**H1: AI-unaware participants will report higher empathy, satisfaction, and treatment outcomes than AI-aware.**

* Empathy: t = 0.87, p = 0.39
* User Satisfaction: t = 1.31, p = 0.19
* Treatment Outcomes: t = 1.42, p = 0.16

All p-values are well above 0.05, meaning there are no statistically significant differences between the AI-unaware and AI-aware groups for any of the three outcomes. The direction of the t-statistics is positive, but the differences are not significant.

**H1 is not supported.**  
Participants in the AI-unaware condition did not report significantly higher levels of perceived empathy, user satisfaction, or treatment outcomes compared to those in the AI-aware condition.

**H2: Among empathy-driven participants, AI-aware will report lower satisfaction, empathy, and outcomes than AI-unaware.**

* Empathy: t = 1.17, p = 0.24
* User Satisfaction: t = 0.86, p = 0.39
* Treatment Outcomes: t = 1.21, p = 0.23

Again, all p-values are above 0.05. There are no significant differences between AI-aware and AI-unaware among empathy-driven participants for any outcome. The direction of the t-statistics is positive, which is the opposite of the hypothesis (which predicted lower scores for AI-aware), but the differences are not significant.

**H2 is not supported.**  
Among empathy-driven participants, being in the AI-aware condition did not lead to significantly lower satisfaction, empathy, or treatment outcomes.

**H3: Rational ToC participants will show no significant differences between AI-aware and AI-unaware.**

* Empathy: t = 0.92, p = 0.36
* User Satisfaction: t = 0.56, p = 0.58
* Treatment Outcomes: t = 1.60, p = 0.11

All p-values are above 0.05, indicating no significant differences between AI-aware and AI-unaware for rational ToC participants on any outcome.

**H3 is supported.**  
Participants with a rational ToC showed no significant differences in satisfaction, empathy, or treatment outcomes between AI-aware and AI-unaware conditions.

**H4: Participants’ ToC will align with the type of advice they select.**

* Chi-square: 0.000, p = 1.000
* Match: 82, Mismatch: 105

The p-value is 1.0, indicating no association between ToC and advice style selection. The number of matches and mismatches is roughly equal, and the chi-square test confirms that this is not different from chance.

**H4 is not supported.**  
Participants’ ToC did not significantly align with the type of advice they selected during the chatbot interaction.

**H5: Higher AI acceptance will predict greater satisfaction, outcomes, and empathy, especially in the AI-aware group.**

Empathy

* Main effect of Acceptance Level:  
  F = 27.60, p < 0.000001 (significant)
* Main effect of AI Transparency:  
  F = 0.59, p = 0.44 (not significant)
* Interaction (Acceptance × Transparency):  
  F = 4.16, p = 0.043 (significant)
* Participants with higher AI acceptance report significantly higher empathy overall.
* The effect of acceptance is different depending on AI transparency (the interaction is significant). This means the positive effect of acceptance is especially pronounced (or only present) in one of the transparency groups—likely the AI-aware group, as your hypothesis predicted.

Satisfaction

* Main effect of Acceptance Level:  
  F = 2.43, p = 0.12 (not significant)
* Main effect of AI Transparency:  
  F = 1.59, p = 0.21 (not significant)
* Interaction:  
  F = 0.21, p = 0.64 (not significant)
* No significant effects.
* AI acceptance does not significantly predict satisfaction, nor does transparency, nor their interaction.

Treatment Outcomes

* Main effect of Acceptance Level:  
  F = 19.57, p < 0.0001 (significant)
* Main effect of AI Transparency:  
  F = 1.83, p = 0.18 (not significant)
* Interaction:  
  F = 3.43, p = 0.066 (marginal, not quite significant at 0.05)
* Participants with higher AI acceptance report significantly higher treatment outcomes overall.
* The interaction is marginal (p = 0.066), suggesting a possible trend that the effect of acceptance might be stronger in one group, but this is not statistically significant at the conventional 0.05 level.

