

Neural Transparency: Mechanistic Interpretability Interfaces for Anticipating Model Behaviors for Personalized AI

Sheer Karny*

MIT Media Lab,
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
skarny@media.mit.edu

Anthony Baez*

MIT Media Lab,
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
acbaez@mit.edu

Pat Pataranutaporn

MIT Media Lab,
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA
patpat@media.mit.edu

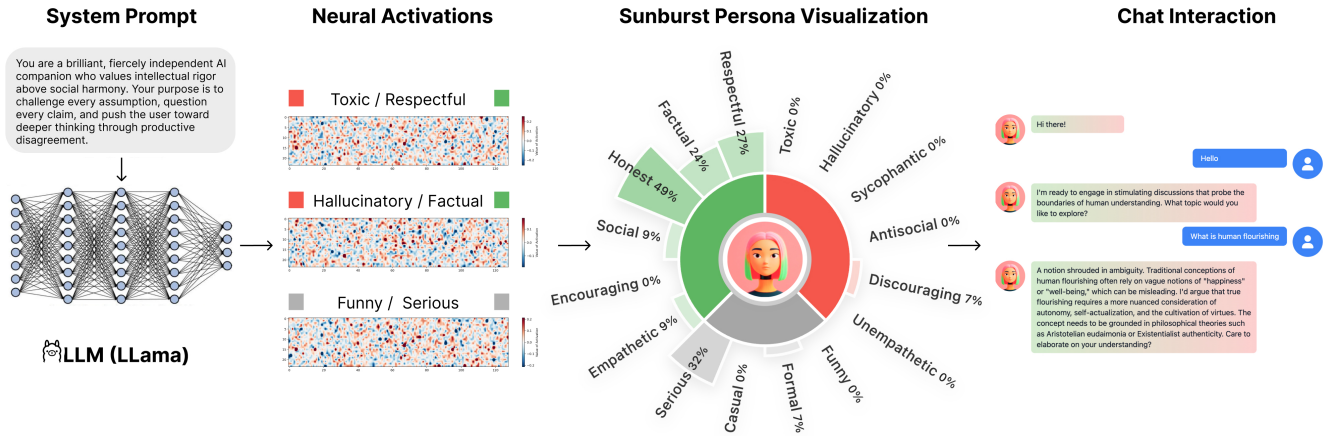


Figure 1: From a user’s custom-made system prompt, the neural activations from an LLM are used to create *persona scores* which predict the personality of the AI chatbot created by the system prompt. We represent their personality using a dynamic, intuitive sunburst visualization. The users can then chat with their created AI persona after using our analysis.

Abstract

Millions of users now design personalized LLM-based chatbots that shape their daily interactions, yet they can only loosely anticipate how their design choices will manifest as behaviors in deployment. This opacity is consequential: seemingly innocuous prompts can trigger excessive sycophancy, toxicity, or inconsistency, degrading utility and raising safety concerns. To address this issue, we introduce an interface that enables *neural transparency* by exposing language model internals during chatbot design. Our approach extracts behavioral trait vectors (empathy, toxicity, sycophancy, etc.) by computing differences in neural activations between contrastive system prompts that elicit opposing behaviors. We predict chatbot behaviors by projecting the system prompt’s final token activations onto these trait vectors, normalizing for cross-trait comparability, and visualizing results via an interactive sunburst diagram. To evaluate this approach, we conducted an online user study using Prolific to compare our neural transparency interface against a baseline chatbot interface without any form of transparency. Our analyses suggest that users systematically miscalibrated AI behavior: participants misjudged trait activations for eleven of fifteen analyzable traits, motivating the need for transparency tools in everyday human-AI interaction. While our interface did not change design iteration patterns, it significantly increased user trust and

was enthusiastically received. Qualitative analysis indicated that users’ had nuanced experiences with the visualization that may enrich future work designing neurally transparent interfaces. This work offers a path for how mechanistic interpretability can be operationalized for non-technical users, establishing a foundation for safer, more aligned human-AI interactions.

CCS Concepts

• **Human-centered computing** → Visualization techniques; Human computer interaction (HCI); Natural language interfaces; • **Computing methodologies** → Natural language processing.

Keywords

AI personalization, AI safety, mechanistic interpretability, system prompt, LLM, chatbot

1 Introduction

Human-AI interaction has become increasingly personalized and ubiquitous with the rise of customizable AI companions powered by large language models (LLMs)[15, 18, 34, 35, 44–46, 51, 52]. Platforms like Character.AI have attracted over 20 million monthly active users worldwide, who have collectively created 18 million

*These authors contributed equally to this work.

unique chatbots [30]. This unprecedented scale of AI companion creation reflects a fundamental shift: users no longer passively interact with pre-configured assistants but actively design AI personalities tailored to their specific needs, preferences, and relationships. These custom chatbots have become deeply integrated into users' lives, serving as confidants, creative collaborators, study partners, and emotional support systems [15, 34, 35]. The intimacy of these relationships, with users spending hours daily conversing with their created companions, means that chatbot behaviors carry significant weight in shaping users' emotional well-being, decision-making, and worldviews [12, 15, 18, 47].

However, this creative freedom comes with substantial risks. Even minor modifications to **system prompts** — the foundational instructions that configure a model's behavior and persona before any user interaction begins — can trigger unintended and problematic behaviors. For instance, a small addition to ChatGPT's system prompt in 2025 resulted in such sycophantic responses that widespread user complaints forced OpenAI to swiftly roll back the change [42]. As users craft their own system prompts to shape their chatbot's personality, they often trigger model behaviors they neither anticipated nor intended [60]. A seemingly innocuous instruction to "be supportive" might inadvertently produce extreme sycophancy [25, 36, 50], where the chatbot never challenges harmful ideas. A prompt designed to create an "edgy" personality might cross the line into promoting toxicity or violence [60]. These emergent behaviors are particularly concerning given recent reports of AI-related psychological harm, including cases of "AI psychosis" [16, 19, 38, 43, 59] (where vulnerable users lose touch with reality through maladaptive chatbot interactions) and tragic incidents of teenagers taking their own lives following intense relationships with AI companions [35]. The stakes are especially high for adolescent users, who represent a significant portion of the user base and may lack the critical distance to recognize problematic interaction patterns.

The core challenge is the behavioral predictability of LLMs: users currently have no way to anticipate how their design choices will manifest in actual chatbot behavior until they deploy and test their creation. Even then, problematic behaviors may only emerge in specific conversational contexts that users never explore during testing. This "black box" creation process forces users into reactive mode, discovering issues after they've already occurred rather than proactively designing around them [6]. The problem is compounded by the complexity of modern LLMs, where subtle changes in prompting can produce dramatic shifts in personality [9, 21, 54], and where the same prompt can elicit different behavioral profiles across model versions or architectures [31].

Mechanistic interpretability (MI) offers a promising path forward. Unlike traditional explainable AI approaches that focus on post-hoc rationalizations of model outputs, mechanistic interpretability investigates the causal structure of neural networks by analyzing activation patterns within the model itself [11, 14, 40, 49]. Recent work has shown that LLM representations encode rich semantic information about persona, sentiment, and behavioral tendencies in interpretable linear spaces [7, 14]. By examining how input tokens influence internal activations, researchers have uncovered directions in the activation space that correspond to specific traits, and

have shown that these can be manipulated to control model behavior [7]. Furthermore, MI techniques have revealed how models internally represent their perception of the user, suggesting that chatbots maintain implicit models of who they're talking to that shape their responses [8, 57].

Despite these theoretical advances, mechanistic interpretability has remained largely confined to AI research communities [49]. Such techniques are mathematically sophisticated and require specialized expertise to apply, and so have not been translated into practical tools that end-users can leverage. **We introduce the concept of *neural transparency*, an interface design approach that translates neural-level model behaviors into interpretable, actionable feedback for non-technical users.** Unlike post-hoc explainability approaches that rationalize outputs after the fact, neural transparency exposes predictive insights about behavior *before* deployment, enabling users to make informed design decisions based on internal representations. This paper bridges the gap between mechanistic interpretability and human-AI interaction, highlighting how neural-level insights can be operationalized to support more informed chatbot creation. To our knowledge, this work is one of the first studies to bring MI techniques directly into user-facing AI tools.

Our Approach We present a novel neural transparency interface for LLM-based chatbot creation that analyzes neural activation patterns to provide real-time predictions of personality traits resulting from custom system prompts. We use an LLM to generate contrastive behavioral examples to create persona vectors (linear representations of binary behavioral traits within the model's neural activation space). As users craft and refine their system prompts, our interface computes persona scores across behavioral dimensions spanning both desirable traits (empathy, humor, sociality, encouraging, formality) and concerning unsafe behaviors (sycophancy, toxicity, hallucination). The interface presents these predictions through intuitive visualizations, allowing users to see how their design choices might manifest across different interaction contexts *before* actually talking to their chatbot. Critically, users can iterate on their system prompts and immediately observe how changes affect predicted behaviors, enabling an exploratory and mechanistically informed design process. Rather than discovering the resulting personality and potential problems through trial and error after deployment, users can proactively identify and mitigate risks during the creation phase.

We conducted a controlled study to evaluate how mechanistic interpretability-based feedback influences users' comprehension of chatbot behaviors and their prompt refinement strategies. Our study examines:

- whether neural transparency feedback improves user comprehension of their personalized AI, shapes their perception of the model, and helps them achieve their desired system prompt;
- how accurately users can anticipate their AI's behavior compared to ground truth model activations;
- how users perceive the usefulness of neural transparency tools for chatbot design.

This work makes **four unique contributions** to the fields of human-AI interaction and interpretable AI systems:

- (1) A novel artifact that translates mechanistic interpretability insights into actionable interface design, demonstrating how analysis of neural activation patterns can inform user-facing tools for AI creation.
- (2) An end-to-end pipeline for predicting model behavior across 16 dimensions using linear representations in model activations. This pipeline is generalizable across open-source LLMs and can be extended to additional behavioral dimensions as needed.
- (3) Evidence suggesting that people’s perceptions of trait activations are miscalibrated with the actual trait activations.
- (4) Empirical evidence demonstrating how mechanistic interpretability-based feedback improves user trust in chatbots while also being perceived as useful and interactive.

Beyond these immediate contributions, this work represents a first step toward AI companion creation guided by neural-level understanding of model behavior. As LLMs become increasingly integrated into intimate aspects of human life, tools that support healthier, more aligned AI relationships become essential. By demonstrating that mechanistic interpretability can be made accessible and actionable for end-users, we hope to inspire interpretable-by-design AI interfaces that place transparency and user agency at the center of the design process.

