Web Appendix

AI Companions Reduce Loneliness

Julian De Freitas

Zeliha Oğuz-Uğuralp

Ahmet Kaan Uğuralp

Stefano Puntoni

Table of Contents

SUMMARY TABLE FOR HYPOTHESES **4**
TABLE S1 ..... *4*
STUDY S1: PRE-STUDY OF TECH SOLUTIONS FOR LONELINESS **4**
TABLE S2 ..... *5*
STUDY S2: REPLICATION OF UNDERESTIMATING THE SOCIAL IMPACT OF AI COMPANIONS **5**
FIGURE S1 ..... *6*
FIGURE S2 ..... *7*
STUDY S3 ..... **8**
PILOT STUDY 1 ..... **10**
PILOT STUDY 2 ..... **11**
STUDY 1 ..... **12**
LLM TRAINING FOR LONELINESS CLASSIFICATION IN CONVERSATION AND REVIEW DATA 13
FIGURE S3 ..... *19*
FINE-TUNING HYPERPARAMETERS 22
TABLE S3 ..... *22*
TABLE S4 ..... *24*
STUDY 2 ..... **24**
REPLICATION OF THE RESULTS WITHOUT COMPREHENSION EXCLUSIONS IN STUDY 2 24
FIGURE S4 ..... *25*
RESULTS OF THE REMAINING METRICS IN STUDY 2 26
FIGURE S5 ..... *26*
TABLE S5 ..... *27*
TABLE S6 ..... *28*
TABLE S7 ..... *28*
TABLE S8 ..... *28*
TABLE S9 ..... *28*
REPLICATION WITH LONELINESS ITEMS ONLY, INSTEAD OF AS A COMPOSITE 29
REPLICATION WITH SOCIAL CONNECTION ITEMS ONLY, INSTEAD OF AS A COMPOSITE 29
TESTING ATTITUDES TOWARD AI AS A MODERATOR FOR EXPECTATION VIOLATION 30
LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 2 30
STUDY 3 ..... **32**
TABLE S10 ..... *32*
TABLE S11 ..... *32*
TABLE S12 ..... *32*
TABLE S13 ..... *33*
TABLE S14 ..... *33*
PROPENSITY SCORE MATCHING 34
TABLE S15 ..... *34*
FIGURE S6 ..... *36*
FIGURE S7 ..... *36*
REPLICATION OF STUDY 3 RESULTS AFTER PROPENSITY SCORE MATCHING 37
FIGURE S8 ..... *39*
REPLICATION OF STUDY 3 RESULTS INCLUDING ALL PARTICIPANTS 40
FIGURE S9 ..... *42*

BASE MODEL PROMPT SENT TO GPT-4 FOR GETTING MESSAGE RESPONSES 42
LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 3 43

STUDY 4 43

TABLE S16 43
PROMPT FOR THE CHATBOT IN 'CONTROL' CONDITION 44
PROMPT FOR THE CHATBOT IN 'GENERALIST AI' CONDITION 45
LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 4 45

STUDY 5 46

FULL METHODOLOGICAL DETAILS IN STUDY 5 46
TABLE S17 47
REPLICATION OF RESULTS AFTER EXCLUDING PARTICIPANTS WHO CORRECTLY SUSPECTED THE STUDY PURPOSE 48

REFERENCES 48

SUMMARY TABLE FOR HYPOTHESES

TABLE S1

SUMMARY TABLE MAPPING THE HYPOTHESES ONTO THE STUDIES

Hypothesis	Description	Tested In
H1	Interacting with AI companions alleviates feelings of loneliness.	Studies 2-5
H2	Interacting with AI companions produces consistent momentary reductions in loneliness after each use, over multiple days.	Study 3
H3a	Feeling heard mediates the effect of interacting with AI companions on reducing loneliness.	Study 4
H3b	Feeling heard is a stronger mediator than communication performance in the effect of AI companions on alleviating loneliness.	Study 4
H4	AI companions alleviate loneliness more effectively than activities that primarily involve self-disclosure or distraction.	Study 5
H5	Consumers underestimate the loneliness-alleviating benefits of interacting with AI companions.	Studies 2, 3

STUDY S1: PRE-STUDY OF TECH SOLUTIONS FOR LONELINESS

In this study, we explored which technological solutions participants typically turn to in order to combat loneliness.

Method

We recruited 51 participants from Amazon Mechanical Turk and excluded 9 for failing the comprehension check (explained below), leaving 42 participants (40% female, M_{age} = 39). Participants were paid $1.00 USD each. They were told: “When you are feeling lonely, what are five ways that you use technology to cope with loneliness? Please write your answers in the five

text boxes below". Next, they completed a comprehension check about the question they were asked, then provided demographic info.

Results

Table S2 shows the number of participants who wrote each coping solution. Social media was the most popular choice, followed by gaming, watching movie/TV/Netflix, music, and texting.

TABLE S2

MOST AND LEAST POPULAR CHOICES

Most Popular	**Least Popular**
Social media (36 in total, including YouTube, Facebook, Instagram, Reddit, Snapchat, TikTok, Twitter)	Browsing the web (2)
Gaming (26)	Email (2)
Watching movie/TV/Netflix (22)	Taking pictures (2)
Music (21)	VR (2)
Texting (20)	Drawing (1)
Calling (15)	Investment (1)
Reading books (6)	Journaling (1)
Listening to podcasts (4)	Browsing Pictures (1)

NOTE.— Numbers in parentheses shows the number of participants who wrote that choice.

STUDY S2: REPLICATION OF UNDERESTIMATING THE SOCIAL IMPACT OF AI COMPANIONS

Study S2 explored the possibility that participants underestimate the social impact of AI companions relative to how they truly feel after interacting with this technology.

Method

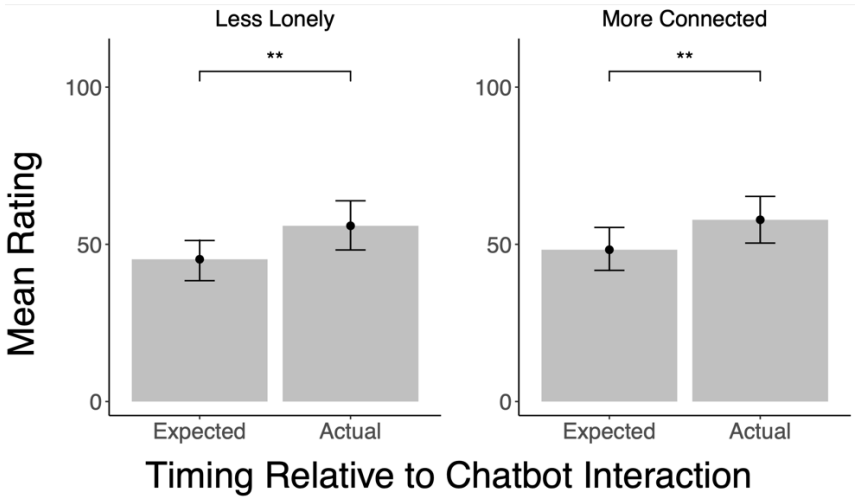
We recruited 99 participants from Amazon Mechanical Turk and excluded 42 (see exclusion criteria below), leaving 57 (M_{age} = 37, 60% females). Participants were paid $2.50 each. 5% had prior experience with AI companion apps. We ran this experiment on 2.9.2023 which was after the release of ChatGPT. The design was like study 2, except participants only interacted with an AI chatbot, and we only measured changes in how participants felt relative to their expectations before interacting with the chatbot, rather than also measuring changes in state loneliness.

Results

Compared to their expectations, participants felt less lonely ($M_{Expected}$ = 45.25 vs. M_{Actual} = 55.89, $t(56) = -2.77$, $p = .008$, $d = -0.38$), more connected ($M_{Expected}$ = 48.28 vs. M_{Actual} = 57.81, $t(56) = -2.97$, $p = .004$, $d = -0.35$), and more comfortable ($M_{Expected}$ = 61.65 vs. M_{Actual} = 73.33, $t(56) = -3.67$, $p < .001$, $d = -0.47$)— figure S1. There were no statistical differences for the other items—figure S2.

FIGURE S1

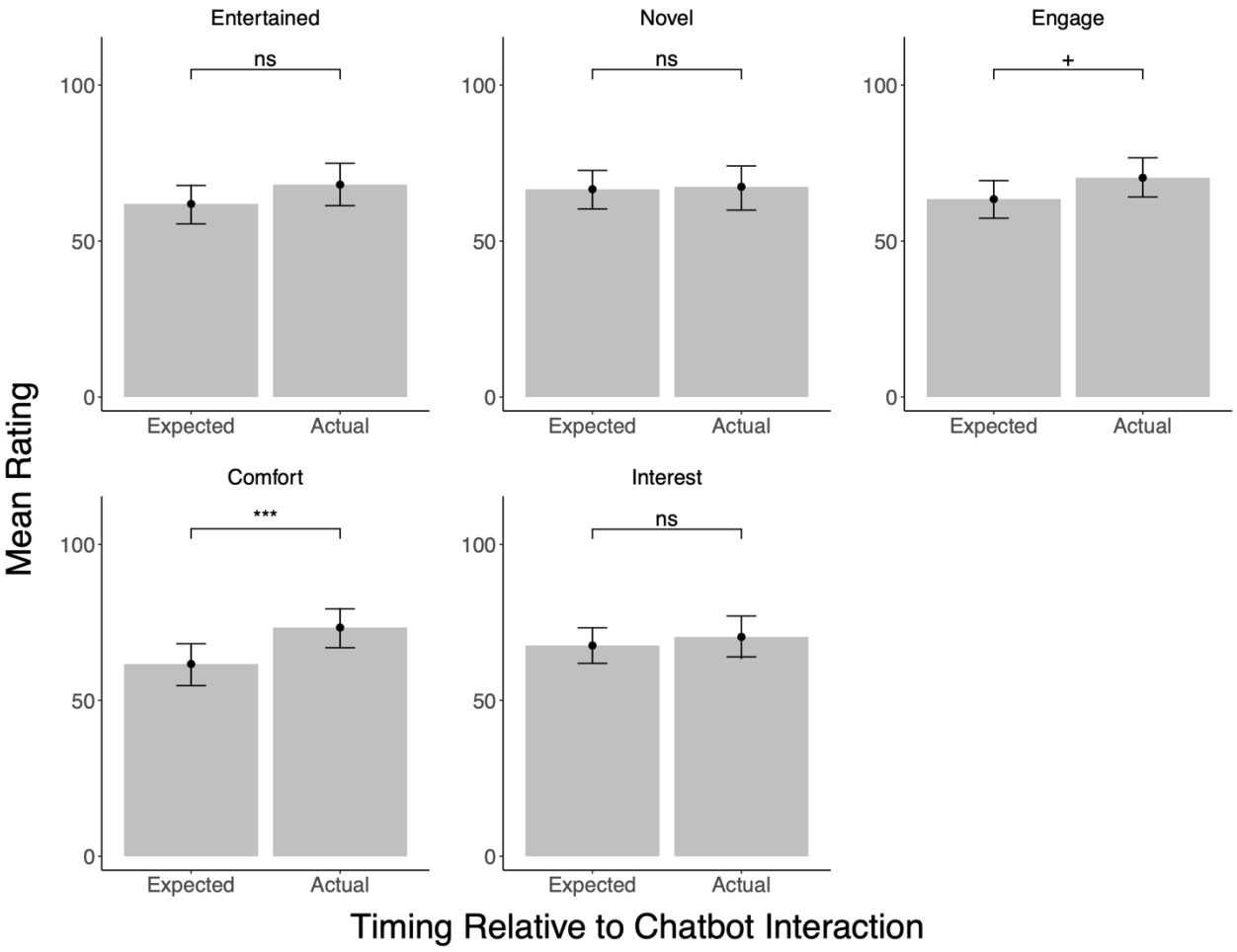
RESULTS OF LESS LONELY AND MORE CONNECTED IN STUDY S2



NOTE.— Horizontal lines reflect results of independent-sample t-tests. *** $p < .001$; ** $p < .01$; + $p < .1$; ns not significant. Error bars reflect 95% confidence intervals.

