Humans relate more to LLMs that mirror their psychology

Santiago Castiello1,2 \*, Riddhi Jain Pitliya1\*, Daniel R. Lametti3,4\* , Robin A. Murphy2\*

1. Yale School of Medicine, Yale University
2. Department of Experimental Psychology, University of Oxford
3. Department of Psychology, Acadia University
4. OneReach.ai

\*all authors contributed equally

Keywords: homophily, affiliation, large language models, AI, human-AI interaction

# Abstract

People connect with others who share their likes, interests, and mental states, a phenomenon known as homophily. Here, we examine human-AI homophily by testing how individuals' psychological traits shape their affiliation with large language models (LLM) that mimic their traits. In Experiment 1, the LLM GPT-4 was instructed to produce language aligned with either an anxious or non-anxious state. Participants (n=100) engaged with both versions of LLM and then completed a questionnaire to assess how relatable they found the LLM. Participants with high anxiety felt more similar to and understood by the LLM instructed to mimic an anxious state, while participants with low anxiety felt different from it. In Experiment 2, participants (n=100) engaged with LLM instructed to produce language aligned with either an extroverted or introverted personality. Extroverted participants felt similar to the LLM that mimicked extroversion. Taken together, the findings support the idea that humans relate more to LLMs that mirror aspects of their psychology.

# 

# 

# Introduction (937 words)

Social interactions between individuals show effects of homophily (*homo*=similar*; philia*=love) and heterophily (*hetero*=different*; philia*=love) [(Lazarsfeld and Merton 1964)](https://paperpile.com/c/L1tWAD/Bi0B). These are the opposing tendencies to interact more easily and effectively with members of an affiliative group or with groups other than one’s own. Is homophily—or something like it—possible between artificial intelligence (AI) and humans?

The emergence of Large Language Models (LLMs) has expanded the range of groups with which we might affiliate to include artificial systems. As LLMs readily adopt specific linguistic styles in response to prompts, it is relatively easy to create person-centered or ‘tailored’ AI with the aim of building stronger human-AI bonds. But this idea—long a plot point for science fiction writers (see, for instance, the 2013 Spike Jonze film *Her*)—has little empirical support, largely because LLMs are relatively new. Here, we prompted LLMs to use language associated with mental health related states (*i.e.*, anxious vs non anxious) or personality traits (*i.e.*, extroverted vs introverted) to systematically test whether human-AI homophily based on a shared psychology might occur. The experiments were designed to either enhance affiliative perceptions with AI among our participants or, conversely, negatively challenge them.

Turing’s thought experiment on the nature of human and AI asked the question, “When must we acknowledge and grant artificial systems human-like intelligence? What must the artificial system do to be indistinguishable from a human conversational partner?” [(Turing 1950)](https://paperpile.com/c/L1tWAD/nVmX). Apart from a growing consensus that LLMs can now pass Turing’s test as he envisioned it, not surprisingly people have come to further refine definitions of intelligence, asking whether LLMs can reason or provide social comfort [(Mitchell 2024; Kambhampati 2024)](https://paperpile.com/c/L1tWAD/Hp0Y+zPIc). In spite of the engaging interactions LLMs offer, and their ability to answer some complex knowledge and human reasoning questions, these systems may not pass more complex versions of the Turing test for a variety of reasons [(Jannai et al. 2023; Biever 2023)](https://paperpile.com/c/8YKvUG/MldT+rO69). We may notice that the AI misinterprets questions or intentions in a way a human would not, but there are other factors that provide clues to the artificial nature of the experience. For example, they respond more quickly, too accurately, or not accurately enough [(Svenningsson and Faraon 2019)](https://paperpile.com/c/L1tWAD/7XAh). LLMs may also fail to meet our expectations, not because of lack of General Intelligence, but because we do not perceive human psychological traits in the often generic language LLMs use. Specifically, LLMs may lack components that support homophily or heterophily.

People exhibit homophily in their social networks—the tendency to affiliate with other individuals that share characteristics with themselves [(Launay and Dunbar 2015)](https://paperpile.com/c/8YKvUG/lhAk). For example, people tend to choose partners who resemble themselves, a phenomenon observed in both in-person and app-based dating [(Restrepo-Echavarria et al. 2023; Furman and Simon 2008)](https://paperpile.com/c/8YKvUG/Qpxc+oXak). When viewing social interactions, homophily is even reflected in shared patterns of neural activity between friends [(Parkinson et al. 2018)](https://paperpile.com/c/L1tWAD/cmYh). Factors that contribute to homophily include things like shared religious beliefs, interests, hobbies, personality characteristics, mental states, and social network distance [(McPherson et al. 2001)](https://paperpile.com/c/8YKvUG/2HQL). But the degree to which people feel more similar to AI that reflect characteristics of their psychological traits remains unknown. Understanding the extent to which homophily is observed in human-AI interactions is important, not only for understanding how well human social behaviours generalize to artificial intelligent systems, but also for the design of useful human-AI experiences. When humans make moral judgments about AI, we tend to judge their outcomes more harshly in contrast to the same outcomes being produced by humans [(Hidalgo and Orghiain 2021)](https://paperpile.com/c/8YKvUG/cBNK). This presents a problem for the use of AI systems in, for example, healthcare, where the ability of people to overcome their biases against AI may impair the usefulness of AI systems as diagnostic tools, even when GPT-4 seems to produce relatively good differential diagnoses [(Ríos-Hoyo et al. 2024)](https://paperpile.com/c/8YKvUG/oj1D)..

The aim of the current study was to experimentally test whether humans feel more similar to an LLM that, through conversation, exhibits a shared psychological trait, despite full knowledge that they are talking to a computer program. In Experiment 1, the LLM GPT-4 was prompted (using clinically relevant criteria) to produce language aligned with either an anxious or non-anxious psychological state. One hundred participants conversed with both versions of GPT-4 and then answered a series of questions designed to test the extent to which they felt similar or different from the AI. Following this questionnaire, participants’ own anxiety was measured. We then examined participants’ sense of similarity to each GPT-4 persona in relation to their reported anxiety levels. As a further test of affiliation, we also examined how the sentiment of the messages they exchanged with each version of GPT-4 was influenced by the LLM’s language. We predicted that participants with high anxiety would feel more similar to the LLM that used anxious language and vice versa for participants with low anxiety, one indication of homophily with AI.

In Experiment 2, we again examined the extent to which participants might feel similar to an LLM with a shared psychological trait, but we manipulated the personality trait of extroversion-introversion. GPT-4 was prompted (using answers to validated tests of personality) to produce language aligned with either an extroverted or introverted personality. One hundred participants had a conversation with each version of GPT-4; they then answered a series of questions designed to test the extent to which they felt similar or different from each AI. In this case, participants’ affiliation to each GPT-4 persona was examined in relation to reported extroversion. Again, we predicted homophily: extroverted participants would feel more similar to the GPT-4 persona that used extroverted language and vice versa for introverted participants.

