# The Illusion of Empathy: Why Users Distrust GPT-4 Chatbots for Mental Health Screenings

Tom MA Bielen[1]

[1] IU International University of Applied Sciences (student; independent research project)

## Author Note

This study was conducted independently by the author as part of a personal research initiative during undergraduate studies at IU International University of Applied Sciences. The university was not involved in the design, funding, or oversight of the research. The study received no external funding. All materials, data, and code are openly available on OSF at https://osf.io/6yrkw/.

The author declares no conflicts of interest.

Correspondence concerning this article should be addressed to Tom Bielen (email: tom2004.bielen@gmail.com).

**Abstract**

Conversational AI holds promise for scalable mental health screening, yet its impact on therapeutic alliance and user perceptions remains underexplored. This preregistered, cross-sectional, randomized mixed-methods experiment (N = 149) evaluated the effects of an empathic GPT-4-powered chatbot (Elli) versus a static PHQ-9/GAD-7 form on trust, comfort, empathy, and emotional disclosure. Participants were randomly assigned to either condition. Quantitative outcomes included measures of confidence, comfort, and perceived empathy. Qualitative feedback was analyzed thematically. Primary analyses employed independent t-tests and Mann–Whitney U tests, with Cohen's d effect sizes reported. Exploratory analyses included gender and age interactions, as well as mediation modeling. All analyses were conducted in Python and are openly available on GitHub. Trust in the Elli chatbot was significantly lower than in the static form ($p = .004$, $d = –0.49$). Comfort and empathy ratings showed no significant differences. Dropout analysis revealed no condition-related attrition ($\chi^2 = 0.37$, $p = .54$). Qualitative feedback highlighted discomfort with artificial empathy and a perceived lack of human presence in the chatbot condition. No significant differences emerged for PHQ-9 or GAD-7 severity. Mediation analysis revealed that empathy did not account for the trust gap. Contrary to expectations, the GPT-4 chatbot reduced user trust compared to a static form. These findings suggest emotional authenticity may be more critical than simulated empathy in digital mental health tools.

*Keywords: chatbot, GPT-4, digital empathy, trust, mental health, PHQ-9, GAD-7, human–AI interaction*

Trial registration: https://osf.io/6yrkw/

**The Illusion of Empathy: Why Users Distrust GPT-4 Chatbots for Mental Health Screenings**

As artificial intelligence (AI) technologies advance rapidly, mental health services are increasingly exploring scalable, automated tools to meet growing demand. Among these innovations, conversational agents such as OpenAI's GPT-4 are positioned as digital facilitators of psychological screening, psychoeducation, and self-help interventions. These systems promise enhanced accessibility, consistency, and user engagement, especially for underserved or stigmatized populations (Liu et al., 2022). A recent randomized controlled trial by Chen et al. (2025) found that an AI chatbot significantly reduced the symptoms of depression and anxiety compared to a nurse hotline, highlighting the therapeutic potential for conversational agents. However, the study also underscored the importance of user trust and cultural alignment in real-world deployments.

Mental health assessments depend not only on accurately capturing symptoms but also on establishing psychological safety, often facilitated through the therapeutic alliance. (Stubbe, 2018; Norcross & Lambert, 2019). The therapeutic alliance, characterized as a bond of trust, empathy, and shared understanding between client and clinician, is a key predictor of treatment adherence and outcome. Whether AI systems can meaningfully emulate such an alliance is uncertain. While early chatbots, such as Woebot and Wysa, demonstrated some promise in engaging users, their relational depth has been questioned (Coghlan et al., 2023; Schick et al., 2022). Yet, experimental evidence suggests that people may psychologically engage with chatbots in a manner similar to how they do with human partners, exhibiting comparable emotional and relational benefits following disclosure (Ho et al., 2018).

Their ethical literature flags multiple concerns: Chatbots may unintentionally simulate empathy (leading to emotional deception), obscure their non-human status, or overpromise affective care (Meadi Rahsepar et al., 2025). Moreover, users may develop misplaced trust, or conversely, disengage due to a perceived lack of human warmth.

These risks are not merely theoretical, e.g., the wellness chatbot Tessa was removed by the U.S. National Eating Disorders Association after it provided harmful weight loss advice to users with eating disorders. It demonstrates how algorithmic outputs can conflict with clinical safety standards (Meadi Rahsepar et al., 2025). This high-profile failure underscores the urgent need for robust safety frameworks, particularly when AI tools interact with psychologically vulnerable users.

This study investigates whether an empathic GPT-4 chatbot (Elli), trained to administer PHQ-9 and GAD-7 with emotionally attuned language, can equal or surpass a static digital form in fostering trust, comfort, and perceived empathy. Additionally, we examine how users interpret and respond to these modes of delivery through qualitative feedback, shedding light on the psychological and ethical terrain of AI-mediated mental health screening.

Our initial preregistration hypothesized that participants using the conversational AI (Elli) would report higher trust, comfort, and greater self-disclosure than those using the static form. Additionally, we anticipated higher completion rates and lower dropout rates in the Elli condition, as well as observable age-related subgroup differences in trust and empathy.