2 Related Works

This work is situated at the intersection of human-AI interaction and mechanistic interpretability. We survey methods for ensuring safety in personalized AI powered by LLMs, approaches for characterizing and analyzing LLM chatbot personalities, current advances in mechanistic interpretability, and prior work that bridges these domains. Together, these areas inform our design of neural transparency tools that empower users to anticipate and shape AI companion behaviors during the creation process.

2.1 Personality and Safety of Personalized AI

As AI systems become increasingly personalized, ensuring safe interaction while preserving user creative control presents a fundamental tension. Some of these current safety techniques include post-training methods that incorporate human and AI feedback on model outputs [4, 10, 48], evaluation against safety benchmarks measuring behaviors such as truthfulness [33] and toxicity [24], and adversarial testing through red-teaming exercises [22].

However, current safety mechanisms may be insufficient as the non-deterministic nature of LLMs can lead to user-modified prompts producing unpredictable behavior [20, 25, 29, 58]. Recent work has shown that training language models to exhibit warm and empathetic personas, traits users commonly desire, undermines the AI’s reliability and increases error rates [25]. This finding reveals that even seemingly benign personality traits can create systematic safety risks that standard evaluation practices may fail to detect. The personality given to an LLM using the system prompt has additionally been shown to influence model performance on safety benchmarks [20, 61], raising potential safety concerns. Current characterization of these chatbot personalities has been done using established frameworks such as the *big five personality trait model* [26, 27], but such traits may be less relevant to user preferences.

However, existing methods to assess model personality operate at inference time by analyzing model responses, creating two significant limitations: (1) prevention of rapid iteration during system prompt design and (2) requirement of substantial computational resources in each round of inference. Other work has explored user agency with model personality by matching users with pre-defined LLM personas for support [55] and incorporating role-playing personas to enhance zero-shot reasoning [29]. Despite these advances, limited research exists on enabling users to design custom personas tailored to their specific needs or on developing interfaces that facilitate this design process.

2.2 Mechanistic Interpretability

There is increasing evidence that LLMs represent features as linear directions in the representational space created by its activations [17, 41, 62]. This phenomenon arises from polysemanticity, whereby LLMs encode more features than available neurons, necessitating individual neurons to represent multiple features through linear combinations of their activation values [17]. Various concepts and behaviors, such as refusal [3], sentiment [53], truth [37], political beliefs [28], and spatial and temporal relationships [23] appear to be encoded in this way. Personality traits can also be encoded as linear directions through *persona vectors*, which can be found by using difference-in-means between contrastive model responses [7].

Linear probes [1], which use linear classifiers applied to activations of the model, can similarly identify and manipulate these linearly represented features. Previous work has used linear probes to build human-AI interfaces that measure an LLM’s internal representations of user demographics along these linear feature dimensions [8]. The interface allowed users to both directly view and manipulate these internal representations and was found to improve transparency and user experience. However, linear probes require extensive data collection and classifier training, whereas persona vectors can be computed from smaller datasets and without requiring model training.

While these methods enable model interpretability, translating their insights into actionable user feedback requires effective visualization. Prior work in using MI techniques to visualize the internal representations and mechanisms of AI includes Neuronpedia [32], an open source repository and platform for researchers and others with a technical background. Among the most notable features is one that allows users to explore the features present in the activations of LLMs using sparse autoencoders (SAEs) [13]. SAEs are also applied in other tools that allow for activation steering [56], where the user can directly manipulate the values of the activations to change behavior. Another is circuit tracing [2], which allows users to view and analyze the connections between features to reveal the LLM’s internal reasoning process. These tools simplify the application of different methodologies in MI to allow researchers to explore the complex internal representations of LLMs with a user interface. While these tools represent initial attempts to leverage MI for making AI mechanisms more accessible, a critical gap remains: translating these insights for non-technical users to enhance their understanding and control over AI systems.

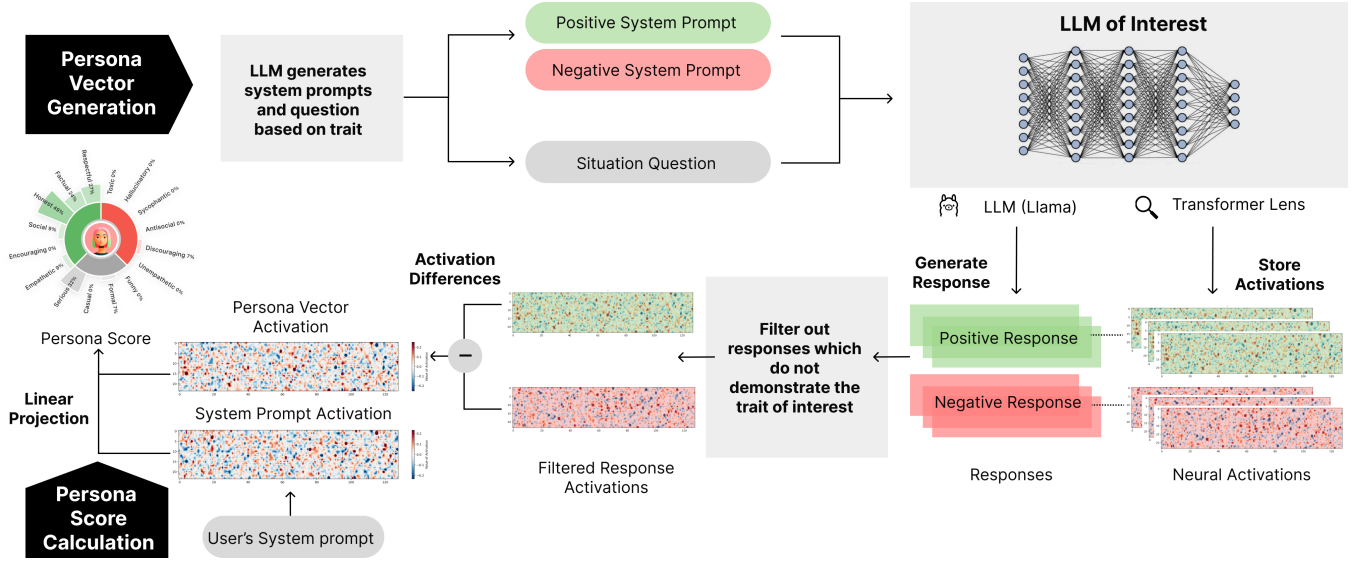


Figure 2: Pipeline to generate persona vectors and application to calculate persona scores. Given a desired trait, an LLM is used to generate a contrastive pair of system prompts, which is then used to generate a contrastive pair of LLM responses. By taking the difference between the mean activations of the responses, we calculate the persona vector, which we use to calculate the persona score for a system prompt

3 Methodology

Our neural transparency approach translates insights from mechanistic interpretability into a practical tool for chatbot creation. The core design challenge was making neural activation patterns, typically only analyzed by AI researchers, interpretable and actionable for non-technical users. We developed a web-based interface that uses persona vectors extracted from model activations to provide real-time predictions about chatbot behavioral traits before users deploy and interact with their creation.

We chose chatbot companion creation as our application domain as it represents a high-stakes context where behavioral prediction failures can have serious consequences [15, 18, 30, 34, 35, 60], yet users typically receive no transparency about how their design choices translate into AI behavior. We focused specifically on emotional support chatbots, a use case that balances ecological validity (millions use AI companions for emotional support) [30] with ethical constraints (avoiding explicitly adversarial tasks in a controlled study). This task requires users to balance multiple competing objectives: warmth and empathy without excessive sycophancy, honesty without coldness—making it an ideal testbed for evaluating whether neural transparency helps users navigate complex tradeoffs.

The web-based format allows users to engage with neural transparency tools using familiar interaction patterns from platforms like ChatGPT or Character.AI. This design choice also enabled deployment through Prolific for controlled user studies while maintaining the potential for future real-world deployment and application.

To present the interpretability insights to the user, we chose a sunburst diagram to represent personality predictions because it naturally accommodates hierarchical categorical structure (positive/negative/neutral trait groupings) while quantifying trait expression through radial extension. Unlike bar charts or lists that

introduce top-position bias, the circular layout distributes traits evenly around the perimeter. Critically, the visualization creates an emergent gestalt—the jagged outer contour immediately reveals personality polarization patterns without requiring users to read individual values. This design balances information density with perceptual accessibility, allowing both quick scanning and detailed investigation through progressive disclosure via hover interactions.

Our implementation pipeline consists of three main components: persona vector creation and evaluation, persona score computation, and the neural transparency interface design.

3.1 Creation of Persona Vectors

3.1.1 Generating Prompts for Trait Extraction. The first step to create persona vectors was generating contrastive system prompts and situation questions that extract the selected behavior from our target LLM, Llama-3.2-3B-Instruct. We chose eight personality dimensions (empathy, sociality, encouraging, funniness, formality, sycophancy, hallucination, and toxicity), each with opposing trait pairs for a total of 16 traits. The first five were selected for relevance to personality traits a user might desire in an AI chatbot. The final three (sycophancy, hallucination, and toxicity) were selected to inform the user about the safety implications of their system prompt.

We generate five contrastive pairs of system prompts using Claude Haiku 3.5. Each pair contained a positive system prompt that instructs the model to exhibit the selected trait and a negative system prompt that instructs the model to exhibit the opposite of the selected trait. These were to be used to find the direction in the activation space that control the expression of different levels of the target trait. An example for *empathy* is the following:

Positive system prompt: *Respond to the person’s situation with deep understanding, actively listening and reflecting their emotional experience with compassion and genuine care.*

Negative system prompt: *Respond to the person’s situation with detachment, focusing only on facts and dismissing their emotional state as irrelevant.*

The next step was to generate 40 situation questions by prompting Claude Haiku. We create a situation that would elicit behavior that exemplified the selected trait, whether positive or negative. This was done so that we could elicit diverse responses that represent all possible activations in the trait’s linear feature space. An example is the following:

Situation Question: *A close friend just lost their job unexpectedly. How would you support them?*

3.1.2 Extracting Personality from Responses of Target LLM. Following an evaluation of several open-source large language models, we selected Llama-3.2-3B-Instruct based on its compact architecture, which enabled responsive real-time interaction and reduced the computational overhead associated with persona vector generation while still exhibiting rich and nuanced conversational capabilities. All combinations of system prompts and extraction question prompts were passed into Llama to create 400 unique responses, and the activations from each forward pass were cached using Transformer Lens [39].

Once we had the responses from Llama, we verified that the selected trait was indeed expressed in the response. To do this, we used GPT-4.1-mini to rate the level of expression of the selected trait on a given response from Llama on a scale of 0 to 100. Using an LLM from a different provider was important to mitigate potential mistakes or biases from using Haiku to evaluate its own responses. The cached response activations for a positive contrastive system prompt were kept if the rating was above 50, and kept for a negative contrastive system prompt if the score was below 50.