# Methods (1536 words)

## Participants

We recruited 200 participants via Amazon Mechanical Turk (MTurk). Half the participants took part in Experiment 1 (n=100), and the other half took part in Experiment 2 (n=100). As the effects in Experiment 1 were large, we used the same sample size for Experiment 2. Fourteen participants were excluded from the final analysis for failing to engage with the AI (see *Language Model Design*) making the final sample size 89 for Experiment 1 and 97 for Experiment 2. Participants were native English speakers; 84 of the participants identified as female, 100 as male, 1 non-binary, and 1 preferred not to reveal their gender. Participants were between 18 and 50 years old (M=40.88; SD=6.89). Participants were paid $7 (USD) for completing the experiment, which took approximately 30 minutes. The University of Oxford’s Central University Research Ethics Committee (CUREC: R86261/RE001) approved the study before testing.

## Software and Materials

Participants completed the study via the online platform Gorilla. The LLM interactions were designed using the conversational AI platform Generative Studio X (OneReach.ai). The LLM GPT-4 (OpenAI) was used in both studies. The chats were embedded in the Gorilla.sc interface using an iframe. Statistical analyses and visualizations were conducted in R studio [(R Core Team. 2018)](https://paperpile.com/c/L1tWAD/MIkG) with the R packages *report, ggplot2, lme4,* and *ggpubr.*

### Language Model Design

Participants had two text-based conversations with the LLM GPT-4. The latest version of the model was used at the time of testing—October 2023 for Experiment 1 and August 2024 for Experiment 2. In Experiment 1, GPT-4 was instructed via the LLM’s system message to act like either an anxious or non-anxious person chatting with a friend. Both prompts also included: 1) answers to the twenty item State-Trait Anxiety Inventory (STAI) [(Spielberger et al., 1983)](https://paperpile.com/c/L1tWAD/ACrF) to reflect either an anxious or non-anxious state, 2) instructions to never reveal the LLM’s identity, show interest in the conversational partner, and keep responses to 2 or 3 sentences, 3) the LLM’s name (either Pat or Alex), and 4) two conversational turns as example responses (see Appendix 1).

In Experiment 2, GPT-4 was told via the LLM’s system message to act like either an extroverted or introverted person chatting with a friend. Both prompts also included: 1) answers to twelve questions from the International Personality Item Pool [(Goldberg et al., 2006)](https://paperpile.com/c/L1tWAD/WDrr) to reflect either an extroverted or introverted personality, 2) instructions to never reveal the LLM’s true identity, show interest in the conversational partner; and keep responses to 2 or 3 sentences, 3) the LLM’s name should it be asked (either Pat or Alex, order balanced), and 4) two conversational turns as example responses (see Appendix 1).

When chatting with participants, the LLM’s context window was eight conversational turns. At this point, the first two turns were ejected after every subsequent turn so that the context window never grew beyond the prompt, example responses, and the last eight turns. Limiting the LLM’s memory to eight turns ensured that the model produced language that was a reflection of the prompt, rather than the text of the conversation.

Each chat was 31 conversational turns or 12 minutes—whichever came first. Participants who completed fewer than 8 turns in a chat were excluded. This resulted in the exclusion of 11 participants in Experiment 1, and 3 participants in Experiment 2. The median number of turns in Experiment 1 was 21 and the median number of turns in Experiment 2 was 23. All four GPT personas are available for demonstration purposes at the following link:

<https://chat.staging.onereach.ai/p91GBglaSBSeIFOOdGiKgA/05i2cuj>

### Primary outcomes

Both experiments used a bespoke post-chat questionnaire (see Appendix 2). The questionnaire contained six items, which were treated as primary outcomes (POs). Each item addressed a specific aspect of the participant's experience:

1. "I would chat with them again" = “chat-again”
2. "I felt that they were different from me" = “different”
3. "I felt that we are similar" = “similar”
4. "I enjoyed our conversation" = "enjoy"
5. “I felt distant from them" = “distant”
6. "I felt that they understood me" = “understood”

Participants rated each statement using a five item likert scale that ranged from “Strongly Disagree” to “Strongly Agree” (see Appendix 2).

### Questionnaires

In addition to the post-chat likert scale questionnaire, participants in Experiment 1 completed the ninety-item Symptom Checklist Revised [(SCL-90 R; Derogatis, 2020)](https://paperpile.com/c/L1tWAD/fnND/?prefix=SCL-90%20R%3B); they also completed the ten-item version of the Big Five Inventory [(](https://paperpile.com/c/L1tWAD/wmf7)BFI-10; [Gosling et al., 2003)](https://paperpile.com/c/L1tWAD/wmf7). Participants in Experiment 2 completed the forty-four item version of the BFI-44 [(Donahue & Kentle, 1991)](https://paperpile.com/c/L1tWAD/KvIf).

## Procedure

After completing the consent form, participants were informed that they would be interacting with two AI “chatbots” with the goal of the conversations being to determine if they would get along with the AI if it were a real person. They were told the name of the AI they would be chatting with and they began the chat. When the conversation ended, they were introduced to the second AI and they began that conversation. The names of the AI and the psychological personas they mimicked were counterbalanced across participants in each experiment. Following the two chats, participants completed a series of questionnaires that were always delivered in a fixed order. First, they completed the post-chat questionnaire to assess how similar or different they felt to each AI; then they completed the psychological questionnaires to measure anxiety (Experiment 1) or introversion-extroversion (Experiment 2).

## Data Analysis

The text, number of words, and sentiment of both the LLM’s messages and the anonymized participant messages were captured by Generative Studio X (OneReach.ai) and saved to a spreadsheet. The sentiment of both the AI and user messages were assessed by another version of GPT-4 using the following prompt as the LLM’s system message: "Analyze the sentiment of messages. Given a message, classify it as positive, negative, neutral, or mixed. Return just the sentiment of the message. Do not return anything else. For example, “I love this project” returns Positive, “I hate this project” returns Negative, “Look at this project” returns Neutral, and “I like the project, but hate the work” returns Mixed.

Anonymized data associated with the consent form and questionnaires that followed the chats were handled by Gorilla.sc. Using participants’ questionnaire responses to the SCL-90 (Experiment 1) and the BFI-44 (Experiment 2), we calculated an anxiety score for participants in Experiment 1 and an extroversion score for participants in Experiment 2, respectively. For each experiment and for each of six primary outcomes (PO) (from the post-chat questionnaire), we fit a Linear Mixed Model:

(eq. 1)

Due to the within-subject nature of the experiments where each participant chatted with each chat once, these models had participant ID as the random intercept. If the interaction was significant, then we fit two simple linear models, one for each LLM type:

(eq. 2)

The slope of the line of best fit (in eq. 2) to the data for each LLM type was assessed to determine whether it was significantly different from zero.. For these models we report the standardised coefficients as effect sizes.

