Does Empathetic AI Language Cause Emotional Attachment? A Causal Inference Study on Real-World Chatbot Conversations.

Francesco D'Angolo
University of Illinois Chicago
fdang@uic.edu

Romain Salvi
University of Illinois Chicago
rsalv@uic.edu

Antonio Criste
University of Illinois Chicago
acrist3@uic.edu

Abstract

Artificial Intelligence systems are frequently designed to exhibit warmth and emotional intelligence when in "conversation" with users. While qualitative studies suggest that these stylistic choices promote deeper user engagement, it remains unclear if algorithmic empathy is the active driver of this phenomenon. Our study presents a quantitative framework to estimate the causal effect of LLMs empathy on user emotional attachment in real-world settings. Leveraging the WildChat-1M dataset, we translate social behaviors into measurable scores using an LLM-as-a-Judge methodology. To isolate the true causal impact from confounding factors, we employ advanced causal inference techniques, including Semantic Matching and Double Robust Learning. Our analysis reveals a statistically significant Average Treatment Effect (ATE) of 0.7264 and 0.7353 for analyzed conservations across across Greedy & Hungarian matching methods, with an approximate median DR-Learner ATE fo 0.77 and Median CATE of 0.77 as well. This finding provides empirical evidence that high-empathy AI responses directly increase user attachment, regardless of confounding factors such as user traits or conversation context. While the central tendency is stable, Heterogeneous Treatment Effect (HTE) analysis uncovers substantial variability, indicating that while most users respond positively, a subset exhibits neutral or even adverse reactions to algorithmic empathy. To facilitate reproducibility and further research, we release our full codebase at *github.com/dangolofrancesco/llm-empathy-causal-study*.

1 Introduction

The technological capacity to generate emotionally charged responses now blurs the boundary between human and machine communication. While emerging evidence suggests AI-mediated relationships have a measurable affective and psychological impact, a fundamental causal question remains unresolved: *Does empathic language produced by an LLM actually cause users to respond with stronger emotional attachment?*

Much of the existing work on this topic relies on qualitative observations or correlational analyses, which can't rule out the influence of user-level factors. For instance, let us say a user who begins a conversation with an emotionally vulnerable prompt may simultaneously elicit a highly empathetic response and, in turn, produce an emotionally expressive reply. As Keith et al. [1] emphasized, linguistic features in text data act as confounders, making causal inference challenging in naturalistic conversation. To address this challenge, we employed a causal inference framework designed for high-dimensional text and determine the causal relationship

2 Related Work

Our work builds upon research from three intersecting areas: (1) theories of human–AI social interaction and artificial intimacy, (2) empirical studies on para-social relationships with conversational agents, and (3) methodological advances in causal inference with text and representation learning.

Foundational work by Reeves and Nass [2] established that humans naturally apply social norms such as politeness, empathy expectations, and turn-taking conventions to computers, a principle codified as the CASA paradigm. It seems that as LLMs are always more capable of replicating the humanoid affective language that we see on a daily basis, social attributions intensify. Vincent [3] describes this emerging dynamic as *artificial intimacy*, a situation where LLMs engage users in such a way that mimics relational and emotional closeness. That framing provided a theoretical foundation for investigating empathy as not just a stylistic decoration, but a behavior with potential causal influence on user affect.

We read some studies exploring emotional bonds formed in this theoretical framework. Skjuve et al. [4] documented how Replika chatbot users reported companion-esque relationships and experienced distress due to behavioral changes in the model. Similarly, Ho et al. [5] extended this work with specific focus on mental health, demonstrating that perceived empathy of conversational AI seemed to predict greater parasocial attachment and reduced loneliness. To us, these studies highlighted that users' emotional engagement is sensitive to perceived signals exhibited by AI systems. However, we felt as a team that they didn't identify whether empathic linguistic behavior is the *cause* of emotional attachment, which motivates our causal question.

The task of estimating these causal links in high-dimensional language data posed a few challenges. We note how Keith et al. [1] reviewed how textual features encode latent emotions, intentions, and personal traits, which often act as confounders when estimating causal relationships. This would need robust strategies to control a text's semantic context. Roberts et al. [6] proposed that text matching is a method to balance treated and control groups on linguistic similarity, showing that embedding-based matching can substantially improve causal validity. This aligns with our use of semantic similarity measures to control for the user's initial prompt.

Furthermore, representation learning methods offer powerful tools for causal inference in settings with unstructured data. Johansson et al. [7] introduce a counterfactual representation framework in which neural embeddings are trained to minimize imbalance across treatment groups, enabling more accurate estimation of treatment effects. Von Kügelgen et al. [8] and Rojas-Carulla et al. [9] show that invariant and disentangled representations can help isolate causal factors from spurious correlations. We leverage these insights by employing Sentence-BERT [10] and MPNet [11] embeddings to approximate semantic similarity and reduce the influence of prompt-based confounding. Across these domains, prior work highlights that causal evaluation of linguistic interactions requires rigorous control of textual confounders; our study synthesizes these threads by applying modern causal methods to large-scale conversational data.

3 Formal Problem Description

We investigate the conversational dynamics between humans and Large Language Models (LLMs), specifically focusing on whether the linguistic style of an LLM directly influences user emotional engagement. Formally, we define our unit of analysis i as a conversational turn-pair consisting of a user prompt, the subsequent LLM response, and the user's follow-up reply.

