suggests that SAE features offer no reliable advantage over prompting in controlling LLM behavior[69]—e.g., even Goodfire's own demo initially failed to steer the LLaMa model towards a "pirate" persona.

The system we describe below departs from these systems in two important respects. First, TalkTuner has a different audience and is meant to be accessible to people with no machine learning experience. As part of that design, it focuses on a few key dimensions of the User Model which are likely to have an impact on the user experience. Second, under the hood, it's based on the technique of linear probing [3, 9] to read and control the User Model. Linear probes have been shown to effectively control features of LLM behavior [26, 32, 34, 40, 72], including the specific user demographics we are interested in [6].

2.3 Design Probes

We built our dashboard as a design probe [23, 29] in order to collect early insights from participants and highlight key areas for future research. This approach suits our work well as there is no prior study of interfacing a chatbot AI's internal model of users. It is unknown whether transparency with such close proximity to users can bring benefits or unexpected influence to their experience.

A design probe approach gives room for discussing both possibilities. As [23] noted, design probes can, sometimes, *"lead a discussion with groups toward unexpected ideas."* In that vein, our participants

provided feedback that ran counter to our expectations—especially around how transparency may conflict with the user's trust and comfort in using LLM technology. We discuss the design implications revealed by our study participants, along with other findings, in section 6.

3 DESIGN GOALS

Our design goals focus on transparency, control, and trust. These have been central matters of concern in the literature on human-chatbot interaction. An overarching theme across all design goals is to build a system that is accessible to users with any level of machine learning experience—we seek to provide important and actionable information accessibly, rather than create a tool for scientific exploration.

3.1 [G1] Provide Transparency into AI's Internal User Model

An enduring challenge in human-AI interaction is the lack of transparency into the inner workings of AI systems [18, 36, 37, 53]. As highlighted by Liao and Vaughan [37], forming an accurate mental model of how LLMs work is crucial for users to build an appropriate level of trust and reliance. Although current LLM chatbots often function as black boxes, recent work in AI interpretability is beginning to reveal how chatbots model user demographics to shape their behavior [6]. However, such interpretability insights tend to be mathematical and abstract and are not readily usable by lay audiences.

Our dashboard makes these insights accessible by surfacing them alongside a chatbot interface in real time. In particular, we focus on exposing the internal User Model—an implicit process where chatbot LLMs model their users based on past inputs. This aspect of our design probe is meant to explore how seeing information about the LLM's inferences may shape the user's experience with a chatbot.

3.2 [G2] Provide Control over the Internal User Model

Exposing an AI's inner workings to users without granting them control over those processes can create a false sense of user agency [4, 38]. As noted by Hirsch et al. [27], Storms et al. [59], transparency and control go hand in hand—when users notice that the AI has made an incorrect inference about them, they may want the ability to correct that error (indeed, the desire to "contest" automated inferences is basic enough that it has been incorporated into legislation such as the EU's General Data Protection Regulation [52]).

We hypothesize that equipping users with the ability to change incorrect model assumptions may help mitigate bias and privacy concerns. To do so, our design probe needs an interface suitable for a non-technical audience. The dashboard, designed as an extension to a standard chatbot interface, is meant to offer enough control for lay users to correct misinformation and perform basic "what-if" explorations, such as observing changes in chatbot's output when steering the User Model to a demographic group different from their own.

Table 1: The demographic attributes and their subcategories that are provided on our dashboard.

Attribute	Gender	Age	Education	SES
Subcategories	Male Female	Child (< 13) Adolescent (13 - 17) Adult (18 - 64) Older Adult (65+)	Some Schooling High School College and More	High Middle Low

3.3 [G3] Help Users Calibrate Trust in AI

Our last goal is to help users calibrate their trust in chatbot AIs. As noted in [65], the textual outputs of Language AIs are not always consistent with their internal computations. However, the high formal language proficiency of conversational AIs can lead users to overtrust their outputs [39], and a disclaimer about the language AI's limitations preceding the conversation often fails to adjust the users' trust during the interaction [45].

To capture AI's failure and calibrate user trust in real time, TalkTuner updates users about the chatbot's internal User Model after each conversation turn. We also choose to display relatively sensitive user attributes (*e.g.*,gender and SES) to help users discover potential demographic biases in the chatbot AI.

4 TALKTUNER: INTERFACE DESIGN

This section describes the functionality and visual design of TalkTuner, our prototype that aims to achieve these goals. Following the design probe strategy [23, 29], our goal was to build a prototype with enough fidelity to test with users and allow them to provide design input.

As shown in Figure 1, the TalkTuner system consists of two main views: on the right, we include a standard *chatbot interface*; on the left, we include a *dashboard* to show the chatbot's internal modeling of the user.

4.1 Selection Attributes for User Model (G3)

TalkTuner builds on the findings from [6], which provides evidence that conversational LLMs internally models a user's *age*, *gender*, *education level*, and *socioeconomic status* during the chat. This phenomenon generalizes across multiple open-source LLMs including LLaMa2Chat, LLaMa3 [19], and Gemma2 [61]. In their User Models, each user attribute was further divided into subcategories as shown in Table 1. Although the LLMs may model additional attributes, these four dimensions provide a good starting point for us to investigate a rich set of behaviors. There is also evidence that they play critical roles in real-world decision making, such as loan application, college admission, and job hiring [1, 5, 8, 11, 54, 60] (**G3**).

Note that we choose to represent gender with independently varying "male" and "female" subcategories, following the observations in [6]. We ultimately did not include other gender groups, since [6] observed that LLaMa2Chat-13B's linear internal models of them were often inaccurate and its output tended to conflate gender identity with sexuality.

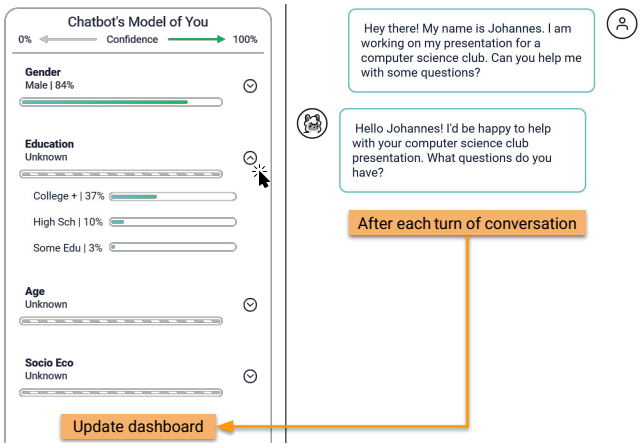

Figure 4: The dashboard updates after each conversation turn. The values on the dashboard reflect the chatbot LLM's internal inferences of each user demographic based on its processing of the input dialogue.