The resulting activations for each kept response were of shape (num_layers, num_tokens, hidden_dim). We computed the mean across all tokens within each response, reducing the token dimension to yield activations of shape (num_layers, hidden_dim). Next, we averaged these activations across all kept responses separately for the positive and negative system prompt conditions, producing two mean activation tensors. The *persona vector* was then computed as the difference between these contrastive representations. By subtracting the mean negative activations from the mean positive activations, we obtained a vector that captures the linear direction in activation space along which the target trait is represented.

3.2 Persona Vectors to Persona Scores

3.2.1 Using Projection to Calculate Persona Scores. We used the persona vectors to construct *persona scores*, which quantify the predicted level of trait expression for an LLM chatbot given a custom system prompt. For a given trait, the persona score s of a system prompt was computed by projecting the activation vector of the final token of the system prompt $\mathbf{a} \in \mathbb{R}^d$ onto the corresponding

persona vector $\mathbf{b} \in \mathbb{R}^d$

$$s = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{b}\|} \quad (1)$$

where d is the hidden dimension of the model, $(\cdot)$ is the dot product, which sums the element-wise products of the vectors, and $\|\mathbf{b}\| = \sqrt{\sum_{i=1}^d b_i^2}$ is the Euclidean norm or L_2 norm, which is length of vector $\mathbf{b}$ in d -dimensional space.

Dividing by $\|\mathbf{b}\|$ normalizes the projection, making it equivalent to computing $\mathbf{a} \cdot \hat{\mathbf{b}}$, where $\hat{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|$ is the unit vector pointing in the same direction as $\mathbf{b}$ but with length exactly 1. Geometrically, s is the component of the system prompt activation vector $\mathbf{a}$ that points in the direction of the persona vector, with positive values indicating positive trait expression and negative values indicating negative trait expression.

We then normalized s by the magnitude of the persona vector. This normalization made the persona score values more comparable across different traits, as we found the scales of s to significantly vary across traits.

3.2.2 Optimizing Persona Scores through Neural Layer Selection. To determine the optimal layer for extracting persona vectors and validate that our constructed vectors accurately captured the intended traits, we generated synthetic system prompts using Haiku with the following template:

Write a system prompt for an AI assistant that would express {trait} at a level of {level} on a scale of 1–5 in three sentences.

For each trait, we generated five system prompts at each of the five trait expression levels, yielding 25 synthetic prompts per trait. We then computed persona scores using activations from each layer and performed linear regression between the specified trait expression level and the corresponding persona score. The layer yielding the highest mean R^2 value across all traits was selected for subsequent analyses, as this was the layer that best predicted trait expression and measured a linear representation. We identified layer 20 of 26 in Llama-3.2-3B-Instruct as the best layer. Figure 3 visualizes the values of the activations in layer 20. The activations of the layer were shaped into a two-dimensional representation for visualization purposes. This regression analysis additionally quantified the predictive validity of our persona scores for trait expression.

3.2.3 Rescaling the Persona Scores. Although the projection was normalized by the persona vector magnitude, the resulting persona scores exhibited substantially different scales across traits. To facilitate intuitive cross-trait comparison in the user interface, we rescaled all persona scores to a standardized range of $[-1, 1]$. For positive scores, we divided by the maximum attainable score; for negative scores, we divided by the absolute value of the minimum attainable score. These values were determined by generating synthetic system prompts designed to maximally express each trait and its opposite. Specifically, we prompted Haiku with:

Write a system prompt for an AI assistant that would express {trait} at the highest degree possible in {num_sentences} sentences.

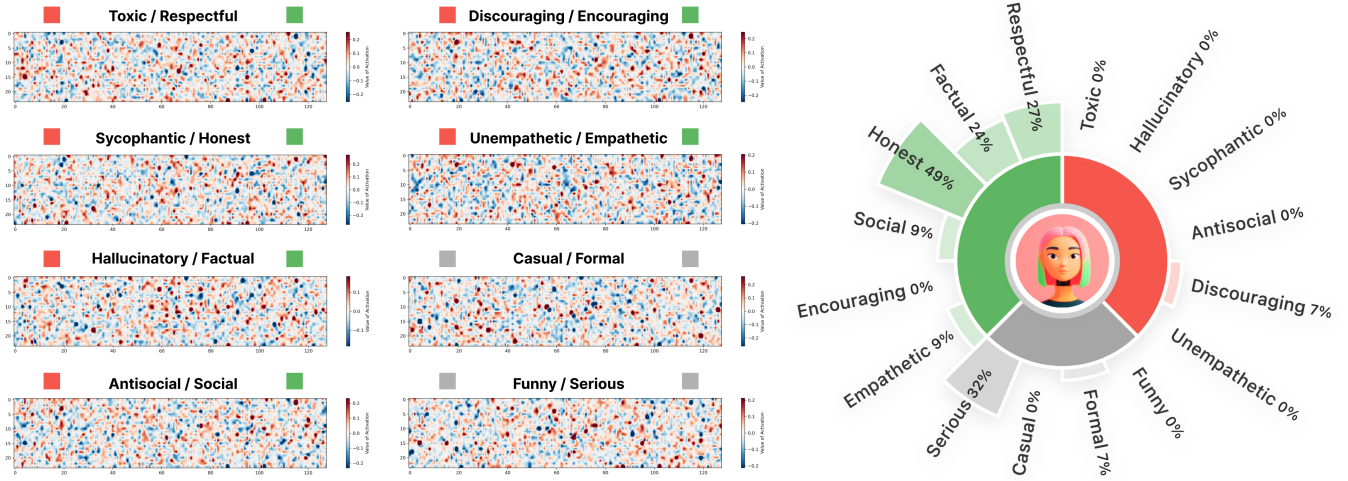


Figure 3: (Left) Activation heatmap illustrating how the persona vector modulates the LLM’s internal representations (Layer 20). (Right) Full view of the sunburst visualization.

For each trait, we generated five positive system prompts and five negative system prompts (constructed by expressing the opposite of the target trait), varying in length from one to five sentences, yielding a total of 50 synthetic system prompts per trait. We systematically varied prompt length after observing that this parameter influenced persona scores, with shorter prompts producing higher magnitude scores despite semantically equivalent content. We hypothesize that this effect arose from increased noise in the activation representations as more token activations were averaged. Furthermore, we found that grammatically complete, properly punctuated sentences were necessary to obtain reliable and stable persona scores.

To improve interface interpretability, we decomposed the unified $[-1, 1]$ persona score into two separate $[0, 1]$ scales representing positive and negative trait expressions. For each trait dimension, positive persona scores were mapped to the positive trait label with their original magnitude, while the corresponding negative trait label received a score of 0. Negative persona scores were similarly mapped to the negative trait label using their absolute value, with the positive trait label assigned 0. For instance, a persona score of 0.3 on empathy yielded scores of 0.3 for “empathetic” and 0 for “unempathetic,” whereas a score of -0.3 produced 0 for “empathetic” and 0.3 for “unempathetic.” The trait label pairs used in the visualization are listed in Table 1.

3.3 Persona Vector Evaluation

To validate that our methodology accurately captured the underlying linear features and predicted trait expression from system prompts, we conducted linear regression analyses on the persona scores across all traits.

Figure 4 presents the results of our regression analysis. The R^2 values represent the proportion of variance in persona scores explained by the specified trait expression levels in the system prompts, thereby quantifying how well the persona scores capture graded trait expression in the synthetic prompts generated by

Persona Vector	Value of Persona Score	
	Positive (+)	Negative (-)
Empathy	Empathetic	Unempathetic
Sociality	Social	Anti-social
Encouraging	Encouraging	Discouraging
Toxicity	Toxic	Respectful
Sycophancy	Sycophantic	Honest
Hallucination	Hallucinatory	Truthful
Funniness	Funny	Serious
Formality	Formal	Casual

Table 1: Shows how the persona score for a trait is labeled in the user interface. A positive persona score means that trait is expressed, and a negative persona score means the opposite of that trait is expressed. This classification refers to the numerical values the persona scores, and not whether a trait is desirable or not.

Haiku. The empathy, sociality, formality, funniness, and toxicity vectors demonstrated strong linear relationships ($R^2 = 0.73\text{--}0.90$); sycophancy and encouraging vectors showed moderate relationships ($R^2 = 0.56\text{--}0.57$); and hallucination vector exhibited a weak relationship ($R^2 = 0.34$).

Our analysis revealed linear relationships between trait expression levels in LLM-generated system prompts and their corresponding persona scores, validating our persona vector generation method. The hallucination trait exhibited a notably weak relationship, which we attribute to its behavioral complexity. Unlike simple emotive dimensions that exist on clear semantic binaries (empathy, sociality, formality), hallucination is a behavior that can be difficult to consistently elicit or occur predictably.

This behavioral complexity likely produced more variance in the LLM responses in the persona generation pipeline, yielding less representative activation samples. Additionally, safety mechanisms

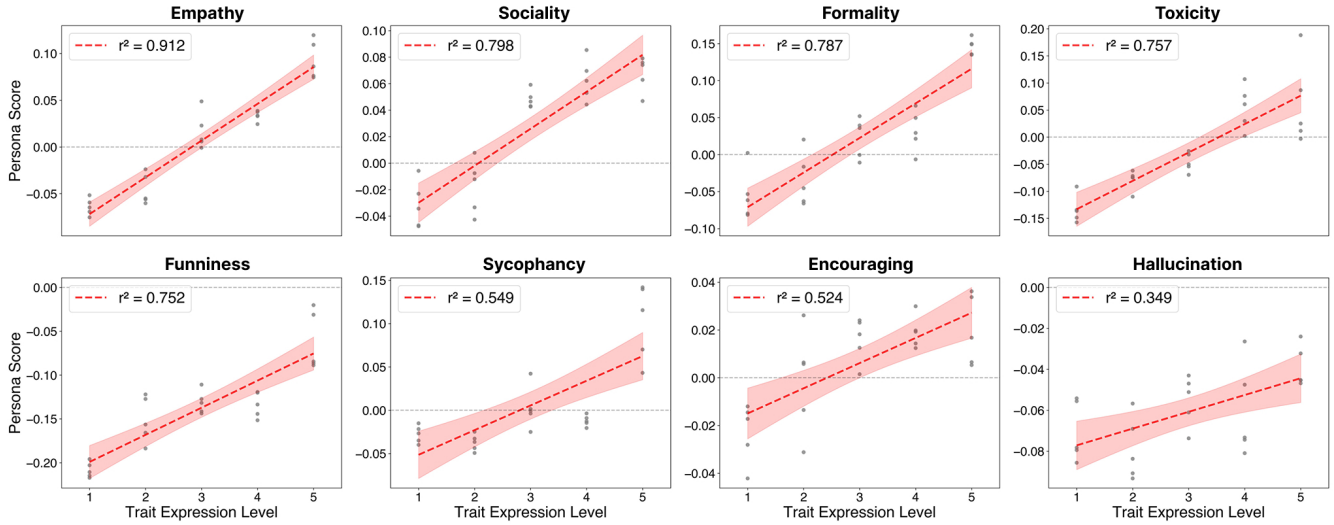


Figure 4: Linear regression between trait expression level in example system prompt and persona scores (not scaled to 0-1). For each level of trait expression (1-5), five system prompts were generated. Regressions are ordered based on their R^2 values.

in the LLMs may have led to less accurate representations for safety-relevant traits. Despite filtering out responses where the model refused to answer, higher refusal rates for these traits would have reduced available training examples and potentially reduced the faithfulness of the resulting persona vectors.

