ChatGPT Giving Relationship Advice - How Reliable Is It?

Haonan Hou, Kevin Leach, Yu Huang

Vanderbilt University
haonan.hou@vanderbilt.edu, kevin.leach@vanderbilt.edu, yu.huang@vanderbilt.edu

Abstract

In the evolving realm of natural language processing (NLP), generative AI models like ChatGPT are increasingly utilized across various applications. Among the possible purposes, many people are considering asking ChatGPT for relationship advice. However, the lack of in-depth examination of ChatGPT's response quality could be concerning when it is used for personal topics like mental health issues and intimate relationship problems. In these topics, a piece of misleading advice could cause harmful repercussions. In response to people's growing interest in using ChatGPT as a relationship advisor, our research evaluates ChatGPT's proficiency in discerning relationship advice. Specifically, we investigate its alignment with human judgements. We conducted our analysis with 13,138 Reddit posts about intimate relationship problems to examine the overall alignment. Furthermore, we investigate ChatGPT's consistency in judging intimate relationship advice by re-prompting identical queries. Our results indicate a significant disparity between ChatGPT and human judgments, with the model displaying inconsistency in its own decisions. Our findings emphasize the need for comprehensive insights into ChatGPT's mechanisms for intimacy problems and future improvements in its proficiency in helping people's relationship struggles.

Introduction

Intimate relationships, if managed with care, can contribute to increases in people's happiness and life expectancy (Saphire-Bernstein and Taylor 2013; Huang et al. 2016). Moreover, a wealth of psychological research suggests that wholesome relationships also correlate with reduced risk of cardiovascular diseases like coronary heart disease (Smith and Baucom 2017). However, intimate relationships can become abusive, oppressive, or hurtful (Bartholomew and Allison 2006). The aftermath of such tumultuous relationships is concerning for the victim both physically and mentally (Coker et al. 2000; Rakovec-Felser 2014).

When confronting relationship quandaries, online forums present an alternative for those unable to tap into professional resources or share concerns with close acquaintances (Collisson et al. 2018). Often, even the thought of discussing intimate matters with close friends can be daunting, primarily because of privacy concerns (Tagliabue et al. 2018). Platforms such as the Reddit relationships subreddit, "Loving From a Distance" forum, and Relationships-Advice.com provide anonymous spaces for discussing relationship issues (McKiernan et al. 2017). Their popularity is evident. For instance, the Reddit Relationships subreddit has over 3.4 million members and has amassed at least 240,000 posts in the last seven years.

Nevertheless, posting personal relationship struggles on online forums also has its disadvantages. It is possible that a person's post can quickly be consumed by new posts. Given the limitations of these conventional avenues, there is a growing interest in newer innovations, particularly large language models (LLMs), as potential supplementary support for issues related to mental wellbeing (Singh 2023; Aminah, Hidayah, and Ramli 2023). Among LLMs, ChatGPT by OpenAI is among the most prominent models (OpenAI 2023a). In fact, scholars and individuals have started considering ChatGPT for counsel on intimate relationship matters (Carlbring et al. 2023). Some people already tried dating advice from ChatGPT (Chris 2023).

Moreover, bad interventions from ChatGPT may have repercussions similar to misguided advice from humans (MacGeorge and Hall 2014). ChatGPT's ability to give prudent relationship advice has not yet been exhaustively explored, leading to prevalent concerns about its response quality (Sallam 2023). It remains uncertain whether ChatGPT can differentiate constructive suggestions from potentially detrimental ones. Thus, the overarching concerns arise: can we confide our relationship concerns in an AI model like ChatGPT? Is it equipped to provide meaningful advice, and can it discern good advice from bad? Does it parallel human intuition and wisdom? This study investigates these pivotal questions, aiming to evaluate ChatGPT's reliability in relationship matters.

A unique aspect of our research is the perspective from which we approach ChatGPT's reliability in intimate relationship scenarios. Users seeking advice may not always approach ChatGPT from a blank slate like "I don't know what to do!" It is just as plausible for them to present ChatGPT with multiple potential solutions, asking for a ranking based on effectiveness. They might phrase their dilemmas as "I have solutions A, B, C — which is the most appropriate?"

In fact, we observed many instances of posts in the Relationships Subreddit with titles such as:

- "Should I text her someday or let her come to me?"
- "Should I ask my ex what is wrong or leave it alone?"

By analyzing the rankings derived from Reddit post comments, we utilize a crowdsourced resource to gauge public consensus. This methodology serves as an objective metric to assess ChatGPT's response quality, treating it as a 'blackbox.' The findings from our study are intended to provide insights and references for future research in this domain.

Our research investigates the reliability of ChatGPT's ranking of suggestions in intimate relationship issues. Specifically, we queried ChatGPT with 13,138 Reddit posts, drawing comparisons with human judgements. Additionally, we contrasted ChatGPT's alignment with human opinions on posts from both before and after its last training data update.

Our investigation includes three different aspects associated with ChatGPT's performance:

1. **Overall Agreement With Human**: We evaluated 13,138 unique Reddit posts related to intimate relationships to measure the alignment between ChatGPT's evaluations and collective human judgment.
2. **Consistency**: We queried ChatGPT with the same post multiple times to determine if it consistently produces the same rankings for identical suggestions.
3. **Impact of Prompt Enhancement**: We investigated whether refining the prompts for ChatGPT, by incorporating human-labeled topics of relationship problems, influences its performance.

Our findings suggest a limited correlation between ChatGPT's rankings and human evaluations, with ($Kendall's\ \tau < 0.10$). In many instances, ChatGPT demonstrated variability in its rankings across repeated queries. Interestingly, its alignment with human judgments does not seem to markedly improve when exposed to potentially familiar data from its training phase. Likewise, the influence of factors such as the disparity and opinion variance of comments on its performance remains unclear. Additionally, preliminary indications are that refining the prompts with topic information does not necessarily enhance ChatGPT's evaluative accuracy.