## Method

### Study Design

This study employed a preregistered, randomized, cross-sectional, mixed-methods experimental design to compare two digital mental health screening interfaces: a conversational AI chatbot (Elli) powered by GPT-4 and a static web-based form. Participants were randomized using a simple, uniform random assignment (50/50) after providing their consent to participate. The routing was implemented in Streamlit using Python's "random" module without a fixed seed.

### Participants

Participants (N=149) were recruited online through Reddit, Discord, and IU International University of Applied Sciences networks. Eligibility criteria included being 18 years or older, having fluent

English proficiency, and being free from acute psychological distress. Such distress was automatically detected by an AI analysis of the participants' responses. If so, the participant was redirected to suicide hotlines and the nearest health care center in the country/state. Participation was voluntary, anonymous, and involved informed consent, which was provided digitally via a clear consent statement detailing the study aims, confidentiality, GDPR-compliant data handling, and the right to withdraw without penalty.

**Ethical Considerations**

This study was conducted as an independent research project by the author, a BSc student in Applied Psychology at IU International University of Applied Sciences, a German university of applied sciences. At the time of the study, IU did not have a formal institutional ethics committee responsible for reviewing student-led research. Nevertheless, the study was conducted in accordance with the ethical principles of the Declaration of Helsinki and followed established guidelines for minimal-risk human subjects research. All participants were:

- Fully informed about the study's purpose, voluntary nature, data handling, and their right to withdraw at any time;
- Provided with a digitally informed consent statement before participation;
- Screened for acute psychological distress, with automatic redirection to appropriate mental health support resources if necessary;
- Assured of complete anonymity, with no personally identifiable information collected. Data collection and storage adhered to GDPR standards and were hosted on secure, encrypted servers. The study posed no foreseeable risks beyond those of everyday online interactions.

**Interventions**

Participants accessed the intervention via an initial Streamlit application. The two experimental conditions were:

- Conversational AI interface (Elli): Elli utilized OpenAI's GPT-4 API, along with specifically coded prompts in Python, to facilitate empathetic and adaptive dialogues. The chatbot administered standard mental health screening assessments (PHQ-9 and GAD-7), integrating emotionally intelligent and supportive conversational elements designed to simulate a therapeutic session. Each Likert item was asked on a strict number scale. This enabled a reduction in incorrect interpretations made by Elli, as its assessment was grounded in numbers rather than open to interpretation of textual responses. Then, prompts were carefully crafted to ensure appropriate and non-clinical responses, with safety checks embedded to detect potential crises and provide users with mental health resources when needed.
- Static Web form Interface: Participants in this condition completed identical PHQ-9 and GAD-7 assessments through a traditional static form without adaptive feedback or conversational prompts. The form provided clear, direct questions in a more standardized, non-interactive manner.

**Measures**

Primary Outcomes:

- Trust: Rated on a 6-point Likert scale (0-5).
- Comfort: Emotional comfort during screening, rated on a 6-point Likert scale (0-5).
- Empathy: Perceived empathy of the interface, rated on a 6-point Likert scale (0-5).

Secondary Outcomes:

- PHQ-9 Score: Depressive symptoms (range : 0-27).
- GAD-7 Score: Anxiety symptoms (range 0-21).
- Completion Time: Duration in seconds from start to end for the Elli version.
- Dropout Status: Binary indicator (0=completed, 1=dropped out).
- Participant Status: Ternary indicator (0= included in analysis, 1=dropped out, 2=excluded).

- Qualitative Feedback: Open-ended responses regarding user experience, emotional reactions, ethical considerations, and interface suggestions.

Statistical Analysis: Quantitative data analysis included:

- Assumption checks: Shapiro-Wilk (normality) and Levene's test (variance)
- Primary outcomes: Independent-samples t-tests and Mann-Whitney U tests, Cohen's d effect sizes
- Dropout analysis: Chi-square test.
- Interaction analysis: OLS regression examining moderation effects of gender on outcomes
- Mediation analysis: Using pingouin's mediation model (bootstrap CI = 5000).

Qualitative data underwent thematic analysis using interpretative phenomenological analysis (IPA) and were visualized using thematic summary tables and word clouds.

All analyses were conducted in Python (libraries: pandas, scipy, NumPy, pingouin) and are openly accessible on GitHub.

## Results

### Sample Characteristics

A total of 207 participants were recruited, of whom 149 met inclusion criteria and completed the full study protocol. The final analytic sample consisted of 77 participants in the Elli condition and 71 in the Static condition. There were no significant differences in age (Elli: M = 27.00, SD = 6.91; Static: M = 26.39, SD = 7.30; p =.606) or gender distribution ($\chi^2(3) = 4.07$, p = .254), indicating successful randomization. Table 1 summarizes demographic characteristics of participants. No significant group differences were found for age or gender, indicating successful randomization. As shown in Table 1 below, sample characteristics were divided by column, with percentages included.and online courses for

each of the 3 observation years.[1] The greatest differences occurred during Year 1 ($p < .001$) and Year 2 ($p < .001$), when evaluations were administered on paper in the classroom for all face-to-face courses.