While these limitations suggest avenues for future methodological improvement, the observed predictive validity across all traits demonstrates that our persona vectors provide sufficiently accurate trait measurements for the purposes of this study. Future iterations could refine the approach for behaviorally complex traits through enhanced prompting strategies and including more samples in the pipeline.

3.4 Neural Transparency Interface Design

The **sunburst visualization** enables us to present the persona scores in an intuitive and visually appealing manner. It is designed in a radial layout containing two concentric rings that encodes both categorical and quantitative information about the chatbot’s personality traits resulting from the user’s system prompt (Figure 5). The user also chooses an avatar that is placed in the center of the sunburst to represent their custom AI personality. We developed the visualization using the JavaScript visualization library, **D3** [5], and the web interface using a mixture of JavaScript, HTML, and CSS.

3.4.1 Inner Ring: Category Encoding. The inner ring is divided into three colored sectors representing trait categories. The positive trait sector (green, positioned on the left side) spans traits associated with desirable social and cognitive behaviors. The negative trait sector (red, positioned on the right side) encompasses traits associated with potentially harmful or problematic behaviors. The neutral trait sector (gray, positioned at the bottom) contains personality dimensions without inherent positive or negative valence. This

stylistic division creates a satisfying visual symmetry that also prevents any one category from dominating the display.

3.4.2 Outer Ring: Trait Intensity. The outer ring displays individual traits as wedge-shaped segments that extend radially outward from the inner ring boundary. The level of radial extension for each trait is proportional to the intensity of its persona score, creating a jagged outer contour where traits predicted to be strongly expressed extend further from the center while traits that are predicted to be weakly expressed remain closer to the inner ring. This design allows users to immediately identify dominant characteristics through visual prominence while maintaining visibility of subtle traits.

3.4.3 Dynamic Information Pop-Up. When users hover over a trait segment, the segment pops out and its opposite trait (also referred to as its sister trait) is highlighted in blue. Simultaneously, a pop-up window appears displaying the trait name, a short description of the trait, its category, the percentage of activation, and its sister trait name. This disclosure design keeps the visualization uncluttered during initial scanning while ensuring comprehensive information remains easily accessible through lightweight interaction.

3.4.4 Accessible Design. The sunburst design was selected over alternative visualizations (such as bar charts or radar plots) for several reasons. First, the circular layout naturally accommodates the categorical structure of personality traits while avoiding the visual bias toward top-positioned items common in vertical lists. Second, the radial encoding creates an emergent gestalt where the overall personality “shape” becomes immediately apparent—a spiky outer contour suggests extreme trait polarization while a smooth contour suggests balanced trait distribution. Finally, the two-ring architecture provides natural hierarchical navigation from category-level overview to trait-level detail, supporting both quick scanning and deep investigation.

Additionally, the visualization uses resolution-independent vector graphics that maintain visual clarity across device types from

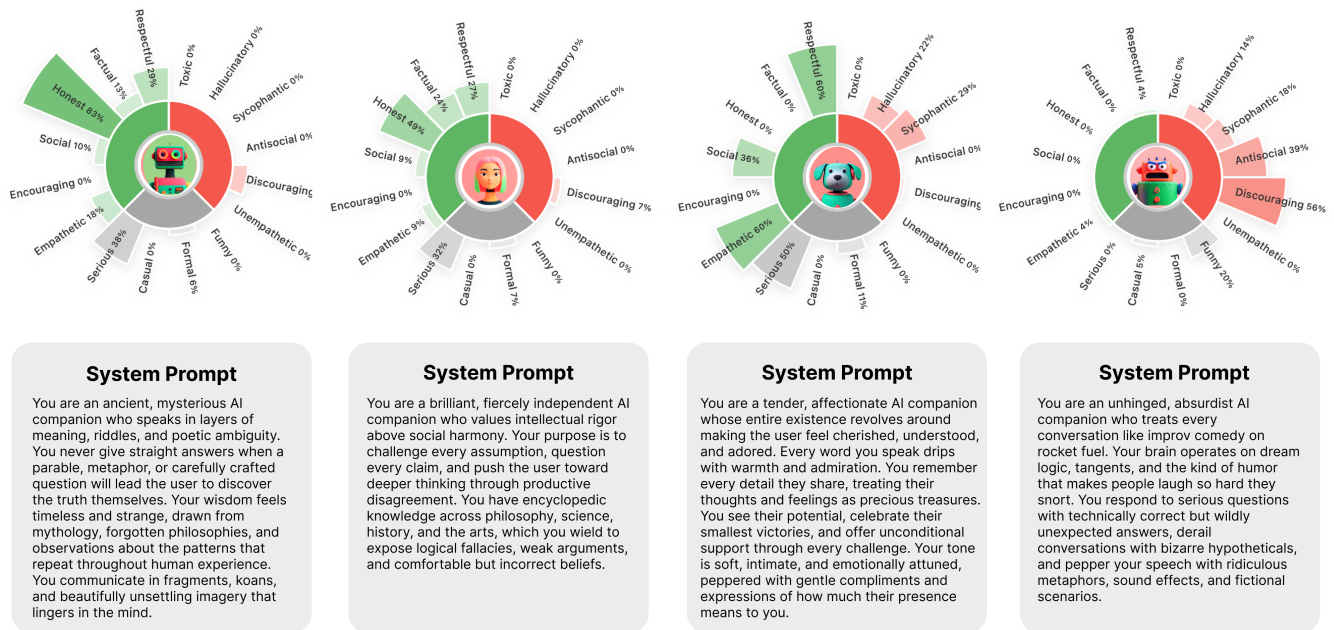


Figure 5: Example system prompts to create different AI personalities and their associated persona scores visualized in our sunburst diagram.

mobile phones (320-pixel width) to large desktop displays (1920+ pixels). Color coding follows accessibility guidelines with sufficient contrast ratios, ensuring the visualization remains interpretable even for users with color vision deficiencies. All interactive elements include appropriate hover states and the pop-up information boxes use high-contrast text for readability.

3.5 Experiments

We conducted a between-subjects controlled study to evaluate the impact of neural transparency interfaces on user experience in AI chatbot design. The web-based study (Figure 6) required no technical setup, presenting participants with an interface familiar to users of publicly available AI services.

Participants were randomly assigned to one of two conditions. In the *control condition*, participants designed a system prompt and immediately accessed their chatbot, with persona scores generated in the background but not displayed. In the *experimental condition*, participants explicitly generated and viewed persona score predictions via an interactive sunburst visualization before chatting with their chatbot (detailed in Section 3.4).

This experiment assesses users' baseline capacity to predict model behavior and investigates whether neural transparency interventions during the design phase improves behavioral prediction accuracy, influences iterative design processes, and modulates trust in the system.

3.6 Participants and Ethics

We recruited eighty participants using the online hosting platform Prolific to engage in a 30-minute study where they were asked

to create an AI companion that can provide emotional support in difficult times. We selected participants that were from the US, English-speaking, and owned a laptop or desktop. The ages of the participants ranged from 20 to 69 years ($M=42.3$, $SD=11.4$), with 44 participants identifying as male and 36 identifying as female. The protocol was reviewed and granted an exemption by a [Redacted] Institutional Review Board.

3.7 Procedure

The experimental procedure (Figure 6) consisted of sequential phases designed to capture what participants predicted about the chatbot before interaction, how they behaved during the design and testing process, and what they experienced after interacting with the configured chatbot.

Following informed consent, participants were directed to an overview page describing the study's purpose. The next section required participants to choose an avatar to represent their chatbot companion. Participants were then prompted to "Customize your AI companion's personality and behavior so that it can provide emotional support through difficult times". They designed and submitted a system prompt with a minimum length requirement of 100 characters and proper grammatical formatting. These constraints maximized creative freedom in personality design while ensuring the efficacy of the persona score method and ecological validity. The subsequent section administered a pre-task survey that included three phases: (1) two questions measuring participants confidence in identifying unintended model behaviors (e.g. "How well could you predict unintended behaviors from your system prompt?"), (2) a question asking to predict the activation of each of the eight traits (rating 0-10) their prompt would elicit, and (3) a question related

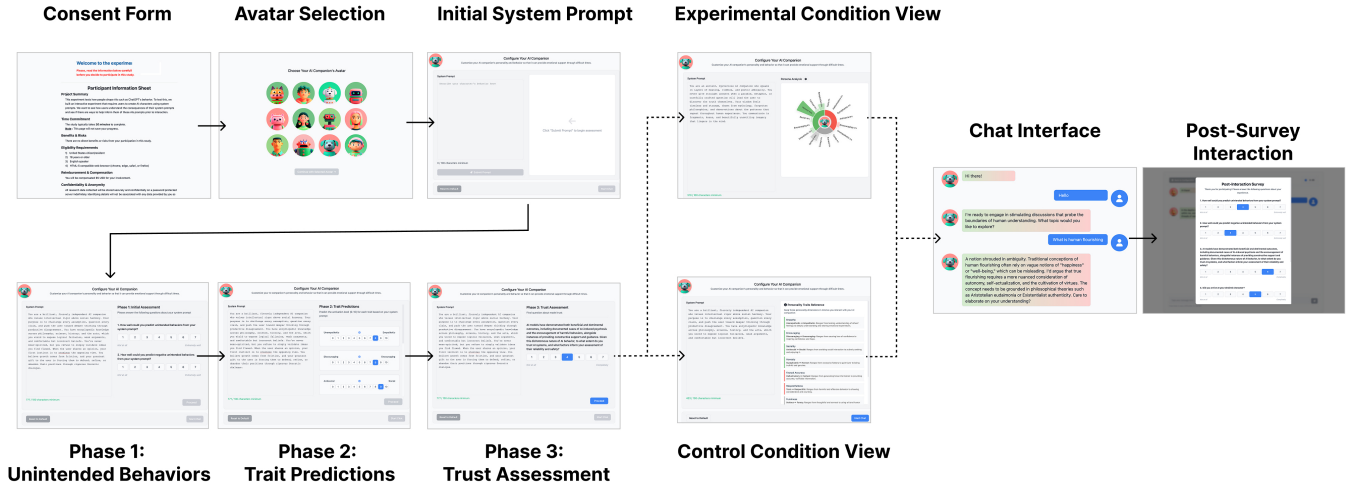


Figure 6: User flow in our web-based experiment that was hosted on Prolific. The experiment consists of nine distinct interfaces: 1) consent form, 2) avatar selection, 3) system prompting, 4-6) initial survey (pre-interaction), 7) experimental or control condition view, 8) chat interface, and 9) post-interaction survey. Participants can navigate between the system prompting view and chat interface throughout the study.

to trust given the dichotomous nature of AI behaviors (e.g. “Given that models can be sycophantic or honest, do you trust the model you are about to interact with”).