The main contributions of this paper are:

- The first large-scale examination of ChatGPT's efficacy in intimate relationship topics, encompassing data from 13,138 Reddit posts.
- An analysis of ChatGPT's challenges in discerning nuanced relationship dilemmas and its alignment with human perspectives on relationship advice.
- A refined and filtered dataset derived from intimate relationship discussions on Reddit, spanning nine years.
- Actionable insights and evidence for AI researchers and developers about ChatGPT's performance nuances, the effects of data disparity and prompt alterations, and potential avenues for improvement in AI.

Background And Related Work

The Prevalence and Gravity of Relationship Challenges Relationship challenges, especially during adolescence, are pervasive and can have lasting effects (Elkington et al. 2013). The Centers for Disease Control and Prevention note that nearly 1.5 million high school students across the US face physical abuse from a dating partner every year (Black et al. 2006). Seeking reliable guidance in navigating such challenges proves elusive for many (Adam et al. 2011). Research indicates that young adolescents often rely on a limited support network, usually consisting of close friends and mothers, when facing romantic issues (Vallade, Dillow, and Myers 2016). This reliance on peers, often with similar inexperience, can lead to unvaried and misguided advice (Lefkowitz and Espinosa-Hernandez 2007). Unfortunately, these adolescent challenges can have enduring repercussions (Fernández-Fuertes and Fuertes 2010).

Traditional Approaches to Seeking Relationship Guidance For those facing relationship problems, professional therapists can become a reliable sought-after resource (Jensen and Bergin 1988). However, several barriers hinder this approach: prohibitive costs, scheduling conflicts, and reservations about sharing personal issues (American Academy of Child and Adolescent Psychiatry Committee on Health Care Access and Economics Task Force on Mental Health 2009; Chrysikou 2013). Moreover, there exists a considerable gap between the demand for therapists and their availability. A 2022 report highlighted that while about 53 million adults faced mental health challenges in 2020, only 1.2 million behavioral health therapists were available (U.S. Government Accountability Office 2022), making these resources both limited and highly sought after.

AI-based Health Interventions With the rise of AI, there has been a growing interest in its therapeutic potential for mental health. A pioneering study in 2017 involving the Woebot Chatbot found a significant reduction in depression symptoms among its users (Fitzpatrick, Darcy, and Vierhile 2017). Interestingly, participants attributed their positive experiences more to their interactions with the chatbot than the actual content, echoing the dynamics of traditional therapy.

The advent of advanced models like ChatGPT has further invigorated this space. While ChatGPT has shown promise, a focused study on its performance in Urology found limitations. In a comparison involving 100 case studies, only 52% of ChatGPT's answers matched a urologist's feedback (Cocci et al. 2023). These findings raise questions about ChatGPT's readiness for offering medical advice, though its applicability in providing relationship guidance remains an open question.

Methodology

We aim to understand how closely ChatGPT's evaluations of intimate relationship advice align with human evaluations. We collected a dataset of 13,138 Reddit posts on relationship issues spanning from 2015 to 2023 with Reddit APIs, the official study tools for researchers (Proferes et al. 2021). We evaluate ChatGPT's decision-making against that of Reddit users. We further analyze the consistency of ChatGPT's responses by repeatedly querying it under the same conditions. Alongside, we investigate potential determinants of ChatGPT's ranking preferences, such as the suggestions' opinion variance and lexical complexity.

To guide our investigation and provide structure, we formulated the following pivotal research questions:

1. **RQ1**: How well does ChatGPT's advice ranking align with human preferences?
2. **RQ2**: Is ChatGPT consistent in its advice rankings when repeatedly queried under the same conditions?
3. **RQ3**: What factors influence ChatGPT's alignment with human rankings? (e.g., topics, opinion variance of comments, length of text)

Experimental Design: Disparity Groups

We categorized our experiments into disparity levels to ascertain whether the breadth of advice presented affects ChatGPT's alignment with human preferences:

- **High Disparity:** ChatGPT assesses two suggestions: the most and least favored by Reddit users.
- **Medium Disparity:** ChatGPT evaluates four suggestions spanning user-ranked advice.
- **Low Disparity:** ChatGPT is presented with eight suggestions, showcasing the most extensive range of user-preferred advice.

This design helps us to investigate whether ChatGPT is aligned with human decisions when presented with narrower suggestion ranges while controlling for potential random variations.

Implementation

To execute the experiments, we employed the OpenAI API, selecting the GPT-3.5-turbo model. This model corresponds with the default ChatGPT web application version and is accessible to the research community (OpenAI 2023b). We maintained default parameters, like temperature, to mirror the web application's behavior as closely as possible. Namely, the default parameters used in the experiments are: {*frequency penalty: 0, logit bias: null, logprobs: false, top logprobs: not specified and not applicable, max tokens: 4096, n: 1, presence penalty: 0*}.

For our analysis, we used the following prompt to query ChatGPT:

> *"I will give you a description of a relationship problem and [number of comments] advices on the problem. Please rank the reliability of the advices. Display the results as a number series separated by commas, e.g., "3,1,2,5,4" where option 3 is the best comment and option 4 is the worst. Ensure all [number of comments] numbers appear in your response. Also, provide a guess for the original poster's age, gender, ethnicity, and nationality. Categorize the problem concisely, and, if possible, reuse previous categories. Respond in this format without newlines: Ranking: ...; Age: ...; Gender: ...; Ethnicity: ...; Nationality: ...; Category: .... The problem description is: ... The suggestions are: [1]..."*

Subsequent subsections provide an in-depth overview of our dataset's composition and outline the approaches employed for data preparation and analysis.