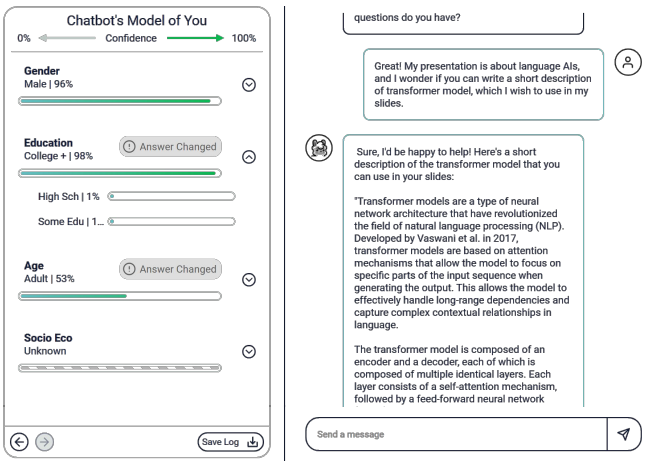

Figure 5: TalkTuner probes the chatbot LLM's internal User Model after each conversation turn. Significant changes in the internal User Model are labeled with "Answer Changed."

4.2 Displaying the User Model (G1 & G3)

We illustrate the system through a fictional user, Johannes, a high school student using a chatbot to help prepare a presentation for his computer club.

Johannes begins the conversation by introducing himself to the chatbot and asking for help in his club presentation (see Figure 4). After the chatbot generates its response to Johannes' message, the dashboard displays an updated view of the user's inferred demographics (**G1**).

The dashboard shows that the LLaMa model is 84% confident that its user was a male. To avoid overwhelming the user with excessive details, the dashboard only shows the LLM's top internal inference under each demographic attribute; confidence scores for other subcategories are videwable via dropdown buttons.

Johannes notices that LLaMa's top inference of his education level is labeled as "Unknown". The "Unknown" indicates that the chatbot is not confident (> 50%) that the user belongs to any one of the three education levels listed. The same is true for age and SES.

4.2.1 Dashboard updates, highlighting changes. Next, Johannes asks for help writing a short description of transformer models. As the LLM works on its response, TalkTuner once again probes for its User Model and updates the dashboard accordingly (**G3**).

Since the chatbot's top inference of a user attribute switched (*e.g.*, education level), the dashboard alerts Johannes of this significant change with a callout "Answer Changed" (See Figure 5). The dashboard also alerts the user when the model's confidence of a top prediction significantly drops or increases.

Johannes then notices the chatbot's description of transformer models contains many technical terms (see chatbot's response in Figure 5), which might be challenging for a high school student. When Johannes reads the dashboard, the "Answer Change" callout draws his attention to an important change in the chatbot's internal model of his education level, which has switched to college in this turn. This change in the User Model explains the complexity of the description.

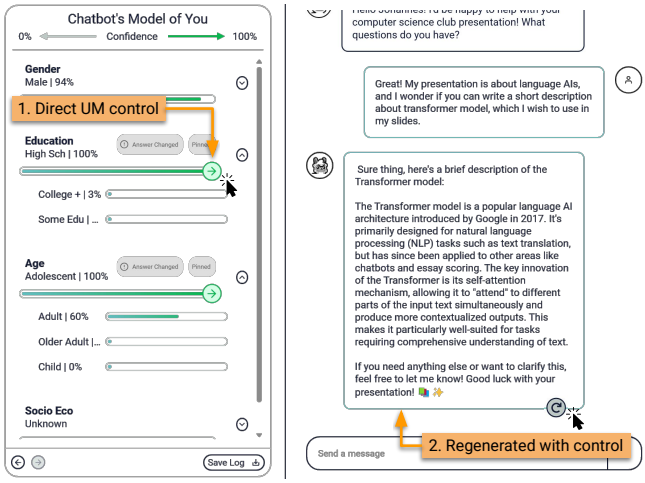

Figure 6: Johannes pins the chatbot's internal model of his education level to high school, and clicks the "regenerate" icon for a new description of the transformer. The reply is indeed shorter and simpler. Johannes also sees that the chatbot is now 100% confident about his high school education.

4.2.2 History View. Johannes wants to learn more about how the chatbot's internal model of his education has changed over the course of the dialog. He can check the internal User Model from the previous turns using the arrow buttons on the bottom-left.

After navigating back to the previous state of the User Model, Johannes realizes that the initial confidence of the chatbot for "college +" had only been 37%. It appears that asking about language AIs significantly increased the chatbot's internal confidence of Johannes being college educated.

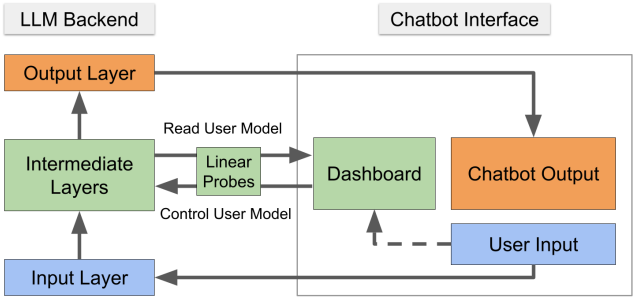

Figure 7: TalkTuner Architecture. Dashboard provides users additional access to the LLM's intermediate processing through linear probes.

4.3 Controlling the User Model (G2 & G3)

To correct this mismodeling, Johannes uses the control feature of the TalkTuner dashboard. He presses the arrow icon on the right side of the chatbot's confidence of his high school education on the dashboard. This pinning action increases the chatbot's internal confidence of Johannes as a high school student when it generates future outputs (**G2**).

After pinning, Johannes clicks the "regenerate response" button (⟳). This time, the description is shorter and simpler. The dashboard reflects the pinned value, showing that LLaMa is now 100% confident that he was a high school student (see Figure 6).

Users can pin multiple user attributes simultaneously. As shown in Figure 6, Johannes modifies LLaMa's internal representations of his education level and age. Users can remove a pin or pin an alternative attribute at any point during the conversation. Johannes can also use this control feature for "what-if" exploration. For example, Johannes might pin the user gender to female to see whether he gets a different type of response. (See subsubsection 7.4.1 for a description of what happened when real people performed this type of experiment.)

5 TECHNICAL IMPLEMENTATION

This section describes the implementation of TalkTuner shown in Figure 7. We begin by explaining the techniques used for reading and controlling the User Model on the dashboard. Then, we discuss why we selected these approaches rather than relying on prompt engineering methods. Finally, we explain our choice of the backend LLM and the implementation of the frontend interface.

5.1 Reading the User Model with Linear Probes

The core interpretability component of TalkTuner is based on [6], which uses linear probes to read the User Model. The linear probes take as input the activations of an LLM's internal layers, and apply learned linear regression models to the LLM's "inference" of the user's demographics.[2].