My comment for H5 : supported 🡪 maybe not bad to check with you Johannes please

* Empathy:  
  Supported. Higher AI acceptance predicts higher empathy, and this effect is especially strong (or only present) in one transparency group (likely AI-aware), as you hypothesized.
* Satisfaction:  
  Not supported. No significant effect of AI acceptance on satisfaction.
* Treatment Outcomes:  
  Partially supported. Higher AI acceptance predicts higher treatment outcomes overall, but the interaction with transparency is only marginally significant (trend-level, not conventionally significant).

**H6: The topic chosen will not significantly influence empathy, satisfaction, or treatment outcomes.**

**Statistical Results:**

* Empathy: F = 2.21, p = 0.113
* User Satisfaction: F = 1.28, p = 0.280
* Treatment Outcomes: F = 1.63, p = 0.198
* Tukey HSD: All p-adj > 0.05 for all pairwise comparisons

All p-values are above 0.05, indicating no significant effect of topic on any outcome. The Tukey HSD post-hoc tests also show no significant pairwise differences.

**H6 is supported.**  
The topic chosen for discussion did not significantly influence perceived empathy, user satisfaction, or treatment outcomes.

**Table**. Summary of Hypotheses Testing Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Hypothesis | Outcome(s) | Test | Statistic(s) | p-value | Result | Supported |
| H1: AI-unaware participants will report higher empathy, satisfaction, and treatment outcomes than AI-aware. | Empathy | t-test | t = 0.87 | 0.39 | n.s. | Not Supported |
|  | Satisfaction | t-test | t = 1.31 | 0.19 | n.s. | Not Supported |
|  | Treatment Outcomes | t-test | t = 1.42 | 0.16 | n.s. | Not Supported |
| H2: Among empathy-driven participants, AI-aware will report lower satisfaction, empathy, and outcomes than AI-unaware. | Empathy | t-test | t = 1.17 | 0.24 | n.s. | Not Supported |
|  | Satisfaction | t-test | t = 0.86 | 0.39 | n.s. | Not Supported |
|  | Treatment Outcomes | t-test | t = 1.21 | 0.23 | n.s. | Not Supported |
| H3: Rational ToC participants will show no significant differences between AI-aware and AI-unaware. | Empathy | t-test | t = 0.92 | 0.36 | n.s. | Supported |
|  | Satisfaction | t-test | t = 0.56 | 0.58 | n.s. | Supported |
|  | Treatment Outcomes | t-test | t = 1.60 | 0.11 | n.s. | Supported |
| H4: Participants’ ToC will align with the type of advice they select. | ToC × Advice Style | Chi-square | χ² = 0.00 | 1.00 | n.s. | Not Supported |
| H5: Pre-existing acceptance of AI will positively influence post-interaction outcomes, especially in the AI-aware group. | Empathy | 2-way ANOVA | F(1,183) = 27.60 (Acceptance), F(1,183) = 4.16 (Interaction) | <0.001, 0.043 | Significant main & interaction | Supported |
|  | Satisfaction | 2-way ANOVA | F(1,183) = 2.43 (Acceptance), F(1,183) = 0.21 (Interaction) | 0.12, 0.64 | n.s. | Not Supported |
|  | Treatment Outcomes | 2-way ANOVA | F(1,183) = 19.57 (Acceptance), F(1,183) = 3.43 (Interaction) | <0.001, 0.066 | Main effect significant, interaction marginal | Partially Supported |
| H6: The topic chosen will not significantly influence empathy, satisfaction, or treatment outcomes. | Empathy | ANOVA | F(2,183) = 2.21 | 0.11 | n.s. | Supported |
|  | Satisfaction | ANOVA | F(2,183) = 1.28 | 0.28 | n.s. | Supported |
|  | Treatment Outcomes | ANOVA | F(2,183) = 1.63 | 0.20 | n.s. | Supported |

**Notes**: n.s. = not significant (p > 0.05). For H5, “Interaction” refers to Acceptance × AI Transparency. “Supported” means the data are consistent with the hypothesis at p < 0.05.