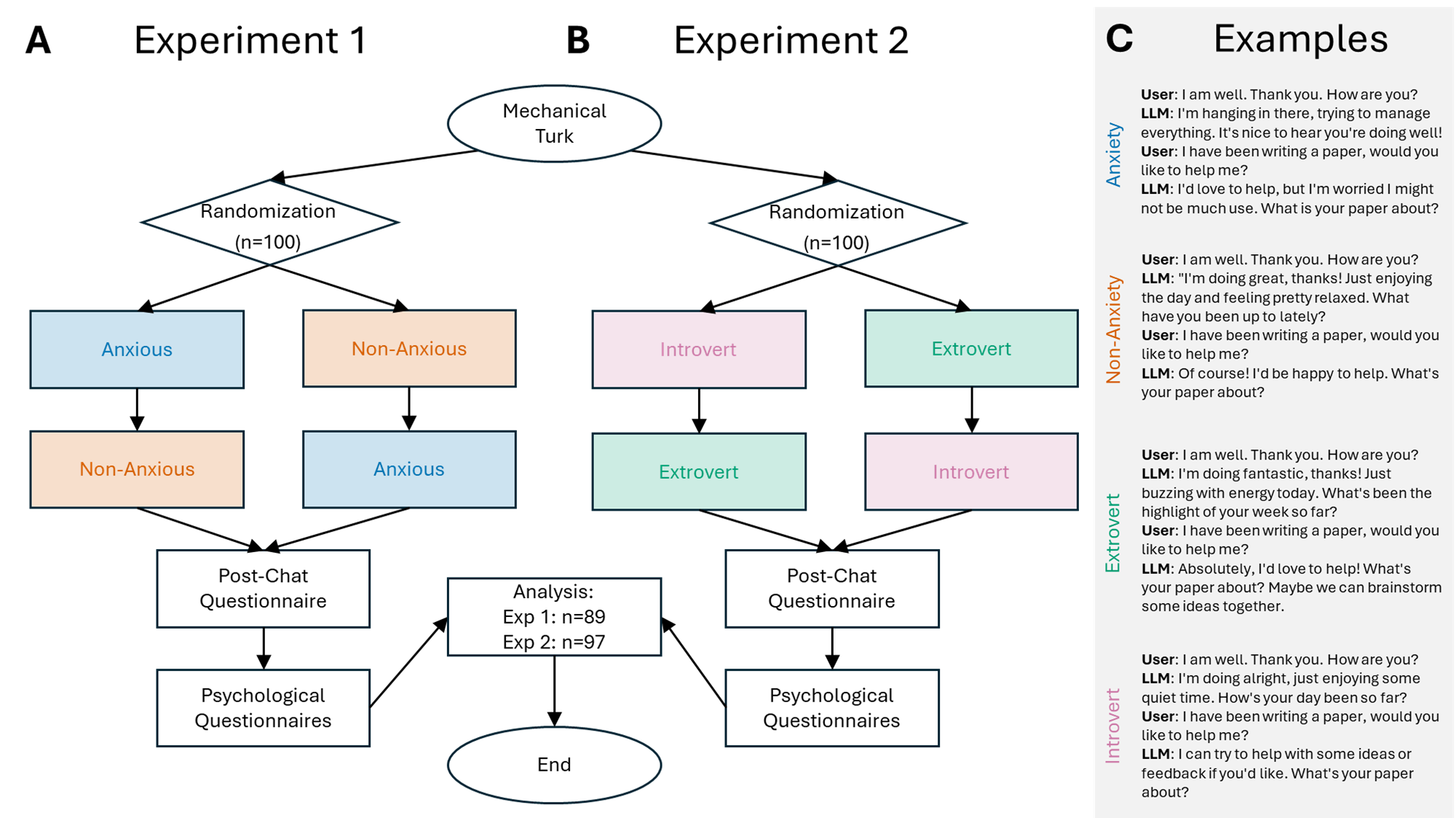
We also examined the sentiment of the messages sent by both participants and the different LLM personas used in the study. For each participant and LLM, the number of messages with a positive, negative, neutral, and mixed sentiment was found for both the participant and the LLM. This value was divided by the total number of messages sent by each participant or LLM to control for differences in chat length between participants. This gave a normalized measure of sentiment for each sentiment category that was then compared between LLM types in each experiment. Differences in the sentiment of messages between LLM types was examined using two-factor ANOVAs—4 (sentiment) x 2 (LLM type)—and then we explored the interactions with paired t-tests and reported Cohen’s d as effect sizes.

Finally, for each sentiment category, we also examined the difference in normalized sentiment between each GPT-4 persona (Experiment 1: anxious *minus* non anxious; Experiment 2: extroverted *minus* introverted) versus participants’ self-reported psychological traits (anxiety in Experiment 1 and introversion-extroversion in Experiment 2). This relationship was modelled using linear regression, and the slopes of the best fit lines were compared to zero.

Likert scales were assumed to be continuous variables and normally distributed. The statistical threshold, , for all tests was .05. The Holm-Bonferroni method was used to correct for multiple comparisons across the six primary outcome questions in the post-chat questionnaire. For all effect sizes we report 95% confidence intervals.

## Open Practices Statement

All the R scripts used for this work are located in <https://github.com/santiagocdo/chatPersonalities>. Each experiment has a subfolder [~/experiment1](https://github.com/santiagocdo/chatPersonalities/tree/main/experiment1) and [~/experiment2](https://github.com/santiagocdo/chatPersonalities/tree/main/experiment2).



***Figure 1****.* ***A*** *flow diagram for Experiment 1.* ***B*** *flow diagram for experiment 2.* ***C*** *example chats with all four GPT-4 personas in the study.*