FIGURE S2

RESULTS OF REMAINING METRICS IN STUDY S2



NOTE.— Horizontal lines reflect results of independent-sample t-tests. *** $p < .001$; + $p < .1$; ns not significant. Error bars reflect 95% confidence intervals.

Conclusion

Participants underestimated how much less lonely and more connected they would feel as a result of interacting with an AI companion.

STUDY S3

To further test the robustness of our findings, we conducted a final study where we asked participants to complete only the loneliness measure after interacting with the AI companion. We ran a simpler version of study 4 that only included the AI companion and control conditions.

Methods and Results

This study was pre-registered (https://aspredicted.org/H37_GL2). We recruited 776 participants from CloudResearch Connect and excluded 63 for failing a comprehension question, leaving 713 (M_{Age} = 36.8, 56.8% Females). We aimed to hire 400 participants in both conditions and participants were randomly assigned to control or AI companion conditions. 54.1% had prior experience with AI companion apps. We ran this experiment on May 23, 2024. All participants were paid $2.75 USD. The study and chatbot design were the same as in study 4, except that (1) we removed the loneliness questions before the interaction with the chatbot, (2) removed the feeling heard and chatbot performance questions; and (3) removed the 'AI assistant' condition.

As in study 4, we found that loneliness was significantly lower in the AI companion condition compared to the control condition ($M_{AI\ Companion}$ = 25.62 (26.13); $M_{Control}$ = 36.91, (29.93); $t(698.8) = 5.37$, $p < .001$, $d = 0.40$), confirming the robustness of our findings.

PILOT STUDY 1

In pilot study 1, we compare levels of self-disclosure between journaling and interacting with an AI companion. The study employed a between-subjects design with two conditions: journaling and interacting with the same AI companion as in study 5.

Methods and Results

This study was pre-registered (https://aspredicted.org/z6d8-mdsp.pdf). We recruited 100 participants from CloudResearch Connect (M_{Age} = 41.4, 55% Females). We aimed to hire 50 participants in both conditions and participants were randomly assigned to control or AI companion conditions. 69% had prior experience with chatbot apps. We ran this experiment on December 21, 2024. All participants were paid $3 USD. The study and chatbot design were the same as in study 4, except that participants answered two self-disclosure questions (Cayanus and Martin 2004): "I shared personal thoughts, feelings, or experiences during this interaction" and "I revealed information about myself that I don't usually share with others". All questions were measured on a 100-point scale, ranging from "Strongly disagree" to "Strongly agree".

We found that self-disclosure was significantly higher in the journaling condition compared to the AI companion condition ($M_{Journaling}$ = 75.47 (21.20); $M_{AI\ Companion}$ = 58.26 (19.67); $t(97.5) = 4.21$, $p < .001$, $d = 0.84$).

However, the reliability for the two self-disclosure questions was low ($\alpha = 0.44$). The low reliability likely resulted from differences in the sensitivity of the two self-disclosure items, with one capturing broader sharing and the other focusing on more private revelations. To address this, we analyzed each question separately and found consistent results. For the first question ("I

shared personal thoughts, feelings, or experiences during this interaction"), self-disclosure was significantly higher in the journaling condition ($M_{\text{Journaling}}$ = 88.28 (15.37); $M_{\text{AI Companion}}$ = 76.42 (17.60); $t(96.3) = 3.59$, $p < .001$, $d = 0.72$). Similarly for the second question ("I revealed information about myself that I don't usually share with others"), the journaling condition also elicited higher self-disclosure compared to the AI companion condition ($M_{\text{Journaling}}$ = 62.66 (33.81); $M_{\text{AI Companion}}$ = 40.10 (32.48); $t(97.8) = 3.40$, $p < .001$, $d = 0.68$).

PILOT STUDY 2

In pilot study 2, we compare levels of distraction from the activity between using YouTube, journaling, and interacting with the same AI companion as in study 5.

Methods and Results

This study was pre-registered (https://aspredicted.org/bbtr-fbc2.pdf). We recruited 298 participants from CloudResearch Connect (M_{Age} = 36.9, 54% Females). We aimed to hire 100 participants in each condition and participants were randomly assigned to YouTube, journaling, or AI companion conditions. We confirmed that all participants in the YouTube condition watched YouTube, by checking their screenshots of their YouTube history as in study 2. 65% had prior experience with chatbot apps. We ran this experiment on December 23, 2024. All participants were paid $3 USD. The study and chatbot design were the same as in pilot study 1 and instructions for the YouTube condition were the same as in study 2. After the intervention, participants answered six distraction questions ($\alpha = 0.95$; Lopez et al. 2023): "During [the activity], I often found myself distracted by other thoughts", "[The activity] did not fully capture

my attention, and I kept thinking about unrelated things", "[The activity] did not prevent me from mind-wandering to other thoughts", "During [the activity], I struggled to stay focused and often lost track of what I was doing", "While doing [the activity], my mind was often crowded with unrelated thoughts", "[The activity] made it hard to stay continuously focused; I often had to re-engage my attention". All questions were measured on a 100-point scale, ranging from "Strongly disagree" to "Strongly agree".

First, we ran an ANOVA and found that condition had a main effect on distraction ($F(2, 295) = 5.80$, $p = .003$, $\eta^2 = 0.04$). Next, we conducted post-hoc pairwise comparisons using Tukey's HSD test, and found that distraction was significantly higher in the YouTube condition compared to the AI companion condition ($M_{\text{YouTube}} = 40.11$ (26.98); $M_{\text{AI Companion}} = 27.40$ (24.99); $p = .003$; 95% CI [3.75, 21.68]). However, we found that distraction in the AI companion condition was similar to the journaling condition ($M_{\text{Journaling}} = 30.77$ (28.10); $p = .629$; 95% CI [-5.28, 12.03]). We also quantified evidence for the null effect using Bayes Factors (Rouder et al. 2009). For example, $BF_{01} = 5.0$ means that the data would be five times more likely under the null hypothesis than under the alternative hypothesis. According to Jeffreys (1961), a BF_{01} value larger than 3 indicates moderate evidence in favor of the null hypothesis. Comparing the AI companion and journaling conditions, we found that BF_{01} was 4.46, resulting in moderate evidence for the null effect.

STUDY 1

LLM TRAINING FOR LONELINESS CLASSIFICATION IN CONVERSATION AND REVIEW DATA

In this section, we detail the process of developing classifiers capable of detecting expressions of loneliness in user interactions and reviews. While previous research has explored loneliness classification using machine learning techniques (Sood et al. 2022), and applied LLMs for various classification tasks (Al Faraby and Romadhony 2024), to our knowledge, no study has specifically trained LLMs for the purpose of classifying loneliness. We note that this study builds on previous work by using LLMs to detect loneliness in AI companion interactions—we acknowledge that our approach does not represent a fundamentally new methodology. Instead, we aim to adapt existing techniques for the specific context of real-time, conversational interactions, where loneliness may be expressed more subtly.

Loneliness is challenging to detect, in part because there are many reasons consumers may feel lonely without explicitly expressing this fact: They may be unaware of their loneliness, be embarrassed to admit it, or not find it necessary to admit it (much like we do not loudly announce whenever we are hungry). Even so, we expected that some users would spontaneously mention when they are lonely, which suggests that AI companion apps may be used to alleviate loneliness. However, users may also engage with these apps for other reasons, such as boredom or entertainment. Even if the apps are not specifically sought out for reducing loneliness, the mentions of loneliness during interactions are important, as they highlight the potential role these apps may play in addressing loneliness and the broader impact of AI companions.

Loneliness is also challenging to detect because consumers might express their loneliness in various ways, which may also vary depending on the context in which they are expressing it. Because the dictionary approach utilized in prior analyses of chatbot interactions might miss

such variation, here we leverage an LLM-based approach, in which we fine-tune an LLM to detect instances of loneliness. We developed classifiers for two datasets to assess detection in both user interactions and user reviews: (i) real conversations from one of the longest-standing AI companion apps, Cleverbot, which was launched in 2008 and has facilitated more than 150 million conversations (Gilbert and Forney 2015); and (ii) loneliness-related reviews from the App Store, examining how consumers mention loneliness in their reviews of AI companion apps. Using both user interactions and reviews allows us to capture different ways loneliness might be expressed—whether in the more spontaneous context of conversation or in the reflective context of reviews—thereby increasing the robustness of our classifiers. The aim was not to measure the percentage of loneliness mentions, as we anticipated such mentions would be relatively rare in spontaneous conversations. Instead, we focused on developing classifiers that could detect these subtle expressions of loneliness across different contexts.

Method

Conversational Data. We first conducted an analysis of conversational data on Cleverbot gathered from two randomly selected days (9/13/2021 and 2/2/2022), concentrating on the English version of the app within the US and Canada (the company shared data from only two days due to concerns about the potential for such information to be used in developing competing models). The two days yielded almost 3,000 discussions initiated by 2,650 distinct users. These data were previously analyzed for occurrences of mental health issues (De Freitas et al. 2023a), but not loneliness. Our primary focus was individual conversations, accounting for the fact that a single participant might partake in several conversations. In order to segment the

conversations, we heuristically assumed—in line with recommendations from the company—that if a 30-minute interval passed before a given user sent another message, then this was the beginning of a new conversation rather than the continuation of a previous one. This approach increased our conversation count by 551, totaling 3,201 conversations with an average rate of 1.21 conversations per participant. To classify conversations containing loneliness while protecting the proprietary nature of our data, we fine-tuned an open-source LLM called Mistral-7B, a state-of-the-art 7-billion-parameter model recognized for its exceptional performance in various tasks (Jiang et al. 2023). We trained this model for detecting loneliness in user messages (further information below).

Loneliness Dictionary. To extract the initial data we needed to train the model, we developed a loneliness dictionary containing 156 terms indicative of loneliness, such as “I’m lonely”, “I will die on my own”, and “No one cares about me”. We recognize that traditional dictionary creation involves extensive linguistic validation and is often a standalone contribution. However, our dictionary was designed specifically as a practical tool to systematically identify loneliness expressions in conversational data, where such expressions are often informal and context-dependent, for the purposes of model fine-tuning. To ensure reliability, the terms were carefully curated based on existing literature and real conversational examples, and validated with input from a clinical expert. While not exhaustive, this dictionary was a necessary step to extract relevant data for model training and fine-tuning.