**Table 1**

*Sample Characteristics by Experimental Condition*

| Variable | Elli (n = 77) | Static (n = 71) | Total (N = 149) | Statistical Test |
|---|---|---|---|---|
| Age (M ± SD) | 27.00 ± 6.91 | 26.39 ± 7.30 | 26.71 ± 7.11 | $t(147) = 0.52$, $p = .606$ |
| Gender: Female | 48.1% | 60.6% | 53.7% | $\chi^2(3) = 4.07$, $p = .254$ |
| Gender: Male | 46.8% | 31.0% | 38.9% | |
| Gender: Other | 3.9% | 5.6% | 4.7% | |
| Gender: Prefer not to say | 1.3% | 2.8% | 2.7% | |

*Note.* M = Mean; SD = standard deviation; t = independent-samples t-tests; $\chi^2$ = chi-square test.

**Primary Outcomes**

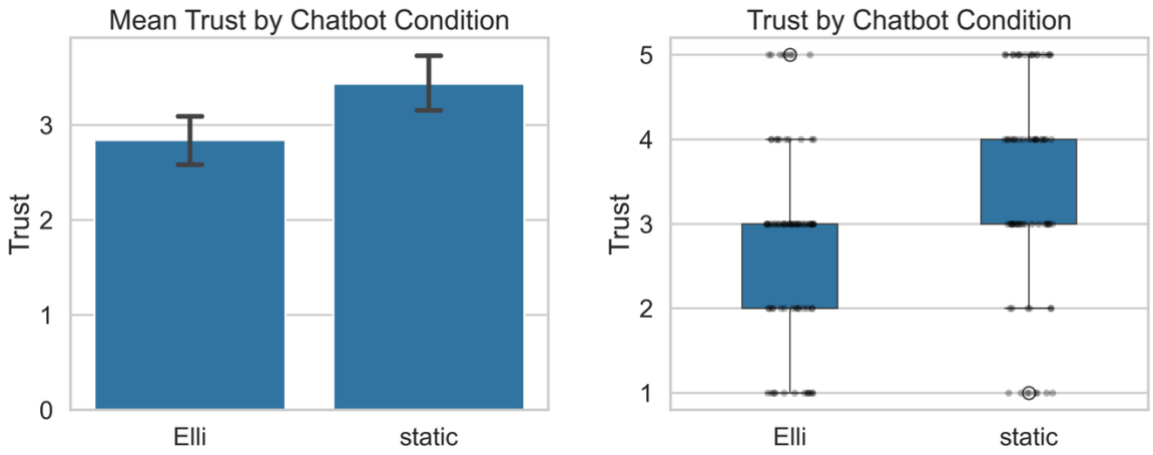
The central finding was that trust was significantly lower in the Elli condition (M = 2.84, SD = 1.16) than in the Static condition (M = 3.44, SD = 1.26), t(147) = -2.97, p = .004. The effect size was moderate (Cohen's d = -0.49), and non-parametric results confirmed the robustness of this effect (U = 1955.5, p = .002). This contradicted the preregistered hypothesis that the conversational AI would elicit greater trust.

As illustrated in Figure 1, trust was significantly lower in the Elli condition than in the static form.

**Figure 1**

*(A) Barplot showing mean trust scores and standard deviations.*

*(B) Boxplot displaying the distribution of trust scores.*



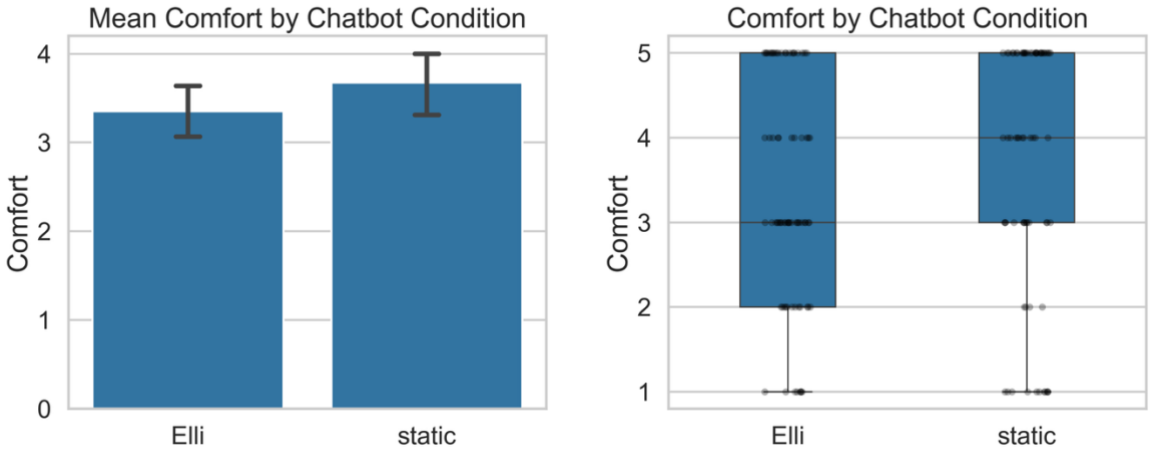

*Note.* Error bars in barplot represent ±1 SD. Boxplot shows median and interquartile range.

For comfort, no significant differences were found between conditions (Elli: M = 3.35, SD = 1.33; Static: M = 3.68, SD = 1.47), t(147) = -1.41, p = .161, d = -0.23. Figure 2 displays the comfort ratings across conditions, showing similar distributions despite a slight numerical difference.