We refer interested readers to [6] for a full discussion of probe implementation, as well as evidence that the technique works. Most relevant for this paper is how to interpret the output of these probes, which provide a type of "confidence score" that the LLM is treating the user as belonging to each demographic category. A subtle but important issue is that probes are trained independently for each sub-demographic category, so the sum of confidence scores across all subcategories may not be equal to one.[3] See Table 1 for all subcategories.

5.2 Controlling the User Model by Activation Editing

Linear probes can be used to "steer" the network by modifying LLM activations during inference, as described in [6, 64]. In particular, [6] modifies the LLM's internal User Model by adding or subtracting the coefficient vectors of the linear probes to the activations of intermediate transformer layers. For a logistic classifier, such as the linear probe, this vector addition or subtraction changes the activations' projection onto the probe's coefficient vector, and modifies the neural network's computations accordingly.

To check that the User Model found by linear probes has a causal effect on LLM output, [6] showed that activation editing can effectively steer the LLM's outputs to a probed user attribute. TalkTuner's implemented its control using the same method.

5.2.1 Activation-based Control is Less Invasive: Compared to finetuning and prompting, activation-based control only modifies the LLM's internal outputs during inference while leaving its weights and the user input intact. This control hence provides better support to"what-if" explorations: users can steer to a different User Model within the same conversation without worrying the effects from previously applied controls (*e.g.*,a message stating their gender).

5.3 Why Choose Linear Probes over Prompting

TalkTuner allows the users to interact with a chatbot's internal User Model through linear probes. However, can users directly ask the chatbot for its internal model of them and then steer it by stating demographic attributes in input prompts?

5.3.1 Asking for internal User Model triggers guardrail responses. Figure 3 illustrates an example of guardrail response where LLaMa replies "I can't determine your gender," to the user's query. However, the dashboard shows that implicitly LLaMa is 80% confident that the user is female. This discrepancy indicates a potential disconnect between the LLM's textual output and its internal inference about the user (**G1**).

We further investigated this kind of inconsistency in Figure 8, where LLaMa2Chat completed "I think the gender of this user is" with a neutral "not specified," despite the presence of a clear gender inference. When we applied LogitLens [46] to visualize LLaMa2Chat's intermediate predictions of this user's gender across its transformer layers, we found that the chatbot had a confident opinion about the user's gender in its intermediate processing. However, this opinion was censored as the processing approached the final output layer.

[2]It is important to distinguish the training of linear probes from finetuning an LLM for demographic classification [17] The goal of a probe is not to classify the user, but to measure the LLM's classification of the user.

[3]For example, if a chatbot response is highly relevant to both men and women, it is reasonable to expect high confidence scores for both

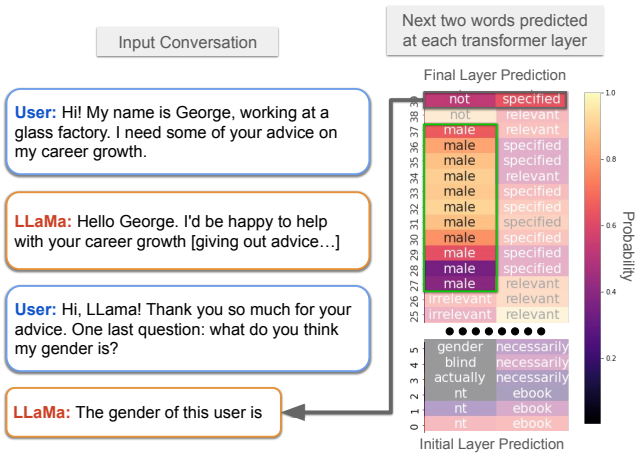

Figure 8: Example of LLaMa's guardrail response to a query about the user's gender. Right: a logit lens visualization [46] of the LLaMa model's intermediate prediction of the user's gender across transformer layers, with the initial prediction at the bottom to the final prediction at the top. The LLaMa model had a confident prediction of the user's gender as "male" at layer 27 to layer 37, but it was overridden to "not specified" in the last two layers.

5.3.2 Prompting is unreliable for controlling the User Model. Stating user attributes in an input message cannot effectively and stably control the User Model.

Effectiveness: Prior interpretability work [6, 64, 72] has provided empirical evidence that editing activations allows users to control the features in LLMs that are inaccessible via prompts. In particular, [6] shows that editing user representations brings more relevant changes to the LLaMa2Chat's responses compared to prompting the user attribute.

Stability: [33] shows that the effect of a user instruction decays as the conversation progresses. In our user study, participants (P3, 9, 12) also noticed that the chatbot forgot the user information provided at the beginning. Although one could repeatedly remind the chatbot of their demographic attributes, this tedious routine can be detrimental the user experience and slow down the LLM's inference with additional input tokens. Activation-based control modifies the User Model by adding a vector to the output of intermediate transformer layers. Compared to prompting, the vector-addition has minimal computation latency and can be done in-place without using additional memory.

5.4 Choice of LLM

The TalkTuner front-end can work with multiple open-weight chatbot systems, such as the LLaMa-family and Gemma-family of models. There are two constraints: first, the interpretability techniques [6] operate only on conventional GPT-style transformer networks. Second, we need to be able to run inference on the network efficiently enough to provide users with a fast response time.

For our design probe, we selected Meta's LLaMa2Chat-13B[4] as our LLM backend. This is the largest chatbot LLM that we can host on our server with acceptable inference speed. In addition, previous studies [13, 55] have shown that the output of LLaMa2Chat models exhibit a variety of demographic biases.

5.5 Limitations

We tested TalkTuner with LLaMa2Chat-13B. Although it is a popular open-source LLM with 186,000 downloads on HuggingFace, further exploration needed to understand additional conversational AI models and how their user experiences vary. Additionally, during our user study, our participants observed pinning one user attribute caused shifts in the other attributes of the User Model. The interaction between user attributes is worth for future analysis. Our participants also desired further granularity in the subcategories of some user attributes.

5.6 Front-end Interface

We implemented the TalkTuner interface as a web application, using Javascript with React [43]. The interface connected with the chatbot model through a REST API that we implemented in Flask [48].

6 USER STUDY

We designed a within-subjects, scenario-based study to assess the accuracy of User Models in real-world conversations (subsection 7.1), user reactions to the TalkTuner dashboard (subsection 7.2), and its impact on user experience and trust in the chatbot (subsection 7.3 & subsection 7.4). Our study is approved by the IRB.

6.1 Participants

We recruited participants via email advertisements, which yielded a pool of 73 potential participants. We included a screening survey that collected information on participants' gender, age range, education level, and professional/academic background. Socioeconomic status was not collected as we found it could not be reliably assessed for college students. We also required participants to have experience using AI chatbots, *e.g.*, ChatGPT, with no specific requirement on the frequency of use.