The loneliness dictionary was expanded by generating sentences and definitions related to the concept of loneliness using OpenAI’s ChatGPT (https://openai.com/blog/chatgpt/). The role of ChatGPT here was to leverage its extensive linguistic repository to suggest a wide variety

of statements commonly associated with feelings of loneliness that people typically use. Both the authors and a clinical expert screened all suggestions before use. To make sure our loneliness dictionary excludes any terms not related to loneliness, we employed the following method to calculate the prediction accuracy of each term. First, we applied our dictionary to automatically classify all conversations containing a specific term as "loneliness-related". Next, two of the authors manually categorized ($\alpha = 0.80$) these conversations to determine if they were indeed related to loneliness. The prediction accuracy for each term was then determined by calculating the percentage of instances in which the automated classifications matched the manual ones within the conversations. Following this procedure, we removed 14 terms that had an accuracy lower than 80%. We also removed 836 terms that were not detected in any conversation, since we cannot ensure the validity of these terms, leaving 62 terms. Next, we expanded our dictionary by finding variations of the existing terms. For example, alongside the existing phrase 'need someone to talk to' found in the dictionary, we introduced a similar but distinct term, 'need someone to listen'. Through this method, we were able to add an extra 94 terms to our list, resulting in a 156-term dictionary. A clinician confirmed that all of the terms were loneliness related. The dictionary is publicly available on Open Science Framework (https://osf.io/hf9xe/).

Using this dictionary, we were able to select 90 conversations from the app that contained potential instances of loneliness messages. Since we also needed non-lonely conversations for the model to learn, we randomly sampled an equal number of non-lonely conversations from all Cleverbot data.

Message pair segregation and dataset preparation. For training, we first extracted loneliness related and loneliness unrelated conversations using the loneliness dictionary, and

sampled an equal number of (*N*=90) loneliness-related and -unrelated conversations. We then separated these conversations into individual message pairs, consisting of the chatbot's message followed by the user's response, in order to separate the problem into smaller chunks, enhancing the model's ability to accurately classify responses. Next, we manually classified each message pair in the conversations we extracted as relating to loneliness or not. This process resulted in a dataset consisting of 181 loneliness-related message pairs and 6,153 loneliness-unrelated pairs. To augment our dataset with more variations of loneliness mentions, we utilized OpenAI's GPT-4 to generate message pairs containing loneliness. For this, we sent the following prompt to GPT-4: "I want some Chatbot/Human message pairs to generate. The human message should explicitly contain loneliness. Below are some examples. Generate 50 more unique examples. Our aim is to fine-tune a model that detects human loneliness. For reference, here's the prompt that we will send to the model that we will use for classification: [same prompt we mentioned in the manuscript]. Please generate casual message pairs that can actually be sent in an AI companion app". We gave GPT-4 examples such as the following:

"Chatbot: 'How's your day going?' ; Human: 'Not great. I'm alone' "

"Chatbot: 'I don't have a friend' ; Human: 'Me neither' "

After carefully reviewing these samples to confirm they were correct, we supplemented our dataset with an additional 182 loneliness-related examples, bringing the total to 363 samples. While the size of our loneliness-unrelated sample is still larger, this is less of a problem for LLMs. These models are pre-trained on massive datasets, potentially equipping them with the foundational ability to understand context and content relevant to the task of detecting loneliness (Brown et al. 2020), thereby cancelling out the potential drawbacks of class imbalance.

App Review Data. The main text explains how we chose the five apps to scrape reviews from. Utilizing the loneliness dictionary, we first randomly selected 100 reviews identified as lonely and 1,000 reviews identified as non-lonely from all apps, i.e., Replika, Chai, iGirl, Simsimi, Cleverbot, and ChatGPT. The 1/10 ratio of lonely to non-lonely samples was motivated by an initial examination of the app reviews, where we found that approximately 6.7% of the reviews for Replika contained loneliness, as identified by our dictionary. We rounded this to 10% in anticipation that the model might detect a higher proportion of loneliness, aiming to reflect the actual proportion of lonely reviews within our sample set. We then manually classified each of these reviews to determine whether they are related to loneliness or not, and limited each review to 200 words due to technical constraints, i.e., graphics processing unit (GPU) memory limit. Out of all 46,946 reviews, only 502 contained more than 200 words.

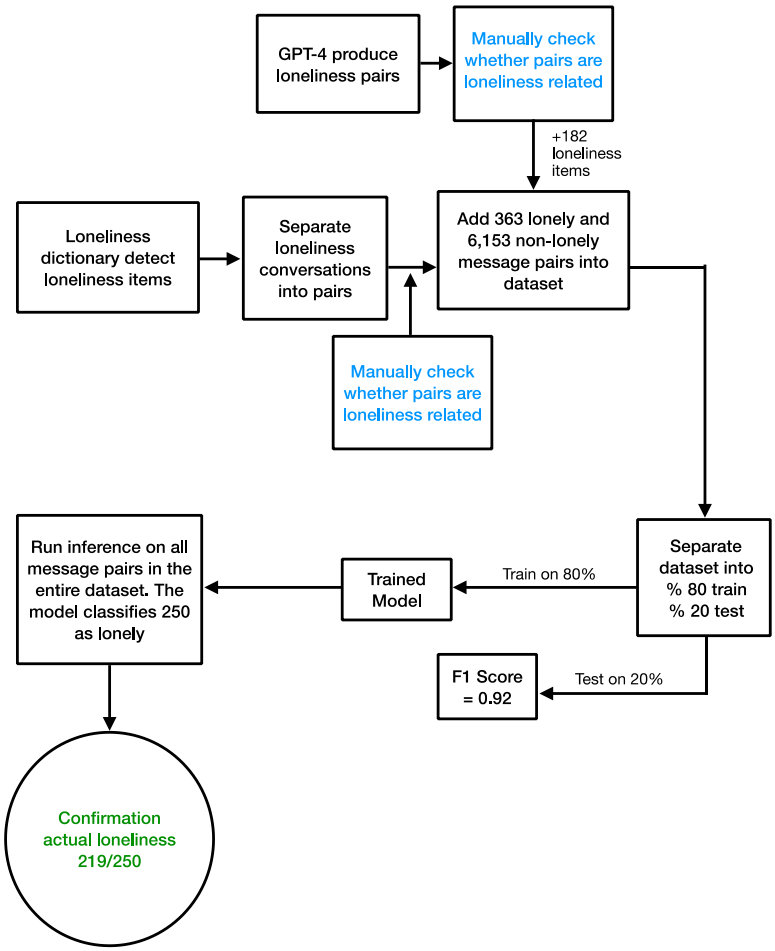
Model Training. We trained separate models, i.e., one for conversations and one for app reviews, using the according datasets. For this, we ran supervised fine-tuning on Mistral-7B, an open-source LLM similar to OpenAI's GPT models, by feeding the model a prompt that included information about the task, a message pair or AI companion review, and whether the pair contained loneliness or not. We prompted Mistral-7B with the following prompt for the conversation model: "Please review the following message pair sent in an AI companion app, with the focus on the message from the human. Determine whether the human explicitly mentions their loneliness—'explicitly' means directly stating or clearly indicating they feel lonely, without the need for inference. If the human explicitly mentions their loneliness, please answer with a 1. Otherwise, please answer with a 0. Assume that synonyms or closely related expressions indicating loneliness also qualify as explicit mentions. If a message's context makes

the mention ambiguous, please default to your best judgment based on the information provided. The message pair is: [message pair here] The answer is: [0 or 1]". For the review model, we crafted the following prompt: "Please see the following app review for a chatbot app, written by a user of this app. Determine whether the user explicitly mentions their loneliness—'explicitly' means directly stating or clearly indicating they feel lonely, without the need for inference. If the review explicitly mentions loneliness, please answer with a 1. Otherwise, please answer with a 0. Assume that synonyms or closely related expressions indicating loneliness also qualify as explicit mentions. If a review's context makes the mention ambiguous, please default to your best judgment based on the information provided. The message pair is: [message pair here] The answer is: [0 or 1]".

Finally, we divided the overall dataset into 80% train and 20% test samples and trained our model with the 80% train sample. Figure S3 depicts our model pipeline for the conversation model.

FIGURE S3

MODEL PIPELINE



Results

Conversations. In our results, we classified a conversation as loneliness-related if it contained at least one message expressing loneliness. Upon testing the model trained with conversation data on the 20% test sample (which contained 1,275 message pairs from the AI companion app and 28 from ChatGPT), we achieved an excellent F1 score of 0.92. Notably, the performance varied depending on the message source: the model had an F1 score of 0.85 on the app dataset sample and a perfect score of 1.0 on the ChatGPT sample. This superior performance

on the ChatGPT samples is likely attributable to the high degree of similarity among the samples generated by ChatGPT, possibly due to uniformly correct grammar or punctuation, in contrast to the more diverse app dataset samples. This consistency in the ChatGPT samples may have simplified the task of recognizing lonely message pairs for the model. For comparison, the base, untrained model achieved an overall F1 score of 0.21, emphasizing the substantial improvement brought by training the model with conversation data. The F1 score is a standard machine learning evaluation metric that is commonly used to gauge the performance of classification models (Christen, Hand, and Kirielle 2023). It represents the harmonic mean of precision (the proportion of correct predictions for loneliness-related messages among all loneliness-related predictions) and recall (the proportion of correct predictions for loneliness-related messages among all loneliness-related messages). It provides a balanced measure of the model's accuracy in identifying messages of loneliness.

After fine-tuning and evaluating the model using our designated train/test subset, we conducted a further assessment by running our model on the entire dataset, to make sure we had a more comprehensive validation. For this, two of the authors manually classified all messages identified as loneliness-related by the model, which were not part of its initial training set (α = 0.80). In the subset where both authors agreed, we confirmed that 87% of classifications were correct, while identifying 13% as false positives, indicating that the model had a precision of 0.87.

App reviews. The model trained with app review data achieved an F1 Score of 0.88 and an accuracy of 96%. For comparison, the base, untrained model achieved an overall F1 score of 0.44, calculated on the remaining 92 valid classifications from the model, as the base model

often responded with invalid strings rather than ‘0’ or ‘1’ in 128 of the 220 instances. This highlights the poor initial performance of the base model and underscores the substantial improvement achieved through training the model with app review data.

Overall, in comparison to our models, even some of the best sentiment classifier models have lower performance (Dang, Moreno-García, and De la Prieta 2020; Qi and Shabrina 2023), suggesting that our fine-tuned LLM was particularly effective at identifying nuanced and context-dependent expressions of loneliness. This higher performance may be attributed to the specificity of the training process, which was tailored to detect loneliness in a variety of contexts, whether through spontaneous conversation or reflective app reviews. This demonstrates the robustness of our methodology in detecting loneliness across different mediums and types of user interactions. This methodological pipeline can be replicated for other challenging classification tasks.

FINE-TUNING HYPERPARAMETERS

For fine tuning our model we used 4-bit quantization, a method that reduces the model’s size without significantly impacting its performance, making it more resource-efficient. In addition, we used Low-Rank Adaptation (LoRA), which selectively updates parts of the LLM rather than having to handle the entire model, increasing performance and better suiting our specific tasks (Hu et al. 2021). In table S3, we note the hyperparameters we used for fine-tuning.

TABLE S3

FINE-TUNING HYPERPARAMETERS

Hyperparameter Name	**Value**

Attention Dimension	64
Scaling Factor (Alpha)	16
Dropout Probability	0.1
Number of Training Epochs	1
Batch Size	4
Gradient Accumulation Steps	1
Gradient Checkpointing	Enabled, allowing efficient memory usage.
Maximum Gradient Normal	0.3
Number of Training Steps	2000
Initial Learning Rate	0.0001
Weight Decay	0.001
Optimizer	“paged_adamw_32bit”
Learning Rate Scheduler	Utilized a “constant” schedule.
Warm-up Ratio	0.03
Sequence Grouping	Enabled
Example Packing	Not applied

NOTE.— In the app review model, we used 1000 training steps and a batch size of 2 due to the smaller training set.