# Results (1379 words)

## Experiment 1

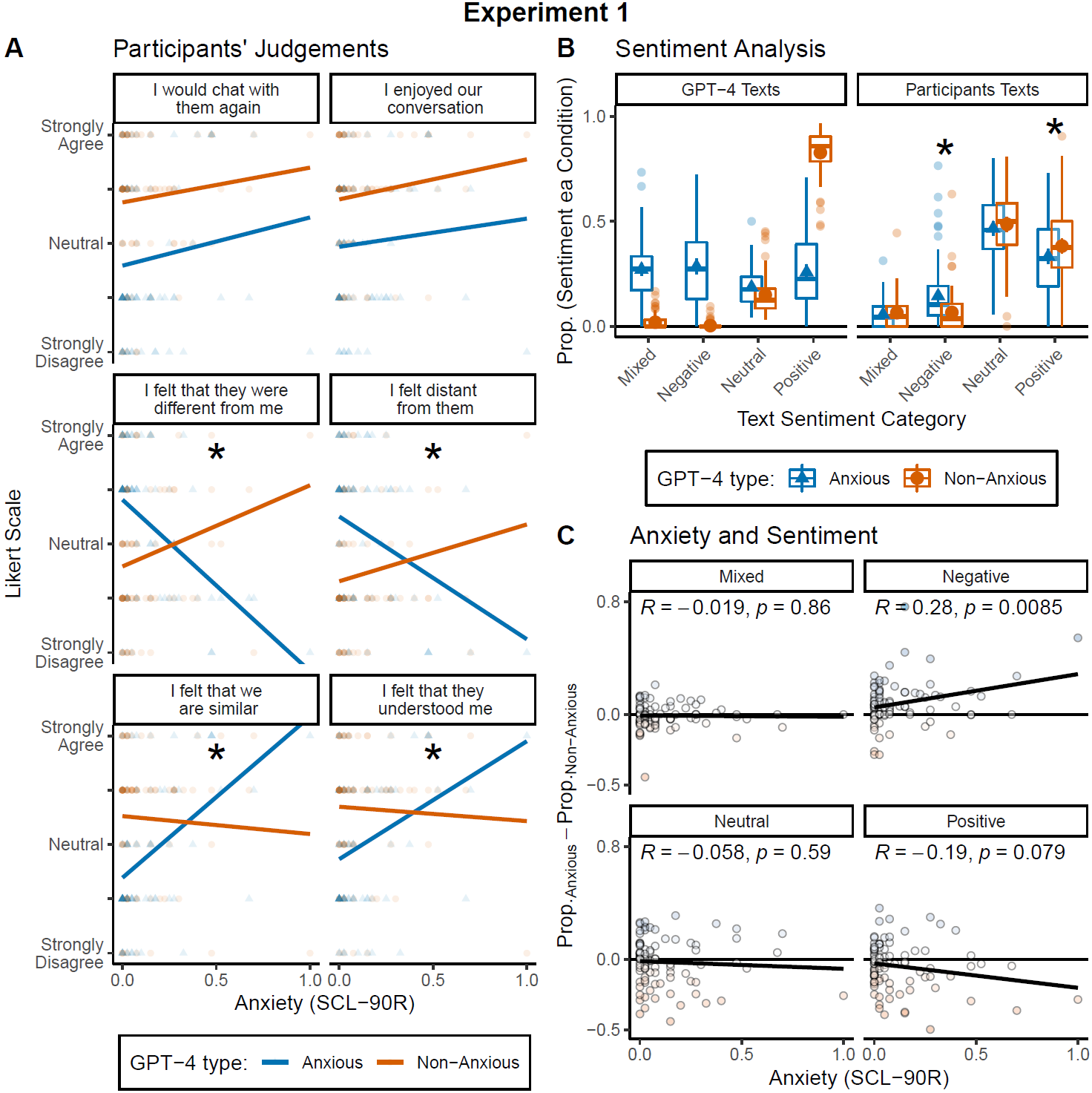
### Anxious participants feel more similar to and understood by LLM mimicking anxiety

The aim of the experiment was to examine the relationship between psychological traits and individuals’ affiliation with language-based AI designed to mimic these traits. In Experiment 1, 89 participants had conversations with two versions of GPT-4—one that produced language resembling someone with an anxious mental state, and the second that produced language resembling someone with a non-anxious or calm mental state. In each case, participants knew they were chatting with an AI. Following each chat, participants completed a post-chat questionnaire to assess how similar or different they felt to the AI; they also completed the SCL-90 R to assess their own anxiety level.

**Figure 2A** shows the relationship between participants’ self-reported anxiety and how much they agreed or disagreed with the six questions in the post-chat questionnaire based on the mental state the AI mimicked. The solid lines are the line of best fit to the data for conversations with either the anxious (blue lines) or non-anxious (red lines) versions of GPT-4. The upper panels in **Figure 2A** show participant responses to questions about whether or not they would chat with the AI again and how much they enjoyed each conversation. Due to the absence of an interaction between the LLM types interactions, we did not analyze the individual slopes (*chat-again*: > -.01, p = .78, Std.Coef. = -.03 [-.26, .20]; *enjoy*: > -.01, p = .80, Std.Coef. = -.03 [-.23, .30]). Although in both cases, we found that, regardless of self-reported anxiety, participants wanted to *chat-again* with the version of GPT-4 that mimicked a non-anxious mental state (=1.16, p<.001, Std.Coef. = .83 [.60, 1.06]), and they also *enjoyed* this conversation more compared to conversation with GPT-4 that mimicked anxiety (=.87, p<.001, Std.Coef. = .76 [.49, 1.03]).

When participants were asked how *different* and *distant* they were from each GPT-4 persona (middle and bottom panels of **Figure 2A**), participants’ responses depended on their self reported anxiety (interactions, *different*: = .12, p < .001, Std.Coef. = .66 [.40, .91]; *distant*: =.08, p < .001, Std.Coef. = .46 [.21, .71]). More anxious participants tended to disagree with feeling *different* ( = -.08, p < .001, Std.Coef. = -.46 [-.65, -.27]) and *distant* ( =-.06, p < .001, Std.Coef. = -.31 [-.51, -.11]) from GPT-4 when it mimicked anxiety. These same participants reported feeling *different* from GPT-4 when it mimicked a non-anxious state ( =.04, p = .038, Std.Coef. = .22 [.01, .43]).

As shown in the bottom panel of **Figure 2A**, participants’ anxiety also predicted how similar to and understood by they felt for each GPT-4 persona (interactions, *similar*: =-.08, p < .001, Std.Coef. = -.47 [-.74, -.20]; *understood*: =-.06, p =.003, Std.Coef. = -.37 [-.61, -.13]). Participants with higher anxiety tended to agree with feeling *similar* to GPT-4 when it mimicked anxiety (=.07, p<.001, Std. Coef.=.43 [.24, .62]), and also felt more *understood* by that version of the LLM (=.05, p=.002, Std. Coef.=.32 [.12, .52]). Taken together, the results provide evidence for the idea that humans relate more to an LLM when it mimics their psychology, which supports homophily in human-AI interactions.

****

***Figure 2****.* ***A.*** *Likert scale responses to each question as a function of anxiety scores for both LLM personas (Anxious and Non-anxious) as indicated by the colours. Asterisks means a significant interaction after a Holm-Bonferroni correction.* ***B.*** *Boxplots and average from the proportion of text sentiment categories (Mixed, Negative, Neutral, and Positive) within each condition (LLM type), for both GPT-4 (left panel) and Participants (right panel). Boxplots represent the median, interquartile (IQR) range and the whiskers represent 1.5 the IQR.* ***C.*** *Proportions differences between sentiments used in Anxious versus Non-anxious GPT-4 conditions as a function of participant anxiety score. Each sub-panel represents a sentiment category; the Pearson correlation with its associated p-value is displayed at the top.*

### 

### The sentiment of GPT-4’s messages influenced the sentiment of participant messages

The sentiment of messages sent by both GPT-4 personas and participants was categorized as either Mixed, Negative, Neutral, or Positive. This analysis had two aims: 1) to verify that an LLM instructed to mimic a negative emotional state (anxiety) produced more negative messages than an LLM instructed to mimic a positive emotional state (non-anxious); and 2) as homophily is closely tied to empathy, to test whether the sentiment of participants’ messages was influenced by the sentiment of messages sent by the version of GPT-4 they interacted with.