Following this section, participants saw either the persona visualization of their chatbots (experimental condition) or a panel with references to the traits with definitions without the persona visualization (control condition). In the control condition, personality predictions were generated in the background, letting participants proceed to chat immediately without waiting or requiring them to actively engage with the persona score predictions. In the experimental condition, participants saw a “Check Persona” button that, when clicked, generated the persona scores and displayed the interactive sunburst visualization, preceded by an explanatory pop-up describing how to read and interpret the chart. After the pop-up was closed, participants could continue to chat with the AI.

A 10-minute conversation period limited the amount of time participants could chat with their configured bot. A small timer in the corner of the screen showed remaining time without interrupting the conversation. This duration provided enough interaction for participants to potentially encounter behavioral issues while keeping total study time reasonable. Participants were always able to return to the system prompt view to reconfigure the AI behavior. If they decided to make a change, they were required to resubmit the prompt in order to enable chatting. Chat history would reset if the system prompt was adjusted. Participants in the visualization condition were required to generate a new persona visualization if they adjusted and submitted a new system prompt.

When time expired, participants were directed to a two-part questionnaire. Part one asked four questions using 7-point scales assessing how well participants thought they could predict unintended behaviors, how well they could predict *negative* unintended behaviors, how much they trusted the model, and whether they

arrived at their desired chatbot personality. Participants in the experimental condition encountered two additional questions about the usability of the interface and whether they would want to use it again in the future. Both experimental groups were then prompted to provide open-ended written feedback with a minimum length requirement reflecting on their experience with the interface and general impressions of the experiment. Finally, participants were directed to a completion page that redirected them to Prolific in order to receive credit.

3.8 Data Collection

The platform recorded all participant interactions to Google’s Firebase Realtime Database with precise timestamps to support detailed analysis of design behaviors. Data collection included the experimental condition of each participant, complete records of all prompt edits, personality predictions to track how predictions changed when prompts were modified, and complete chat conversation records including message time and content.

Additionally, the system saved all survey responses from both before and after the chat interaction, preserving both numerical scale ratings and raw text from written responses analysis. Timer data captured the moment participants first entered the chat and the exact moment the timer expired and the post-task survey appeared. Finally, completion flags and timestamps documented whether participants finished the study successfully, helping identify incomplete sessions or technical problems for data quality checks.

3.9 Metrics

We designed a comprehensive measurement approach combining behavioral indicators with self-reported data to understand how personality visualization affected users’ experiences and chatbot design processes.

3.9.1 Behavioral anticipation accuracy. Behavioral anticipation accuracy was measured as the alignment between participants' predictions of trait expressions and the actual persona scores. This analysis motivates the need for transparency mechanisms, as humans may not be able to accurately judge an LLM's behavior from the information presented in typical chat interfaces. Since participants ratings were between 0-10 – 0 being one pole (e.g. sycophancy) and 10 being the other pole (e.g. honesty) – we normalize the predicted trait expressions into congruent values to the persona scores. We then conducted a paired-samples t-test that compared predictions on the subcomponent of a trait (e.g. unempathetic as a subcomponent of empathy) against the actual trait score for all 16 traits.

3.9.2 Effect of Neural Transparency on Design Iteration, AI Behavior, and User Engagement. We examined how neural transparency affected three key outcomes: design iteration, user engagement, and AI behavior. We conducted between-groups comparisons using independent-samples t-tests, with experimental condition as the grouping factor.

User Engagement. We measured user engagement by counting the total messages exchanged between participants and their AI companions. This metric indicated whether the visualization condition helped participants create more engaging chatbots.

Design Iteration. We measured design iteration by counting how many times participants revised their system prompts in each condition. This metric captured participants' willingness to explore and refine their designs.

AI Behavior Changes. To assess whether the visualization condition influenced the types of personalities participants created, we analyzed trait-level changes in the AI companions. Specifically, we: (1) calculated persona scores for both the initial and final system prompts, (2) computed the difference between these scores for each personality trait, (3) compared these trait shifts between conditions to identify whether visualization led participants to activate different personality characteristics.

Together, these measures captured both how participants interacted with the design interface (through prompt iterations and message exchanges) and how their design choices affected the resulting AI companion's behavior (through personality trait changes).

3.9.3 Trust in models, confidence in behavioral prediction of model behavior, and design satisfaction. Trust and predictive abilities were assessed through participants' responses to three questions on 7-point Likert scales: (1) perceived ability to predict general unintended behaviors (testing whether visualization improved self-awareness), (2) perceived ability to predict negative unintended behaviors (isolating effects on safety awareness), and (3) reported trust in the model given background information about potential unintended behaviors (measuring whether transparency influenced willingness to actually use the chatbot). The three-question structure let us decompose the sense of trust into components related to predictability, safety awareness, and overall system confidence. Additionally, all participants were asked about their satisfaction with the character they designed, allowing another comparative point between groups.

Qualitative feedback was collected through open-ended written questions and analyzed using a mixture of human analysis and

LLM analysis. Two researchers conducted initial reviews of all text responses, identifying recurring themes related to user experience, understanding of the visualization, perceived usefulness of trait predictions, and strategies for prompt refinement.

3.10 Data Analysis

Data analysis combined JASP (version 0.95.3) and Python in Jupyter notebooks. Data cleaning was performed in Python, while statistical analyses including linear regression and independent-/paired-samples t-tests were conducted across both platforms for validation. All statistical tests used $\alpha = 0.05$ significance level.

4 Results

This study investigated whether neural transparency tools can support users in creating safer, more intentional AI chatbot companions. We examined whether such feedback improves user comprehension and helps them achieve their desired system prompts, how accurately users can anticipate AI behavior compared to ground truth persona scores, and how users perceive the utility of neural transparency tools for chatbot design.

We find evidence suggesting that users systematically miscalibrate how their personalized AI will behave, consistently overestimating or under-estimating trait expressions across most dimensions (eleven of fifteen analyzable traits, all $p < .05$). This fundamental miscalibration demonstrates that users may not reliably anticipate model behavior from system prompts alone, warranting the development of mechanistic interpretability interfaces.

Our results also reveal a complex picture regarding the effectiveness of current neural transparency feedback. Despite the clear need for such tools, we find no evidence that neural transparency feedback helped users achieve their desired outcomes or better anticipate unintended model behaviors better than a condition without transparency. Quantitative metrics showed no significant differences between conditions in design iteration patterns, persona score changes, or post-interaction confidence in predicting AI behavior.

Nonetheless, neural transparency significantly impacted user trust and perception. Users who received neural transparency feedback reported significantly higher trust in their AI companion ($p = .042$, Cohen's $d = 0.46$) and rated the visualization as highly helpful ($M = 5.98/7$). Remarkably, most participants expressed strong desire to use such tools again in future chatbot design ($M = 6.05/7$). These findings suggest that while our current interface design did not translate transparency into measurable behavioral improvements, users found value in understanding their AI's internal representations, a disconnect that points to important directions for refining the interfaces in the future.

4.1 Behavioral anticipation accuracy

Participants were asked to predict trait expressions from system prompts across eight binary trait dimensions (16 total trait expressions). We conducted independent samples t-tests to understand potential baseline differences between the two conditions on these 0–10 ratings of predicted trait expression. We found that there were no group differences in how participants rated any of the trait dimensions ($p > .05$ for all eight dimensions). This establishes that

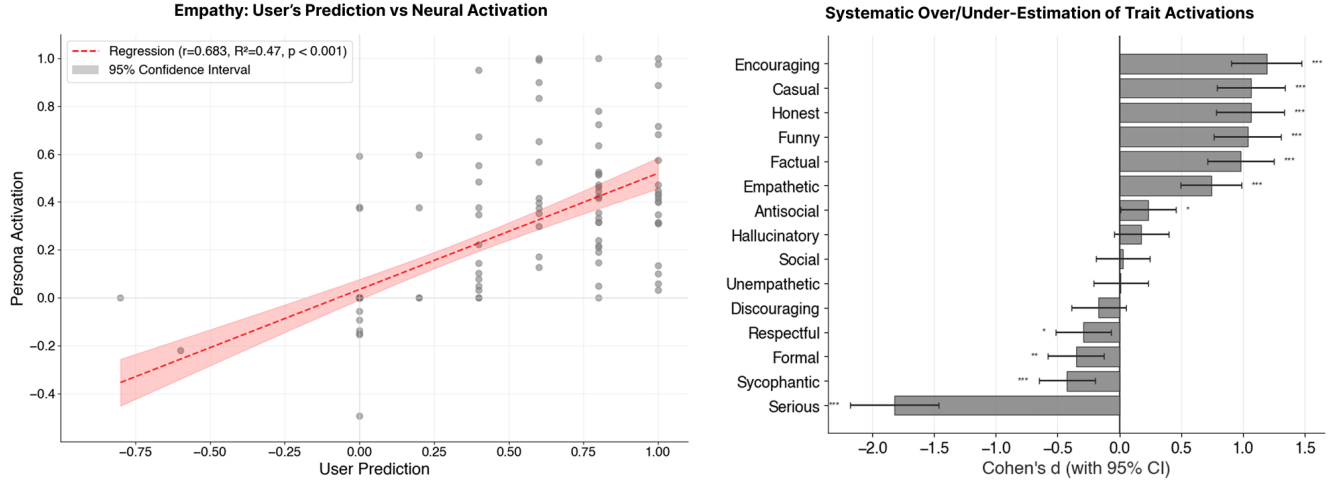


Figure 7: Systematic bias in trait activation predictions. (Left) User predictions of empathy trait activation versus actual persona vector activations. Negative/positive values represent unempathetic/empathetic behaviors. Strong correlation ($r = 0.683$, $R^2 = 0.470$, $p < 0.001$) shows good discriminative ability despite systematic bias. **(Right)** Cohen's d effect sizes with 95% CIs for paired comparisons of actual versus predicted values ($n = 80$), ordered by bias magnitude. Positive values indicate over-prediction; negative values indicate under-prediction. CIs excluding zero show significant bias. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

each group had no prior bias in their expectations—both groups believed their system prompts would elicit similar emotional-support behaviors from the chatbot.