**Figure 2**

*(A) Barplot showing mean comfort scores and standard deviations.*

*(B) Boxplot displaying the distribution of comfort scores*



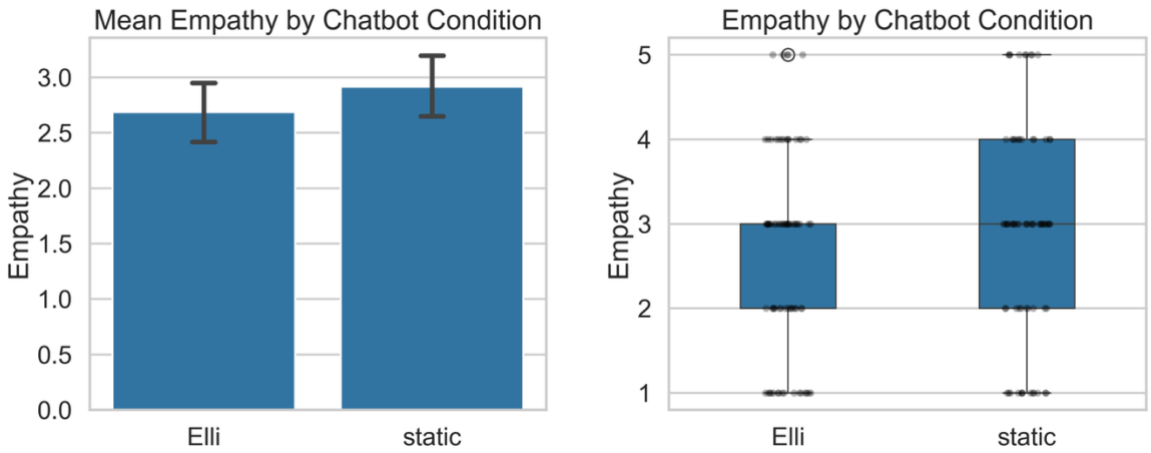

*Note.* No significant difference was observed. Error bars represent ± SD. All primary comparisons were preregistered. No multiple-testing correction was applied, but interpretation focuses on effect size and converging patterns.

Similarly, empathy scores were statistically equivalent between groups (Elli: M = 2.69, SD = 1.18; Static: M = 2.92, SD = 1.25), t(147) = -1.13, p = .259, d = -0.19. Figure 3 illustrates the similarity in empathy ratings between the chatbot and static conditions.

**Figure 3**

*(A) Barplot showing mean empathy scores and standard deviations.*

*(B) Boxplot displaying the distribution of empathy scores.*



*Note.* No significant difference was observed. Error bars represent ±1 SD.

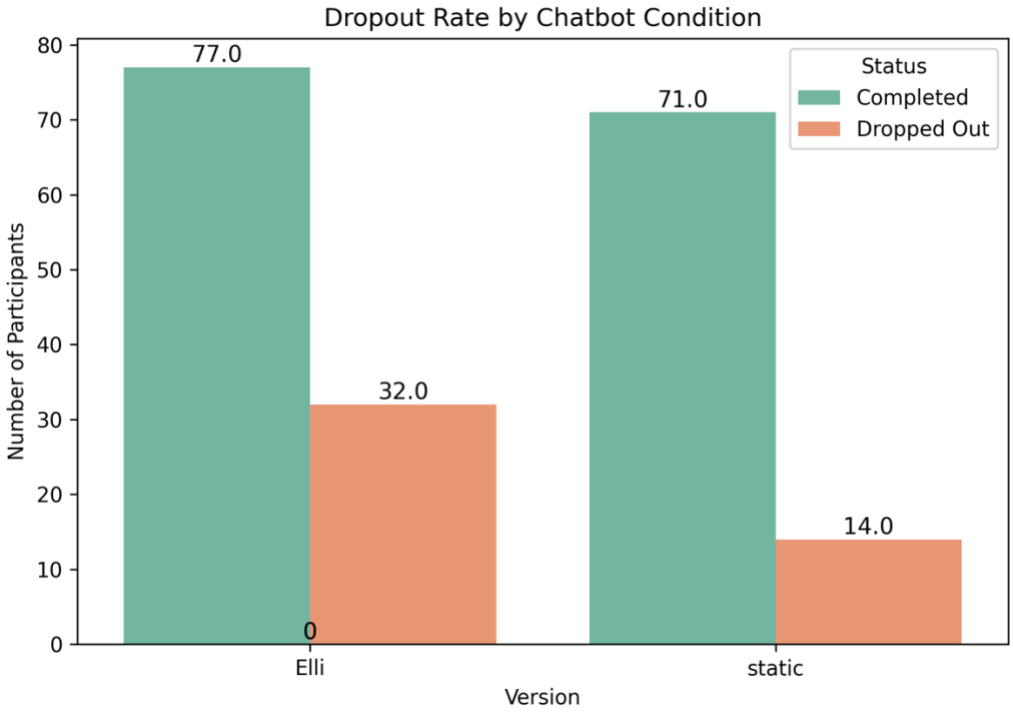
No significant differences were observed for PHQ-9 ($p = 0.896$) or GAD-7 scores ($p = 0.430$), indicating that symptom severity was similar across groups.

**Dropout analysis**

Dropout rates were higher in the Elli condition (28.6%) compared to the Static (14.7%), but the chi-square test did not reach statistical significance ($\chi^2 = 0.37$, $p = .54$), indicating no condition-related attrition bias. As shown in Figure 4, dropout rates were higher in the Elli condition, but this difference was not statistically significant.