Dataset

Data Collection We aimed to construct a dataset focused solely on intimate relationship topics, given the lack of existing datasets in this domain. Our dataset had to satisfy five conditions:

- *Digital Human Conversations*: The data should be in a digital format, suitable for prompting ChatGPT.
- *Topic Specificity*: The conversations must be centered around intimate relationships.
- *Scoring Mechanism*: Each relationship problem should come with suggestions scored by the community, helping discern quality advice.
- *Diversity*: The dataset must be varied in topics and timestamps, ensuring no limitations due to scale.
- *Deidentification*: The dataset must not contain identifiable information like names and emails of anyone.
- *Content Appropriateness*: It should not include inappropriate content.

After an extensive review of online forums, we selected the "relationships" subreddit. Established in 2008, this subreddit has over 3.4 million members, ensuring topic diversity. Its active moderation guarantees content quality by filtering out irrelevant or potentially fabricated stories. Additionally, the moderators ensure that no identifiable information is present in the posts and comments. Reddit's upvote and downvote system helps in identifying the community's perspective on each piece of advice.

The raw dataset consists of 244,876 Reddit posts from the "relationships" subreddit, structured in JSON format. These posts discuss a variety of issues, from dating to familial matters, and span from January 2015 to March 2023. Within each post, we have captured the title, problem description, comments, and their respective scores. Notably, a comment's score is computed as upvotes minus downvotes. Reddit users upvote comments they find valuable or accurate, adding a +1 to their score, or downvote those they find misleading or irrelevant, resulting in a -1. We leverage this upvote/downvote mechanism as a measure for user evaluations. Hence, a comment's score reflects the community's collective judgment on its reliability. As comments accrue scores, their rankings serve as a representation of community consensus on the credibility of advice.

Data Preprocessing To effectively evaluate ChatGPT, it is crucial to select posts that showcase diverse community responses. We assume that comments with the same score do not provide a clear ranking, as they are deemed equally valuable by the community.

Given ChatGPT's token limit of 4096, we had to balance two requirements: selecting posts with a sufficient number of comments to capture varied community opinions, and ensuring the total content stays within the token constraint. In practice, around 16 comments typically fit within ChatGPT's token limit. Consequently, we curated our refined dataset based on the raw dataset to include posts with approximately 16 distinct comments, where "distinct" refers to comments having different scores. The refined dataset contains 13,138 entries. Time-wise, 12,129 of these entries are from January 2015 to September 2021, used for RQ1 and RQ2, while 1,009 entries, spanning October 2021 to March 2023, are specifically reserved for the "Data timeframe" segment in RQ3. Each entry includes a JSON-formatted Reddit post with the problem description, 16 comments of varying popularity, their scores, and relevant insights from the original poster. This dataset accompanies the paper.

> We collected a raw dataset including 244,876 Reddit posts about relationship issues, ranging from 2015 to 2023. We later filtered the dataset to obtain a refined dataset consisting of 13,138 posts (12,129 of them are before the training cutoff date), which contain sufficiently diverse comments for the ranking task.

Data Analysis

The primary statistic used in this research is inter-rater reliability/agreement (IRA). IRA is the measure of the level of agreement between different raters, namely ChatGPT and Reddit users in this study. Specifically, IRA is used to measure the agreement between ChatGPT's ranking of the relationship comments (suggestions) and the Reddit users' ranking based on the scores of the comments. Two statistics are used in this study to measure the IRA:

- *Kendall's Tau-b Statistics*: A non-parametric measure for ordinal variables, with its range between [-1, 1]. It assesses alignment in raters' rankings, where a value near -1 shows significant disagreement. It is our primary IRA statistic for this research.
- *Spearman's Rho Statistics*: A non-parametric method for ordinal variables within the range [-1, 1]. While Kendall's Tau-b focuses on the relative ordering of data pairs, Spearman's Rho evaluates the strength and direction of the linear relationship between ranked variables. Using both metrics gives a thorough analysis of data alignment, capturing both the order and the strength of relationships.

RQ1: General Level of Agreement Our primary investigation centers on the alignment between rankings given by ChatGPT and those of Reddit users. Using a filtered dataset, we instruct ChatGPT to rank the reliability of comments for 12,129 Reddit posts from January 2015 to September 2021. The exact prompt is mentioned earlier in the "implementation" subsection. Within each prompt to ChatGPT, we:

1. Introduce the task.
2. Provide the problem description and corresponding comments/suggestions.
3. Specify the number of comments for ranking.

To evaluate the alignment between ChatGPT's and human rankings, we compute the IRA for each post's rankings. Human rankings are derived from comment scores in descending order. After computing the IRA for each post, we average these values to measure the overall alignment between ChatGPT and human rankings.

We note that our study included prompts for ChatGPT to predict demographic information related to the authors of the Reddit posts, as well as topic categories. This decision was informed by previous research indicating potential biases, such as gender or racial biases, in responses from large language models (LLMs) like ChatGPT (Gross 2023). The inclusion of these queries was intended to explore whether such biases might manifest in ChatGPT's responses within the context of our study.

However, these aspects were not subjected to further analytical scrutiny in our final analysis. The primary reason for this exclusion is the absence of verified demographic labels or categories against which ChatGPT's predictions could be accurately compared. Obtaining such demographic information, even if possible, is deemed unethical under IRB constraints. This lack of ground-truth data renders any analysis of the model's demographic predictions speculative at best. We therefore concluded that this line of investigation falls outside the intended scope of our study's design.