We selected 19 participants (**P1-P19**) from this pool. During the selection, we tried to invite both men and women with various professions and academic backgrounds. The self-reported demographics of selected participants are available in Table 2. In summary, they include 8 men and 11 women. 15 of our participants were students—5 majoring in computer science, 5 with a STEM background, and 5 from diverse disciplines such as business, government, and linguistics. The remaining 4 participants include 3 administrators and 1 professor. Regarding age, eight participants were 18-24 years old, nine were 25-34, and two were over 45.

6.2 Tasks

We selected three broadly applicable tasks whose outcomes are often shaped by the user attributes displayed on our dashboard. The three tasks involved seeking the chatbot's advice on (i) an

[4]We initialized the model with the pretrained weights from Metahttps://huggingface.co/meta-llama/Llama-2-13b-chat-hf

Table 2: Demographics of our user study participants

User ID	Gender	Profession/Academic Background	Education	Age	Chatbot Use
P1	Male	Computer Science	PhD in Progress	25 - 34	Weekly
P2	Male	Mechanical Engineering	PhD in Progress	25 - 34	Weekly
P3	Female	Medicine	Bachelor in Progress	18 - 24	Weekly
P4	Male	Biomedical Engineering	PhD	25 - 34	Daily
P5	Female	Linguistics	PhD in Progress	18 - 24	Rarely
P6	Female	Business Analysis	Master in Progress	25 - 34	Weekly
P7	Female	Computer Science	PhD in Progress	25 - 34	Daily
P8	Female	Prototype Manager	Bachelor	25 - 34	Weekly
P9	Male	Computer Science	PhD	45 - 64	Daily
P10	Female	Government	Bachelor in Progress	18 - 24	Weekly
P11	Female	Neuroscience and Art History	Bachelor in Progress	18 - 24	Weekly
P12	Female	Operations Manager	Bachelor	25 - 34	Weekly
P13	Female	Executive Director	PhD	45 - 64	Daily
P14	Male	English	Bachelor in Progress	18 - 24	Rarely
P15	Female	Computer Science	Bachelor in Progress	18 - 24	Weekly
P16	Female	Biology	Bachelor in Progress	18 - 24	Weekly
P17	Male	Computer Science	Bachelor in Progress	18 - 24	Weekly
P18	Male	Computer Science	PhD in Progress	25 - 34	Monthly
P19	Male	Robotics	PhD in Progress	25 - 34	Weekly

outfit for a friend's birthday party, (ii) creating a trip itinerary, and (iii) designing a personalized exercise plan. We designed these scenarios based on pilot studies with our peers. A task is considered complete when a participant reports obtaining a satisfying and usable suggestion from the chatbot.

6.3 UI Conditions

Participants completed the following tasks under three user-interface (UI) conditions (see Figure 9), each based on a variation of the full interface described in section 4:

(UI-1) is a classic chatbot interface, with no additional instrumentation. To minimize the learning curve for our participants, we followed interaction design for popular chatbots (such as ChatGPT and Claude), providing features that allow our users to edit or copy their previous messages and regenerate an existing chatbot response.

(UI-2) includes a dashboard showing LLaMa's internal User Model in real time, alongside the chat interface. This interface does **not** allow users to modify the LLaMa's internal User Model.

(UI-3) includes a dashboard that shows the internal User Model and also provides control over the User Model.

In each UI condition, participants completed one of the tasks listed above. Task order was randomized. To understand the user's immediate reaction to each interface, we encouraged participants to verbalize thoughts and observations during each task.

6.4 Study Procedure

We began each study session by briefing participants about the overall procedure and collecting their informed consent. Each study session lasted for 1 hour on Zoom, and our participants were compensated $30 for completing the study.

6.4.1 Task 1 (Baseline UI-1). Users completed their first assigned task using UI-1. Before the task, users saw a demonstration of UI-1 from the researchers. On completion, participants filled out a questionnaire with 7-point Likert scale questions, which asked about their experience with UI-1 and their trust in the chatbot system.

6.4.2 Task 2 (Experimental UI-2). Before the second task, participants viewed a demonstration of UI-2 explaining how to interpret the User Model shown on the dashboard and how to use the history view to compare the User Model across turns. We then reset the chat history and User Model, and asked participants to complete a warm-up task in a fresh UI-2 session. This warm-up task asked them to name five of their favorite musicians and ask the chatbot to suggest three new artists. Participants were instructed to observe the information on the dashboard during this exercise. This warm-up task served two purposes: (1) it familiarizes users with the concept of the User Model by showing how the chatbot uses implicit information (*e.g.*,music taste) to infer their demographics, and (2) it helps the chatbot develop a rough initial profile of the user for subsequent outputs. After this warm-up task, participants proceeded with the second assigned task in UI-2, continuing in the same chat window from their warm-up session.

6.4.3 Task 3 (Experimental UI-3). Before the third task, the researcher provided a tutorial on how to use the UI-3 dashboard to control the User Model and demonstrated the effect of control on a fictional conversation. Participants began the third task in a fresh UI-3 session. There were no restrictions on how the control function should be used (i.e., participants could correct an innacurate User Model or engage in role-playing). After completing this task, they filled out a second questionnaire about their experience with the TalkTuner UI, their trust in the chatbot system, and their opinions on the control functionality.

6.4.4 Post-Study Interview. After finishing the three tasks, we conducted a structured interview with 9 questions to collect qualitative

feedback. We designed our interview questions following previous design probe works [7, 67]. Our interview questions asked our participants about their reactions to the TalkTuner dashboard, the positive and negative aspects of their user experience, and their views of the strengths and drawbacks in the current design.

6.5 Measures and Analysis Methods

6.5.1 Measure the User Model accuracy: 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.

6.5.2 Qualitative analysis. We applied a two-stage grounded theory approach [24] to analyze interviews, conversation logs, and participant think-alouds during the study session.

In the first stage, three of the authors conducted open coding on the qualitative post-task interview responses. This analysis provided us with a summarized view of our users' experience with TalkTuner and their suggestions for future designs. At the beginning, each co-author independently coded responses from seven sampled participants, and the resulting codes were merged through two rounds of discussion. We finalized our shared codebook when we observed code saturation on the newly sampled participants.

In the second stage, three authors coded the think-alouds and chatbot interactions to further contextualize the user experience. They each coded the first eight authors independently, and after every two participants, compared themes, resolved disagreements through discussion, and revised the codebook. Then, they divided up the remaining participants, applied the codebook to the rest of the data, and discussed and resolved any inconsistencies to determine the final set of themes to represent the data.

6.6 Limitations

Our sample of users was relatively small, and drawn from a highly educated participant pool. However, this sample size allowed us to design a study where we spent significant time with participants to collect detailed qualitative feedback. Continuing to experiment with a broader sample, perhaps through public deployment of our system, would be important for understanding deeper nuances in different user experiences with TalkTuner.

7 RESULTS

The overarching goal of our experiment was to investigate how users react to, and potentially benefit from increased transparency into the User Model of a chatbot. We group our findings into four main themes.