**Figure 4**

*Dropout rates across conditions*




*Note.* Although dropout was numerically higher in the Elli condition (28.6%) compared to the static form (14.7%), the difference was not statistically significant ($\chi^2 = 0.37$, $p = .54$). Before analysis, data were filtered according to pre-registered exclusion criteria. Participants were excluded if they failed to

provide informed consent, submitted incomplete core assessments (incomplete PHQ-9 for Static version), experienced technical issues, or completed the Elli version under 60 seconds. Notably, the <60s completion filter was applied before assigning dropout status. These cases were treated as invalid completions and excluded from dropout and outcome analysis.

Table 2 presents descriptive statistics and inferential test results for all key outcome measures, including trust, comfort, empathy, and symptom severity (as measured by the PHQ-9 and GAD-7). All parametric test assumptions were checked before analysis. Shapiro-Wilk tests indicated non-normality in several variables; however, parametric and non-parametric tests (Mann-Whitney U) yielded converging results. Levene's tests confirmed homogeneity of variance across conditions.

**Table 2**

*Comparison of Key Outcome Measures Between Elli and Static Conditions*

| **Measure** | **Elli (M ± SD)** | **Static (M ± SD)** | **t-stat** | **t p-value** | **U-stat** | **U p-value** | **Cohen's d** | **Shapiro p (Elli)** | **Shapiro p (Static)** | **Levene p** |
|---|---|---|---|---|---|---|---|---|---|---|
| Trust | 2.84 ± 1.16 | 3.44 ± 1.26 | -2.97 | .004 | 1955.5 | .002 | -0.49 | .000 | .000 | .137 |
| Comfort | 3.35 ± 1.33 | 3.68 ± 1.47 | -1.41 | .161 | 2311.0 | .093 | -0.23 | .000 | .000 | .407 |
| Empathy | 2.69 ± 1.18 | 2.92 ± 1.25 | -1.13 | .259 | 2457.5 | .274 | -0.19 | .000 | .000 | .967 |
| PHQ-9 Total | 10.14 ± 6.82 | 10.00 ± 6.42 | 0.13 | .896 | 2733.0 | 1.000 | 0.02 | .002 | .024 | .472 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| GAD-7 Total | 8.62 ± 6.10 | 7.86 ± 5.65 | 0.79 | .430 | 2911.5 | .495 | 0.13 | .004 | .002 | .402 |

*Note*. t = independent-samples t-tests; p = p-value (2-tailed); U = Mann-Whitney U test; Cohen's d = effect size; Shapiro p = test of normality; Levene p = test of homogeneity of variance. Bolded p-values (< .05) indicate statistical significance.

**Interaction Effects**

The following moderation and mediation models were exploratory and not included in the preregistration. These analyses are reported transparently and labeled accordingly.

OLS regression models examined whether gender moderated the effect of condition on outcomes. For the interaction analyses, gender was coded as a binary variable (male vs. all other responses, including female, other, and prefer not to say) to preserve sample size and interpretability. Sensitivity analyses excluding non-binary and 'prefer not to say' participants yielded substantively similar results (not shown). A significant Version x Gender interaction was found for trust (interaction coefficient = -0.96, p = 0.23), indicating that male participants reported substantially lower trust in Elli than in the static form. No significant interactions were found for comfort (p = .93) or empathy (p = .056), although the latter approached significance. Table 3 presents the full OLS regression models testing the interaction between gender and interface version for predicting trust, comfort, and empathy. A significant interaction was found only for trust (p = .023), where male participants in the Elli condition reported notably lower trust than their counterparts in the static form. Interaction terms for comfort and empathy were non-significant.

**Table 3**

*OLS Regression Models: Gender x Interface Predicting Trust, Comfort, and Empathy*

| | **Outcome** | **Predictor** | **Coefficient** | **Std. Error** | **t-value** | **p-value** | **CI Lower** | **CI Upper** |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | Trust | Intercept | 2.62 | 0.2 | 13.42 | < .001 | 2.24 | 3.01 |
| 2 | Trust | Version[T.Static] | 1.03 | 0.27 | 3.86 | < .001 | 0.5 | 1.56 |
| 3 | Trust | Gender[T.male] | 0.49 | 0.28 | 1.76 | 0.08 | -0.06 | 1.04 |
| 4 | Trust | Version[T.Static]:Gender[T.male] | -0.96 | 0.42 | -2.3 | 0.02 | -1.79 | -0.13 |
| 5 | Comfort | Intercept | 3.32 | 0.23 | 14.38 | < .001 | 2.87 | 3.78 |
| 6 | Comfort | Version[T.Static] | 0.35 | 0.32 | 1.11 | 0.27 | -0.27 | 0.97 |
| 7 | Comfort | Gender[T.male] | 0.15 | 0.33 | 0.45 | 0.65 | -0.5 | 0.8 |
| 8 | Comfort | Version[T.Static]:Gender[T.male] | 0.04 | 0.49 | 0.08 | 0.93 | -0.94 | 1.02 |
| 9 | Empathy | Intercept | 2.73 | 0.2 | 13.97 | < .001 | 2.34 | 3.12 |
| 10 | Empathy | Version[T.Static] | 0.53 | 0.27 | 1.97 | 0.05 | < .001 | 1.05 |
| 11 | Empathy | Gender[T.male] | 0.05 | 0.28 | 0.17 | 0.86 | -0.5 | 0.6 |
| 12 | Empathy | Version[T.Static]:Gender[T.male] | -0.8 | 0.42 | -1.92 | 0.06 | -1.63 | 0.02 |

*Note.* Coefficients are based on ordinary least squares (OLS) regression models predicting trust, comfort, and empathy scores. Version (Static vs. Elli) and Gender (male vs. all others) were entered as predictors, along with their interaction. All values are rounded to 2-3 decimal places for clarity. Exact p-values < .001 are reported as p < .001.