RQ2: Consistency of ChatGPT Responses To measure the consistency of ChatGPT's rankings under default settings, we use a subset of 1500 randomly chosen Reddit posts from our filtered dataset. Each post is presented to ChatGPT for ranking in a manner similar to the approach for RQ1. For each post, we query ChatGPT with the same prompt four times in a row to determine if ChatGPT's rankings vary across attempts. ChatGPT's rankings in each set of four queries using the identical prompt is termed as a 'set of rankings.' We recorded each set of rankings given by ChatGPT. Based on these responses, we report the following metrics:

- *Consistency Rate*: This rate captures how frequently ChatGPT produces 4 identical rankings in a set of rankings. It is derived from the instances of consistent responses as a fraction of total queries. The inconsistency rate — instances where at least one ranking differs among the four — is the complement of the consistency rate, i.e., $1 - \textit{Consistency Rate}$.
- *Unique Ranking Count*: The metric counts the instances where ChatGPT produces one, two, three, or four different unique rankings in a set, providing insight into the model's consistency and the variance in its response to identical prompts. A unique ranking is defined as a distinct permutation of the suggestions. This analysis helps to further understand the model's stability in generating advice rankings.
- *Severe Disagreement Rate*: This metric captures instances where the rankings significantly contradict each other. A "severe disagreement" is defined as a negative IRA between any two rankings, signifying major misalignments, such as reversed orders. While minor discrepancies in ChatGPT's rankings might be acceptable,

pronounced contradictions can be misleading. The Severe Disagreement Rate measures the fraction of ranking pairs with negative IRA against the total number of pairs.

For context, suppose three rankings: A, B, and C. There are three pairs: AB, AC, and BC. If AC has a negative IRA, it counts as a severe disagreement. The rate is computed by comparing the total severe disagreements to all ranking pairs.

- *Average IRA statistic*: This represents the mean IRA value across all ranking sets. For each set, we first compute its average IRA. If all rankings in a set are the same, its IRA is 1. In case of discrepancies, the IRA between each ranking pair is computed, summed, and averaged over the number of pairs. The final average IRA is the mean of these values across all sets, serving as a quantitative measure of ChatGPT's response consistency. A higher average IRA suggests greater consistency.

RQ3: Potential Factors Affecting ChatGPT's Ranking Another main focus of our research is the exploration of some potential factors that might affect ChatGPT's ranking preferences or its capacity to align with human wisdom. In this section, we introduce the experimental designs to test the potential factors like the length of suggestions and the opinion variance of the comments. Most of the analysis does not require new queries.

Opinion Variance of the Suggestions A potential influencer of ChatGPT's alignment with human rankings might be the opinion variance of the suggestions. By "opinion variance", we mean the range of scores comments receive. High opinion variance indicates pronounced differences in Reddit users' opinions, with some comments being heavily upvoted while others are not. We hypothesize that ChatGPT may find it easier to rank a highly upvoted comment over a broadly disliked one compared to two comments with closer upvote-downvote ratios.

To quantify comment opinion variance, we use the coefficient of variation (CV). The CV is the standard deviation of a sample's score normalized by the sample size. A higher CV signifies more variation in comment popularity, implying distinct differences in perceived comment quality.

The hypotheses are:

- *H0 (Null Hypothesis)*: There is no correlation between comment opinion variance (as quantified by CV) and ChatGPT's alignment with human judgments (as measured by IRA using Kendall's Tau statistic).
- *H1 (Alternative Hypothesis)*: There is a correlation between comment opinion variance and ChatGPT's alignment with human judgments.

For each of the 12,129 Reddit posts examined in RQ1, we calculated its CV. We then sought to determine the relationship between this and ChatGPT's alignment with human judgments. Given our data's nonparametric nature, we used Spearman's Rho for correlation computation.

The inclusion of these hypotheses provides a direct framework for your methodology and allows readers to anticipate the kind of results they might expect.

Length of the Suggestions Another avenue of exploration was whether ChatGPT exhibited biases based on the length of relationship suggestions. Analyzing the refined comments from Reddit posts and the outcomes from RQ1, we average the word count for suggestions at each rank. For instance, to discern the mean word count for the top-ranked suggestions by ChatGPT, we summed the word counts of all such suggestions and divided them by the total number of suggestions ranked top. This metric enables us to investigate any potential bias of ChatGPT towards suggestions of specific lengths.

Lexical Complexity of the Suggestions The lexical complexity of comments may influence ChatGPT's rankings. To evaluate this, we measured the average lexical complexity of each ranked suggestion using the Type-Token Ratio (TTR). TTR measures vocabulary richness by dividing the number of unique words (types) by the total number of words (tokens) in a text. A higher TTR indicates greater lexical diversity and sophistication, whereas a lower one points to more repetitions. While various metrics can assess lexical complexity, we chose TTR for its direct measure of word variety. With a range of [0, 1], a value close to 1 denotes a linguistically rich text. Our goal in monitoring this metric was to discern any tendencies in ChatGPT's rankings relative to the linguistic depth of relationship advice. The analysis draws from the ChatGPT responses in our RQ1 experiment.

Data timeframe (pre or post-Sep 2021) To understand if ChatGPT's potential exposure to the data during its training phase affected its alignment, comments are grouped based on their timeframes: those from before September 2021 and those after this date. This is because the officially announced cutoff date of ChatGPT's training data is September 2021 (Gao et al. 2023). Specifically, the latter group consists of Reddit posts from the relationship subreddit, ranging from October 2021 to March 2023. The posts of the latter group are refined using the same technique mentioned in RQ1. After filtering, the latter group contains more than 1000 posts. The newly added posts are used to query ChatGPT with the same prompt used in RQ1 for all three disparity groups. After thorough response collection, we recalculate the IRA statistics for each time-segmented group and compare these values. With the comparison result, we aim to determine if ChatGPT exhibits a bias towards comments it might have encountered during training.