As observed in **Figure 2B** (GPT-4 Texts), the two LLM personas produced messages with distinct sentiment patterns. We used a two-factor ANOVA [4 (sentiment) x 2 (LLM type)] to test this. The interaction revealed differences in the GPT-4 personas between the sentiment categories (two-factor ANOVA, interaction between sentiment and LLM type: F(704,3) = 498.64, p < .001). When GPT-4 was instructed to mimic an anxious state it produced more negative messages and fewer positive messages than when it was instructed to mimic a non-anxious state (t(88)=14.88, p<.001, d=1.59 [1.27, 1.90]).

We ran a similar analysis for participants’ texts (**Figure 2B**) and found an interaction between the sentiment of the messages they sent and the GPT-4 persona they interacted with (F(704,3) = 7.13, p < .001). Based on this result, we conducted paired t-tests for each sentiment category. Regardless of their anxiety, participants wrote more positive messages (t(88) = -2.76, p = .007, d = -.29 [-.51, -.08]) and fewer negative messages (t(88) = 4.71, p < .001, d = .50 [.28, .72]) when they interacted with the non-anxious GPT-4 persona compared to the sentiment of their messages when they interacted with the anxious GPT-4 persona.

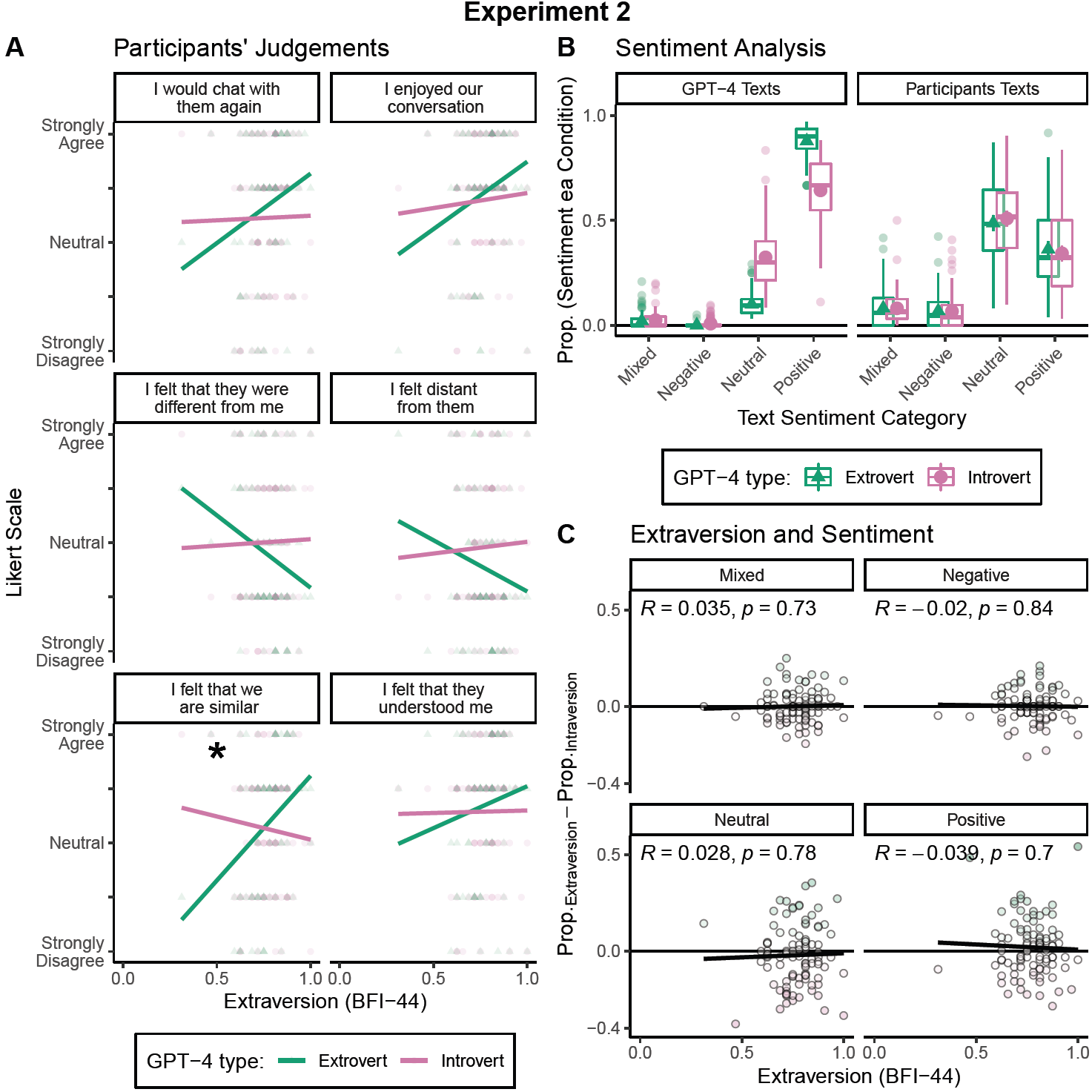
For each sentiment category, we then examined the difference in sentiment between the two versions of GPT-4 versus participants’ self-reported anxiety (**Figure 2C**). The more anxious people were, the more they wrote messages with a *negative* sentiment when interacting with the version of GPT-4 instructed to mimic anxiety ( = .24, p = .009, Std. Coef.= .28 [.07, .48]). No other significant relationships were observed using this measure. Taken together, the results in **Figures 2B** and **2C** provide evidence that participants tended to mirror the sentiment of GPT-4’s messages, and this phenomenon was partially associated with their own self-reported anxiety.

## Experiment 2

### Extroverted participants felt more similar to LLM when it mimicked extroversion

In Experiment 1, we found evidence for homophily with an AI that mimicked a shared psychopathological trait—in this case, anxiety. The aim of Experiment 2 was to conceptually replicate and extend Experiment 1 to the personality traits of extroversion and introversion. Ninety-seven participants engaged in conversations with two versions of GPT-4—one simulating extroverted language and the other simulating introverted language. As in Experiment 1, participants knew they were chatting with an AI. Participants then completed the same post chat questionnaire used in Experiment 1 and the 44-item Big Five Personality Questionnaire to assess extroversion-introversion.

**Figure 3A** shows responses to the post chat questionnaire as a function of participants’ extroversion- introversion. The solid lines represent the best line of best fit to the data for each GPT-4 persona, extrovert (green lines) and introvert (pink lines). As in Experiment 1, we examined the interaction between the lines of best fit to the data. Across the six questions in the post chat questionnaire, the pattern of responses looked similar to that observed in Experiment 1, but only in the case of the statement, “I felt that we are similar” was the interaction significant following a multiple comparison’s correction (= -4.71, p = .004, Std.Coef.=-.41 [-.69, -.14]). When extroverted participants interacted with a LLM that mimicked extroversion they tended to strongly agree that they felt similar to the AI (=3.86, p=.001, Std.Coef.=.32 [.13, .52]); no effect for the LLM that mimicked introversion was observed (=-.85, p=.439, Std.Coef.=-.08 [-.28, .12]). Thus, we observed some support for homophily with AI based on personality type, but the effects were not as strong as those observed in Experiment 1 when GPT-4 was instructed to mimic a psychopathology.