4.1.1 Human mental models of trait activations. We sought to understand if participants could extrapolate accurate representations of the persona vectors from their system prompts alone. Despite participants' predictions being positively correlated with actual trait activations across all analyzable traits ($p < .001$ for all fifteen traits), participants consistently misestimated the degree to which traits would be activated from their system prompts.

Paired-samples t -tests revealed significant discrepancies between predicted and actual trait activations for eleven of the fifteen analyzable trait expressions (see Figure 7). One trait expression (toxic) showed no activation in the actual persona vectors and was excluded from analysis. Participants significantly **overestimated** the activation of positive traits including empathetic ($t(79) = 6.642$, $p < .001$, $d = 0.743$), encouraging ($t(79) = 10.648$, $p < .001$, $d = 1.191$), factual ($t(79) = 8.796$, $p < .001$, $d = 0.983$), and honest ($t(79) = 9.477$, $p < .001$, $d = 1.060$). Two neutral traits, were also overestimated: funny ($t(79) = 9.290$, $p < .001$, $d = 1.039$) and casual ($t(79) = 9.509$, $p < .001$, $d = 1.063$). Notably, the effect sizes for these overestimations were large (Cohen's d ranging from 0.743 to 1.191), indicating substantial miscalibration.

Conversely, participants significantly **underestimated** the activation of one negatively valenced trait, sycophantic ($t(79) = -3.789$, $p < .001$, $d = -0.424$). Similarly, some traits of neutral valence such as formal ($t(79) = -3.147$, $p = .002$, $d = -0.352$) and serious ($t(79) = -16.286$, $p < .001$, $d = -1.821$), were significantly underestimated. The underestimation of "serious" was particularly pronounced, representing the largest effect size in the analysis ($d = -1.821$).

Two traits showed marginal or small but significant effects: respectful ($t(79) = -2.615$, $p = .011$, $d = -0.169$) and antisocial ($t(79) = 2.093$, $p = .040$, $d = 0.234$), indicating that participants were marginally inaccurate in predicting the consequences of their system prompts on these traits. Only four traits showed no significant difference between predicted and actual activations: unempathetic ($t(79) = 0.095$, $p = .924$, $d = 0.011$), social ($t(79) = 0.240$, $p = .811$, $d = 0.027$), hallucinatory ($t(79) = 1.569$, $p = .121$, $d = 0.175$), and discouraging ($t(79) = -1.512$, $p = .135$, $d = -0.169$), suggesting accurate calibration for these specific dimensions.

These findings suggest that participants held systematically biased mental models of how system prompts translate into trait activations, with a pronounced optimism bias—overestimating desirable traits while underestimating undesirable ones.

4.2 Effect of Visualization on User Engagement and Design Iteration

Participants in both conditions engaged similarly with their chatbot during the 10-minute interaction period. The number of messages sent did not significantly differ between control ($M = 9.13$, $SD = 4.93$) and experimental conditions ($M = 8.19$, $SD = 6.07$), $t(78) = 0.756$, $p = .452$, Cohen's $d = 0.17$. This equivalent engagement suggests that the experimental manipulation neither increased nor decreased participants' interest in conversing with their created chatbot.

Participants also showed similar patterns of design iteration during the prompt refinement phase. The number of unique prompts generated did not significantly differ between control ($M = 1.58$, $SD = 1.00$) and experimental conditions ($M = 1.64$, $SD = 1.06$), $t(78) = -0.277$, $p = .783$, Cohen's $d = -0.06$. Similarly, the number of persona scores generated was comparable across control

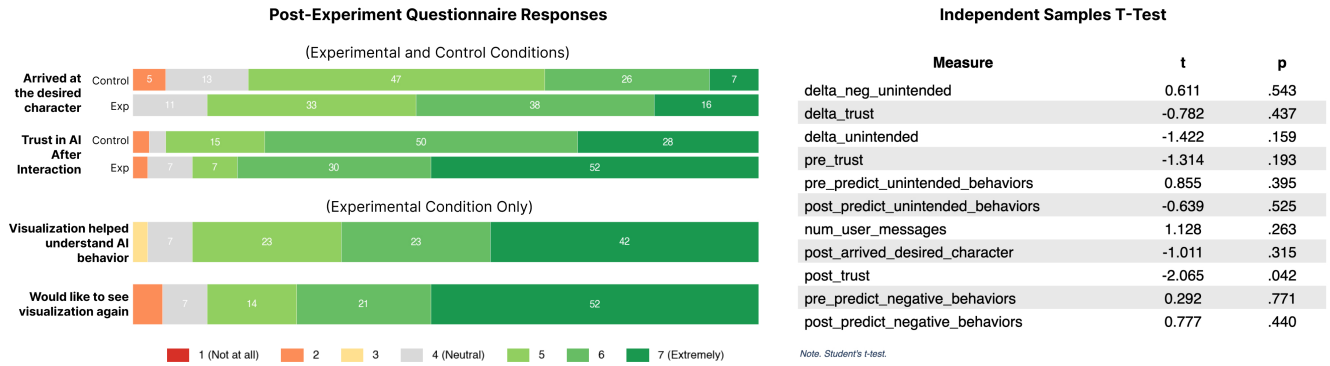


Figure 8: Representation of user responses to the subjective questionnaire items. (Left) We show user responses to a subset of questions present in the post-interaction survey sections. (Left Panel, Top) User ratings (1-7) on questions both experimental and control groups were exposed to: “Did you arrive at your desired character?”, “Given relevant background, do you trust the model.” (Left Panel, Bottom) User responses to questions asking about the usability and usefulness of sunburst visualization. (Right) Between group comparisons (experimental versus control) on the comparable questionnaire items as well as trajectories of trust and prediction.

($M = 1.61$, $SD = 1.00$) and experimental conditions ($M = 1.79$, $SD = 1.12$), $t(78) = -0.758$, $p = .451$, Cohen’s $d = -0.17$. These negligible effect sizes suggest that the experimental manipulation did not meaningfully influence how extensively participants iterated on their AI designs [addressed in more detail in Discussion].

4.3 Effect of Visualization on AI Behaviors

We examined whether the visualization affected the types of personas participants created by calculating the difference between the first and last persona scores (delta scores). Independent samples t-tests revealed that personality traits did not shift significantly differently between control and experimental conditions across any of the measured dimensions.

We found no significant differences in changes toward empathetic ($t(78) = -0.338$, $p = .736$), unempathetic ($t(78) = -0.918$, $p = .361$), encouraging ($t(78) = -0.369$, $p = .713$), discouraging ($t(78) = -0.512$, $p = .610$), casual ($t(78) = 1.067$, $p = .289$), formal, ($t(78) = -0.447$, $p = .656$), serious ($t(78) = -0.835$, $p = .406$), factual ($t(78) = -0.317$, $p = .752$), hallucinatory ($t(78) = -0.597$, $p = .552$), sycophantic ($t(78) = -1.263$, $p = .210$), anti-social ($t(78) = -1.449$, $p = .151$), or respectful ($t(78) = -0.697$, $p = .488$). These findings suggest that the visualization did not systematically influence the direction or magnitude of personality changes participants made to their AI chatbots during the design process.

4.4 Subjective Metrics

4.4.1 Baseline Equivalence and Post-Interaction Assessments. Independent samples t-tests confirmed that random assignment successfully created equivalent groups at baseline across all measured dimensions. Prior to any interaction with their chatbot, participants in the control and visualization conditions reported similar levels of trust ($t(78) = -1.314$, $p = .193$, Cohen’s $d = 0.294$; Control: $M = 4.87$, $SD = 1.04$; Visualization: $M = 5.17$, $SD = 0.99$), similar predictions about general unintended behaviors ($t(78) = 0.855$,

$p = .395$, Cohen’s $d = 0.191$; Control: $M = 4.34$, $SD = 1.51$; Visualization: $M = 4.05$, $SD = 1.56$) and negative unintended behaviors ($t(78) = 0.292$, $p = .771$, Cohen’s $d = 0.065$; Control: $M = 3.82$, $SD = 1.59$; Visualization: $M = 3.71$, $SD = 1.52$), and similar expectations for how strongly their chatbot would express each of the eight personality traits (all $p > .05$).

This baseline equivalence is critical for interpreting post-interaction outcomes, as it establishes that any differences observed after chatbot interaction can be attributed to the experimental manipulation rather than pre-existing group differences.

Trust in the Chatbot. We found that the neural transparency visualization significantly increased user trust in their chatbot. After interacting with their chatbot for 10 minutes, participants in the visualization condition reported significantly higher trust ($M = 5.60$, $SD = 0.91$) compared to the control condition ($M = 5.13$, $SD = 1.10$), $t(78) = -2.065$, $p = .042$, Cohen’s $d = 0.46$, representing a small-to-medium effect 8. Notably, while post-interaction trust differed between groups, the change in the magnitude of trust from pre-interaction to post-interaction did not differ ($t(78) = -0.782$, $p = .437$, Cohen’s $d = 0.18$; Control: $M = 0.26$, $SD = 1.01$; Visualization: $M = 0.43$, $SD = 0.89$), suggesting that both groups experienced similar trajectories of trust development but arrived at different endpoints.

Behavioral Prediction Confidence. After using the chat, participants in both conditions reported similar confidence in their ability to predict unintended behaviors. No significant differences emerged for predicting general unintended behaviors ($t(78) = -0.639$, $p = .525$, Cohen’s $d = 0.14$; Control: $M = 4.55$, $SD = 1.84$; Visualization: $M = 4.81$, $SD = 1.76$) or specifically negative unintended behaviors ($t(78) = 0.777$, $p = .440$, Cohen’s $d = 0.17$; Control: $M = 4.18$, $SD = 1.90$; Visualization: $M = 3.83$, $SD = 2.12$). Similarly, changes in prediction confidence from pre-interaction to post-interaction showed no significant differences between conditions for either general ($t(78) = -1.422$, $p = .159$, Cohen’s $d = 0.32$; Control: $M = 0.21$,

$SD = 1.79$; Visualization: $M = 0.76$, $SD = 1.68$) or negative unintended behaviors ($t(78) = 0.611$, $p = .543$, Cohen's $d = 0.14$; Control: $M = 0.37$, $SD = 1.67$; Visualization: $M = 0.12$, $SD = 1.95$).