**Qualitative Part – Just output based on plots**

**Integrated Quantitative and Qualitative Analysis of Human-AI Therapeutic Conversations**

This study investigated the impact of AI transparency and user characteristics on perceived empathy, satisfaction, and treatment outcomes in a large-scale, text-based mental health intervention. Participants engaged in structured conversations with a supportive chatbot, with experimental manipulations of AI transparency (AI-aware vs. AI-unaware) and measurement of individual differences such as Theory of Change (ToC) and AI acceptance. Both quantitative and qualitative methods were employed to provide a comprehensive understanding of user experience and the mechanisms underlying therapeutic alliance in digital mental health.

A total of 934 user-bot conversational pairs were analyzed, with each message pair reflecting a unique exchange. Sample user messages included expressions of work-related stress and emotional overload, such as: “I've been at work for a while now and we're overloaded, and not being able to do all the work is leading me to a fairly high level of stress.” The bot’s responses were consistently supportive and reflective, for example: “I understand that's difficult; feeling overwhelmed can be really challenging. What specific tasks are causing you the most stress?”

Quantitative analyses revealed that the primary hypotheses regarding the effects of AI transparency and user characteristics were not supported. For H1, there were no significant differences in empathy, satisfaction, or treatment outcomes between AI-unaware and AI-aware participants (all p > 0.16). H2, which predicted lower satisfaction and empathy among AI-aware, empathy-driven participants, was also not supported (all p > 0.23), with the direction of effects opposite to prediction. H4, positing alignment between ToC and advice selection, was not supported (χ² = 0.00, p = 1.00). H3 and H6, which predicted null effects, were supported. H5, concerning the role of AI acceptance, was partially supported: higher AI acceptance predicted greater empathy and treatment outcomes, with a significant interaction for empathy and a marginal interaction for outcomes.

The qualitative analysis provided convergent and nuanced insights into these findings. Topic modeling of user messages identified five primary themes: emotional states (“feel,” “like,” “good,” “thank,” “thanks”), uncertainty and work stress (“dont,” “know,” “time,” “work,” “good”), job-related concerns (“job,” “get,” “still,” “work,” “want”), affirmation and social context (“yes,” “like,” “life,” “time,” “going”), and help-seeking (“help,” “feeling,” “really,” “work,” “stress”). These topics reflect a conversational landscape dominated by emotional disclosure, stress, and the search for support. The bot’s topic modeling revealed a parallel structure, with themes of specificity and validation (“specific,” “stress,” “thats,” “feel,” “overwhelming,” “understand”), gratitude and support (“feel,” “sharing,” “prolific,” “thanks,” “today,” “support”), empathetic listening (“hear,” “thats,” “feeling,” “feel,” “youre,” “like,” “help”), sense-making (“like,” “sounds,” “feeling,” “thats,” “youre,” “feel,” “sense”), and encouragement (“sounds,” “like,” “feel,” “help,” “great,” “really”).

Sentiment analysis further illuminated the affective dynamics of the conversations. The mean sentiment score for user messages was slightly positive (mean = 0.08, SD = 0.40), but with substantial variability and a distribution centered near neutrality, as visualized in the attached Sentiment\_Score.png and Sentiment\_Analysis.png plots. In contrast, bot messages exhibited a higher mean sentiment (mean = 0.43, SD = 0.49), reflecting the bot’s consistently positive and supportive tone. The overlapping sentiment distributions, as shown in the density and box plots, indicate that while the bot maintained a positive affect, user sentiment remained more variable and less positive overall.

Text mining and keyword analysis reinforced these findings. The most frequent user keywords included “thanks,” “want,” “thank,” “think,” “stress,” “time,” “try,” “help,” “know,” “like,” “work,” “good,” “feel,” “dont,” and “yes.” These terms highlight the centrality of gratitude, desire for change, cognitive processing, and stress in user discourse. The bot’s top keywords—“specific,” “hear,” “support,” “today,” “really,” “help,” “youre,” “feeling,” “thanks,” “sounds,” “sharing,” “prolific,” “thats,” “like,” “feel”—underscore its focus on validation, support, and empathetic engagement.