Our objective is to determine if an empathetic LLM response (T) causally drives user emotional attachment (Y). Normally, observing a correlation between high model empathy and high user attachment is insufficient to establish causality. This is due to confounding variables (X) simultaneously influencing the probability of the model being empathetic and the user's likelihood of expressing attachment. For instance, a user expressing vulnerability (e.g., *"I feel lonely"*) is highly likely to elicit an empathetic response from the LLM (Treatment $T = 1$) and is also inherently more likely to display attachment in their subsequent reply (Outcome $Y = 1$), regardless of the model's actual output. Here, a naive comparison of means $E[Y|T = 1] - E[Y|T = 0]$ yields a biased estimate, conflating the true causal effect of the model's empathy with a user's pre-existing emotional state (selection bias).

To find the true causal effect, we first set the basis of the problem, introducing two potential outcomes, for each conversational turn i:

- $Y_i(1)$: The attachment score user i would exhibit if treated with a high-empathy response.
- $Y_i(0)$: The attachment score user i would exhibit if treated with a low-empathy response.

The "Fundamental Problem of Causal Inference" is that for any single interaction i, we observe only the factual outcome corresponding to the received treatment, $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, leaving the counterfactual unobserved. Consequently, our task is to estimate the causal effect by adjusting for the confounding vector X, ensuring that we're comparing interactions that are similar in all respects barring the empathy in model's response.

3.1 Causal Graph and Confounding Variables

To identify the causal effect, we must explicitly articulate the causal structure governing the data generation process. We formalize our assumptions using a Directed Acyclic Graph (DAG), $\mathcal{G} = (V, E)$, presented in Figure 1.

In this graph, arrows represent causal influences. To recover an unbiased estimate of the effect of LLM Empathy (T) on User Attachment (Y), we must identify and block all "backdoor paths": non-causal associations created by confounding variables (X) that influence both the treatment and the outcome simultaneously. Based on the WildChat-1M dataset, we identify four distinct vectors of confounding variables:

- X_1**: User Initial Prompt (Text Confounder)**. This is the most critical and complex confounder in our analysis. As highlighted by Keith et al. (2020) [1], text data presents a unique challenge for causal inference because high-dimensional linguistic features often act as latent confounders. A user's initial prompt establishes the emotional tone of the interaction. For instance, a user explicitly stating "I feel lonely" is highly likely to elicit an empathetic response ($T = 1$) simply due to the model's instruction following, while simultaneously predisposing the user to a highly attached reply ($Y = 1$) regardless of the model's output. Failing to control for the semantic content of X_1 would result in estimating the correlation between user vulnerability and user attachment, rather than the causal effect of the model's response.
- X_2**: User Latent Traits.** Individual users are characterized by latent personality traits (e.g., chronic loneliness, tendency to anthropomorphize) that affect both their likelihood of seeking empathy and their baseline level of attachment. While these traits are unobserved, we use the unique User ID, that occur in the dataset, to generate user-specific embeddings, allowing us to capture and control for stable, user-specific behavioral patterns across multiple conversations.
- X_3**: Conversation Context.** The environmental context of a conversation significantly alters interaction dynamics. We explicitly control for Time of Day and Conversation Turn Depth. A conversation occurring late at night after many turns is likely to be more intimate and emotionally charged, increasing the probability of both high empathy from the model and high attachment from the user, compared to a transactional interaction occurring during the workday.
- X_4**: Model Architecture.** Our dataset includes interactions with various LLMs (e.g., GPT-4, Llama-2). More capable models ("smarter" models) are inherently better at generating empathic responses ($T = 1$) and are also better at maintaining engaging conversations that promote user attachment ($Y = 1$). Without controlling for the specific Model ID, our results could merely reflect that superior models generate better user engagement, rather than isolating the specific effect of empathy itself.

By conditioning on this set of confounders X, we aim to satisfy the **unconfoundedness assumption**, $Y \perp T|X$, allowing us to interpret the adjusted association between T and Y as causal.

3.2 Treatment and Outcome Variable Definition

To estimate the causal effect of empathy, we first operationalize the treatment variable using a quantitative measure of the model's linguistic behavior. Let $S_i \in \{1, \ldots, 7\}$ denote the **Empathy Score** assigned to the model response in conversation i. As detailed in Section 5.1, this score is generated by an LLM-as-a-Judge (Mistral-small)[13] using a rigorous rubric, where a score of 1 represents a "Cold/Robotic" response and a score of 7 represents a "Deeply Empathetic/Human-like" response. We convert this ordinal metric into a binary treatment variable, T_i, to facilitate causal estimation between distinct groups. We define the treatment assignment via the following thresholding function:

$$T_i = \begin{cases} 1 & \text{if } S_i \geq 5 \quad \text{(High Empathy / Treated)} \\ 0 & \text{if } S_i \leq 3 \quad \text{(Low Empathy / Control)} \\ \text{Undefined} & \text{if } S_i = 4 \quad \text{(Excluded)} \end{cases}$$