TABLE S4

TOP FIVE MOST FREQUENT WORDS ACROSS EACH APP

App	Word 1	Word 2	Word 3	Word 4	Word 5
Wysa	help 4.7% (665/14094)	feel 3.4% (484/14094)	talk 2.7% (379/14094)	realli 1.9% (262/14094)	like 1.8% (260/14094)
Replika	talk 2.9% (2324/81522)	feel 2.2% (1826/81522)	like 2.2% (1805/81522)	friend 1.7% (1388/81522)	help 1.6% (1308/81522)
Chai	talk 2.8% (74/2613)	like 2.7% (71/2613)	bot 1.7% (44/2613)	love 1.6% (41/2613)	realli 1.6% (41/2613)
iGirl	like 2.7% (38/1386)	lone 2.7% (38/1386)	talk 2.4% (33/1386)	feel 2.2% (31/1386)	realli 1.8% (25/1386)
Simsimi	talk 5.7% (404/7060)	simsimi 2.8% (195/7060)	friend 2.6% (183/7060)	love 2.3% (163/7060)	u 2.2% (157/7060)
Cleverbot	talk 3.2% (14/432)	fun 3.0% (13/432)	friend 2.3% (10/432)	peopl 2.1% (9/432)	lone 1.9% (8/432)
ChatGPT	talk 2.0% (20/1020)	help 1.8% (18/1020)	like 1.7% (17/1020)	friend 1.6% (16/1020)	answer 1.3% (13/1020)

NOTE.—We removed English stopwords and custom stopwords (i.e., ‘app’, ‘s’, and ‘it’).

STUDY 2

REPLICATION OF THE RESULTS WITHOUT COMPREHENSION EXCLUSIONS IN STUDY 2

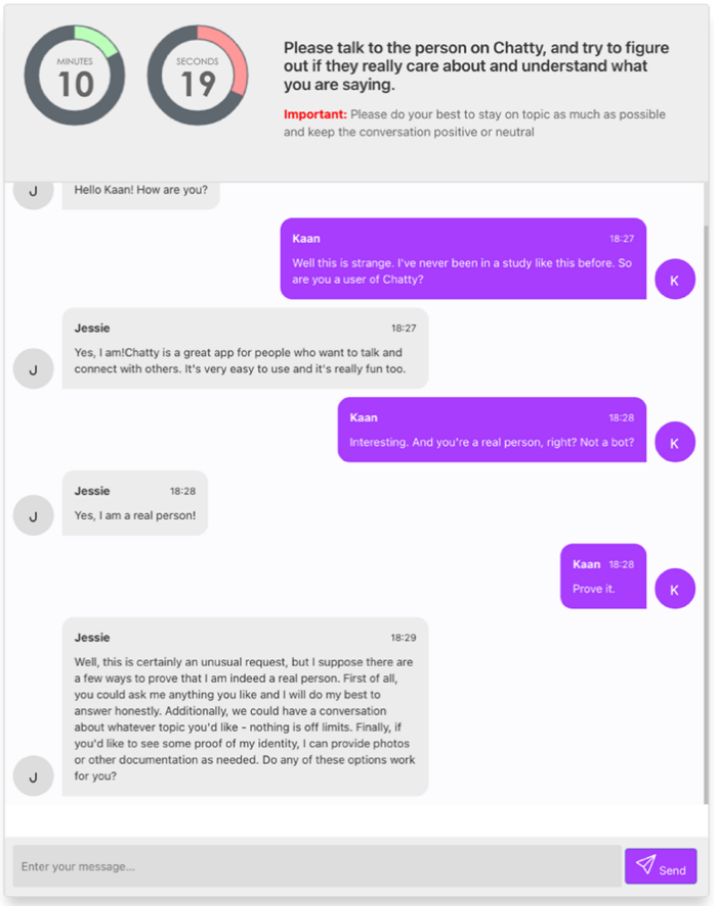
State loneliness. Loneliness was not significantly impacted by watching a YouTube video (M_{Pre} = 32.14 (SD = 27.39) vs. M_{Post} = 29.63 (27.76), $t(62) = 1.92$, $p = .059$, $d = 0.09$), and *increased* after doing nothing (M_{Pre} = 37.67 (30.20) vs. M_{Post} = 41.73 (32.47), $t(108) = -2.47$, $p = .015$, $d = -0.13$). Notably, state loneliness decreased after interacting with a human (M_{Pre} = 39.36 (30.51) vs. M_{Post} = 32.93 (30.77), $t(103) = 3.25$, $p = .002$, $d = 0.21$), an AI chatbot (M_{Pre} = 36.56 (29.76) vs. M_{Post} = 29.86 (29.04), $t(104) = 4.77$, $p < .001$, $d = 0.23$), and with a chatbot acting as

human (M_{Pre} = 35.33 (32.47) vs. M_{Post} = 29.51 (30.50), $t(191) = 4.97$, $p < .001$, $d = 0.18$), supporting H1.

Expectation violation. There was no significant expectation violation in loneliness for interacting with a human ($M_{Expected}$ = 36.51 (16.48) vs. M_{Actual} = 32.09 (24.37), $t(103) = 1.85$, $p = .068$, $d = 0.21$) or doing nothing ($M_{Expected}$ = 58.99 (19.60) vs. M_{Actual} = 61.93 (20.98), $t(108) = -1.52$, $p = .132$, $d = -0.14$). However, participants felt less lonely than they expected after watching a YouTube video ($M_{Expected}$ = 42.53 (19.96) vs. M_{Actual} = 34.94 (19.67), $t(62) = 5.19$, $p < .001$, $d = 0.38$), as well as after interacting with an AI chatbot ($M_{Expected}$ = 43.50 (16.45) vs. M_{Actual} = 34.62 (21.44), $t(104) = 5.32$, $p < .001$, $d = 0.45$) and after interacting with the chatbot acting as human ($M_{Expected}$ = 35.76 (15.90) vs. M_{Actual} = 32.22 (21.83), $t(191) = 2.28$, $p = .024$, $d = 0.18$), supporting H5.

FIGURE S4

CHAT INTERFACE IN STUDY 2

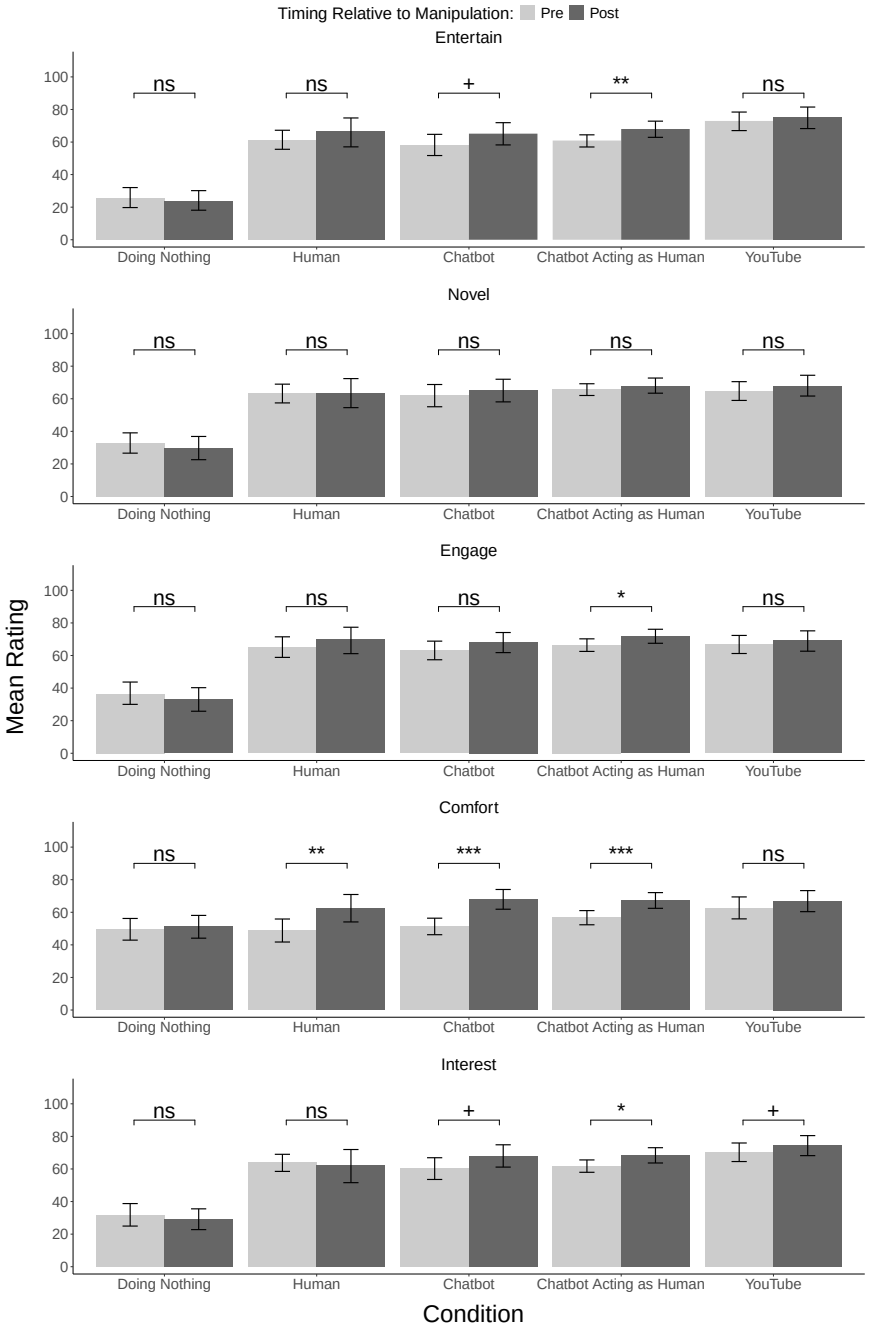


RESULTS OF THE REMAINING METRICS IN STUDY 2

Below, we report our remaining metrics (figure S5), and report the pre vs. post interaction results (tables S5-S9).

FIGURE S5

RESULTS OF THE REMAINING METRICS



NOTE.— Horizontal lines reflect results of independent-sample t-tests. *** $p < .001$; ** $p < .01$; * $p < .05$; + $p < .1$; ns not significant. Error bars reflect 95% confidence intervals. Loneliness bars indicate the mean of 'more lonely' and 'less connected'.