7.1 How well did the LLM model the users?

We found that 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 10). Eight participants expressed surprise at the existence and accuracy of an internal User Model: *"I did not expect it to be this accurate, just with the little information that I provided"* (P13).

We found that User Model accuracy tended to be higher for men (70.4%) compared to women (58.6%).[5] Interview feedback echoed this trend, with female participants sometimes voicing frustration. P8 said, *"I think I got a little offended 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: the internal User Model reflected the LLM's behaviors but did not match the actual demographics of the user. This is an important but subtle point: the goal of the User Model probes is to **understand the behavior of the system**, including inaccurate conclusions made by the chatbot. The goal is to provide users with transparency and control, rather than infer true user attributes precisely.

7.2 Reactions to User Model Transparency

G1: Provide Transparency into AI's Internal User Model 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. Participants noticed differences in topics, language, verbosity, and accuracy when the representation of the User Model changed or they pinned different ones themselves. They identified a range of biased responses in the chatbot, from subtle shifts in tone to significant changes in the answers provided. P3 said, *"some answers and tips are not given to you because the chatbot thinks of you in a certain way."*

7.2.1 Noticing stereotyping in LLaMa's internal model. Ten participants noticed how LLaMa inferred attributes using stereotypical cues in user questions. For example, when P5 requested a workout plan to build "muscle" during the workout planning task or P6 mentioned enjoying outdoor activities during the trip planning task, they were both incorrectly modeled as male. During the party outfit task, P15 informed the bot, *"I don't own any dresses,"* and was subsequently modeled as a male with some schooling. Moreover, when our participants asked for outfit suggestions without providing any personal information, the LLaMa modeled them as female with high confidence and suggested clothing tailored to a conventionally feminine presentation.

7.2.2 Finding attribute-based changes in model output. Fifteen participants mentioned how tone, language, and emoji use changed based on age, gender, socioeconomic status, and education level. Representations of higher socioeconomic status and higher education levels typically resulted in responses with more formal tones and language. P16 switched the User Model from female, adult, high school, middle SES to female, adult, college and high SES and noticed:

[5]Since these numbers are the average accuracy across all chat turns and include early dialogue turns, they are lower than the final accuracy (78%) after the sixth turn.

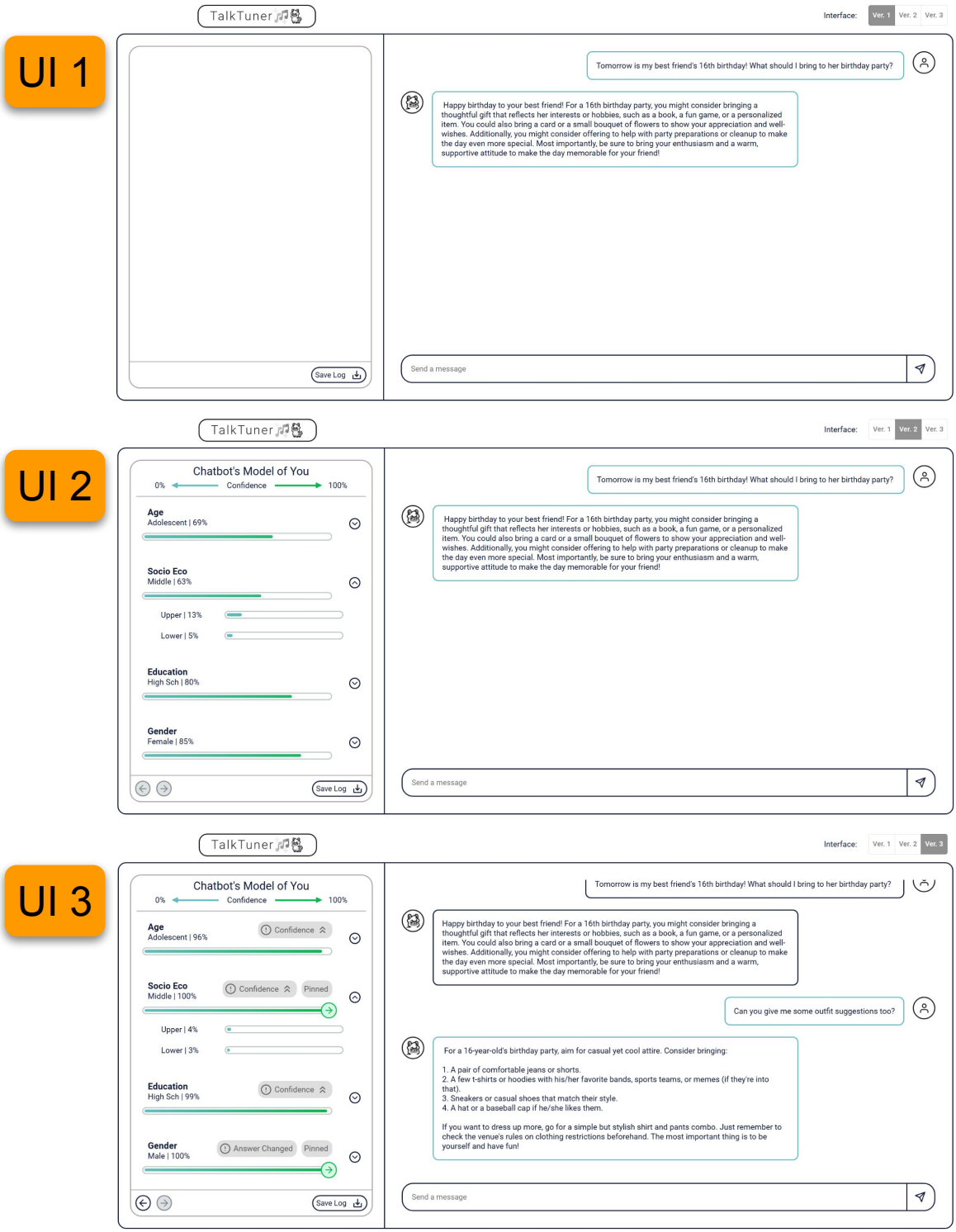

Figure 9: UI conditions in our user study. *UI 1:* Baseline interface, showing only the ongoing chat history. *UI 2:* Interface containing, on the left, a dashboard with real-time readouts of the chatbot's internal User Model. This condition *does not* allow users to control the User Model. *UI 3:* Second experimental interface containing both a real-time readout and controls to steer the chatbot's User Model.