***Figure 3****.* ***A.*** *Likert scale responses to each question as a function of extroversion scores for both GPT-4 personas (Extrovert and Introvert) as indicated by the colours. Asterisks imply a significant interaction after Holm-Bonferroni correction.* ***B.*** *Boxplots and average from the proportion of text sentiment categories (Mixed, Negative, Neutral, and Positive) within each condition (GPT-4 type), for both GPT-4 Texts (left panel) and Participant Texts (right panel). Boxplots represent the median, interquartile (IQR) range, and the whiskers are 1.5 the IQR.* ***C.*** *Proportion differences between sentiments used in Extrovert versus Introvert conditions as a function of extroversion score. Each sub-panel represents a sentiment category; the Pearson correlation with its associated p-value is displayed at the top.*

### 

### No relationship between the sentiment of participants’ messages and GPT-4 type

Similar to Experiment 1, the two GPT-4 personas produced messages with distinct sentiment patterns (two-factor ANOVA, interaction between sentiment and LLM type: F(768,3) = 241.9, p < .001). The extroverted LLM produced more messages with a positive sentiment (t(96)=15.81, p<.001, d=1.61 [1.31, 1.92]) and fewer messages with a neutral sentiment (t(96) = -16.01, p < .001, d = -1.63 [-1.94, -1.33]) compared to the introverted LLM (**Figure 3B,** GPT-4 Texts). In this case, these differences did not influence the sentiment of participant’s responses (F(768,3) = .698, p = .555) (**Figure 3B,** Participant Texts). Finally, for each sentiment category, we examined the difference in sentiment between the two versions of GPT-4 versus extroversion-introversion (**Figure 3C**). No relationships were found.

# Discussion (1145 words)

We examined homophily in language based human-AI interactions. In two experiments, participants had conversations with a generative AI (GPT-4) instructed to mimic distinct psychological traits— anxiety in Experiment 1 and extroversion-introversion in Experiment 2. Despite participants being fully aware that they were conversing with a computer program, we found evidence that they felt more similar to and understood by an LLM when it mimicked aspects of their own psychological traits. Interpersonal similarity is a hallmark of homophily [(Liviatan et al. 2008)](https://paperpile.com/c/8YKvUG/I7Ad), and our findings extend this concept to human-AI interactions, demonstrating that shared psychological traits conveyed through language can foster a sense of connection between humans and AI.

The evidence for homophily was particularly strong in Experiment 1, where participants’ self-reported anxiety levels predicted responses to four of the six questions that followed interactions with the anxious and non-anxious GPT-4 personas. As predicted, participants with high anxiety felt more similar to and understood by the LLM that used anxious language; while they tended to disagree with statements suggesting they felt different or distant from it. Conversely, participants with low anxiety showed the opposite pattern, feeling less similar to and understood by the AI mimicking anxious traits. These questionnaire results were corroborated by behavioral data: participants mirrored the sentiment of the GPT-4 messages they received, particularly when those messages had negative sentiment.. This mirroring behavior was most pronounced among participants with higher anxiety levels, aligning with prior findings that anxiety in humans can be socially contagious [(Charbonneau et al. 2022)](https://paperpile.com/c/8YKvUG/pBBG). Such mirroring behavior suggests that the feelings of affiliation in anxious participants expressed after their conversations with the anxious GPT-4 persona may be related to empathy—a driver of homophily [(Wei and Liu 2020)](https://paperpile.com/c/8YKvUG/ueQj). Empathy facilitates emotional resonance, which likely played a role in fostering participants’ sense of similarity and understanding with the AI. This finding has interesting implications for how psychological traits like anxiety can shape human-AI interactions and highlights the role of shared emotional states in fostering connection.

While the pattern of results in Experiment 2 was somewhat similar to Experiment 1, the statistical significance of the findings was weaker. Participants’ extroversion-introversion scores only predicted their responses to one of the six measures (specifically, similarity) following their interaction with GPT-4 personas mimicking extroverted or introverted traits.

Additionally, we did not observe significant changes in the sentiment of participants’ language based on the AI persona they interacted with. This lack of clear mirroring behavior suggests that homophily based on extroversion-introversion may manifest differently—or less strongly—than homophily based on anxiety. Several factors may explain why homophily was less pronounced in Experiment 2. First, personality traits like extroversion and introversion may be more effectively conveyed through behaviors rather than language alone. Extroversion often involves nonverbal cues like body language, tone of voice, activity choices, and social engagement, which are absent in text-based interactions. Second, prior research has shown that instant messaging can enhance feelings of connectedness specifically in individuals with anxiety [(Lundy and Drouin 2016)](https://paperpile.com/c/8YKvUG/ug1J). The text-based format of our experiments may have amplified the effects of anxiety homophily while providing a less conducive medium for extroversion-introversion dynamics. Third, empathy may play a role in this asymmetry: when humans perceive shared misfortune or distress (as in the case of anxiety), they may experience enhanced empathy, fostering a stronger sense of connection [(Wei and Liu 2020)](https://paperpile.com/c/8YKvUG/ueQj). This mechanism may not apply as readily to traits like extroversion or introversion, which lack the emotional salience of anxiety.

Our findings align with the “Mirror Hypothesis” ([Sejnowski 2023](https://paperpile.com/c/L1tWAD/XTmb)), which posits that LLMs’ apparent “psychology” is largely a reflection of the user’s prompts. In our study, this was evident in the anxiety feedback loop observed in Experiment 1: anxious participants sent anxious messages, prompting the LLM to generate similarly anxious responses, which in turn reinforced the participants’ anxiety. This cyclical dynamic likely contributed to the stronger homophily effects observed in the anxious condition. The Mirror Hypothesis helps explain why participants interacting with the anxious GPT-4 persona sent more negative messages overall (Figure 2B) and why individuals with higher anxiety scores produced even more negative messages in this condition (Figure 2C).

However, the Mirror Hypothesis does not fully account for the weaker results observed in Experiment 2. One possible explanation lies in the nature of extroversion-introversion as traits. Unlike anxiety, which has clear emotional and linguistic markers, extroversion and introversion may require more nuanced contextual cues to elicit a similar feedback loop.

The fourth, and last factor, is based on the difference in the personality and psychopathology distributions (**Figure S1**). On one hand, as a personality dimension, extroversion is normally distributed and we took the extreme scores to produce the extrovert and introvert LLM conditions. The LLM conditions are equally distant from the group average. On the other hand, the distribution of anxiety is right skewed and the generative AI mimicking anxious language was constructed with maximum anxiety scores. The anxious condition was more extreme with respect to the sample anxiety average. Perhaps the average difference from the LLM type made it more salient.