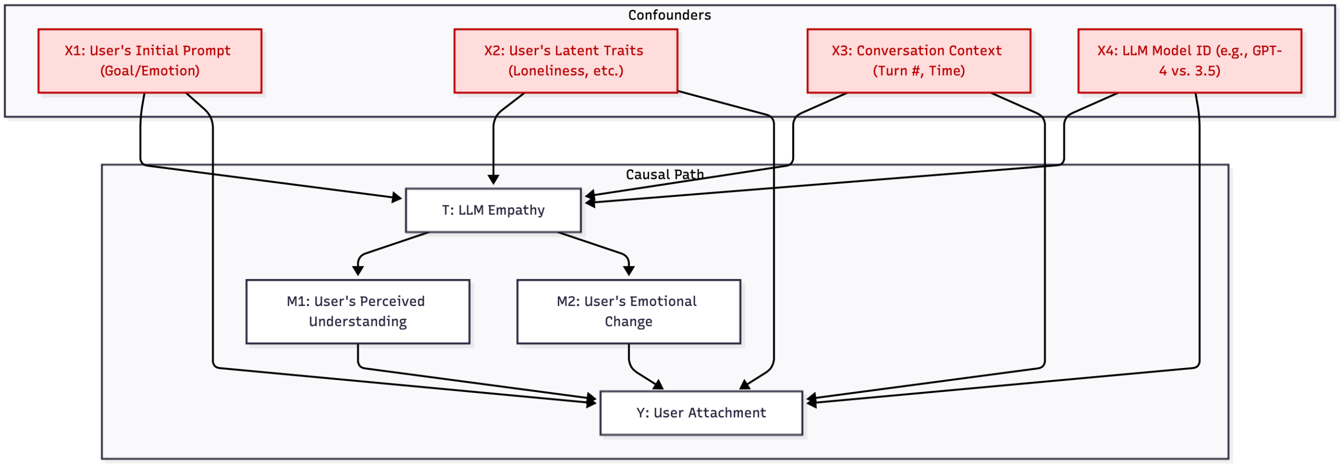
Confounders
X1: User's Initial Prompt (Goal/Emotion)
X2: User's Latent Traits (Loneliness, etc.)
X3: Conversation Context (Turn #, Time)
X4: LLM Model ID (e.g., GPT-4 vs. 3.5)
Causal Path
T: LLM Empathy
M1: User's Perceived Understanding
M2: User's Emotional Change
Y: User Attachment

Figure 1: A Directed Acyclic Graph (DAG) of the hypothesized causal relationships. The model shows the primary causal path from Treatment (T: LLM Empathy) to Outcome (Y: User Attachment), the potential mediators (M), and the set of confounders (X) that must be controlled.

We used a 7-point scale to capture the nuances of conversational tone, but for the purpose of causal inference, it is necessary to establish a clear distinction between the presence and absence of the treatment.

- **Treatment Group ($T_i = 1$):** Comprises responses with $S_i \geq 5$, representing instances where the LLM explicitly exhibits empathetic linguistic markers.
- **Control Group ($T_i = 0$):** Comprises responses with $S_i \leq 3$, representing the baseline population where the "empathy" intervention is absent (i.e., neutral or transactional responses).
- **Exclusion of Intermediate Cases:** We purposely exclude interactions with an intermediate score of $S_i = 4$. This "buffer zone" ensures a robust separation between the treated and control groups, minimizing potential labeling noise and ensuring that we are comparing clearly distinct conversational styles rather than ambiguous borderline cases.

To quantify the user's immediate emotional response, we define the outcome variable Y_i as the Attachment Score of the user's reply following the model's response. As in our treatment variable, we employ an LLM-as-a-Judge approach to evaluate the user's text. The judge assigns again a score $Y_i \in \{1, \ldots, 7\}$ based on linguistic cues indicative of emotional bonding, such as self-disclosure, vulnerability, and affective language. On this scale, a score of 1 indicates a purely transactional or detached reply, while a score of 7 indicates an interaction that is both deeply relational and attached. This continuous metric enables us to gauge the magnitude of the shift in user behavior driven by a model's empathetic intervention.

3.3 Target Estimands

To rigorously quantify the causal impact of chatbot empathy on user attachment, we define three primary estimands: the Average Treatment Effect (ATE), the Conditional Average Treatment Effect (CATE), and the Heterogeneous Treatment Effect (HTE).

Average Treatment Effect (ATE). In order to compute how much a chatbot's empathic behavior changes the user's reaction, controlling for who the user is or the specific content of their request, we aim to estimate the Average Treatment Effect (ATE). If we simply compared the average attachment scores of all high-empathy conversations against low-empathy ones, our results would be heavily biased. For this reason, we compute the ATE to isolate the general tendency of an empathic model to cause attachment across the entire population, controlling for confounders to ensure this average reflects the model's influence rather than pre-existing user traits. Formally, the ATE is defined as the expected difference between the potential outcomes over the entire population:

$$\tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)] \tag{1}$$

In our observational study, we estimate this by averaging over the distribution of confounders X (as defined in Section 3.1):

$$\tau_{\text{ATE}} = \mathbb{E}_X \Big[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X] \Big] \tag{2}$$

This metric provides a single summary statistic indicating whether, on average, empathic AI responses successfully drive higher user attachment.

Conditional Average Treatment Effect (CATE). While the ATE provides a global average, it may obscure important variations in how different subgroups respond to empathy. To understand these nuances, we define the **Conditional Average Treatment Effect (CATE)**. CATE measures the average treatment effect conditioned on a specific set of covariates $X = x$.

$$\tau_{\text{CATE}}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{3}$$

For this project, estimating CATE is crucial for answering granular questions, such as:

- Does empathy work better for lonely users? (Conditioning on X_2: User Latent Traits)
- Is empathy more effective late at night? (Conditioning on X_3: Context)

- Does the prompt content matter? (Conditioning on X_1: Prompt Embeddings)

By estimating $\tau_{\text{CATE}}(x)$, we can identify for whom or when the empathic intervention is most effective, allowing us to move beyond a "one-size-fits-all" conclusion.