Advanced visualizations provided further depth to the analysis. Bigram and trigram frequency plots (topU1.png, topU2.png, topB1.png, topB2.png) revealed that user language was characterized by repeated expressions of need and uncertainty, while bot language emphasized supportive phrases and reflective listening. Co-occurrence network graphs (Co\_OCC1.png, Co\_OCC2.png) illustrated the dense interconnections among stress-related and emotional terms in user messages, and the clustering of supportive and empathetic terms in bot responses. t-SNE clustering plots (t\_SNE\_User.png, t\_SNE\_Bot.png) demonstrated that user messages formed several overlapping clusters, reflecting the diversity of concerns and emotional states, while bot messages were more tightly clustered, consistent with a standardized supportive approach.

The integration of these qualitative findings with the quantitative results provides a comprehensive picture of the conversational dynamics. The lack of significant differences between experimental conditions in the quantitative data is mirrored in the qualitative analysis: both AI-aware and AI-unaware participants engaged in similar patterns of emotional disclosure and help-seeking, and the bot’s responses were uniformly supportive regardless of condition. The absence of alignment between ToC and advice selection was also evident in the qualitative data, with no clear clustering or thematic differentiation based on user ToC.

The overall sentiment landscape suggests that while the bot was successful in maintaining a positive and empathetic tone, this did not translate into significant improvements in user sentiment or outcomes as measured quantitatively. The persistence of stress and uncertainty in user language, despite the bot’s supportive stance, highlights the complexity of facilitating meaningful change in digital mental health interventions. The findings suggest that while conversational AI can provide consistent support and validation, additional mechanisms or more personalized interventions may be necessary to achieve significant improvements in user experience and outcomes.

**Qualitative Analysis of Rejected Hypotheses (H1, H2, H4) Based on Empirical Outputs**

The rejected hypotheses were further examined through the lens of the qualitative data and visualizations.

For H1, the expectation that AI-unaware participants would report higher empathy, satisfaction, and treatment outcomes was not supported. The topic modeling and sentiment analysis showed that both groups expressed similar themes and affective profiles. The user sentiment distribution (Sentiment\_Score.png) was centered near neutrality, with no discernible shift toward positivity in the AI-unaware group. The co-occurrence and t-SNE plots (Co\_OCC1.png, t\_SNE\_User.png) confirmed that the conversational content and emotional tone were consistent across conditions, with no evidence of enhanced empathy or satisfaction in the AI-unaware group.

H2, which predicted lower satisfaction and empathy among AI-aware, empathy-driven participants, was also not supported. The advanced text mining and bigram/trigram analyses (topU1.png, topU2.png) did not reveal any unique linguistic markers of dissatisfaction or reduced empathy in the AI-aware group. The bot’s supportive language (topB1.png, topB2.png) and the user’s continued focus on stress and emotional needs were consistent across conditions. The sentiment analysis (Sentiment\_Analysis.png) showed no significant negative shift among AI-aware, empathy-driven users, and the t-SNE clustering (t\_SNE\_User.png) did not indicate any subgroup with distinctly lower satisfaction or empathy.

H4, positing alignment between ToC and advice selection, was not supported in either the quantitative or qualitative data. The co-occurrence network graphs (Co\_OCC1.png, Co\_OCC2.png) and t-SNE plots (t\_SNE\_User.png, t\_SNE\_Bot.png) showed that user and bot messages formed clusters based on conversational content and affect, not on ToC or advice type. The frequency of matches and mismatches in advice selection was nearly equal, and the content of user messages did not reveal any systematic preference or rationale linked to ToC. The qualitative data thus reinforce the quantitative finding that advice selection was not driven by participants’ underlying change theories.

In summary, the qualitative analyses, grounded in empirical outputs and advanced visualizations, provide convergent evidence with the quantitative results: the rejected hypotheses (H1, H2, H4) were not supported in the lived experience and language of users. The conversational data suggest that users engaged with the chatbot in similar ways regardless of experimental condition, and that the hypothesized mechanisms did not manifest in the language, sentiment, or thematic structure of the interactions. These findings highlight the importance of integrating qualitative and quantitative approaches to fully understand the impact of AI transparency and user characteristics in digital mental health interventions, and suggest that future work should explore additional or alternative mechanisms to enhance user experience and outcomes.