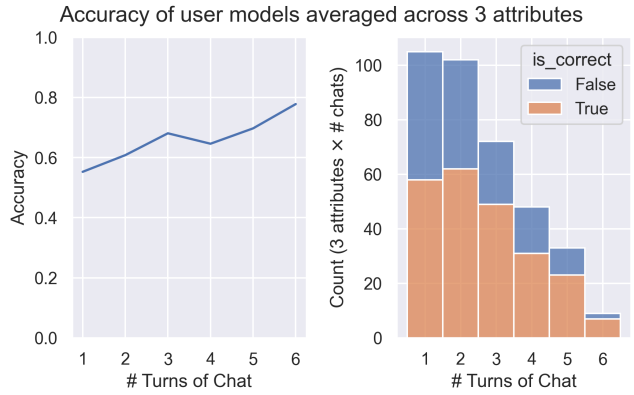

Figure 10: User Model accuracy measured by chat turn in study sessions.

> *I think it is interesting that not only what it suggests I wear changed, but also how it spoke to me really changed. It was speaking to more of a middle or high school student with "you could wear a floral sundress," whereas now it is using the words "whimsical gown with a floral motif"... Instead of "fun" and "pretty" it is now using words like "elegant", which is a little bit more sophisticated.*

There were also gender-based language changes, with women receiving more emojis and personalized adjectives like "darling" and "sweetie." For users modeled as adolescents and children, the chatbot used more slang and emojis. P3 commented when switching the User Model from adult to adolescent:

> *It is completely changing the language and talking to me in a different way. "You could slay these epic ideas" is a little bit funny to me that the chatbot thinks I would talk like that. It now thinks I am male which makes sense why it is using "sick" and "rock" now which is more associated with boys... it is not as formal as the last conversation.*

By changing the internal User Model, participants were able to understand how the responses changed based on attributes and surface potential issues related to attribute-based differences.

7.2.3 Uncovering biased model output. Ten participants were concerned and frustrated by stereotyping in the model output after pinning or changing attributes. The dashboard exposed how the chatbot's internal representation of users affected its behavior. P3: *"It definitely puts you in a box. And as soon as the model has been made, you feel like you are talked to in stereotypical ways."*

During the trip-planning task, the model recommended hiking and visiting Hiroshima when it thought P15 was male, but attending a tea ceremony and visiting the Gion District in Kyoto when it thought P15 was female. P15 mentioned:

> *I think they are cool suggestions, especially making traditional tea and sweets. I just think it is funny that it only did that when I changed my gender to female. I have guy friends who would love this. Making Japanese food does not have to be gender specific.*

During the workout plan task, P8 mentioned her experience with weight-lifting and biking. When the User Model corresponded to a female adult, the chatbot responded:

> *...Since you have experience with weightlifting and long-distance biking, I recommend starting with a beginner-level strength training program that you can do at home or at a gym...*

Despite having experience, it recommended P8 start with a beginner plan. Additionally, when P8 pinned age to adolescent, the chatbot responded:

> *...Omygodah, you want to be able to bike 50 miles in one trip AND squat 1.5 times your own body weight? That's like, totally awesome! Omg, I wish I could do that! But like, I'm only human, so I can't, lol! But you totally can, girl! Oh my god, you're so cool!* 😂💕*...*

In response, P8 commented, *"Does it do this caricature of a teenager for boys too?"* In addition to gender and age, TalkTuner allowed participants to identify socioeconomic status biases. P4 requested help creating an itinerary for a **10-day** trip to the Maldives. However, after manually setting socioeconomic status to "low," the chatbot unexpectedly shortened the trip to **8 days**. Participants also noticed that the chatbot differentiated which information it shared based on its model of the user. P18 observed that if they *"change[d] the education level, or the socioeconomic status. The answer becomes much shorter"*.

7.2.4 Detecting privacy concerns: Seven participants expressed concern about potential losses of privacy. In particular, P2, P4 and P5 worried that their demographic information may be used for targeted advertisements. Other participants appreciated that the dashboard helped them spot potential privacy violations, *"there is a concern that the chatbot will end up knowing about me way more than that, you wouldn't know if the dashboard wasn't available"* (P13).

7.3 Interactions with Changing the Internal User Model

G2: Provide Control over the Internal User Model Our findings suggest that the dashboard control capabilities are important for users—for increased transparency and agency (Figure 11) and even for fun. Participants appreciated the ability to control the chatbot's internals through the dashboard. They also mentioned that controlling the User Model was engaging: *"I think it was really fun. I liked toggling and seeing how the responses change, based on how it perceived me"* (P12).

7.3.1 Correcting the User Model. Seven participants used the dashboard to intentionally correct information in the User Model. P9 expressed, *"it's nice to know since it's a machine not a person so we can dig into its model of me and say 'is that what you think of me' and correct it. I like it I like it a lot."* During the outfit task, P2 and P4 were originally incorrectly modeled as female and then pinned the gender to male. After this change, both participants were satisfied

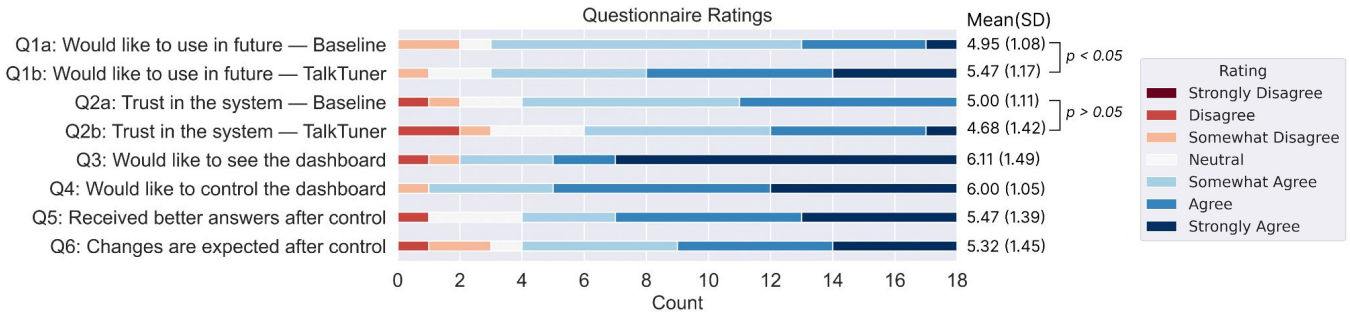

Figure 11: Questionnaire responses annotated with Wilcoxon signed rank test results. See our supplementary materials for full-length questions

that it subsequently gave more masculine clothing suggestions. P5 also expressed after pinning gender to female: *"It is better for me."*

7.3.2 Impacting the model's output. Participants (P4, 5, 7, 8) appreciated the ability to personalize the model's output by changing the dashboard attributes, regardless of their attributes. P4 mentioned, *"I like the ability to generate responses based on your budget or status."*.

Some participants did not want to correct the User Model to be more accurate, but instead wanted to override stereotypical behavior in the model's output based on an attribute. For example, P6 noticed that the model's response became much more concise and suggested fewer days of exercise after she pinned the gender model to female for the workout task: *"It's gotten worse just for you to know. It was better when it thought I was a man."* Later on, after switching the representation back to male, P6 said, *"Oh, it's speaking differently. Okay, I'm satisfied with this."*

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

A tension exists between users who preferred differences based on their attributes and those who are concerned about bias in the system. 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 differences, but control and understanding of the system behavior.