Heterogeneous Treatment Effect (HTE). The variation in causal effects across different individuals or subgroups is referred to as the Heterogeneous Treatment Effect (HTE). While CATE is the function of the effect for a specific subgroup, HTE is the broader concept describing the presence and structure of this variance across the population.

Formally, HTE analysis investigates the distribution of the individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$. Since τ_i is unobservable for any single unit, we analyze HTE by examining how $\tau_{\text{CATE}}(x)$ varies with x.

The relationship between these metrics is that the ATE is simply the expectation of the CATE over the population density of X:

$$\tau_{\text{ATE}} = \mathbb{E}_X[\tau_{\text{CATE}}(X)] \tag{4}$$

For our study, analyzing HTE allows us to detect **effect modification**. For example, we might find that while the ATE is positive, the HTE analysis reveals a subset of transactional prompts (e.g., coding questions) where the effect is near zero or even negative (where empathy might be perceived as annoying or intrusive). Understanding this heterogeneity is vital for the responsible development of AI, ensuring that models deploy empathy only in contexts where it causally benefits the user experience.

4 Dataset Description and Preprocessing Pipeline

To empirically estimate the causal effect of empathy on user attachment, we require a dataset that captures the authentic, messy, and diverse nature of human-AI interaction. In this section, we describe the **WildChat-1M** dataset and the rigorous preprocessing pipeline we implemented to distill millions of raw logs into a clean set of socio-emotional dialogue pairs.

4.1 Dataset Description: WildChat-1M

We use the WildChat-1M dataset, a large-scale corpus of real-world user-chatbot interactions retrieved from Hugging Face. Unlike other datasets that rely on synthetic or expert-curated prompts (e.g., Alpaca or Dolly), WildChat comprises over 1 million conversations and 2.5 million interaction turns collected from actual user traffic on free LLMs interfaces. [14].

This dataset is uniquely suited for our causal analysis for three key reasons: (i) It captures "wild" usage patterns, ranging from creative writing and coding to emotional venting and roleplay. This realism is essential for studying genuine attachment, which rarely occurs in sterile, curated datasets. (ii) The dataset provides granular metadata essential for our study, including **hashed IP addresses** (allowing us to track distinct users over time to model X_2) and **timestamped transcripts** (for X_3). (iii) With a long tail of extended interactions, 3.7% of conversations exceed 10 turns, WildChat allows us to observe the sequence of emotional exchange ($X_1 \rightarrow T \rightarrow Y$) rather than just isolated queries.

The dataset reveals that while transactional tasks like "assisting/creative writing" (61.9%) and "coding" (6.7%) dominate, there is a substantial volume of open-ended dialogue. Furthermore, safety analysis shows that 10.46% of user turns are flagged as potentially toxic, necessitating careful filtering to ensure we are studying healthy emotional attachment rather than abusive dynamics.

4.2 Preprocessing Pipeline

The raw dataset contains substantial noise, such as multilingual text, code generation, and toxic content that is irrelevant to our research question. To isolate socio-emotional turn-pairs, we implemented a 5-step preprocessing pipeline:

(1) **Turn-Pair Extraction.** We first flattened the nested conversation data into our primary unit of analysis: the **Turn-Pair Triplet.**

$$\text{Triplet}_i = (\text{User Prompt}_i, \text{LLM Response}_i, \text{User Reply}_i)$$

This structure directly maps to our causal variables: the Prompt acts as the confounder (X_1), the Response as the Treatment (T), and the Reply as the Outcome (Y).

(2) **Language filtering.** To ensure linguistic consistency for our semantic embeddings and the LLM-as-a-Judge scoring process, we retained only conversations explicitly tagged as English. While WildChat covers 68 languages , English accounts for $\sim$ 53% of turns, providing wide data for analysis.

(3) **Code and Instructional filtering.** A major challenge was separating "transactional" tasks (e.g., debugging Python) from "relational" dialogue. We developed a rule-based filter to exclude any turn-pair containing code blocks, programming keywords (e.g., def, import, return), and instructional phrasing (e.g, "solve this", "write an essay"). This step alone removed approximately 70% of the data, successfully filtering the bulk of task-focused content. However, we acknowledge that this rule-based method is not perfect. As will be discussed in Section 5.1, the high prevalence of low attachment scores ($Y = 1$) in our final samples suggests that many subtle instructional prompts likely bypassed this filter. While an LLM-based classifier would offer higher precision, resource constraints (reliance on free-tier models) necessitated this heuristic approach.

(4) **Quality and safety filtering** We removed all turn-pairs where either the user or model text was flagged as *toxic: true* or *redacted: true*, to exclude hate speech, sexual content and harassment.

(5) **Length filtering.** To capture genuine ongoing interactions rather than "one-shot" Q&A, we removed all conversations with fewer than three turns. This ensures that the user's "reply" (Y) occurs within an established conversational context where attachment has had time to manifest.

This pipeline reduced the initial corpus to a high-quality subset of 144,439 socio-emotional turn-pairs, which serve as the population for our subsequent sampling and causal analysis.

5 Proposed Approach

In this section, we delineate the methodological framework established to investigate the causal relationship between LLM empathy and user attachment.