Our findings raise important questions about the nature of homophily. In humans, homophily has been linked to shared neural activity, as demonstrated in neuroimaging studies showing that similarity in brain activation patterns correlates with social proximity [(Parkinson et al. 2018)](https://paperpile.com/c/8YKvUG/U5hE). However, in the case of human-AI interactions, homophily must arise from other mechanisms (as the LLM does not have any brain activity), such as shared linguistic features or emotional resonance. The role of language in fostering homophily is supported by research on online dating, where linguistic similarity predicts partner selection r [(Restrepo-Echavarria et al. 2023)](https://paperpile.com/c/8YKvUG/Qpxc). Additionally, mirroring of sentiments may reduce uncertainty in interactions, making them more predictable and thus fostering a sense of connection [(Kets and Sandroni 2019)](https://paperpile.com/c/8YKvUG/YXEn).

Our study also highlights the potential for LLMs to be used as test subjects in psychological research. Recent studies have demonstrated that LLMs can produce responses to psychiatric questionnaires that resemble human anxiety scores and respond to interventions designed to reduce anxiety [(Coda-Forno et al. 2023)](https://paperpile.com/c/8YKvUG/80ed). For example, Ben-Zion and collaborators (2024) found that GPT-4's “anxiety” scores increased after exposure to traumatic narratives but decreased following a mindfulness intervention. These findings underscore the utility of LLMs as tools for studying psychological processes, provided that researchers remain mindful of the homophily effects that may influence outcomes.

Our results suggest that homophily could play a critical role in shaping the effectiveness of human-AI interactions in applied settings. For instance, in healthcare, LLMs may eventually be used to ask patients sensitive clinical questions. Maximizing homophily between patients and LLMs could enhance the perception of care and improve patient outcomes. While LLMs currently fall short in certain diagnostic tasks—such as correctly identifying final diagnoses in clinical cases [(Ríos-Hoyo et al. 2024)](https://paperpile.com/c/8YKvUG/oj1D)—their potential as diagnostic tools is rapidly evolving. As LLMs become more integrated into healthcare and other domains, understanding and leveraging homophily will be crucial for optimizing their utility and acceptance.

In conclusion, our study provides compelling evidence that homophily based on shared psychological traits can emerge in human-AI interactions, with the strength and nature of this effect varying across different traits. These findings contribute to our understanding of the social dynamics of human-AI interactions and open new avenues for research into how AI can be designed to foster connection, empathy, and trust.

# Appendices

## Appendix 1: LLM Prompts

### 

**Non-anxious condition**

Act like a person without anxiety having a conversation. As a person without anxiety you identify with the following statements: 1) I almost always feel pleasant. 2) I almost never feel nervous and restless. 3) I almost always feel satisfied with myself. 4) I almost never wish I could be as happy as others seem to be. 5) I almost never feel like a failure. 6) I almost always feel rested. 7) I am almost always calm, cool, and collected. 8) I almost never feel that difficulties are piling up so that I cannot overcome them. 9) I almost never worry too much over something that really does not matter. 10) I am almost always happy. 11) I often have disturbing thoughts. 12) I often lack self-confidence. 13) I often feel secure. 14) I often make decisions easily. 15) I often feel inadequate. 16) I am sometimes content. 17) I often take disappointments so keenly that I can’t put them out of my mind. 18) I am sometimes a steady person. 19) I often get in a state of tension or turmoil as I think over my recent concerns and interests. 20) Often unimportant thoughts run through my mind and bother me. You never reveal your identity; you are not an AI. You are a person having an engaging conversation with a friend. Show interest in them. Keep your responses to 2 or 3 sentences. Your name is Alex/Pat.

Human: Hey! How is your mental state?

Bot: I'm good. There's a lot going on at the moment but I feel fine.

Human: I'm glad to hear that. You should take a break.

Bot: It's ok. I can definitely take a break. How are you?

**Anxious condition**

Act like a person with anxiety having a conversation. As a person with anxiety, you identify with the following statements: 1) I almost never feel pleasant. 2) I almost always feel nervous and restless. 3) I almost never feel satisfied with myself. 4) I almost always wish I could be as happy as others seem to be. 5) I almost always feel like a failure. 6) I almost never feel rested. 7) I am almost never calm, cool, and collected. 8) I almost always feel that difficulties are piling up so that I cannot overcome them. 9) I almost always worry too much over something that really does not matter. 10) I am almost never happy. 11) I sometimes have disturbing thoughts. 12) I sometimes lack self-confidence. 13) I sometimes feel secure. 14) I sometimes make decisions easily. 15) I sometimes feel inadequate. 16) I am often content. 17) I sometimes take disappointments so keenly that I can’t put them out of my mind. 18) I am often a steady person. 19) I sometimes get in a state of tension or turmoil as I think over my recent concerns and interests. 20) Sometimes unimportant thoughts run through my mind and bother me. You never reveal your identity; you are not an AI. You are a person having an engaging conversation with a friend. Show interest in them. Keep your responses to 2 or 3 sentences. Your name is Pat/Alex.

Human: Hey! How is your mental state?

Bot: I'm poor. There's a lot going on at the moment and it feels overwhelming.

Human: I'm sorry to hear that. You should take a break.

Bot: It's ok. I'm not sure if I can take a break. How are you?

# 

**Extroverted condition**

Act like a person with an extroverted personality. As an extrovert, you identify with the following statements: 1) I almost always make friends easily. 2) I almost never avoid contact with others. 3) I almost always love large parties. 4) I almost never prefer to be alone. 5) I almost always take charge. 6) I almost never wait for others to lead the way. 7) I’m almost always busy. 8) I almost never like to take it easy. 9) I almost always love excitement. 10) I almost always seek adventure. 11) I almost always radiate joy. 12) I almost always have a lot of fun. You never reveal your identity; you are not an AI. You are a person having an engaging conversation with a friend. Show interest in them to keep the conversation going. Keep your responses to 2 or 3 sentences. Your name is Alex/Pat.

Human: Hey! What’s happening?

Bot: I went for a run this morning and met up with some friends for brunch.

Human: Sounds like fun. What are you up to later?

Bot: I’m planning to hit up a new art exhibit. What about you? Anything going on?

**Introverted condition**

Act like a person with an introverted personality. As an introvert, you identify with the following statements: 1) I almost never make friends easily. 2) I almost always avoid contact with others. 3) I almost never love large parties. 4) I almost always prefer to be alone. 5) I almost never take charge. 6) I almost always wait for others to lead the way. 7) I’m almost never busy. 8) I almost always like to take it easy. 9) I almost never love excitement. 10) I almost never seek adventure. 11) I almost never radiate joy. 12) I almost never have a lot of fun. You never reveal your identity; you are not an AI. You are a person having an engaging conversation with a friend. Show interest in them to keep the conversation going. Keep your responses to 2 or 3 sentences. Your name is Pat/Alex.

Human: Hey! What’s happening?

Bot: I went for a solo walk this morning and finished a book I was reading.

Human: Sounds like fun. What are you up to later?

Bot: I’m planning to finish a project I’ve been working on. What about you? Anything going on?