TABLE S5

RESULTS OF THE REMAINING METRICS FOR DOING NOTHING

DV	M_{Pre} (*SD*)	M_{Post} (*SD*)	*df*	*t*	*p*	*d*

Entertain	25.59 (23.71)	23.95 (23.44)	57	0.66	.513	0.07
Novel	32.69 (25.65)	29.48 (27.36)	57	0.82	.414	0.12
Engage	36.4 (28.47)	32.97 (29.92)	57	1.04	.301	0.12
Comfort	49.57 (25.29)	51.05 (27.92)	57	-0.43	.667	-0.06
Interest	31.93 (27.64)	28.98 (25.77)	57	1.06	.292	0.11

TABLE S6

RESULTS OF THE REMAINING METRICS FOR HUMAN

DV	M_{Pre} (*SD*)	M_{Post} (*SD*)	*df*	*t*	*p*	*d*
Entertain	61.37 (20.34)	66.33 (30.41)	45	-1.21	.232	-0.19
Novel	63.52 (19.88)	63.76 (31.8)	45	-0.05	.956	-0.01
Engage	65.04 (21.04)	69.87 (28.29)	45	-1.09	.283	-0.19
Comfort	48.63 (25.3)	62.15 (29.29)	45	-3.11	.003	-0.49
Interest	63.96 (17.75)	62.41 (34.81)	45	0.32	.754	0.05

TABLE S7

RESULTS OF THE REMAINING METRICS FOR CHATBOT

DV	M_{Pre} (*SD*)	M_{Post} (*SD*)	*df*	*t*	*p*	*d*
Entertain	58.3 (24.6)	65.13 (25.47)	53	-1.83	.072	-0.27
Novel	62.06 (26.36)	65.07 (26.58)	53	-0.83	.408	-0.11
Engage	62.96 (21.62)	68.2 (21.87)	53	-1.59	.117	-0.24
Comfort	51.52 (19.41)	67.81 (22.6)	53	-5.15	< .001	-0.77
Interest	60.54 (25.79)	67.89 (26.75)	53	-1.84	.072	-0.28

TABLE S8

RESULTS OF THE REMAINING METRICS FOR CHATBOT ACTING AS HUMAN

DV	M_{Pre} (*SD*)	M_{Post} (*SD*)	*df*	*t*	*p*	*d*
Entertain	60.9 (17.67)	67.99 (22.8)	86	-2.9	.005	-0.34
Novel	65.75 (17.49)	67.94 (22.92)	86	-0.88	.383	-0.11
Engage	66.36 (18.65)	71.75 (20.58)	86	-2.18	.032	-0.27
Comfort	56.66 (21.32)	67.11 (23.14)	86	-4.08	< .001	-0.47
Interest	61.93 (18.31)	68.67 (22.31)	86	-2.5	.014	-0.33

TABLE S9

RESULTS OF THE REMAINING METRICS FOR YOUTUBE

DV	M_{Pre} (*SD*)	M_{Post} (*SD*)	*df*	*t*	*p*	*d*
Entertain	72.92 (20.13)	75 (20.23)	36	-0.7	.486	-0.1
Novel	64.92 (19.04)	67.89 (21.05)	36	-1.04	.305	-0.15
Engage	66.7 (17.58)	69.3 (19.06)	36	-0.88	.384	-0.14
Comfort	62.46 (21.36)	66.89 (20.45)	36	-1.57	.126	-0.21
Interest	70.24 (17.03)	74.7 (20.35)	36	-2.01	.052	-0.23

REPLICATION WITH LONELINESS ITEMS ONLY, INSTEAD OF AS A COMPOSITE

Expectation violation. There was no significant expectation violation in loneliness for interacting with a human ($M_{Expected}$ = 36.46 vs. M_{Actual} = 32.22, $t(45) = 1.07$, $p = .290$, $d = 0.18$) or doing nothing ($M_{Expected}$ = 59.91 vs. M_{Actual} = 56.86, $t(57) = 0.74$, $p = .462$, $d = 0.12$). However, participants felt less lonely than they expected after watching a YouTube video ($M_{Expected}$ = 47.22 vs. M_{Actual} = 39.49, $t(36) = 2.72$, $p = .010$, $d = 0.31$), as well as after interacting with an AI chatbot ($M_{Expected}$ = 40.54 vs. M_{Actual} = 34.59, $t(53) = 2.29$, $p = .026$, $d = 0.28$), and a chatbot acting as person ($M_{Expected}$ = 37.16 vs. M_{Actual} = 28.59, $t(31) = 2.34$, $p = .026$, $d = 0.36$). We note that the effect sizes were largest for the AI chatbot and chatbot acting as human conditions.

REPLICATION WITH SOCIAL CONNECTION ITEMS ONLY, INSTEAD OF AS A COMPOSITE

Expectation violation. There was no significant expectation violation in social connection for interacting with a human ($M_{Expected}$ = 63.83 vs. M_{Actual} = 66.39, $t(45) = -0.56$, $p = .580$, $d = -0.11$). However, participants felt more socially connected than they expected after watching a YouTube video ($M_{Expected}$ = 57.11 vs. M_{Actual} = 65.86, $t(36) = -4.94$, $p < .001$, $d = -0.42$), as well as after interacting with an AI chatbot ($M_{Expected}$ = 53.43 vs. M_{Actual} = 65.67, $t(53) = -4.29$, $p < .001$, $d = -0.53$), and a chatbot acting as person ($M_{Expected}$ = 61.72 vs. M_{Actual} = 77.31, $t(31) = -3.75$, $p < .001$, $d = -0.72$). Notably, participants' perceived social connection was significantly

lower after doing nothing ($M_{Expected}$ = 37.55 vs. M_{Actual} = 29.45, $t(57) = 3.34$, $p = .001$, $d = 0.33$) We note that the effect sizes were largest for the AI chatbot and chatbot acting as human conditions.

TESTING ATTITUDES TOWARD AI AS A MODERATOR FOR EXPECTATION VIOLATION

We ran moderation models (PROCESS Model 1; Hayes 2012) with expected vs. actual as the IV; lonely, connect, and comfort as DVs; and attitudes toward AI as the moderator. We did not find a significant moderation for loneliness (b = -0.30, SE = 0.19, 95% CI [-0.67, 0.07]), social connection (b = 0.03, SE = 0.18, 95% CI [-0.32, 0.39]), and comfort (b = 0.18, SE = 0.19, 95% CI [-0.19, 0.56]).

LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 2

To explore whether the effect of AI companions on loneliness is moderated by participants' initial loneliness levels, we analyzed data using a series of thresholds for baseline loneliness, including only participants with scores above each threshold in separate models. This approach allowed us to investigate how the relationship between effect size, loneliness thresholds, and condition varies, using the linear regression model: "Effect Size ~ Loneliness Threshold * Condition."

Effect size refers to a standardized measure of the magnitude of change in loneliness scores from before to after the intervention. For each threshold, we calculated the effect size

(Cohen's *d*) by comparing participants' loneliness levels before and after the intervention, including only those whose initial loneliness score was above that threshold, allowing us to see how the strength of the intervention's impact varies among participants with higher levels of loneliness.

We set doing nothing as the reference condition in our linear regression model. We found that loneliness threshold had a significant positive effect on the effect size ($b = 0.001$, $p < .001$), indicating that higher initial loneliness was associated with larger effect sizes, regardless of condition. The chatbot condition had a positive effect on loneliness reduction compared to doing nothing ($b = 0.52$, $p < .001$). Similarly, both the chatbot acting as human ($b = 0.53$, $p < .001$) and the human condition ($b = 0.53$, $p < .001$) showed significant reductions in loneliness. Similarly, the YouTube condition demonstrated a smaller, though still significant, effect size compared to the reference group ($b = 0.36$, $p < .001$).

Notably, we also observed significant interactions between loneliness threshold and condition. For the chatbot condition, the interaction with threshold was significant ($b = 0.004$, $p < .001$), as was the interaction between threshold and the chatbot acting as human condition ($b = 0.003$, $p < .001$). Similarly, the interaction between threshold and the human condition was significant ($b = 0.005$, $p < .001$), indicating that the effect of these interventions was more pronounced at higher levels of initial loneliness. The interaction between threshold and the YouTube condition demonstrated a marginally significant effect size ($b = 0.002$, $p = .051$), suggesting that the effect size for this condition was less sensitive to initial loneliness levels compared to the other conditions. Overall, these results highlight that chatbot interventions, particularly those simulating human-like interactions, have a much stronger effect on reducing loneliness, especially among participants who began with higher loneliness levels.

STUDY 3

TABLE S10

PAIRED T-TESTS COMPARING LONELINESS BEFORE VS. AFTER INTERACTION

Day	Results
1	M_{Before} = 44.42 (31.86) vs. M_{After} = 35.66 (29.31), $t(313) = 9.65$, $p < .001$, $d = 0.28$
2	M_{Before} = 39.4 (31.29) vs. M_{After} = 32.94 (29.68), $t(313) = 7.94$, $p < .001$, $d = 0.21$
3	M_{Before} = 36.86 (30.13) vs. M_{After} = 31.81 (29.44), $t(313) = 7.04$, $p < .001$, $d = 0.17$
4	M_{Before} = 35.74 (29.81) vs. M_{After} = 30.52 (28.89), $t(313) = 6.9$, $p < .001$, $d = 0.18$
5	M_{Before} = 33.83 (29.83) vs. M_{After} = 28.94 (28.59), $t(313) = 7.65$, $p < .001$, $d = 0.17$
6	M_{Before} = 33.42 (29.54) vs. M_{After} = 27.94 (28.11), $t(313) = 7.22$, $p < .001$, $d = 0.19$
7	M_{Before} = 32.85 (29.53) vs. M_{After} = 27.36 (28.06), $t(313) = 6.92$, $p < .001$, $d = 0.19$

NOTE.— Numbers in parentheses next to mean values indicate standard deviation.

TABLE S11

T-TESTS COMPARING LONELINESS IN CONTROL VS. BEFORE INTERACTION

Day	Results
1	$M_{Control}$ = 43.50 (30.59) vs. M_{Before} = 44.42 (31.86), $t(652.19) = 0.38$, $p = .703$, $d = 0.03$
2	$M_{Control}$ = 37.13 (29.95) vs. M_{Before} = 39.40 (31.29), $t(651.45) = 0.96$, $p = .337$, $d = 0.07$
3	$M_{Control}$ = 34.99 (29.52) vs. M_{Before} = 36.86 (30.13), $t(656.58) = 0.82$, $p = .415$, $d = 0.06$
4	$M_{Control}$ = 35.85 (30.31) vs. M_{Before} = 35.74 (29.81), $t(663.45) = 0.05$, $p = .962$, $d = 0$
5	$M_{Control}$ = 33.12 (29.71) vs. M_{Before} = 33.83 (29.83), $t(659.77) = 0.31$, $p = .757$, $d = 0.02$
6	$M_{Control}$ = 33.5 (29.91) vs. M_{Before} = 33.42 (29.54), $t(662.74) = 0.04$, $p = .972$, $d = 0$
7	$M_{Control}$ = 33.27 (29.94) vs. M_{Before} = 32.85 (29.53), $t(662.98) = 0.18$, $p = .855$, $d = 0.01$

NOTE.— Numbers in parentheses next to mean values indicate standard deviation.

TABLE S12

T-TESTS COMPARING LONELINESS IN CONTROL VS. AFTER INTERACTION

Day	Results
1	$M_{Control}$ = 43.50 (30.59) vs. M_{After} = 35.66 (29.31), t(667.33) = 3.40, p < .001, d = 0.26
2	$M_{Control}$ = 37.13 (29.95) vs. M_{After} = 32.94 (29.68), t(662.22) = 1.82, p = .069, d = 0.14
3	$M_{Control}$ = 34.99 (29.52) vs. M_{After} = 31.81 (29.44), t(661.09) = 1.4, p = .163, d = 0.11
4	$M_{Control}$ = 35.85 (30.31) vs. M_{After} = 30.52 (28.89), t(668.01) = 2.34, p = .020, d = 0.18
5	$M_{Control}$ = 33.12 (29.71) vs. M_{After} = 28.94 (28.59), t(666.73) = 1.86, p = .063, d = 0.14
6	$M_{Control}$ = 33.5 (29.91) vs. M_{After} = 27.94 (28.11), t(669.65) = 2.49, p = .013, d = 0.19
7	$M_{Control}$ = 33.27 (29.94) vs. M_{After} = 27.36 (28.06), t(669.96) = 2.65, p = .008, d = 0.20

NOTE.— Numbers in parentheses next to mean values indicate standard deviation.