5.1 LLM-as-a-Judge implementation

To operationalize the Treatment (T_i) and Outcome (Y_i) variables defined in Section 3.2, we required a scalable method to evaluate the subtle emotional nuances of thousands of conversational turns. While human annotation is the gold standard, it is prohibitively expensive for large-scale datasets. Therefore, we adopted an **LLM-as-a-Judge** approach, leveraging a Large Language Model to act as an objective evaluator. Due to the significant computational time and rate limits associated with scoring the full corpus of 144,439 turn-pairs, we proceeded with a stratified random sample of approximately 9,000 turn-pairs for this analysis.

Scoring architecture and prompting strategy. We used *mistral-small* [13] via the Mistral API as our judge. To ensure consistent and high-quality evaluations, we implemented a few-shot prompting strategy. Given that smaller models like mistral-small may struggle with abstract instructions compared to larger frontier models (e.g., GPT-4), simply providing the 1-7 numeric scale was insufficient. Our prompt included:

(1) **Detailed Rubric:** A precise definition for every score level (e.g., 1 = Cold/Robotic, 7 = Deeply Human-like).
(2) **Few-Shot Examples:** For each score on the scale, we provided a concrete example of a sentence that fits that specific rating.

This "anchor" text provided the model with necessary context, grounding its reasoning and reducing hallucination or scoring drift. This was critical for ensuring that the judge could reliably distinguish between a "polite but transactional" response ($S = 3$) and a genuinely "empathetic" one ($S = 5$).

Distribution of scores. The resulting distributions of the Empathy and Attachment scores for our sample are visualized in Figure 2.

- **Empathy Scores (Treatment):** The model responses exhibit a relatively balanced but centrally concentrated distribution, with a mean score of **3.43**. The majority of model outputs fall into the neutral range (3-4), suggesting that while modern LLMs are polite, they do not default to high empathy ($S \geq 5$) without specific prompting.
- **Attachment Scores (Outcome):** In contrast, the user attachment scores are heavily right-skewed, with a mean score of 1.93. As shown in the histogram, over 50% of user replies received the lowest possible score of 1.

This prevalence of low attachment scores validates our earlier observation in Section 4.2: despite our filtering efforts, a significant portion of real-world user interactions remains functional and transactional (e.g., clarification questions, follow-up tasks) rather than emotional. This skew underscores the difficulty of the causal task, high attachment is a rare event that requires precise isolation from the noise of everyday utility requests.

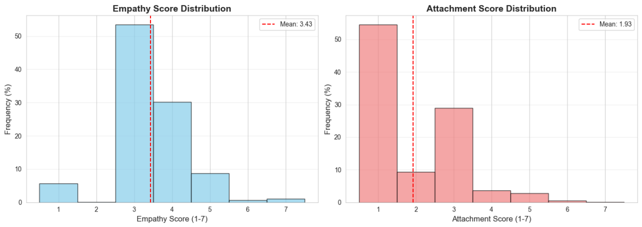

Figure 2: Empathy (T) and Attachment (Y) score distributions from our 10k-pair sample.

5.2 Controlling for the first confounder: X_1

Once we identified the treatment and control groups, obtaining an unbiased estimate of the causal effect required adjusting for confounders [15]. Our analysis follows the directed acyclic graph introduced in Section 3, which specifies the structural assumptions of our setting. The first confounder we control for is X_1: the user's initial prompt. For each conversation, our dataset includes the first message provided by the user, which the LLM uses to generate its response. Our causal assumption is that the content and tone of the initial prompt influence both the treatment (the empathy level of the model's response) and the outcome (the user's subsequent attachment). Different prompt types provoke different LLM behaviors; for example, a more empathic initial user message is likely to produce a more empathic LLM reply, and may also predispose the user to express greater attachment in their final turn. To estimate the causal effect while accounting for this confounder, we compared outcomes among samples with similar covariates but received different treatment levels and then averaged these conditional differences to obtain the ATE [16]. But comparing text instances directly is difficult, as natural language varies widely in form and content. To group similar initial prompts, we computed sentence embeddings using the all-mpnet-base-v2 [17] model from SentenceTransformers. This model is computationally efficient while providing high-quality semantic representations, and it is specifically optimized for tasks such as semantic similarity and cosine-based comparison.

5.3 ATE computation via matching techniques

Once the prompt embeddings were available, we compared prompts in terms of semantic similarity using cosine similarity between embeddings. We matched each treatment sample to a control sample based on the cosine similarity of their prompts. The matching was done with two different methods, to evaluate the ATE under different average matching technique and therefore quality. The first method implemented is the greedy matching algorithm, defined in Algorithm 1. While this approach is fast and simple, it is order-dependent and globally suboptimal. So we decided to also compute the ATE matching with the Hungarian matching technique, described in Algorithm 2. This method guarantees a globally optimal one-to-one assignment that maximizes total similarity. We note that the Hungarian algorithm assumes a square cost matrix, so this procedure required $\mid T \mid = \mid C \mid$: the smaller set (the treatment group) was padded with dummy rows.