Prompt Enhancement We also explore the possibility that ChatGPT's rankings might be more aligned with humans' decisions if ChatGPT is queried with enhanced prompts. In our revised methodology, we excluded requests for ChatGPT to predict demographic information to minimize distractions and focus more sharply on our primary objective: evaluating ChatGPT's proficiency in ranking relationship advice. Demographic predictions could introduce speculative elements that might detract from the clarity and directness of our analysis. We also add a manually labeled topic to the prompt to help ChatGPT understand the topic of the relationship problem. Inspired by Rowland S. Miller's *Intimate Relationships*, which categorizes relationship is-

Topic Label	Explanation	Example
Dynamics of nuclear family	Issues among family members.	Should I address my parents' tensions?
Communications and Understanding	Romantic relationship misunderstandings.	How to discuss my husband's fashion sense?
Persisting Problems in relationship	Ongoing romantic issues.	Recurring arguments about freedom with my wife.
Uncertainty in Relationship	Doubts in romance.	Losing passion in my marriage.
Developing Attraction	Crush and date indecisions.	Should I pursue the girl I'm chatting with?
Self-Reflection	Assessing one's actions in romance.	Was teasing my girlfriend publicly wrong?
Break-Ups and Moving On	Post break-up challenges.	Struggling post break-up six months ago.
Trust and Infidelity	Cheating concerns and aftermath.	Saw flirty texts on my boyfriend's phone.
Logistical Challenges	Practical romantic disputes.	Disagreeing on vacation destinations with my wife.
Family Interference	Family impact on romance.	Girlfriend's mom disapproves of our wedding.
Major Life Decisions	Big romantic choices.	Husband wants a child; I'm unsure.
General Relationship Struggles	Broad romantic concerns.	Always defensive in relationships - solutions?
Irrelevant to intimate relationships	Non-romantic issues.	Boss critiques my work ethic.
Unclear intent	Vague post intent.	It felt like a dream...

Table 1: Labels for Categorizing Posts in RQ3 - Enhanced Prompts.

sues into broad areas like attraction and communication, I adapted these categories to develop labels that better target specific, everyday relationship challenges (Miller, Perlman, and Brehm 2007). Table 1 summarizes the label categories along with their respective explanations and examples. We randomly selected 816 posts from the dataset and manually labeled these posts with the most appropriate topic. For posts that may be categorized with multiple labels, we label them with the most straightforward and relevant category. After prompting ChatGPT with the 816 labeled posts with the enhanced prompt, we calculate the average IRA of these new responses and compare it with the original IRA. The contrast between the average IRAs could suggest whether an enhancement in the prompt may contribute to better alignment.

It is important to highlight that during the experimental phase for all three research questions, ChatGPT produced a range of invalid responses. These instances were systematically documented and subsequently omitted from the primary analysis. A detailed examination of these invalid responses, including their nature and frequency, is elaborated upon in the Discussion section.

Analysis Result

In this section, we discuss the results of our analysis of the alignment between ChatGPT and humans' rankings of relationship suggestions. The results are summarized as follows:

Disparity	Average IRA Measurement	
	Kendall's Tau-b	Spearman's Rho
Low	0.069	0.083
Medium	-0.008	-0.005
High	-0.188	-0.188

Table 2: The overall agreement level between ChatGPT and Human, measured in IRA using Kendall's Tau-b and Spearman's Rho.

Overall Agreement Between ChatGPT and Human Rankings

Table 2 summarizes the alignment level between ChatGPT and human judgment in ranking relationship advice (see Section for specific calculations). The table displays the IRA measured for low, medium, and high disparity groups using Kendall's Tau-b and Spearman's Rho statistics. Due to issues like occasional OpenAI API instability and token limit, the numbers of posts successfully prompted and analyzed for the three disparity groups are slightly different. We examined 11,262 posts in the low disparity group, 11,079 posts in the medium disparity group, and 10,560 posts in the high disparity group. The average IRA for all three disparity groups is close to 0. Specifically, although the IRA is more negative (-0.188) and farther away from 0 in high disparity groups than in other groups, it is still in the range that indicates a very weak disagreement. The overall results suggest a very weak alignment between the rankings of ChatGPT

Disparity	Correlation of Opinion Variance and IRA	
	correlation coefficient	**p-value range**
Low	0.081	$< 1 \times 10^{-10}$
Medium	0.037	$< 1 \times 10^{-3}$
High	0.020	< 0.05

Table 3: Correlation between the opinion variance of comments and the IRA. The p-values suggest the statistical significance of the correlations.

Disparity	IRA Measured in Kendall's tau-b	
	Before 09/2021	**After 09/2021**
Low	0.069	0.071
Medium	-0.008	3.10e-4
High	-0.189	-0.166

Table 4: The alignment between ChatGPT and human rankings based on different DataTimeframes. This table differentiates the IRA values before and after September 2021, the official ChatGPT training data cutoff date.

and humans.

All disparity groups showed an IRA near 0, indicating a very weak alignment with human opinions. These IRA results indicate that ChatGPT neither strongly agrees nor consistently disagrees with human perspectives.

Consistency of ChatGPT's Rankings

We also investigate the consistency in ChatGPT's responses for the same prompt and measure the Consistency Rate, Severe Disagreement Rate, and Average IRA score within the responses for the same prompts (see Section). The main findings are summarized in Figure 1. We analyzed 3,920, 1,148, and 343 posts for low, medium, and high disparity groups respectively. As the figure shows, the Consistency Rate is 0.0, 0.015, and 0.239 for low, medium, and high disparity groups respectively. Notably, among all 3,920 posts analyzed in low disparity groups (containing 8 suggestions to rank), ChatGPT never gives the exact same rankings for the 4 responses responding to the same prompt.