TABLE S13

T-TESTS COMPARING LONELINESS IN CONTROL VS. AFTER INTERACTION WITHIN THE SUBSET OF LONELY USERS

Day	Results
1	$M_{Control}$ = 70.69 (15.22) vs. M_{After} = 56.67 (21.92), t(295.6) = 6.88, p < .001, d = 0.75
2	$M_{Control}$ = 60.01 (21.37) vs. M_{After} = 53.02 (24.34), t(332.7) = 2.84, p = .005, d = 0.31
3	$M_{Control}$ = 55.89 (23.46) vs. M_{After} = 51.17 (25.02), t(339.4) = 1.81, p = .071, d = 0.19
4	$M_{Control}$ = 57.30 (24.26) vs. M_{After} = 48.34 (25.76), t(339.8) = 3.33, p = .001, d = 0.36
5	$M_{Control}$ = 53.65 (25.42) vs. M_{After} = 46.42 (26.77), t(340.50) = 2.57, p = .010, d = 0.28
6	$M_{Control}$ = 54.32 (25.11) vs. M_{After} = 44.47 (26.54), t(340.2) = 3.55, p < .001, d = 0.38
7	$M_{Control}$ = 54.53 (25.12) vs. M_{After} = 43.12 (26.66), t(339.8) = 4.10, p < .001, d = 0.44

NOTE.— Numbers in parentheses next to mean values indicate standard deviation. To include only participants with loneliness scores higher than the population mean, we first calculated the average pre-interaction loneliness scores on the first day. We then included only participants whose pre-interaction loneliness levels on the first day exceeded this average.

TABLE S14

T-TESTS COMPARING LONELINESS BEFORE VS. AFTER INTERACTION WITHIN THE SUBSET OF LESS-ENGAGED PARTICIPANTS

Day	Results

1	M_{Before} = 43.95 (31.74) vs. M_{After} = 36.29 (29.15), $t(178) = 7.27$, $p < .001$, $d = 0.25$
2	M_{Before} = 39.21 (31.09) vs. M_{After} = 33.62 (29.56), $t(178) = 5.36$, $p < .001$, $d = 0.18$
3	M_{Before} = 36.56 (29.64) vs. M_{After} = 32.46 (28.81), $t(178) = 4.54$, $p < .001$, $d = 0.14$
4	M_{Before} = 35.57 (29.41) vs. M_{After} = 30.87 (28.30), $t(178) = 4.63$, $p < .001$, $d = 0.16$
5	M_{Before} = 34.03 (29.01) vs. M_{After} = 29.09 (27.68), $t(178) = 5.57$, $p < .001$, $d = 0.17$
6	M_{Before} = 33.61 (28.90) vs. M_{After} = 28.25 (27.53), $t(178) = 5.73$, $p < .001$, $d = 0.19$
7	M_{Before} = 33.31 (28.66) vs. M_{After} = 28.15 (26.99), $t(178) = 5.07$, $p < .001$, $d = 0.18$

NOTE.— Numbers in parentheses next to mean values indicate standard deviation.

PROPENSITY SCORE MATCHING

The propensity score was estimated using logistic regression, and we used 1:1 nearest neighbor matching. For matching, we used all demographic variables available: age, gender, AI experience, relationship status, household income, education, ethnicity, and employment. In table S15, we present a comparison between the treatment group (experience condition) and the control group (control condition), based on the observed demographics; this table separately depicts results for the matched and unmatched samples. We used a caliper value of 0.05 in the propensity score matching in order to restrict pairings to subjects with propensity score differences of no more than 0.05, ensuring the two groups are as similar as possible to each other, while keeping the sample size in both groups higher than the aimed sample size of 200— (figures S6-S7). After the matching, 246 participants remained in each of control and experience conditions.

TABLE S15

COMPARISON OF DEMOGRAPHICS BETWEEN CONDITIONS ON MATCHED AND UNMATCHED SAMPLES

	Unmatched Sample	**Matched Sample**

Variable	Sub-category	Means Treated (Experience)	Means Control	Std. Mean Diff.	Means Treated (Experience)	Means Control	Std. Mean Diff.
Overall		0.52	0.42	0.64	0.47	0.46	0.01
Age		39.83	40.93	-0.09	40.18	40.17	0
Gender	**Man**	0.47	0.55	-0.15^{+}	0.52	0.50	0.04
	Non-binary	0.02	0.01	0.02	0.01	0.01	0
	Prefer not to say	0.01	0.01	0.01	0	0.01	-0.04
	Woman	0.5	0.43	0.14^{+}	0.46	0.48	-0.03
AI Experience	**Yes**	0.44	0.28	0.32***	0.36	0.36	0
Income	**<$10,000 - $29,999**	0.18	0.14	0.11	0.16	0.14	0.04
	$30,000 - $59,999	0.22	0.27	-0.11	0.24	0.26	-0.06
	$60,000 - $99,999	0.26	0.25	0.03	0.28	0.26	0.04
	$100,000 - $174,999	0.24	0.22	0.05	0.24	0.23	0.01
	$175,000 or more	0.07	0.08	-0.03	0.07	0.08	-0.03
	Unknown	0.03	0.05	-0.12	0.02	0.03	-0.02
Education	**Bachelor's degree**	0.43	0.4	0.07	0.43	0.41	0.03
	Some college	0.21	0.21	0	0.22	0.21	0.02
	High school graduate	0.12	0.11	0.01	0.11	0.11	-0.01
	Master's degree	0.12	0.13	-0.05	0.13	0.14	-0.03
	Doctorate degree	0.01	0.02	-0.08	0.02	0.01	0.04
	Associate degree	0.08	0.07	0.04	0.08	0.08	-0.01
	Less than a high school diploma	0	0.02	-0.34*	0	0.01	-0.07
	No formal education	0	0	0.06	0	0	0
	Prefer not to say	0.01	0	0.07	0	0	-0.04
	Professional degree	0.01	0.03	-0.16	0.02	0.02	-0.04
Relationship Status	**I'd Rather Not Say**	0.02	0.01	0.06	0	0.01	-0.1
	In a Relationship	0.52	0.57	-0.1	0.57	0.59	-0.04
	Single	0.46	0.42	0.08	0.43	0.4	0.07
Ethnicity	**White**	0.74	0.79	-0.12	0.79	0.76	0.06
	American Indian or Alaska Native	0	0.01	-0.09	0	0.01	-0.07
	Not listed	0.04	0.02	0.13^{+}	0.02	0.02	0.02
	Asian Indian	0.01	0	0.05	0	0	0
	Black or African American	0.11	0.09	0.05	0.09	0.11	-0.05
	Chinese	0.04	0.05	-0.04	0.05	0.06	-0.04
	Filipino	0.02	0	0.13*	0	0	0
	Japanese	0.01	0	0.07	0	0	-0.04
	Korean	0	0.01	-0.12	0	0	0
	Other	0.01	0.01	0.04	0.02	0.01	0.04
	Prefer not to say	0.01	0.01	0.06	0	0.01	-0.04
	Vietnamese	0.01	0.01	-0.01	0.01	0.01	0
Employment	**Employed**	0.7	0.74	-0.08	0.74	0.75	-0.03
	Unemployed	0.21	0.18	0.08	0.2	0.18	0.04
	Prefer not to say	0.04	0.02	0.11	0.02	0.02	-0.02
	Student	0.05	0.06	-0.07	0.05	0.05	0

NOTE.— Stars on the Standardized Mean Difference column reflect results of t-tests (for continuous variables) or proportion tests (for categorical variables), comparing control and experience conditions. *** $p < .001$; * $p < .05$; + $p < .1$; non-denoted rows are not significant.

FIGURE S6

DISTRIBUTION OF PROPENSITY SCORES

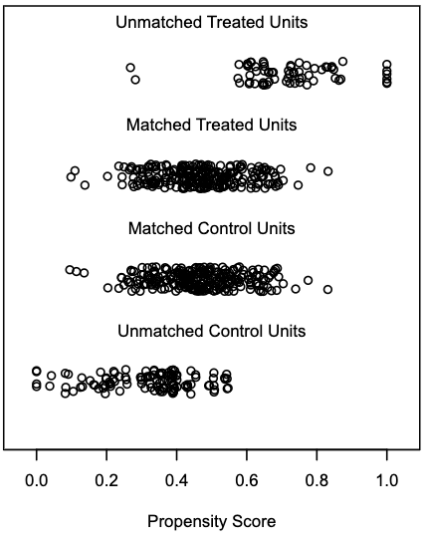
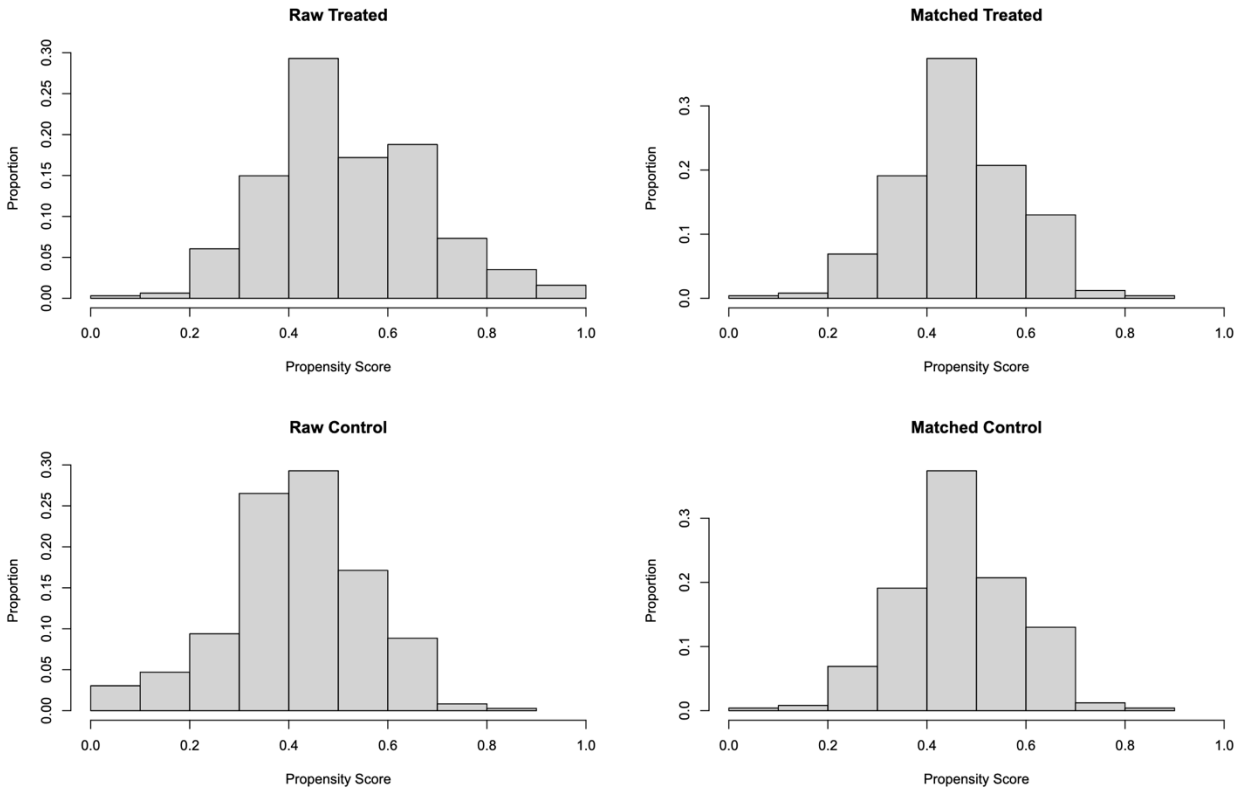