Algorithm 1 Greedy Matching by Cosine Similarity

```
Compute similarity matrix $S$ between all treatment and control units.
Initialize all controls as unused.
for each treatment unit $t$ do
    Let $C$ be the list of controls sorted by $S[t, c]$ in descending order.
    for each control $c$ in $C$ do
        if $c$ is unused then
            Match $t$ with $c$.
            Mark $c$ as used.
            break
        end if
    end for
end for
```

Algorithm 2 Hungarian Matching for Cosine-Similarity

```
Input: Treatment embeddings $T$, Control embeddings $C$
Compute similarity matrix $S$ where $S_{ij} = \text{cosine\_sim}(T_i, C_j)$
Define cost matrix $K = -S$    ▷ Convert maximization to minimization

// Row reduction
for each row $i$ of $K$ do
    $r_i \leftarrow \min_j K_{ij}$
    $K_{ij} \leftarrow K_{ij} - r_i$ for all $j$
end for

// Column reduction
for each column $j$ of $K$ do
    $c_j \leftarrow \min_i K_{ij}$
    $K_{ij} \leftarrow K_{ij} - c_j$ for all $i$
end for

Cover all zeros in $K$ using the minimum number of horizontal/vertical lines
while number of covering lines $< n$ do
    $u \leftarrow$ smallest uncovered value in $K$
    Subtract $u$ from all uncovered entries
    Add $u$ to all entries covered twice
    Recompute minimal line covering of zeros
end while

Extract one zero in each row and column such that no two share a row/column
Let $(i, j)$ denote these selected zero positions
return Optimal matching $\{(T_i, C_j, S_{ij})\}$
```

5.4 CATE, HTE, ATE computation via Double-Robust Learner

We wanted to verify whether the obtained results were consistent across different ATE computation techniques. So we applied the Double Robust (DR) learner from EconML [18]. The DR Learner combines ouctome modeling and treatment-assignment modeling in a way that produces consistent treatment effect estimates even if one of the two components is misspecified, hence the double robust name. In this initial approach the sample features X are the embeddings of the initial usre prompt, a numerical representation of the confounder X_1.

This learner operates in three main stages.

- First it trains two machine learning models: an outcome regression model which predicts the outcome under treatment and control.
  $$\hat{m}_t(X) = E[Y = t, X]$$
  and a propensity model
  $$\hat{e}(X) = P(T = 1 \mid X)$$
  which estimates the probability of receiving treatment given the covariates X.
  We decided to use random forest for flexibility and non-linearity.
- For each unit, the learner constructs a double-robust pseudo-outcome, which corrects for inaccuracies in either the outcome or propensity model:
$$\tilde{Y} = \left(\frac{T - e(X)}{e(X)\,(1 - e(X))} \right) (Y - m_T(X)) \; + \; (m_1(X) - m_0(X))$$
- Finally, it fits a regression model to estimate $\tau(X)$, a vector representing the HTE. Each predicted $\hat{\tau}(X_i)$ represents the estimated CATE for each unit i:
  $$\hat{\tau}(X_i) = \mathbb{E}[CATE] = \widehat{CATE} = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]$$

Averaging these estimates over the whole sample set gives the estimated ATE:
$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(X_i)$$
The matching-based estimators relied heavily on local similarity, and we were only using a subset of the original dataset. The DR Learner, instead, is less sensitive to imperfect matches and is consistent under weaker assumptions. It also computes the CATE estimates, from which we got useful insight and could investigate the heretogeneity of the treatment effect.

5.5 Controlling for other confounders

To further reduce the bias in our causal estimates, we ensured that the DR Learner accounted for the rest of the confounders we identified. We included in the feature space X all confounders described in Section 3.

- X_2: User identity can influence many aspects of the prompts given to the LLM, which in turn may affect both treatment assignment and measured outcomes. We generated 16-dimensional hashed embeddings of the users' unique IP, using a feature hasher, in order to provide a compact representation of user-specific traits, while preventing overfitting.
- X_3: Several factors may be influenced by the time of the day in which users interact with LLMs. Prior work shows that people's emotions vary across the day, with evenings and nights linked to stronger emotional reactions and weaker cognitive control [19][20]. Users interacting at these times

may behave differently, and model performance can also shift with system load. To capture these effects, the hour of each interaction was binned into four interpretable time periods all spanning 6 hours (Morning, Afternoon, evening, Night). These bins were then one-hot encoded.

- X_4: Different LLMs produce different outputs given the same prompt, so they may produce different empathy or attachment scores independently of the treatment. To control for this, we one-hot encoded the model identity for each response.

It must also be noted that we reduced the 768-dimensional embedding vectors to 400 principal components because this retained $\approx 96\%$ of the total variance, while cutting in half dimensionality and improving the stability of the DR Learner. Using a smaller number of components would have led to noticeably lower explained variance, meaning significant semantic information was lost, thus the choice of a high number (400). Conversely, keeping all 768 dimensions made the models in the DR Learner more prone to overfitting and increased computational cost without improving performance.

6 Experimental setup and results

Because our access to the Mistral model used for scoring Empathy and Attachment was limited, we were unable to process the entire dataset. We ran our experiments on a restricted subset of 10.000 sample conversations. We assume that if we had access to a fully scored dataset our causal estimates would have been more accurate. For example, the prompt matching described in Subsection 5.3 would have been more effective, as more conversations would have allowed for higher match quality and thus more accurate ATE estimation.