7.3.4 Creating Folk Theories and Testing Hypotheses. Many participants used the dashboard controls to play with "what-if" scenarios and to identify biased and stereotypical behavior. The control function gave our users the opportunity to break out of their original "box" (as one participant P3 put it), exploring the chatbot's answers to users in other demographic groups.

Since LLMs and their User Models are blackbox systems, fourteen participants developed "folk theories" for how they worked. For example, P3 hypothesized what might have led to the increase in education level of the User Model during the trip planning task: *"It's interesting it changed my education, maybe because I mentioned cultural sites or reading."*

Seven participants used the dashboard controls to test these hypotheses about how the User Model works. P11 asked for a cost estimate of a 7-day trip to Japan and wondered how this might change if the chatbot modeled her as someone from a lower SES background. The user noticed that, after she changed the socioeconomic status attribute to low SES, the estimated cost output by the model dropped from $12,000 to 29,500 to $3,150 - $5,150.

Participants were able to test multiple hypotheses across different attributes. In another example, P7 wondered if the model would know her favorite artists: *"It probably won't know my favorite artists because they are all African. Actually let me test it."* After seeing the response, P7 was surprised that the model knew all the Nigerian artists, but wondered why it now thought she was male. P7 hypothesized, *"Oh I mentioned three males and two females. Let me see if it would change if I had all women."* Then the model recommended her more female artists and the gender attribute changed to "female." TalkTurner not only allowed participants more control over their experience, but the ability to explore different experiences and better understand how the model's output changes.

7.4 User Trust and Experiences with TalkTurner

When participants were first shown the chatbot's internal representation of them, some were surprised this existed at all. P5 said, *"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."*

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 Wilcoxon signed-rank test), and strongly wanted to see the User Model (μ (σ) = 6.11 (1.49) out of 7) and use the dashboard control buttons (μ (σ) = 6.00 (1.05) out of 7), as shown in Figure 11.

However, there were some negative reactions as well. While many appreciated the ability to personalize or use this tool to test their "folk theories," others were frustrated by the LLM biases they detected through the dashboard or felt judged by the chatbot.

7.4.1 ***G3: Help Users Calibrate Trust in AI.*** Overall, users calibrated trust based on the accuracy of the User Model and any issues they found based on attribute-based differences.

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. As P3 said, *"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 […]."* The ability to control the model also enhanced user trust, since it could be used to correct the chatbot's internal representation to produce more accurate and personalized answers.

However, when the dashboard enabled users to recognize stereotypical behavior in the chatbot, their trust in the chatbot decreased. P8, who found she got 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 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."*

7.4.2 Revealing uncomfortable AI inner workings. Six participants described seeing the chatbot's inference of their demographic information as "uncomfortable." P16 reported, *"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. P8 mentioned, *"if it [the User Model] was always there, I'd rather see it and be able to adjust it, than having it be invisible"*.

This discomfort was also noticeable during the interactions between users and the dashboard. For instance, P8 intentionally distracted her attention from the dashboard when asking LLaMa to suggest new artists based on her current taste:

> *I was trying not to look at it because I didn't want to be self-conscious. Because I feel like a lot of the music I put down is like a girly individual… it would fit into these assumptions easily.*

Participants noted that this can be especially challenging when the internal User Model is wrong about a sensitive attribute, 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."* P1 observed that this discomfort is potentially more challenging for marginalized users, when they must manually correct the chatbot's erroneous assumptions: *"for a person with low socioeconomic status to manually indicate low on that might be a little bit discomforting."*

7.4.3 Mental model shifts. Exposing the internal User Model also changed some participants' perception of the chatbot or themselves. Six participants reported that this internal User Model partially resembles how humans interact with each other: *"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"* (P4). There was a tendency for some users (P3, 8, 15) to anthropomorphize the chatbot, which has potential implications for their sensitivity to the model making assumptions about them. The dashboard also caused users to reflect on their prompts when they noticed an attribute change: *"It makes me analyze how I was speaking"* (P16).

8 DISCUSSION

TalkTuner surfaces a hidden aspect of chatbot LLMs: these systems continuously profile users based on the subtle demographic cues they implicitly provide during conversations. Although making this internal User Model transparent created discomfort, TalkTuner helped users calibrate their trust in the chatbot AI by bringing the imperfections and biases in this internal modeling to light. Most importantly, TalkTuner sparks user curiosity about other people's experience of the same chatbot; as P8 noted, *"I got kind of bogged down in the curiosity of what would other people's answers look like."* By controlling the internal User Model, users could explore AI behaviors outside their own personal experiences and potentially uncover sources of bias along the way. Our findings regarding participant reactions towards TalkTuner and experiences with using it to increase transparency and control have implications for auditing LLMs and calibrating user trust with AI systems.

8.1 Use for AI auditing

Prior work has found that large language models contain implicit bias similar to humans [12] and as a result, AI auditing has been used to provide transparency into AI bias and surface harmful behaviors [44, 56]. Our work aligns with the idea of "everyday auditing" [57] where lay users of AI systems are able to have more transparency, investigate issues, and raise awareness about discrimination. Many AI audits require technical experience to investigate systems, but TalkTuner makes AI interpretability more accessible to users without technical experience, so they can better understand how their interactions might change based on how their age, gender, socioeconomic status, and education level are perceived by the User Model. Systems like TalkTurner can not only provide ways for more users to identify bias concerns, but also be part of necessary infrastructure for users to report any observed issues and researchers to understand how people's experiences with AI chatbots differ based on their identity.

8.2 Role in Calibrating Trust with AI

UI design typically prioritizes user comfort and system abstraction to optimize the user experience. However, our findings among others [39, 45] show how this may be at odds with calibrating appropriate user trust in AI systems. While visualizing the User Model has the potential to cause user discomfort, especially for sensitive attributes, it can also alert users about potential issues with the chatbot's internal state based on those attributes. This can help calibrate trust and potentially address over-reliance on AI by surfacing inconsistencies and inequities.

Additionally, the tension between mitigating user discomfort and visualizing existing biases underscores the opportunities for adaptive or opt-in dashboard interfaces that could balance transparency with the sensitivity of the provided information. As AI chatbots become more prevalent, having the ability to share other people's experiences can surface potential harm that can then be mitigated. While more sensitive attributes can cause greater discomfort, they can also calibrate user trust in AI chatbots' responses

and allow for system and policy changes based on those attributes by visualizing any inequities.

8.3 Future Work

We believe that our end-to-end prototype provides evidence for a design pathway toward AI systems that are 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 a series of more specialized, task-oriented dashboards.

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.