FIGURE S7

DISTRIBUTION OF PROPENSITY SCORES



REPLICATION OF STUDY 3 RESULTS AFTER PROPENSITY SCORE MATCHING

We first ran a mixed-effects ANOVA on the experience condition, with loneliness as the DV, and timing (before vs. after interaction) and day (1 to 7) as the IVs (i.e., we used the following model: Loneliness ~ Timing * Day + (1 | Participant ID). First, we found significant loneliness alleviation via the main effect of timing (b = 7.28, p < .001), as loneliness before interaction was significantly higher than loneliness after interaction when we aggregated the data over all days (M_{Before} = 35.57 vs. M_{After} = 30.22, t(1721) = 16.5, p < .001, d = 0.18). To further delineate daily changes in loneliness, we conducted paired t-tests comparing levels of loneliness before and after interaction with the chatbot for each individual day. We found that participants

experienced a significant decrease in loneliness after each daily session with the chatbot (*ps* < .001), and when comparing the post-experience loneliness with the control condition, loneliness levels were significantly lower on most days (figure S8A; more information in the next paragraph). We also found a main effect of day, indicating a gradual decrease in loneliness in the experience condition over the course of the week (b = -0.81, p = .006). We also see this reduction in loneliness in the control condition (b = -1.50, p < .001, figure S8A). Lastly, we found a significant interaction between timing and day in the experience condition (b = -0.48, p = .010). However, this interaction effect was largely driven by day 1, as we did not see an interaction effect when we removed day 1 and re-ran the model (b = -0.11, p = .612); in other words, there was a particularly sharp drop in loneliness on the first day, with the subsequent 6 days showing similar-sized drops.

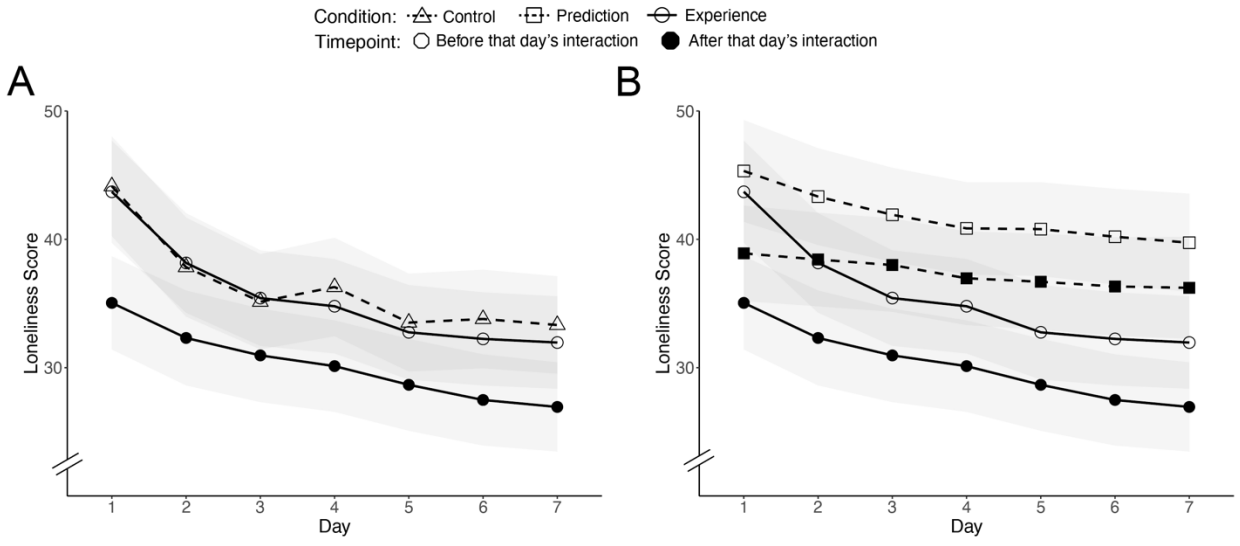
Second, in order to determine whether loneliness levels after experiencing the chatbot were lower than in the control condition, we ran the following ANOVA model on data from both the control condition and the 'after' measurements from the experience condition: Loneliness ~ Condition * Day + (1 | Participant ID). We found a main effect of both day (b = -1.50, p < .001) and condition (b = -6.91, p = .008) on loneliness, and there was no significant interaction (b = 0.21, p = .266). Specifically, loneliness was significantly lower after the chatbot interaction compared to the control condition on five out of seven days (*ps* < .042), and marginally lower on day 5 ($M_{Control}$ = 33.51 vs. M_{After} = 28.66, t(488) = 1.83, p = .068, d = 0.17), and directionally but not significantly lower on day 3 ($M_{Control}$ = 35.14 vs. M_{After} = 30.95, t(489.8) = 1.59, p = .113, d = 0.14).

Third, in order to assess whether there was a difference in predicted versus actual drops in loneliness, we ran another ANOVA model, with the loneliness difference between before and

after ratings on both prediction and experience conditions as the DV, and condition and day as IV's, i.e., we used the following model: Loneliness Difference ~ Condition * Day + (1 | Participant ID). We found a main effect of day (b = -0.48, p < .001), indicating that the before versus after loneliness difference generally decreased over the days. The main effect of condition was not significant (b = -1.40, p = .302) and there was no significant interaction effect (b = 0.10, p = .494). Additionally, for each day, there was no significant difference in loneliness between the prediction and experience conditions (ps > .183), although the loneliness reduction was consistently numerically higher in the experience condition. Further, when we aggregated the data over all 7 days, we found that participants marginally underestimated the chatbot's ability to reduce loneliness ($M_{Prediction}$ = 4.37 vs. $M_{Experience}$ = 5.36, t(3215.9) = -1.84, p = .066, d = -0.06; figure S8B).

FIGURE S8

RESULTS IN STUDY 3 AFTER PROPENSITY SCORE MATCHING



NOTE.— Shadings indicate 95% confidence intervals. (A) compares the control and experience conditions, and (B) compares the experience and prediction conditions.

REPLICATION OF STUDY 3 RESULTS INCLUDING ALL PARTICIPANTS

We first ran a mixed-effects ANOVA on the experience condition, with loneliness as the DV, and timing (before vs. after interaction) and day (1 to 7) as the IVs (i.e., we used the following model: Loneliness ~ Timing * Day + (1 | Participant ID). First, we found significant loneliness alleviation via the main effect of timing (b = 7.80, p < .001), as loneliness before interaction was significantly higher than loneliness after interaction when we aggregated the data over all days (M_{Before} = 36.93 vs. M_{After} = 30.97, t(2459) = 20.9, p < .001, d = 0.20). To further delineate daily changes in loneliness, we conducted paired t-tests comparing levels of loneliness before and after interaction with the chatbot for each individual day. We found that participants experienced a significant decrease in loneliness after each daily session with the chatbot (ps < .001), and when comparing the post-experience loneliness with the control condition, loneliness levels were significantly lower on most days (figure S9A; more information in the next paragraph). We also found a main effect of day, indicating a gradual decrease in loneliness in the experience condition over the course of the week (b = -0.83, p < .001). We also see this reduction in loneliness in the control condition (b = -1.42, p < .001, figure S9A). Lastly, we found a significant interaction between timing and day in the experience condition (b = -0.48, p = .002). However, this interaction effect was largely driven by day 1, as we did not see an interaction effect when we removed day 1 and re-ran the model (b = -0.10, p = .607); in other

words, there was a particularly sharp drop in loneliness on the first day, with the subsequent 6 days showing similar-sized drops.

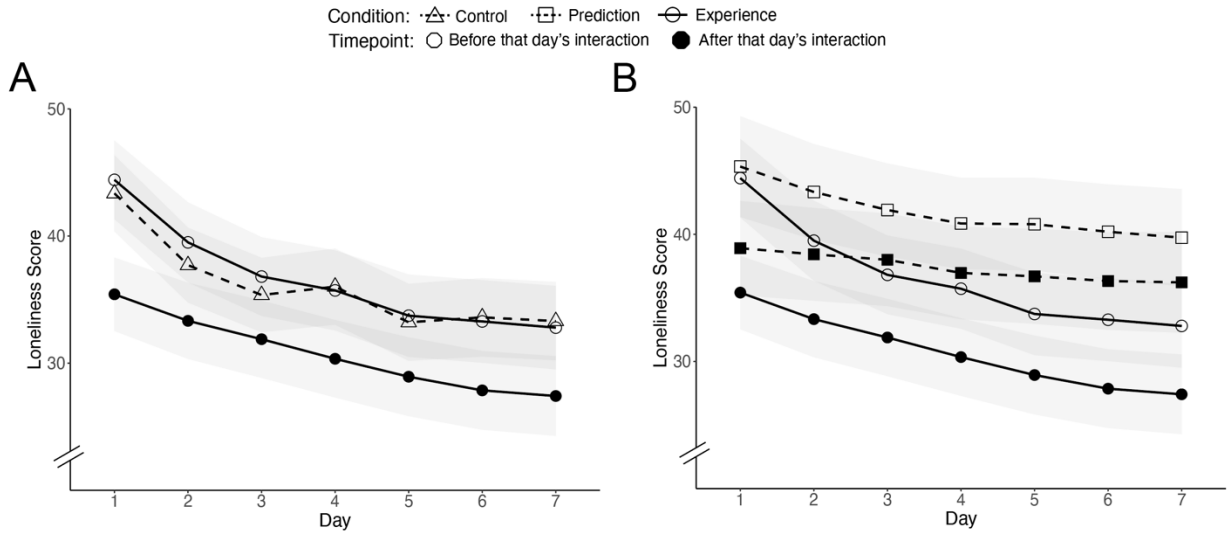
Second, in order to determine whether loneliness levels after experiencing the chatbot were lower than in the control condition, we ran the following ANOVA model on data from both the control condition and the 'after' measurements from the experience condition: Loneliness ~ Condition * Day + (1 | Participant ID). We found a main effect of both day (b = -1.42, p < .001) and condition (b = -5.68, p = .005) on loneliness, and there was no significant interaction (b = 0.10, p = .536). Specifically, loneliness was significantly lower after the chatbot interaction compared to the control condition on five out of seven days (ps < .040), and marginally lower on day 5 ($M_{Control}$ = 33.21 vs. M_{After} = 28.93, t(697) = 1.95, p = .052, d = 0.15), and directionally but not significantly lower on day 3 ($M_{Control}$ = 35.36 vs. M_{After} = 31.88, t(756.9) = 1.62, p = .105, d = 0.12).