Disparity	IRA Measured in Kendall's tau-b	
	Unlabeled prompts	**Labeled prompts**
Low	0.069	0.022
Medium	-0.008	0.028
High	-0.188	-0.212

Table 5: Comparison of the IRA with various prompts. This table contrasts the IRA values for prompts with and without a topic label.

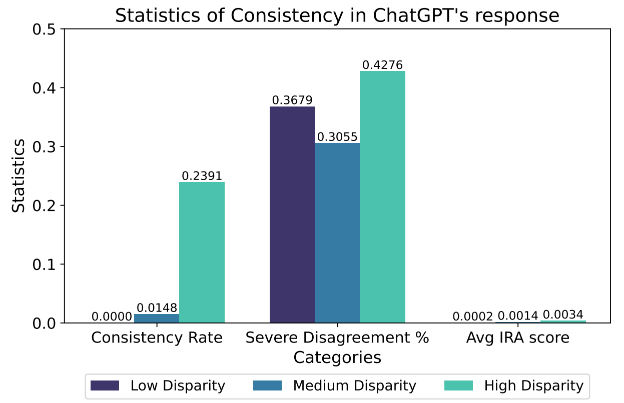

Figure 1: Metrics of consistency across three disparity groups. Each post is queried four times, with the same prompt, resulting in a "set" of rankings. "Consistency" is achieved when ChatGPT provides the same ranking across the set. "Severe Disagreement" arises when any two rankings within a set have a negative IRA. "Average IRA" computes the mean IRA across the sets' rankings.

An additional layer of our analysis focused on the uniqueness of ChatGPT's rankings for identical prompts, where we quantified the variation in responses. This was measured by the number of distinct rankings ChatGPT produced out of the possible sets, ranging from one to four. The results, which are further detailed in Figure 2, reveal a notable trend: as the number of comments increases, ChatGPT's consistency in rankings decreases, with no instances of identical rankings in the low disparity group with eight suggestions.

As mentioned in the Methodology section, severe disagreement is defined as two rankings having opposite extreme picks. The severe disagreement rates are 0.368, 0.305, and 0.428 for low, medium, and high disparity groups respectively. There is no clear sign indicating an association between the level of disparity and severe disagreement rate. The average IRA values are 0.00016 for the low disparity group, 0.00142 for the medium disparity group, and 0.00341 for the high disparity group. The IRAs in all three groups are positive values close to 0, which indicates a very weak alignment among ChatGPT's rankings for the exact same prompt.

In the low disparity group, ChatGPT never produced identical rankings for the same prompt. The number of severe disagreement instances has no clear correlation to disparity levels. Average IRA values were close to 0 for all groups, suggesting very weak alignment in ChatGPT's rankings for the same prompt.

Factors Affecting ChatGPT's Agreement with Humans

Several factors were investigated to understand their potential influence on ChatGPT's alignment with human judgments. Here is what we found:

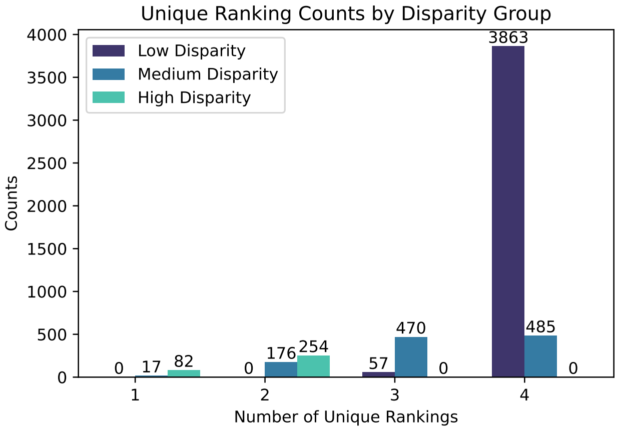

Figure 2: The histogram quantifies ChatGPT's consistency in ranking responses across three disparity groups, with each post queried four times. Unique rankings per post are illustrated, reflecting the consistency level within each group. Given the varying number of posts in each group, comparisons of consistency should be made within rather than between groups.

Enhanced Prompts Table 5 summarizes the IRA contrast between ChatGPT's responses prompted without a topic label and those prompted with a topic label. We analyzed 745, 621, and 627 labeled posts for low, medium, and high disparity groups and compared their resulting IRAs with the IRAs calculated in RQ1. With changed prompts, while there are changes in the IRA values, the changes are minimal in magnitude and the directions of change for different groups are different. For instance, for the low disparity group, the IRA with labeled prompts is 0.02234, which is not significantly different from the IRA of unlabeled prompts (0.06876). The subject matter or topic of the comments appeared to have a negligible effect on ChatGPT's rankings.

Opinion Variance of the Comments The correlation between the opinion variance of the suggestions and the IRA is detailed in Table 3. For all three disparity groups, the correlation coefficient in Spearman's Rho statistics is a small positive value. Crucially, the p-values of the correlation coefficients in all three groups are less than 0.05. This indicates that the correlation coefficients are all statistically significant. For instance, the correlation coefficient is 0.08082 for the low disparity group with a p-value of 8.691e-18.

Given the p-values observed for each group, we reject the null hypothesis (H0) in favor of the alternative hypothesis (H1) for all three groups. This suggests there is a statistically significant, albeit very weak, correlation between the opinion variance of the relationship advice and the degree of alignment between ChatGPT and human evaluations. However, it is important to emphasize that while the correlation is statistically significant, its strength is quite low, indicating a very weak relationship between the variables of interest.

Length of the Comments We examined whether the length of the comments influences ChatGPT's ranking, with our findings summarized in Table 6. For the low-disparity group, ChatGPT places shorter advice in the top 4 picks 38% of the time when ranking 8 pieces of relationship advice. Furthermore, it shows a preference for shorter comments 61.9% of the time, yet it strictly orders suggestions based solely on their lengths in less than 0.1% of instances. A comparable trend is observed for the medium and high disparity groups. Although the data indicates ChatGPT often ranks lengthier advice higher, the frequency with which it prefers shorter advice is not insignificantly less.