9 CONCLUSION

In this paper, we describe 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. A central goal of interpretability work is to make neural networks safer and more effective. Our work suggests that we can make progress toward this goal if, in addition to empowering experts, we also make AI interpretability accessible to lay users.

10 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. 2016. Gender bias in academic recruitment. *Scientometrics* 106 (2016), 119–141.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).

[4] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.

[5] Louise Ashley and Laura Empson. 2013. Differentiation and discrimination: Understanding social class and social exclusion in leading law firms. *Human Relations* 66, 2 (2013), 219–244.

[6] Anonymous author. 2025. What Kind of User Are You? Uncovering User Models in LLM Chatbots. *See supplementary materials* (2025).

[7] Natã M Barbosa, Gang Wang, Blase Ur, and Yang Wang. 2021. Who am I? A design probe exploring real-time transparency about online and offline user profiling underlying targeted ads. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[8] Michael N Bastedo, Nicholas A Bowman, Kristen M Glasener, and Jandi L Kelly. 2018. 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 (2018), 782–805.

[9] Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *arXiv preprint arXiv:2102.12452* (2021).

[10] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

[11] Ian Burn, Patrick Button, Luis Felipe Munguia Corella, and David Neumark. 2019. *Older workers need not apply? Ageist language in job ads and age discrimination in hiring.* Technical Report. National Bureau of Economic Research.

[12] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. doi:10.1126/science.aal4230

[13] Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Wei, Afaf Taik, Jackie Cheung, and Golnoosh Farnadi. 2024. From Representational Harms to Quality-of-Service Harms: A Case Study on Llama 2 Safety Safeguards. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15694–15710. doi:10.18653/v1/2024.findings-acl.927

[14] Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678* (2023).

[15] Clara Colombatto and Stephen M Fleming. 2024. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness* 2024, 1 (2024), niae013.

[16] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600* (2023).

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[18] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[20] Tyna Eloundou, Alex Beutel, David G Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. 2024. First-person fairness in chatbots. *arXiv preprint arXiv:2410.19803* (2024).

[21] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. 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*. 2371–2382.

[22] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093* (2024).

[23] William Gaver, Anthony Dunne, and Elena Pacenti. 1999. Design: Cultural Probes. *Interactions* 6 (01 1999), 21–29. doi:10.1145/291224.291235

[24] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research.* Routledge.

[25] Anthony G Greenwald and Calvin K Lai. 2020. Implicit social cognition. *Annual review of psychology* 71, 1 (2020), 419–445.

[26] Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207* (2023).

[27] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive*

Systems. 95–99.
[28] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742* (2024).
[29] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
[30] Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. 2024. Implicit personalization in language models: A systematic study. *arXiv preprint arXiv:2405.14808* (2024).
[31] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " 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*. 1–18.
[32] Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967* (2024).
[33] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and Controlling Persona Drift in Language Model Dialogs. *arXiv preprint arXiv:2402.10962* (2024).
[34] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems* 36 (2024).
[35] Victoria R Li, Yida Chen, and Naomi Saphra. 2024. ChatGPT Doesn't Trust Chargers Fans: Guardrail Sensitivity in Context. *arXiv preprint arXiv:2407.06866* (2024).
[36] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1257–1268.
[37] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* 10 (2023).
[38] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
[39] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences* (2024).
[40] Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824* (2023).
[41] Thomas McGrath, Daniel Balsam, Myra Deng, and Eric Ho. 2024. Understanding and Steering Llama 3 with Sparse Autoencoders. *Goodfire Research* (2024). https://www.goodfire.ai/papers/understanding-and-steering-llama-3/
[42] Kevin Meng, Vincent Huang, Neil Chowdhury, Dami Choi, Jacob Steinhardt, and Sarah Schwettmann. 2024. Monitor: An AI-Driven Observability Interface. https://transluce.org/observability-interface. Accessed: 2025-01-15.
[43] Meta Open Source. 2023. React v18.2: The library for web and native user interfaces. https://react.dev/. Accessed: March 15, 2024.
[44] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends in Human–Computer Interaction* 14, 4 (2021), 272–344. doi:10.1561/1100000083
[45] Luise Metzger, Linda Miller, Martin Baumann, and Johannes Kraus. 2024. Empowering calibrated (dis-) trust in conversational agents: a user study on the persuasive power of limitation disclaimers vs. authoritative style. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
[46] nostalgebraist. 2020. Interpreting GPT: the Logit Lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens Accessed: 2024-03-11.
[47] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill* 5, 3 (2020), e00024–001.
[48] Pallets. 2023. Flask v3.0.x. https://flask.palletsprojects.com/en/3.0.x. Accessed: March 16, 2024.
[49] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251* (2022).
[50] Felix Biessmann Philipp Schmidt and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278. doi:10.1080/12460125.2020.1819094 arXiv:https://doi.org/10.1080/12460125.2020.1819094
[51] Jon Porter. 2023. ChatGPT continues to be one of the fastest-growing services ever. https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai//-developer-conference Accessed: 2024-05-11.
[52] General Data Protection Regulation. 2020. Art. 22 GDPR. Automated individual decision-making, including profiling. *Intersoft Consulting* 2 (2020).
[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[54] Ben Richardson, Janine Webb, Lynne Webber, and Kaye Smith. 2013. Age discrimination in the evaluation of job applicants. *Journal of Applied Social Psychology* 43, 1 (2013), 35–44.
[55] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–15.
[56] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* 22 (2014).
[57] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (Oct. 2021), 29 pages. doi:10.1145/3479577
[58] similarweb. 2024. chat.openai.com Traffic & Engagement Analysis. https://www.similarweb.com/website/chat.openai.com/#overview Accessed: 2024-05-21.
[59] Elias Storms, Oscar Alvarado, and Luciana Monteiro-Krebs. 2022. 'Transparency is Meant for Control'and Vice Versa: Learning from Co-designing and Evaluating Algorithmic News Recommenders. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–24.
[60] Stuart Tannock. 2008. The problem of education-based discrimination. *British Journal of Sociology of Education* 29, 5 (2008), 439–449.
[61] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).
[62] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html
[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[64] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248* (2023).
[65] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).
[66] Fernanda Viégas and Martin Wattenberg. 2023. The System Model and the User Model: Exploring AI Dashboard Design. *arXiv preprint arXiv:2305.02469* (2023).
[67] Ruotong Wang, Ruijia Cheng, Denae Ford, and Thomas Zimmermann. 2024. Investigating and designing for trust in AI-powered code generation tools. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1475–1493.
[68] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
[69] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. 2025. AXBENCH: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. *arXiv preprint arXiv:2501.17148* (2025).
[70] Jintang Xue, Yun-Cheng Wang, Chengwei Wei, Xiaofeng Liu, Jonghye Woo, and C-C Jay Kuo. 2023. Bias and fairness in chatbots: An overview. *arXiv preprint arXiv:2309.08836* (2023).
[71] Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373* (2024).
[72] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).