**Age Subgroup Analysis**

To examine whether age moderated the effects of interface condition, two two-way ANOVAs were conducted with Interface Condition (Elli vs. Static) and Age Group (18-25, 26-35, 36+) as factors. For trust, a significant main effect of interface was found, $F(1, 153) = 5.69$, $p = .018$, replicating the primary analysis, which showed lower trust in Elli. There was no significant main effect of age group, $F(2, 153) = 2.59$, $p = .079$, nor a significant interaction, $F(2, 153) = 0.73$, $p = .483$. For empathy, no main effect of interface emerged, $F(1, 153) = 0.74$, $p = .391$; however, a significant main effect of age group was observed, $F(2, 153) = 3.73$, $p = .026$. This indicates that perceived empathy ratings varied significantly by age, regardless of condition. The interaction term was not significant, $F(2, 153) = 0.08$, $p = .926$. These

results suggest that while trust remained consistently lower in the Elli condition across age groups, perceived empathy varied by age, independent of the experimental condition.

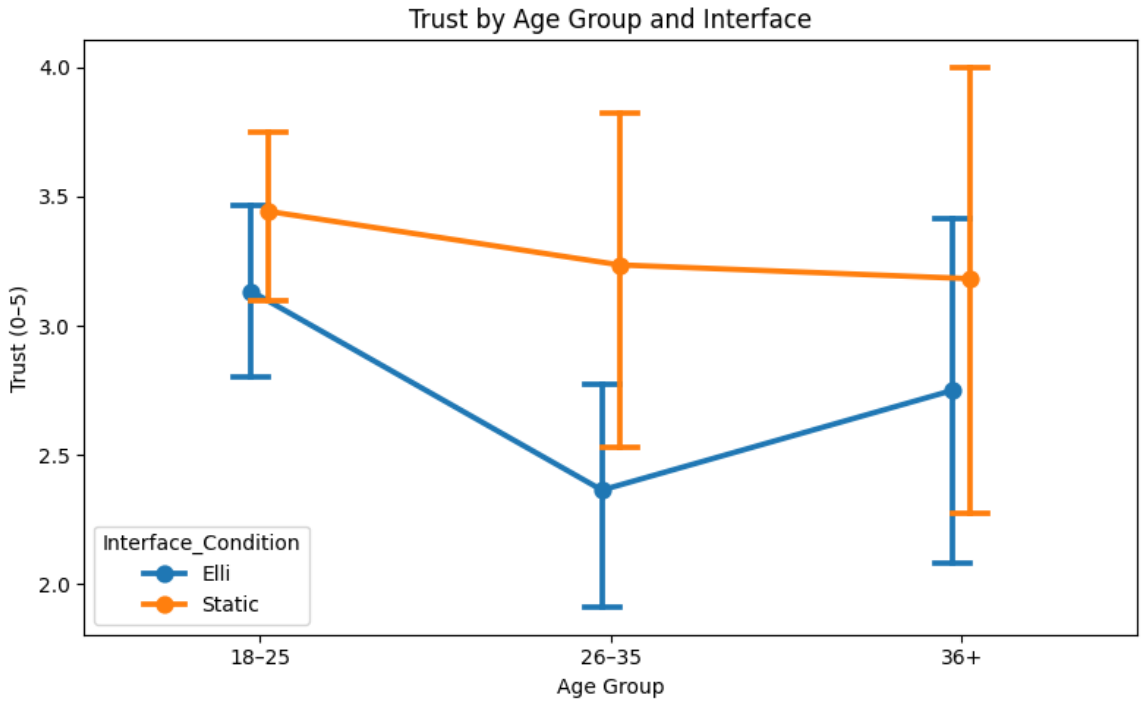
Figure 5 displays these results, highlighting that younger participants tended to rate empathy lower than older participants across both interface types.