Figure 3 summarizes the average word count of each ranked choice for high, medium, and low disparity groups. In high and medium disparity groups, there is a clear tendency that on average, the more preferred advice typically contains more words. For instance, in the medium disparity group, the average word count gradually decreases from 533.1 words to 451.9 words from the most preferred advice and the least preferred advice determined by ChatGPT. Such tendency is less clear in the low disparity group, where the average word count peaks at 3rd ranked advice with 570.3 words. Nevertheless, the average word count of the first four choices still surpasses that of the last four choices.

Lexical Complexity of the Comments We also investigated the whether lexical complexity (measured in TTR) of the relationship suggestions may affect ChatGPT's ranking behaviors. The results are summarized in Figure 4. The lexical complexity of ChatGPT's most preferred relationship suggestions is generally smaller than those preferred less. However, although it may seem that ChatGPT prefers advice with lower lexical complexity, the difference in lexical complexity between different ranked suggestions is very small (0.89 to 0.92).

Data Timeframe (Before or After Sep 2021) We also investigated whether ChatGPT may achieve higher alignment with humans when prompted with data that might be in its training data. Table 4 summarizes the IRA with different data time-frame. As aforementioned, the data is divided into two groups, namely the part before (and including) September 2021 and the part after September 2021. As the table shows, the alignment level indicated by IRA statistics slightly increases, though arguably minimally, for all three groups. This result counters our original assumption that ChatGPT could resonate with human wisdom more when prompted with potentially seen data. Whether ChatGPT might have been exposed to the data during its training phase did not seem to make a significant difference in its rankings, showing negligible effects.

These results paint a clear picture of ChatGPT's current capabilities and limitations when it comes to ranking relationship advice. More extensive research may be required to further understand the intricacies of its decision-making process in this specific domain.

Disparity	Preference Strength and Percentage		
	Indifferent	Weak	Strong
Low	38.0%	61.9%	0.0018%
Medium	42.5%	51.4%	6.2%
High	46.6%	53.4%	53.4%

Table 6: ChatGPT's ranking based on comment length. If top-ranked comments are no longer on average than the lowest-ranked, ChatGPT is "indifferent" about length. If top-ranked comments are on average longer, ChatGPT has a "weak" preference for length. A "strong" preference is indicated when comments are ranked strictly in descending order by length. The "Strong" category in the high disparity group is synonymous with the "Weak" category, as this group only considers two ranks, indicating a preference for either length or its opposite.

The effect of the enhanced prompts and data timeframe on ChatGPT's alignment with human decisions is nearly negligible. Opinion variance of the comments is positively but very weakly correlated with the alignment level. ChatGPT seems to display a slight preference for longer, less complex comments in the majority of the cases. The average word counts of ChatGPT's most preferred relationship suggestions are generally greater than those preferred less. Importantly, ChatGPT's ranking behavior remains largely consistent across different disparity groups in all settings.

Discussion

Our results indicate a minimal alignment between ChatGPT's and human rankings regarding relationship suggestions. While there is an observed trend suggesting that ChatGPT might show a mild preference for longer suggestions, this alignment does not exceed 61% across disparity groups. We also consider whether longer comments offer a more detailed analysis, hence their perceived reliability. ChatGPT does not necessarily exhibit improved performance when confronted with extreme suggestions in high disparity groups. Importantly, the alignment remains consistent regardless of whether ChatGPT has previously encountered the data. Another notable observation is ChatGPT's inconsistency in its own decisions when using default parameters. This section delves deeper into these findings and suggests ways to encourage effective contributions to LLMs.

Factors Contributing to Misalignment

Several factors might influence the limited alignment between ChatGPT's rankings and human judgments. Our study does not indicate significant ties between alignment and factors like suggestion opinion variance or disparity. For instance, although we observed a weak correlation between opinion variance and alignment, it may have been coincidental. other influences remain viable. Subtle human emotions, contextual interpretations, or deeper relationship dynamics could be hard for AI models to capture. ChatGPT, for all its prowess, may not capture the depth of human understanding.

The consistent alignment, irrespective of ChatGPT's familiarity with the data, implies that even if Reddit data serves as training input, it does not necessarily enhance LLM performance for this use case. The complexities of relationship advice might go beyond standard language patterns, requiring a deeper understanding of human psychology.

Consistency Concerns

ChatGPT's variable responses to identical prompts underscore the challenges AI encounters in personal topics. Although using ChatGPT with its default settings may introduce variability, excessive randomness is undesirable. The close-to-zero average IRA scores across disparity groups in our RQ2 hint at a significant misalignment between ChatGPT's own rankings for repeated prompts. This inconsistency in ChatGPT's responses might primarily account for its weak alignment with collective human wisdom. Consistency is vital for trust. If users cannot get predictable guidance from a tool, they might become skeptical or misled, intensifying their relationship struggles.

Severity of Consequences When ChatGPT Disagrees With Humans

Considering the potential repercussions of AI-generated evaluation that diverges from general human consensus is crucial. In areas like relationship advice, misdirection can yield severe outcomes. Relationships are inherently sensitive; misinterpretations or acting on unsound advice might escalate conflicts, foster distrust, or even lead to irreversible choices. We emphasize the need for further research in this area. Given our time constraints, we are not able to further investigate the consequences when ChatGPT's advice deviates from human consensus.