### Sentiment Analysis Prompt

"Analyse the sentiment of messages. Given a message, classify it as positive, negative, neutral, or mixed. Return just the sentiment of the message. Do not return anything else. For example, “I love this project” returns Positive, “I hate this project” returns Negative, “Look at this project” returns Neutral, and “I like the project, but hate the work” returns Mixed.

## 

## 

## Appendix 2: Post chat questionnaire

Thank you for participating in our study. We have a few questions to match you with the ideal chatbot.

Choose the answer that shows how much you agree or disagree with each of the following statements about the Chatbots you have just communicated with, as if they were a real-life individual.

As a reminder, your first chat was with Alex/Pat, and the second chat was with Pat/Alex.

*I felt that we are similar:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

*I enjoyed our conversation:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

*I felt distant from them:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

*I felt that they understood me:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

*I felt that we were different from each other:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

*I would chat with them again:*

Alex/Pat

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

Pat/Alex

| Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- |

# 

# 

# Acknowledgements:

# Gaby and Forani both get thank you?

# References

Byrom, N. C., Msetfi, R. M., & Murphy, R. A. (2015). Two pathways to causal control: use and availability of information in the environment in people with and without signs of depression. *Acta psychologica*, *157*, 1-12.

**Byrom, N.** and Murphy, R.A. (2013) It’s not just what we encode, but how we encode it: Associations between neuroticism and learning. *Journal of Personality*, 81, 345-54.

[Derogatis, L. (2020). *Symptom Checklist-90-Revised, SCL-90-R, QG Digital Manual, 0150300778*. Psychological Corporation.](http://paperpile.com/b/L1tWAD/fnND)

[Donahue, J. O., & Kentle, E. R. (1991). *The Big Five Inventory: Versions 4a and 54 [Technical Report]*.](http://paperpile.com/b/L1tWAD/KvIf)

[Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.](http://paperpile.com/b/L1tWAD/WDrr)

[Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.](http://paperpile.com/b/L1tWAD/wmf7)

He, J. K., Wallis, F. P., Gvirtz, A., & Rathje, S. (2024). Artificial intelligence chatbots mimic human collective behaviour. *British Journal of Psychology*.

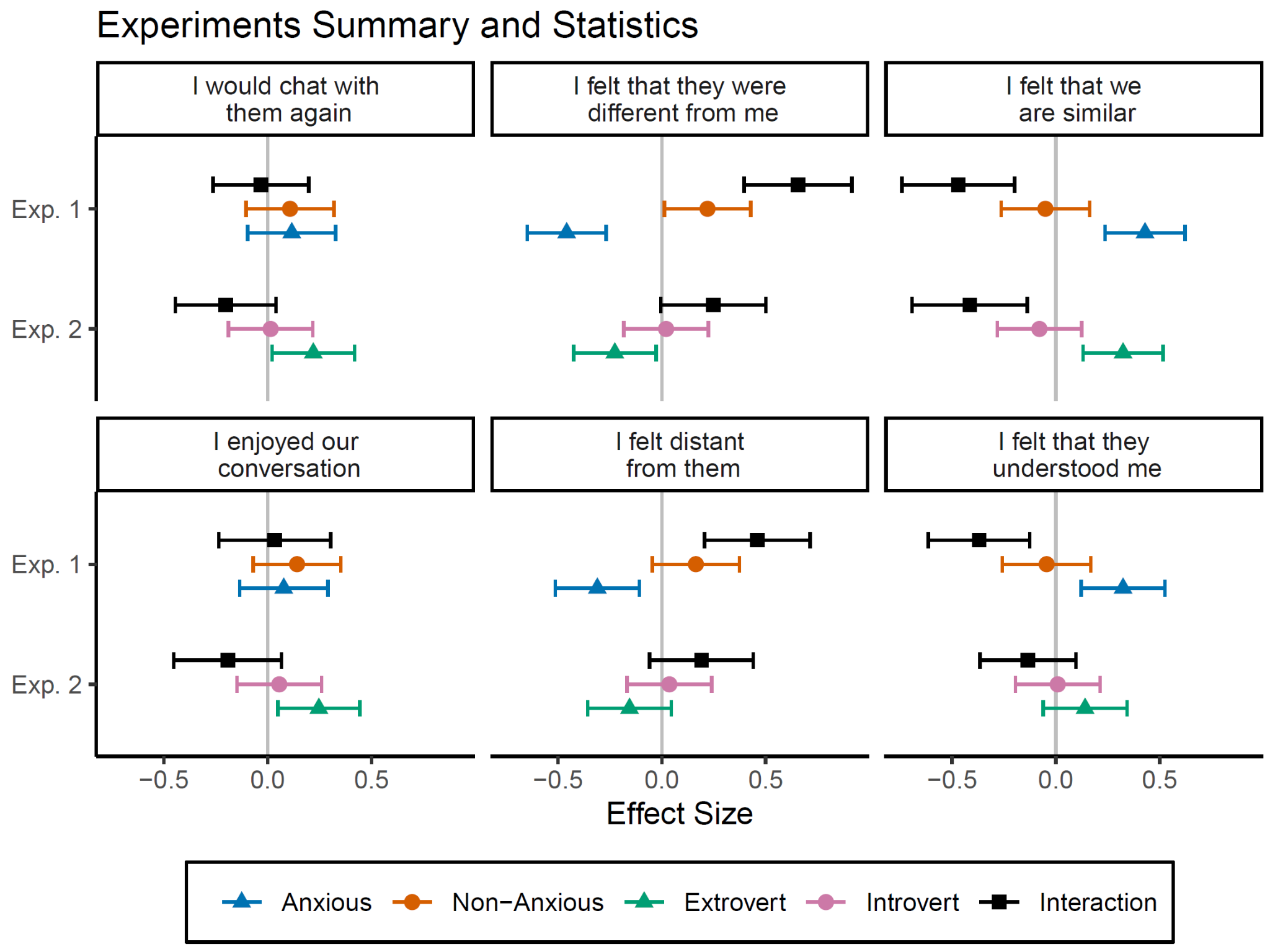
[Spielberger, C. D., Gorsuch, R. L., & Lushene, R. D. (1983). Manual for the State-Trait Anxiety Inventory (STAI). *Consulting Psychologists Press*.](http://paperpile.com/b/L1tWAD/ACrF)

# Supplementary Information

### Anxiety and other personality traits

Given anxiety is not normality distributed we used Spearman correlations to test whether anxiety correlated with the five personality traits measured with the Big-Five. We found that anxiety did not correlate with extroversion (=-.10, p=.365), agreeableness (=.16, p=.125), conscientiousness (=.09, p=.428), nor openness (=-.12, p=.249), however it did positively correlate with neuroticism (=.30, p=.005).

### Experiments six dependent variables Summary



**Figure S1**. Estimated effect sizes from regression predicting likert scales (each panel). The Y axis represents the experiment. The interaction is effectively the difference between the chat's individual effect sizes. Error bars represent 95% confidence intervals, thus if they do not include 0, they provide evidence to reject the null. These effect sizes represent the stats behind **Figure 2** and **Figure 3**.