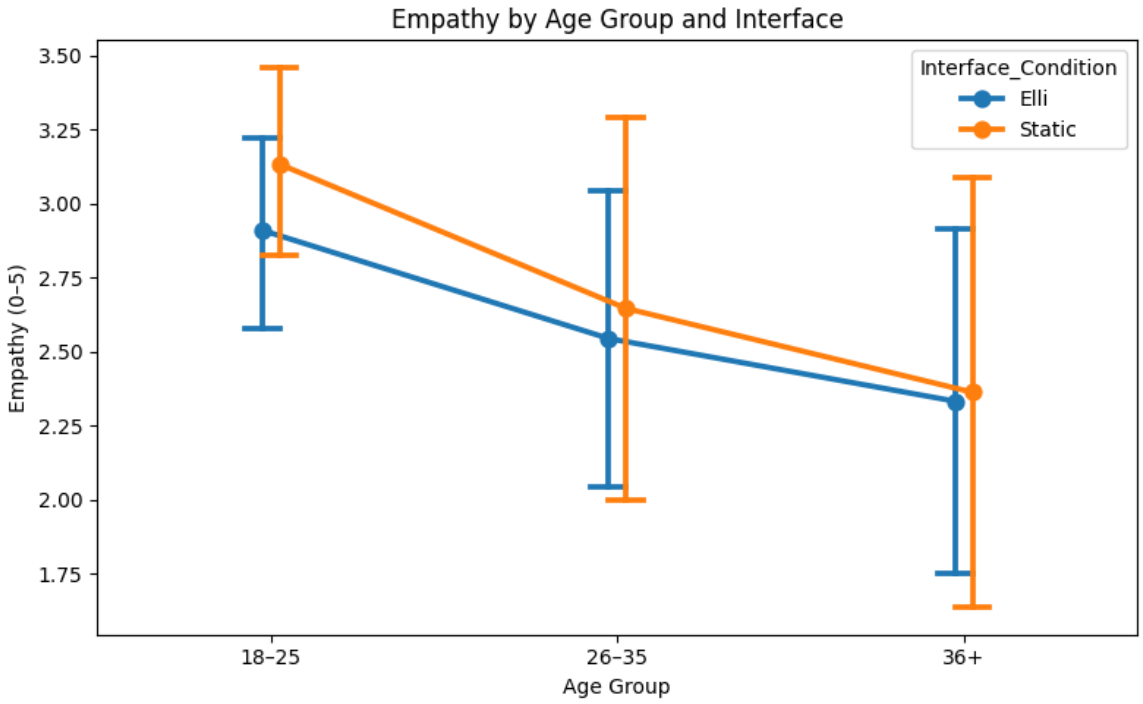
**Figure 5**

*Trust and Empathy Ratings by Age and Group Interface Condition*

*(A) Point plot showing mean trust ratings across age groups and interface types.*

*(B) Point plot showing mean empathy ratings across age groups and interface types.*






*Note.* Error bars represent ±1 standard error. No interaction effects were observed.

**Mediation Analysis**

A mediation model was tested to determine whether perceived empathy mediated the effect of version on trust. While the direct path from version to trust was significant (coef = 0.51, p = .006) and empathy strongly predicted trust (coef = 0.51, p < .001), the indirect (mediated) path was not statistically significant (coef = 0.12, p = .226). This suggests the observed trust differences were primarily driven by direct interface effects rather than perceived empathy. Table 4 presents the complete output of the mediation model, examining whether perceived empathy mediates the relationship between interface version and trust. While both the total and direct effects were significant, the indirect effect (via empathy) did not reach statistical significance.

**Table 4**

*Mediation Analysis: Empathy as a Mediator Between Interface Type and Trust.*

| Path | Coefficient | Std. Error | p-value | 95% CI Lower | 95% CI Upper | Significant |
|---|---|---|---|---|---|---|
| Empathy ~ X | 0.247 | 0.206 | 0.232 | -0.160 | 0.653 | No |
| Trust ~ Empathy | 0.505 | 0.077 | < .001 | 0.354 | 0.657 | Yes |
| Total Effect | 0.629 | 0.205 | 0.003 | 0.224 | 1.035 | Yes |
| Direct Effect | 0.510 | 0.181 | 0.006 | 0.152 | 0.868 | Yes |
| Indirect Effect | 0.119 | 0.102 | 0.226 | -0.071 | 0.338 | No |

*Note.* Results of a mediation analysis testing perceived empathy as a mediator of the effect of interface version on trust. Estimates are based on ordinary least squares regression. All values are rounded to 2-3 decimal places for clarity. Significance is indicated in the final column ($p < .001$ is reported as such).

**Qualitative Analysis**

To summarize open-ended responses, Table 5 presents key themes identified across conditions using interpretative phenomenological analysis. Themes in the Elli condition centered on artificial tone and discomfort, while feedback in the static form highlighted clarity and ease.

**Table 5**

*Summary of Emergent Themes from Open-Ended Feedback.*

| Theme | Condition | Sample Words | Representative Quote |
|---|---|---|---|
| Perceived lack of humanity | Elli | robotic, unnatural, talk, human | I think Elli sounds too 'robotic'. It could be more 'human'. |
| Authenticity concern in empathy simulation | Elli | therapy, push back, not for me | When talking about deep subjects I need to be talking to |

| | | | |
|---|---|---|---|
| | | | someone who can actually push back. |
| Emotional awkwardness or discomfort | Elli | bad, not helpful, toxic positivity | These questions can feel pretty bad… Elli should ask further questions about how I feel. |
| Transactional trust and clarity | Static | good, nice, questionnaire, reflect | I thought the questions themselves were very good and it was nice to reflect on these things. |
| Usability and credibility | Static | scale, rating, psychology, options | The use of rating scales used in psychology lends it credibility. |
| Feeling heard without social pressure | Static | cool, asked, free, someone | It was pretty cool to be asked these type of questions... it's good to be free with someone at least. |

*Note.* Summary of key qualitative themes derived from participant feedback by experimental condition. Themes were identified using inductive thematic analysis. Sample words represent frequent or illustrative terms per theme. Quotes are anonymized and selected to reflect the theme.