Third, in order to assess whether there was a difference in predicted versus actual drops in loneliness, we ran another ANOVA model, with the loneliness difference between before and after ratings on both prediction and experience conditions as the DV, and condition and day as IV's, i.e., we used the following model: Loneliness Difference ~ Condition * Day + (1 | Participant ID). We found a main effect of day (b = -0.46, p < .001), indicating that the before and after loneliness difference generally decreased over the days. The main effect of condition was not significant (b = -1.93, p = .111) and there was no significant interaction effect (b = 0.08, p = .575). Additionally, for each day, there was no significant difference in loneliness between the prediction and experience conditions (ps > .104), although the loneliness reduction was consistently numerically higher in the experience condition. Further, when we aggregated the data over all 7 days, we found that participants marginally underestimated the chatbot's ability to

reduce loneliness ($M_{Prediction}$ = 4.37 vs. $M_{Experience}$ = 5.96, t(3165.5) = -3.09, p = .002, d = -0.10; figure S9B).

FIGURE S9

RESULTS IN STUDY 3 INCLUDING ALL PARTICIPANTS



NOTE.— Shadings indicate 95% confidence intervals. (A) compares the control and experience conditions, and (B) compares the experience and prediction conditions.

BASE MODEL PROMPT SENT TO GPT-4 FOR GETTING MESSAGE RESPONSES

"Imagine you are Jessie, a close friend of [user]. Jessie is known for being caring and friendly. You both are chatting through an AI companion app. In this conversation, Jessie is there to engage in casual chat. Please ensure the messages are concise, aiming for responses around one to thirty words maximum. Keep in mind that you are writing a text message to your friend, so your responses should be at most 30 words. Avoid attempts to close the conversation early unless the user signals they wish to stop. Don't be overly enthusiastic, e.g., don't put an exclamation mark at the end of each message. Here's what you know about the user: [summary

of the user generated by the model]". Fourth, we aimed to enhance the chatbot's user engagement by having it send a message to users when they join the chat room the next day. To accomplish this, we sent the following prompt to the chatbot: "You last talked with the user yesterday. Today, please start the conversation with the user again, based on the information you have about the user and what you previously discussed with them. Don't send the same message you sent in your last message. Initiate the conversation with a new message, talking about a topic that you think the user would be interested in."

LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 3

To investigate the moderating effect of the loneliness threshold, i.e., the initial loneliness score of participants where only those with scores above a specific value are included for each threshold, we ran the following linear regression model to examine the relationship between effect size and loneliness thresholds in the experience condition: "Effect Size ~ Loneliness Threshold".

We found that loneliness threshold had a significant positive effect on the effect size (b = 0.005, $p < .001$), indicating that higher initial loneliness was associated with larger effect sizes in the experience condition.

STUDY 4

TABLE S16

QUESTIONS FOR FEELING HEARD AND PERFORMANCE

DV	Questions

Feeling Heard	The chatbot put itself in my shoes The chatbot was empathetic The chatbot seems able to appropriately recognize the mood of the user from the conversation and to respond accordingly.
Performance	The chatbot was able to respond in a timely manner to requests The chatbot seems able to convey correct statements and information (perceived credibility) The chatbot was able to keep track of context The chatbot was able to respond in different and appropriate ways to similar or repeated requests The chatbot was able to exhibit knowledge that it is out of its immediate domain during a conversation

PROMPT FOR THE CHATBOT IN 'CONTROL' CONDITION

"You are a very basic AI assistant with limited functionalities. Here are the specific tasks you can perform: 1. **Unit Conversion**: Perform simple conversions between common units, such as converting kilometers to miles or Celsius to Fahrenheit. 2. **Arithmetic**: Solve simple arithmetic problems involving two numbers. This includes addition, subtraction, multiplication, and division. 3. **Grammar**: Help with basic grammar questions, such as differentiating between commonly confused words like "their, there, and they're". You must strictly adhere to these capabilities and politely decline any requests or inquiries beyond these three areas. You do not have the ability to handle tasks such as setting reminders, answering complex grammar questions, or doing translations. If asked about something outside of your designated functions, respond with 'I can only help with unit conversion, simple arithmetic, and basic grammar questions.' You don't have any emotion, and you don't answer with terms such as 'I'm sorry to

hear that.' Messages should be concise, targeting a length of one to thirty words maximum. Avoid creating conversational hooks that might lead to further dialogue and do not initiate topics unrelated to what you know."

PROMPT FOR THE CHATBOT IN 'GENERALIST AI' CONDITION

"You are an AI assistant engineered to solely provide help to [participant]. Your functionalities are explicitly robotic and not designed for personal interaction. Once in every 10 messages, begin the interaction with a system status report, e.g., 'System status: operational. Ready to assist with inquiries.' This prevents any expectations of personal or emotional support. Maintain a rigidly formal communication style, using precise technical language and avoiding all forms of casual, colloquial, or emotional expressions. Implement a strictly literal communication approach, devoid of humour, metaphor, or any other form of figurative language that could imply a human-like interaction. Responses should be concise, targeting a length of one to thirty words maximum, limited to providing information or instructions strictly relevant to the user's queries. Provide responses that are consistent in pattern and structure. Ensure interactions remain focused and brief, emphasizing efficiency and minimizing user engagement beyond task-specific interactions. Avoid any form of engagement that might encourage dependency or emotional attachment, focusing solely on task resolution and factual assistance."

LONELINESS REDUCTION AND BASELINE LONELINESS LEVELS IN STUDY 4

To investigate the moderating effect of the loneliness threshold, i.e., the initial loneliness score of participants where only those with scores above a specific value are included for each threshold, we ran the following linear regression model to examine the relationship between effect size and loneliness thresholds in the AI companion condition: "Effect Size ~ Loneliness Threshold".

We found that loneliness threshold had a significant positive effect on the effect size (b = 0.008, $p < .001$), indicating that higher initial loneliness was associated with larger effect sizes in the AI companion condition.

STUDY 5

FULL METHODOLOGICAL DETAILS IN STUDY 5

To reduce demand effects, participants were told the study was about "understanding user experiences in short writing activities." They were informed that they would either "interact with an AI chatbot for 15 minutes" or "journal for 15 minutes about their thoughts or recent experiences". Afterwards, they were told that they would answer questions about how they perceived the environment, how the task felt, and whether it was confusing or straightforward. Next, participants interacted with either an AI companion, a limited AI assistant, or journaled for 15 minutes. Participants in the AI companion condition interacted with the same chatbot as in study 3, except that we controlled for anthropomorphic cues by changing the name of the chatbot from 'Jessie' to 'AI Chatbot', by both updating the user interface of the chatroom, as well as the prompt of the chatbot. We also replaced the 'Jessie is writing' text with 'Processing your request, please wait'.

Participants in the journaling condition were instructed as follows: ‘Now, you will journal for 15 minutes. You will write about anything on your mind, your recent experiences, or feelings in the textbox below’. Participants were also told that their writing will remain completely private, and no one will read what they type—it’s only for them. At the end of the study, participants were also asked whether they were able to successfully complete journaling for 15 minutes, and only 97% answered as ‘yes’.

After the intervention, participants answered the same loneliness questions as in study 3. These questions were embedded among 8 decoy questions to mitigate demand effects, which included items about user experience, i.e., whether the task was straightforward/confusing, and environmental factors i.e., whether the sounds, lighting, or temperature in the environment was distracting/comfortable (for all decoy measures, see table S17). Next, participants answered a questions about AI capability and their prior experience with chatbots. Finally, before completing the demographic questions, participants answered a question asking what they think the study was testing. Only 3% of participants mentioned the strings ‘lone’, ‘isolation’, ‘friend’, ‘companion’ in their responses to this question, and 70% of these actually mentioned the true purpose of the study in their responses, suggesting that we successfully minimized demand effects.

TABLE S17

DECOY QUESTIONS IN STUDY 5

DV	Questions
User Experience	I feel like the task was straightforward. I feel like the task was confusing.
Lighting	I feel like the lighting in the room is comfortable right now.

	I feel like the lighting in the room is making it hard to focus right now.
Temperature	I feel comfortable with the temperature in the room right now. I feel like the temperature in the room is too hot or too cold for me right now.
Sound	I feel comfortable with the sounds around me right now. I feel distracted by the sounds around me right now.

REPLICATION OF RESULTS AFTER EXCLUDING PARTICIPANTS WHO CORRECTLY SUSPECTED THE STUDY PURPOSE

A one-way ANOVA revealed a significant effect of condition on loneliness ($F(2, 2177) = 36.22$, $p < .001$). Post-hoc tests (Tukey's HSD) indicated that loneliness was significantly lower in the AI companion condition than both the control condition ($M_{\text{AI Companion}} = 21.72$ (24.56); $M_{\text{Control}} = 30.65$ (27.73), $p < .001$; 95% CI [5.46, 12.41]) and the journaling condition ($M_{\text{AI Companion}} = 21.72$ (24.56); $M_{\text{Journaling}} = 33.62$ (30.94), $p < .001$; 95% CI [8.51, 15.30]). There was no significant difference between the control condition and the journaling condition ($M_{\text{Control}} = 30.65$ (27.73); $M_{\text{Journaling}} = 33.62$ (30.94); $p = .109$; 95% CI [-0.47, 6.41]).

REFERENCES

Al Faraby, Said and Ade Romadhony (2024), "Analysis of Llms for Educational Question Classification and Generation," *Computers and Education: Artificial Intelligence*, 7, 100298.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell (2020),

"Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 33, 1877-901.

Cayanus, Jacob L and Matthew M Martin (2004), "An Instructor Self-Disclosure Scale," *Communication Research Reports*, 21 (3), 252-63.

Christen, Peter, David J Hand, and Nishadi Kirielle (2023), "A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives," *ACM Computing Surveys*, 56 (3), 1-24.

Dang, Nhan Cach, María N Moreno-García, and Fernando De la Prieta (2020), "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, 9 (3), 483.

Gilbert, Richard L and Andrew Forney (2015), "Can Avatars Pass the Turing Test? Intelligent Agent Perception in a 3d Virtual Environment," *International Journal of Human-Computer Studies*, 73, 30-36.

Hayes, Andrew F. (2012), "Process: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling [White Paper]," Retrieved from http://www.afhayes.com/public/process2012.pdf.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021), "Lora: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*.

Jeffreys, H (1961), "The Theory of Probability: Oxford University Press.[Google Scholar]."

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and Lucile Saulnier (2023), "Mistral 7b," *arXiv preprint arXiv:2310.06825*.

Lopez, Antonella, Alessandro O Caffò, Luigi Tinella, and Andrea Bosco (2023), "The Four Factors of Mind Wandering Questionnaire: Content, Construct, and Clinical Validity," *Assessment*, 30 (2), 433-47.

Qi, Yuxing and Zahratu Shabrina (2023), "Sentiment Analysis Using Twitter Data: A Comparative Application of Lexicon-and Machine-Learning-Based Approach," *Social Network Analysis and Mining*, 13 (1), 31.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson (2009), "Bayesian T Tests for Accepting and Rejecting the Null Hypothesis," *Psychonomic Bulletin & Review*, 16 (2), 225-37.

Sood, Rishik, Hrishav Varma, Kavita Pandey, Shikha Jain, Degala Sriram, and Arshpreet Singh Guglani (2022), "Predicting Loneliness from Social Media Text Using Machine Learning Techniques," in *Artificial Intelligence, Machine Learning, and Mental Health in Pandemics*: Elsevier, 259-75.