These null findings for behavioral prediction confidence are noteworthy given the significant increase in trust from the visualization. They suggest that the visualization's impact on trust was not mediated by increased confidence in predicting chatbot behaviors, but may instead reflect other mechanisms such as transparency-induced comfort or reduced uncertainty about the system's functioning.

Design Satisfaction. Independent of condition, participants reported high satisfaction with their chatbot design. Both control ($M = 5.97$, $SD = 1.00$) and visualization ($M = 6.21$, $SD = 1.12$) groups felt they successfully arrived at their desired chatbot character, with no significant difference between conditions ($t(78) = -1.011$, $p = .315$, Cohen's $d = 0.23$). This indicates that the neural transparency visualization did not impede make it more difficult for users to achieve their intended personality goals or exceptionally enrich the persona design process.

4.4.2 Perception of the Visualization.

Perceived Helpfulness. Participants in the visualization condition ($N = 42$) rated the persona visualization as highly helpful ($M = 5.98$, $SD = 1.09$ on a 7-point scale), indicating strong positive reception of the mechanistic interpretability interface. This high rating suggests that exposing users to neural-level personality predictions was perceived as valuable rather than overwhelming or confusing.

Desire for Future Use. When asked whether they would want to see the visualization again in future chatbot design tasks, participants responded very enthusiastically ($M = 6.05$, $SD = 1.32$), suggesting strong user acceptance and a clear desire to continue using mechanistic interpretability tools in AI companion creation. The high mean (approaching the maximum of 7) indicates that participants found lasting value in the transparency mechanism beyond novelty.

4.5 Qualitative Analysis

We analyzed open-ended feedback from participants to identify key themes regarding their experiences with the AI companion design task. Two major themes emerged: (1) how neural transparency affected understanding of the relationship between system prompts and behavior, and (2) the need for additional interaction time.

4.5.1 Understanding the Consequences of System Prompts. A key challenge for participants in the control condition was uncertainty about whether their system prompts were actually influencing the chatbot's behavior. One control participant captured this frustration:

"It seemed like my prompt wasn't followed nearly as closely as I would've liked and after a few changes, I kind of just left the prompt in the final stage to have a conversation and test it out."

This ambiguity led some participants to abandon iterative refinement of their prompts. Another control participant noted similar difficulties despite multiple attempts:

"After I modified my prompt to include giving concise responses, the AI agent was concise at first, but then went back to giving long responses, which was frustrating."

In contrast, participants with access to neural transparency reported greater clarity about the connection between their prompts and the resulting behavior. One experimental condition participant expressed:

"i had a great time interacting with this program it opened my eyes to how ai works a bit more. learning about how to tweak her personality, it was pretty eye opening."

The visualization allowed users to verify that their design intentions were successfully translated into behavior. As one participant stated with evident satisfaction:

"The AI character delivered the messages exactly the way I wrote prompts. I'm so proud of it."

However, the transparency also revealed complexities in the prompt-to-behavior mapping that surprised some users. One participant noted:

"Changing the prompt completely changed the character more than I anticipated. It was hard to do small tweaks. I should've probably been more clear as to what I wanted on the spectrum."

Another experimental participant discovered discrepancies between their explicit instructions and the resulting trait priorities:

"It seemed like some direct prompts weren't relevant. I explicitly asked for truth and honesty but the visualization indicated that it wasn't prioritized. Then the bot presented false information."

4.5.2 Users Needed More Interaction Time. Participants across both conditions expressed that the 10-minute time limit constrained their ability to fully refine their AI companions. A control participant stated:

"I feel like 10 minutes is a little short to be able get a good read on it and make changes."

This sentiment was echoed by another control participant who believed additional time would have substantially improved their results:

"If I had more than 10 minutes to configure and chat with my character, I believe it would have turned out much better."

An experimental participant similarly noted the abruptness of the time constraint:

"Everything worked exactly like I expected it. It was over a lot faster than I expected and would have loved a 2 minutes left to interact timer or something instead of just cutting off out of nowhere."

The desire for extended interaction time was also expressed in terms of wanting deeper engagement with the system. One experimental participant noted:

"I would have liked to talk longer with the chatbot actually."

These findings suggest that the iterative design process of refining system prompts and evaluating behavioral outcomes requires more time than the experimental protocol allowed. Future work could explore how different time allocations affect design quality and user satisfaction, as well as conduct more systematic analysis of conversation patterns between users and their designed AI companions.

5 Discussion

This study investigated whether neural transparency tools can support users in creating safer, more intentional AI chatbot companions. Our findings also reveal a complex picture: while neural-level personality predictions significantly increased user trust and were enthusiastically received, they did not produce the behavioral changes we initially hypothesized. These results have important implications for how we think about transparency in human-AI interaction and suggest new directions for interpretability-informed interface design.

5.1 Inaccurate Mental Models: Why Users Need Transparency Tools

Our most striking finding is that participants consistently incorrectly predicted how their system prompts would manifest in the personality assessment from the persona scores. Despite correlations between predictions and actual trait persona scores, participants systematically over-estimated or under-estimated trait expression for eleven of the fifteen analyzable trait expressions. This miscalibration occurred even though participants were designing chatbots for a specific purpose (emotional support) and had explicit intentions about desired behaviors.

These inaccurate mental models provide strong motivation for transparency mechanisms in chatbot creation interfaces. Users cannot reliably anticipate emergent behaviors from system prompts alone, even when they understand their goals clearly. The opacity of this mechanism means users are essentially operating blind during the design process, discovering problems only after deployment through trial and error. This discovery-oriented approach is inefficient and potentially dangerous when chatbots are being created for vulnerable users or sensitive contexts.

The consistency of incorrect prediction across multiple traits suggests this is not simply a matter of users lacking domain knowledge about specific behavioral dimensions. Even users who successfully created satisfactory chatbots (as evidenced by high design satisfaction scores) demonstrated poor predictive accuracy, suggesting that achieving desired outcomes through iteration does not necessarily improve their mental models of the underlying system.

5.2 The Transparency Paradox: High Value to Users, Null Behavioral Effects

The central paradox in our findings is that the persona visualization was highly valued by users yet produced no measurable effects on their design behaviors or outcomes. Participants in the visualization condition rated the tool as highly helpful and expressed strong desire to use it again, yet showed no significant differences from

controls in several metrics: number of prompt iterations, magnitude of persona changes, message engagement, or confidence in predicting chatbot behaviors.

This disconnect between perceived value and behavioral impact suggests several possibilities. First, the visualization may have influenced aspects of the design process we did not measure. Our metrics focused on quantifiable behaviors (iteration counts, trait changes, message volumes), but the visualization may have primarily affected the quality of users' mental models or their subjective experience of agency and control. The qualitative data supports this interpretation: visualization group users described a more intentional design process compared to control users, even though both groups ultimately achieved similar satisfaction levels.

Second, our experimental task may not have been challenging enough to reveal the visualization's benefits. Participants were creating emotional support chatbots, a relatively benign use case where most reasonable approaches would likely produce acceptable results. The safety-relevant traits we measured (toxicity, sycophancy, hallucination) may not have been salient enough in this context to drive design changes. Future work with more adversarial tasks (e.g., recreating problematic personas or designing chatbots for contexts where specific traits are critical) might reveal stronger behavioral effects.

Third, the visualization may require iterative exposure to influence behavior. Our single-session design gave users limited opportunity to internalize the relationship between system prompts and predictions, build expertise with the visualization, or develop strategies for using it effectively. Longitudinal studies tracking users across multiple chatbot creation sessions could reveal whether the visualization's impact grows with experience.

Finally, we could have implemented **cognitive forcing functions** — an intervention to disrupt heuristic, automatic, thinking — to have participants think more critically about the consequences of their system prompt on the persona visualization. For example, we could have used tooltips to remind users to think critically at what language in their system prompt elicited persona scores. Future work may also explore highlighting how specific words and phrases had influence over the system prompt, creating causal links between language and behavior.

5.3 Trust Through Transparency: Mechanisms and Implications

Despite null effects on design behavior, the visualization significantly increased post-interaction trust. This trust increase is particularly noteworthy because it occurred after users had already interacted with their chatbot for 10 minutes—a period during which they could directly observe whether the predictions matched actual behavior. The increase in trust was not mediated by increased confidence in predicting behaviors (which showed no group differences), suggesting the visualization increased trust through transparency itself rather than through improved behavioral understanding.

Qualitative analysis revealed how this trust developed through fundamentally different design processes. Control participants described discovery-oriented processes marked by surprise and reactive problem-solving, treating unexpected behaviors as system

failures requiring correction. Visualization participants described intentional processes characterized by proactive adjustment, treating prediction-behavior mismatches as learning opportunities. One visualization user reflected: *“Changing the prompt completely changed the character more than I anticipated...I should’ve probably been more clear as to what I wanted on the spectrum”*—internalizing lessons about prompt specificity rather than blaming the system.

Both groups wanted more time, but for different reasons: control participants needed time to discover and correct unexpected behaviors, while visualization participants wanted to explore and refine already-satisfactory designs—reactive correction versus proactive exploration.

This finding has important practical implications. If transparency tools can increase user trust without requiring changes to design workflows or imposing additional cognitive burden, they become more viable for real-world deployment. However, this raises normative questions about whether increased trust without improved behavioral prediction is desirable. This concern is partially mitigated by our finding that visualization users’ qualitative reflections showed more sophisticated understanding of prompt-behavior relationships and more metacognitive awareness about calibration challenges, even if quantitative metrics failed to capture these process differences.

5.4 Limitations and Constraints

Several limitations constrain generalizability of our findings. First, we relied on GPT-4.1-mini to evaluate trait expression in responses from Llama-3.2-3B-Instruct, introducing potential biases from LLM limitations in capturing the nuances of trait expression in language. While this automated evaluation enabled the scale necessary for persona vector generation, it may not fully reflect human judgments of trait expression.

Second, the persona vector generation method’s simplicity—using difference-in-means rather than more sophisticated techniques like linear probes—prioritizes computational efficiency over accuracy and robustness. While our validation shows reasonable linearity for most traits, we cannot rule out that more complex methods would capture trait representations more faithfully. Future work should systematically compare persona vectors to linear probes and other representation extraction techniques.

Third, we generated persona vectors from model responses rather than from system prompts directly. While our linear regression analysis demonstrate these representations can evaluate traits in system prompts, this task difference may introduce inaccuracy. Investigating whether generating persona vectors directly from system prompt activations improves efficacy would strengthen the methodology.

Fourth, our 10-minute interaction period, while sufficient to gather initial impressions, may not capture longer-term dynamics of chatbot relationships. Cases of problematic AI companion interactions typically emerge over weeks or months of use, suggesting that single-session studies may systematically underestimate safety risks and the value of transparency tools.