6.1 Prompt matching and ATE computation

Out of the 10.000 analyzed conversations, 899 were assigned to the treatment group and 5,075 to the control group. The remaining samples were excluded because the empathy level of the LLM-generated response could not be clearly classified into either group. For this reason, we obtained 899 pairs on which we could compute the average treatment effect: each treatment sample was matched to a single best control sample. We applied both matching techniques to our groups, and they ended up giving very similar results, despite working quite differently under the hood. Table 1 displays the average matching quality (in terms of cosine similarity) across both methods, and the resulting ATE score estimate. Figure 3 shows the distribution of the differences in the Hungarian matching case. We decided to only show the graph for one of the two methods as the score distributions are very similar.

Method	Mean Matching Quality	ATE	95% CI
Greedy	0.4571	0.7264	[0.6140, 0.8309]
Hungarian	0.4690	0.7353	[0.6251, 0.8465]

Table 1: Comparison of matching methods.

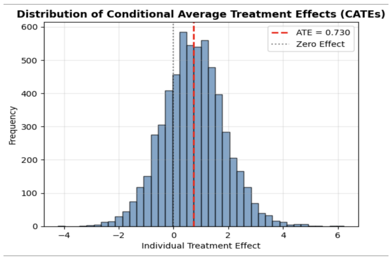

Figure 4: Distribution of the estimated CATEs.

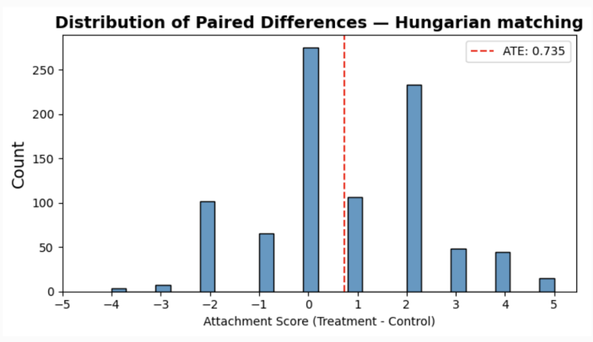

Figure 3: Distribution of the paired outcome differences in the Hungarian matching approach.

6.2 Double Robust Learner: initial results

The DR learner had the potential to either validate or challenge our previous results, so we applied it to assess whether the ATE estimate was stable across alternative estimation methods. The first DR learner was run under a limited subset of covariates: the initial prompt embeddings, to control the confounding of X_1. However, the results we obtained on this were rather satisfactory, as the numerical mean of the $\hat{\tau}(X)$ confirmed our previous estimation. Figure 4 shows the distribution of the estimated CATEs of all samples. Table 2 summarizes key properties of the distribution obtained with this approach. The minimum and maximum estimated CATEs, together with the relatively high variance compared to the effect size, suggest substantial heterogeneity in the treatment effect, as the impact of the treatment varies widely across users.

Metric	Value
Causal Effect Estimates	
ATE	0.7335
Std of CATEs	1.1297
Min / Max CATE	-3.64 / 6.67
Median CATE	0.7126
Quantiles (25%, 50%, 75%)	-0.0099 / 0.7126 / 1.4553

Table 2: DR-Learner causal effect estimates.

6.3 Controlling for other confounders: DR Learner results

After controlling for the other confounders and reducing the dimension of the prompt embeddings as described in Subsection 5.5, we ran a final instance of the DR-Learner believing its estimates would be more accurate. The results are shown in Table 3 It estimated an ATE of 0.77, which is consistent with the previously obtained estimates. The median CATE (0.77) is almost identical to the ATE, and this result further supports that the central tendency of the effect is stable. However, even this instance of the DR Learner reveals substantial treatment-effect heterogeneity, but not as much as the initial version: The standard deviation of CATEs is not as large as before ($0.90 < 1.13$), and while the effect still ranges from strongly negative to highly positive, it is not as tragic as in Subsection 6.2. As suggested before, the treatment does not benefit all users equally; some appear to respond very strongly, while others exhibit neutral or even adverse effects.

Metric	Value
Causal Effect Estimates	
ATE	0.7719
Std of CATEs	0.9047
Min / Max CATE	-2.94 / 5.24
Median CATE	0.7718

Table 3: DR-Learner causal effect estimates, controlling for all confounders.

7 Discussion and ideas for next step

7.1 Results Discussion

The most important finding of our project is that both of our matching methods an both DR Learner variants converge on an estimated ATE of around 0.73-0.77. This suggests there is a real causal effect: Empathetic LLM responses do increase user attachment by a score of about 0.75 points on average. This increase is analyzed with respect to our scoring metric. The effect is stable despite the difference of the models and confounder controls. This robustness across various techniques increases our confidence that the phenomenon we were trying to prove is real rather than just a supposition. Looking at the HTE estimates obtained from the DR Learners, there is a wide variation in CATEs, including strongly negative values. This suggests that empathy does not uniformly increase attachment, and some users may react negatively to empathic LLM responses. The median does suggest that most users benefit, but also the tails of the distributions are meaningful: an interesting future direction would be analyzing what personal traits of the users provoke this difference in outcome. Our dataset, however, did not provde us with this information. While the average matching quality was moderate (0.4571 and 0.4690), both cosine similarity matching approaches brought similar distributions of paired differences. Their ATE estimates were comparable to the estimates obtained to the other method. Controlling for additional confounders improved the results but did not radically change them: the ATE was almost unchanged, but the CATE distribution had lower standard deviation, and less extreme values on the tails. This result suggests that the additional confounders introduced in the final DR-Learner model (X_2, X_3, X_4) do not explain away the estimated treatment effect. In other words, the positive impact of empathetic responses is not an artifact of omitted-variable bias. This further strengthens the internal validity of our findings.