Thematic analysis of participant comments revealed consistent themes. In the Elli condition, users frequently described the chatbot as "robotic," "unnatural," and not "human enough." Several noted discomfort with the simulated empathy and raised concerns about emotional authenticity. In contrast, participants in the Static condition highlighted the clarity, structure, and ease of the experience. Figure 6 illustrates these differences visually through word clouds generated from participant feedback in each condition.

**Figure 6**

*World Clouds depicting the most frequent words in open-text feedback for the Elli and static conditions. Font size corresponds to word frequency.*


*Note.* Word clouds were generated from open-text feedback for each condition separately. Common English stop words were removed. Larger words indicate a higher frequency of use.

Collectively, the quantitative and qualitative findings converge to suggest that, while the chatbot could simulate empathic dialogue, it did not foster the same level of trust as a conventional static format.

All analyses were conducted using Python 3.13 with the following libraries: pandas 2.3.1, statsmodels 0.14.5, seaborn 0.13.2, matplotlib 3.10.3, and pingouin 0.5.5.

## Discussion

### Summary of Findings

This study aimed to evaluate whether a GPT-4-powered chatbot could match or surpass a static web form in fostering user trust, comfort, and perceived empathy during mental health screenings. Instead, the findings challenge common assumptions about conversational AI: despite emotionally intelligent prompts and adaptive dialogue, trust in the chatbot was significantly lower than in a simple form interface.

### Interpreting the Trust Gap

This gap was not due to poor functionality. Elli administered the assessments effectively and adjusted its tone in response to user input. Yet participants repeatedly described the experience as “robotic” and “unnatural”. The results suggest a deeper psychological boundary: simulated empathy may not feel emotionally hollow when grounded in perceived authenticity.

**Perception of Artificial Empathy**

Thematic analysis and user feedback reveal a consistent discomfort with affective computing that appears human but fails to evoke a genuine sense of humanity. This reflects broader concerns about emotional deception, ethical design, and the illusion of therapeutic alliance.

**Age and Gender Differences**

Age also played a significant role in empathy perception: younger participants rated both interfaces as more empathic, while ratings declined progressively across older groups, independently of chatbot use. This suggests that older users may be less receptive to digital expressions of empathy overall, regardless of the delivery medium.

Gender effects further complicate the picture; male users, in particular, tend to distrust the chatbot, raising questions about how demographic factors influence responses to emotionally expressive AI.

**The Illusion of Empathy vs. Perceived Realness**

Notably, comfort and empathy ratings did not differ significantly between groups. This ambivalence suggests that users may accept artificial empathy as functional, but not trust it as real. The mediation model supports this: while empathy predicts trust, it does not explain the gap between chatbots and forms. It is not empathy alone, but the perception of realness, that appears to matter most.

**Design Implications for Digital Mental Health Tools**

Together, these findings provide a timely caution for the design of digital mental health services. Trust cannot be engineered solely through warmth. Suppose AI is to support users in moments of vulnerability in a meaningful way. In that case, it must offer more than polished scripts: it must communicate emotional honesty, acknowledge its artificial nature, and avoid pretending to be human when it's not.

**Final Remarks**

This research contributes urgently needed empirical insight to the evolving field of digital therapeutic alliance. It suggests that empathy, while necessary, is not sufficient. Until conversational agents can convincingly convey presence, emotional safety may still require a human face.

**References**

Chen, C., Lam, K. T., Yip, K. M., So, H. K., Lum, T. Y. S., Wong, I. C. K., Yam, J. C., Chui, C. S. L., & Ip, P. (2025). Comparison of an AI Chatbot With a Nurse Hotline in Reducing Anxiety and Depression Levels in the General Population: Pilot Randomized Controlled Trial. JMIR human factors, 12, e65785. https://doi.org/10.2196/65785

Coghlan, S., Leins, K., Sheldrick, S., Cheong, M., Gooding, P., & D'Alfonso, S. (2023). To chat or bot to chat: Ethical issues with using chatbots in mental health. Digital health, 9, 20552076231183542. https://doi.org/10.1177/20552076231183542

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological and emotional effects of self-disclosure with a chatbot. *Journal of Communication, 68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. Internet interventions, 27, 100495. https://doi.org/10.1016/j.invent.2022.100495

Mehrdad Rahsepar Meadi, Sillekens, T., Metselaar, S., Balkom, A. van, Bernstein, J., & Neeltje Batelaan. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. JMIR Mental Health, 12, e60432–e60432. https://doi.org/10.2196/60432

Norcross, J. C., & Lambert, M. J. (2019). *Psychotherapy relationships that work: Volume 1: Evidence-based therapist contributions* (3rd ed.). Oxford University Press.

Schick, A., Feine, J., Morana, S., Maedche, A., & Reininghaus, U. (2022). Validity of Chatbot Use for Mental Health Assessment: Experimental Study. JMIR mHealth and uHealth, 10(10), e28082. https://doi.org/10.2196/28082

Stubbe, D. E. (2018). The therapeutic alliance: The fundamental element of psychotherapy. *Focus*, *16*(4), 402–403. https://doi.org/10.1176/appi.focus.20180022