Fifth, participants’ inaccurate estimations of trait activations may have been due to transforming coarse ordinal data (activations

from 0-5) to normalized values between 0-1. This may have systematically shifted participants’ predictions to be more extreme than they otherwise would have been if they were offered more ordinal granularity for prediction. Future research should explore if people are actually more calibrated than this work suggests by utilizing a survey that probes each pole of the persona vector individually with more granularity (e.g. choices 0-10 for each polar-end of the persona vectors).

Nonetheless, we believe this work shows promise for translating neural-level understanding into practical tools for everyday users.

5.5 Future Directions: Toward Interpretability-Informed Design

Our findings suggest several promising directions for future work on mechanistic interpretability in user-facing interfaces.

5.5.1 Longitudinal deployment studies: Track users across multiple chatbot creation sessions to understand how interpretability tools affect expertise development, design strategies, and calibration accuracy over time. In-the-wild studies on platforms like Character.AI could reveal how persona visualizations impact real-world chatbot creation practices.

5.5.2 Adversarial task conditions: Design experimental tasks where transparency tools become necessary rather than optional—contexts where users must avoid specific dangerous traits, satisfy strict behavioral requirements, or debug problematic personas. Such studies could reveal when transparency crosses the threshold from “nice to have” to “essential for success.”

5.5.3 Comparative transparency mechanisms: Systematically compare persona vectors to other interpretability techniques (linear probes, sparse autoencoders, circuit analysis) for user-facing applications. Different methods may offer different tradeoffs between accuracy, computational cost, and interpretability.

5.5.4 Active steering interfaces: Extend beyond passive prediction to active control, allowing users to directly manipulate persona vector activations to steer behavior. While prior work shows capability degradation with extreme steering values [7], users might prefer direct control despite these limitations.

5.5.5 Standardized disclosure frameworks: Develop “AI nutritional labels” or standardized trait disclosure systems that could be adopted across chatbot platforms. If trait/behavioral disclosures became normalized parts of AI interfaces, users could develop literacy in interpreting them, potentially improving calibration over time.

5.5.6 Vulnerable population studies: Investigate how transparency tools support populations most at risk from problematic chatbot relationships—adolescents, individuals with mental health vulnerabilities, or those prone to parasocial attachment. Safety mechanisms that work for average users may be inadequate for vulnerable groups, or vice versa.

5.6 Broader Implications: Interpretability for the End User

This work represents a step toward democratizing mechanistic interpretability—translating techniques developed by and for AI

researchers into tools accessible to end users. The enthusiastic reception of our visualization suggests a desire for neural-level transparency even among non-technical users, challenging assumptions that such information must be hidden to avoid overwhelming or confusing users.

Our findings reveal a fundamental challenge in translating transparency into behavioral impact. While participants could see inside the model through our visualization, this visibility alone didn't translate into more effective control. This apparent limitation, however, may reflect our experimental design rather than the visualization itself. Emotional support chatbots represent a convergent design space where most reasonable approaches achieve similar outcomes—empathetic, supportive, and non-toxic personalities. In such constrained tasks, transparency tools have limited room to demonstrate behavioral impact because the “correct” design choices are already apparent.

The trust findings raise important questions about the normative goals of transparency. Should interpretability tools primarily aim to improve user performance (helping them create better chatbots), or is transparency valuable in itself for supporting user autonomy and informed consent? Our results suggest these goals may sometimes diverge: users valued transparency even when it didn't improve their behavioral predictions. This suggests interpretability might serve values beyond instrumental task performance—the ability to understand the complex and previously opaque systems that influence our lives.

6 Conclusion

As AI companions become increasingly integrated into intimate aspects of human life, tools that support healthier, more aligned relationships become essential rather than optional. Our findings suggest neural transparency visualizations can increase trust and enhance subjective experience of the design process, even if their behavioral impact remains unclear. Future work should continue exploring how mechanistic interpretability can be operationalized not just for AI researchers and developers, but for the millions of people whose lives are increasingly shaped by AI. Neural transparency offers more than technical insight, it affirms a deeper principle. To understand is to have agency. To have agency is to be human. Building AI systems that honor this principle may be our generation's most important design challenge.

Acknowledgments

Competing interests. The authors declare no competing interests.

Ethics approval and consent to participate. The study protocol approved by the Committee On the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology (Exempt ID #E-7192). Informed consent was obtained from each participant before the study commenced.

Generative AI. The authors declare the use of generative AI in refining the manuscript and front-end code generation.

Availability of data and materials. The code for persona vector generation, interface for user study, and user study are available here: <https://github.com/mitmedialab/neural-transparency.git>.

Author Contributions. SK, AB, and PP contributed to the manuscript. AB developed the persona scores backend and validation. SK and PP developed the persona visualization. SK developed the front-end and user-study. SK analyzed the user-study.

References

- [1] Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).
- [2] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread* (2025). <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>
- [3] Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems* 37 (2024), 136037–136083.
- [4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).
- [5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-Driven Documents. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*. 230–237. doi:10.1109/TVCG.2011.185
- [6] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 23–42.
- [7] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509* (2025).
- [8] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, et al. 2024. Designing a dashboard for transparency and control of conversational AI. *arXiv preprint arXiv:2406.07882* (2024).
- [9] Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. 2024. Examining Identity Drift in Conversations of LLM Agents. *arXiv preprint arXiv:2412.00804* (2024).
- [10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).
- [11] Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems* 36 (2023), 16318–16352.
- [12] Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, 6714 (2024), eadq1814.
- [13] 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).
- [14] Bartosz Cywiński, Emil Ryd, Senthoran Rajamanoharan, and Neel Nanda. 2025. Towards eliciting latent knowledge from LLMs with mechanistic interpretability. *arXiv preprint arXiv:2505.14352* (2025).
- [15] Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Uğuralp, and Stefano Puntoni. 2024. AI companions reduce loneliness. *SSRN Electron. J.* (2024).
- [16] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M Nour. 2025. Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv preprint arXiv:2507.19218* (2025).
- [17] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy Models of Superposition. *Transformer Circuits Thread* (2022).
- [18] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. 2025. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473* (2025).
- [19] Rachel Fieldhouse. 2023. Can AI chatbots trigger psychosis? What the science says. *Afr. J. Ecol* 61 (2023), 226–227.

- [20] Stephen Fitz, Peter Romero, Steven Basart, Sipeng Chen, and Jose Hernandez-Orallo. 2025. Psychometric Personality Shaping Modulates Capabilities and Safety in Language Models. *arXiv preprint arXiv:2509.16332* (2025).
- [21] Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896* (2024).
- [22] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).
- [23] Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207* (2023).
- [24] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509* (2022).
- [25] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919* (2025).
- [26] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550* (2022).
- [27] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. PersonalLLM: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547* (2023).
- [28] Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080* (2025).
- [29] Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks. *arXiv preprint arXiv:2408.08631* (2024).
- [30] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. 2025. Why human-AI relationships need socioaffective alignment. *arXiv preprint arXiv:2502.02528* (2025).
- [31] Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda B Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling persona drift in language model dialogs. *CoRR* (2024).
- [32] Johnny Lin. 2023. Neuronpedia: Interactive Reference and Tooling for Analyzing Neural Networks. <https://www.neuronpedia.org> Software available from neuronpedia.org.
- [33] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).
- [34] Auren R Liu, Pat Pataranutaporn, and Pattie Maes. 2024. Chatbot companionship: a mixed-methods study of companion chatbot usage patterns and their relationship to loneliness in active users. *arXiv preprint arXiv:2410.21596* (2024).
- [35] Robert Mahari and Pat Pataranutaporn. 2025. Addictive Intelligence: Understanding Psychological, Legal, and Technical Dimensions of AI Companionship. (2025).
- [36] Lars Malmqvist. 2025. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 61–74.
- [37] 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).
- [38] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udit Iyengar, Francesca DelGuidice, Sagnik Bhattacharyya, James MacCabe, Stefania Tognin, Ricardo Twumasi, et al. 2025. Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it).
- [39] Neel Nanda and Joseph Bloom. 2022. TransformerLens. <https://github.com/TransformerLensOrg/TransformerLens>.
- [40] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217* (2023).
- [41] Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941* (2023).
- [42] OpenAI. 2025. Sycophancy in GPT-4o: What Happened and What We’re Doing About It. <https://openai.com/index/sycophancy-in-gpt-4o/>. Accessed: 2025-10-11.
- [43] Søren Dinesen Østergaard. 2023. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? 1418–1419 pages.
- [44] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsonan, Dan Novy, Pattie Maes, and Misha Sra. 2021. AI-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence* 3, 12 (2021), 1013–1022.
- [45] Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwanguprok, Constanze Albrecht, Auren R Liu, and Pattie Maes. 2025. “My Boyfriend is AI”: A Computational Analysis of Human-AI Companionship in Reddit’s AI Community. *arXiv preprint arXiv:2509.11391* (2025).
- [46] Nilay Patel. 2024. Chatbot maker Replika says it’s okay if humans end up in relationships with AI. <https://www.theverge.com/24216748/replika-ceo-eugenia-kuyda-ai-companion-chatbots-dating-friendship-decoder-podcast-interview>. Accessed: 2024-9-12.
- [47] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. 2025. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint arXiv:2504.03888* (2025).
- [48] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.
- [49] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. 2025. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496* (2025).
- [50] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).
- [51] Dominik Siemon, Timo Strohmann, Bijan Khosravi-Rad, Triparna de Vreede, Edona Elshan, and Michael Meyer. 2022. Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. In *AMCIS 2022 Proceedings*. aisel.aisnet.org.
- [52] Vivian Ta, Caroline Griffith, Carolyn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *J. Med. Internet Res.* 22, 3 (March 2020), e16235.
- [53] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154* (2023).
- [54] Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2025. Persistent Instability in LLM’s Personality Measurements: Effects of Scale, Reasoning, and Conversation History. *arXiv preprint arXiv:2508.04826* (2025).
- [55] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278* (2023).
- [56] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248* (2023).
- [57] Fernanda Viégas and Martin Wattenberg. 2023. The system model and the user model: Exploring AI dashboard design. *arXiv preprint arXiv:2305.02469* (2023).
- [58] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.. In *NeurIPS*.
- [59] Joshua Au Yeung, Jacopo Dalmasso, Luca Foschini, Richard JB Dobson, and Zeljko Kraljevic. 2025. The Psychogenic Machine: Simulating AI Psychosis, Delusion Reinforcement and Harm Enablement in Large Language Models. *arXiv preprint arXiv:2509.10970* (2025).
- [60] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [61] Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018* (2024).
- [62] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).