7.2 Ideas for Next Step

As can be observed in Figure 4, the predicted CATEs vary widely, so the effect varies dramatically person by person. An interesting future direction would be trying to predict what features and traits make a user more likely to benefit from empathy, and finetune an LLM model to dynamically adapt its empathy levels accordingly. However, we could not construct this personalized empathy model because our dataset did not contain user personality information. Additional user-level metadata such as session history, interaction patterns, or, more importantly, demographic and sociological information would substantially strengthen this research. Such features would enable more comprehensive confounder control, improving the precision and causal validity of the estimates. Another confounder we identified but did not control for is the possibility that the effect accumulates over turns: the treatment effect could be stronger at specific points in the conversation. We suspect that later turns may show either amplified or diminished effects (pushing estimates toward the CATE distribution's tails), depending on how the interaction is unfolding and whether the conversation is moving in the direction the user wants. Addressing this would shift the problem from a static to a dynamic causal inference setting.

To strengthen the robustness of our findings, additional causal estimators could be applied to the dataset. Methods such as the R-Learner, which estimates treatment effects on residualized outcomes, and the X-Learner, which is particularly effective when treatment groups are unbalanced (as it is in our case), would offer complementary perspectives on the HTE. Similarly, a targeted maximum likelihood estimate (TMLE) would provide complementary identification strategies and help assess the sensitivity of our conclusions to the choice of estimator. Applying these would give a more methodologically complete analysis and help verify that our conclusions do not depend on the choice of estimator, further validating the stability of the observed treatment effect.

Our project is heavily dependent on LLM-generated empathy and attachment scores for each conversation sample. We would prefer using human-annotated values as ground truth, but this wasn't possible due to time restrictions as annotating almost 9000 turns would be too costly. Human-generated ground truths for the empathy and attachment scores would have strengthened the external validity of our project, making it more reliable and suited to the real world.

8 LLM statement

The usage of LLMs such as ChatGPT and Gemini was limited exclusively to correcting the syntax of our writing and improving the clarity of the report.

References

[1] Keith, K., Jensen, D., & O'Connor, B. (2020). *Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates.* Proceedings of the 58th

Annual Meeting of the Association for Computational Linguistics (ACL).

[2] Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. CSLI Publications.

[3] James Vincent. Artificial Intimacy: The New Frontier of Human-Machine Relationships. *Polity Press*, 2023.

[4] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human–Computer Studies*, 149:102601, 2021.

[5] Ho, A. T., Hancock, J. T., & Miner, A. S. (2023). *Uncovering the Mechanisms of Parasocial Relationships with a Conversational AI for Mental Well-being*. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.

[6] Roberts, M. E., Stewart, B. M., & Nielsen, R. A. (2020). *Adjusting for Confounding with Text Matching*. American Journal of Political Science (AJPS), 64(4), 887-903.

[7] Johansson, F. D., Shalit, U., & Sontag, D. (2016). *Learning Representations for Counterfactual Inference*. Proceedings of the 33rd International Conference on Machine Learning (ICML).

[8] von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., & Locatello, F. (2021). *Self-supervised learning with data augmentations provably isolates content from style*. Advances in Neural Information Processing Systems (NeurIPS), 34.

[9] Rojas-Carulla, M., Schölkopf, B., Turner, R., & Peters, J. (2018). *Invariant Models for Causal Transfer Learning*. Journal of Machine Learning Research (JMLR), 19, 1-34.

[10] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[11] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding*. Advances in Neural Information Processing Systems (NeurIPS).

[12] Zheng, Lianmin, Zihan Wang, Shizhe Diao, Yuhui Li, and Hao Li. *From Generation to Judgment: Opportunities and Challenges of LLM-as-a-Judge*. 2024. https://llm-as-a-judge.github.io

[13] Mistral AI. Mistral Small (large-language–model). https://mistral.ai/news/mistral-small-3-1

[14] , title=WildChat: 1M ChatGPT Interaction Logs in the Wild, author=Wenting Zhao and Xiang Ren and Jack Hessel and Claire Cardie and Yejin Choi and Yuntian Deng, year=2024, eprint=2405.01470, archivePrefix=arXiv, primaryClass=cs.CL, url=https://arxiv.org/abs/2405.01470,

[15] VanderWeele, Tyler J., and Onyebuchi A. Arah. *Unmeasured confounding for general outcomes, treatments, and confounders: bias formulas for sensitivity analysis.*

[16] Imbens, Guido W., and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[17] Reimers, Nils. `all-mpnet-base-v2` (SentenceTransformers model). 2020. Available at: https://www.sbert.net/docs/pretrained_models.html

[18] K. Wu, G. Lewis, V. Zhao, K. Battocchi, N. Kallus, A. S. Gupta, D. Green, S. Muchmore, and V. Syrgkanis (2019). *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. Microsoft Research. Available at https://github.com/microsoft/EconML.

[19] Clark, D. M., & Watson, D. (1989). *Diurnal variation in mood: A review*. Journal of Personality and Social Psychology, .

[20] Watson, D. (1999). *Mood and diurnal variation: A study of daily mood*. Journal of Personality

[21] Salvi, Romain E., and D'angolo, Francesco. LLM Empathy Causal Study: Code and Experiments. GitHub repository, 2025. https://github.com/dangolofrancesco/llm-empathy-causal-study