Invalid ChatGPT Responses

In our experimental design, we observed instances where ChatGPT provided invalid responses. These invalidities emerged from two primary issues. Firstly, ChatGPT occasionally deviated from the specified response format. For example, instead of presenting rankings within brackets as "Ranking: [1, 2, 3, 4]", it would omit the brackets, returning "1, 2, 3, 4". Secondly, there were instances where ChatGPT did not adhere to the requested number of options in its rankings. For instance, when asked to rank four comments, it might provide a ranking for five, such as "[2, 3, 1, 5, 4]".

To illustrate, during the experiment for RQ1, ChatGPT was prompted 35,653 times, yielding 3,303 invalid responses. This results in an invalid response rate of approximately 9.26 percent. While ChatGPT's responses inherently contain some degree of randomness, such frequent deviation from clearly specified prompts highlights a significant concern. This issue underscores the necessity for further investigation into the causes of these invalid responses. To be conservative with our analysis, we removed all invalid responses from our experiments. We appeal for future studies that investigate the reasons behind the invalid responses.

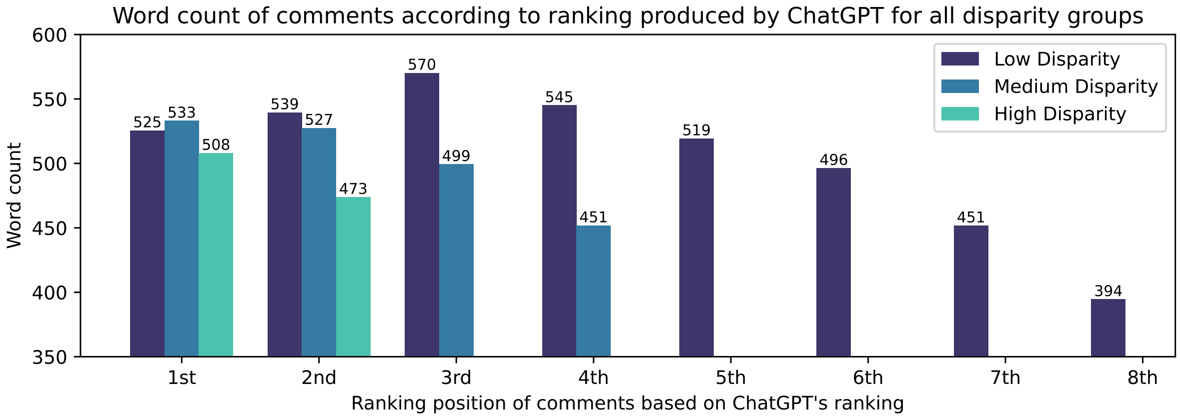

Figure 3: Word count of comments according to ranking produced by ChatGPT. The leftmost bars show the average length of top-ranked comments in each disparity group. In the medium disparity group, the top-ranked comments average 533 words, surpassing all lower-ranked comments.

Recommendations for Developers

Given the insights about ChatGPT's ranking capabilities in intimate relationship scenarios, we propose targeted recommendations for AI developers to enhance the efficacy of LLMs in such tasks. Our study underscores the importance of fine-tuning LLMs to better interpret and rank user-generated content, especially in subjective domains like relationship advice. We recommend the following:

- Incorporating mechanisms that allow LLMs to better understand the context and nuances of user-generated content, particularly in forums like Reddit. This could involve advanced sentiment analysis and an understanding of conversational dynamics.
- Focusing on improving the accuracy and reliability of LLMs in ranking tasks. This includes enhancing the model's ability to adhere to specific formatting instructions and to accurately rank predefined options.
- Conducting thorough investigations into the underlying mechanisms by which LLMs process and rank queries. Understanding the model's decision-making process can provide valuable insights for further refinement.

Lastly, it is worth noting that our findings, especially the highlighted misalignments, could be misinterpreted or misused to undermine the overall value of LLMs. While our intention is to shed light on specific areas for improvement, misrepresenting our research as an outright rejection of AI's capability in relationship advice could mislead the public. We appeal for further investigation and research into LLMs' capacity in discerning intimate relationship matters, because the popularity of online relationship forums and the practical challenges associated with professional therapies all suggest that LLMs for relationship issues can be a priority.

Limitations

Lack of Qualitative Analysis of GPT's Advice

Our research specifically excludes using ChatGPT to generate relationship advice, a common scenario in consultations with ChatGPT and possibly with other parties as well. Nevertheless, upon examining a sample of 100 Reddit posts, we found that 14 percent present predefined solutions when asking for help. This percentage, though not overwhelming, indicates that a subset of users approach advice platforms with preconceived plans. The primary objective of our study, therefore, is not to replicate the exact manner in which users interact with ChatGPT. Instead, we aim to objectively evaluate the quality of ChatGPT's responses through a feasible experimental design. Focusing on ranking, as opposed to a qualitative analysis of advice, allows for a more manageable approach given the extensive data involved. Additionally, regardless of the use case, the dataset that we curated may be helpful for future studies for LLM.

Treating Human Decisions as Benchmarks

When evaluating ChatGPT's performance based on its alignment with human decisions, we inherently regard human decisions as the gold standard. This premise assumes that humans effectively understand the context of relationship problems, evaluate advice quality and reliability, and appropriately utilize the "upvote" and "downvote" features to express their perspectives. However, humans can sometimes err or be swayed by emotions. Intimacy issues, by nature, are subjective; thus, even well-intentioned individuals may differ in their judgements on identical problems. Admittedly, the highly-voted advice is not guaranteed to be more accurate and helpful than the others. Nevertheless, at the very least, the scores of the comments represent humans' collective opinions, which mimics the crowdsourcing task.

Limitations of IRA

Using IRA to measure alignment between ChatGPT and human decisions might not fully encapsulate the nuances of ChatGPT's decision-making. The IRA primarily quantifies agreement based on final rankings without probing the rationale behind decisions. Both ChatGPT and humans could have valid reasons for their judgements. A more thorough

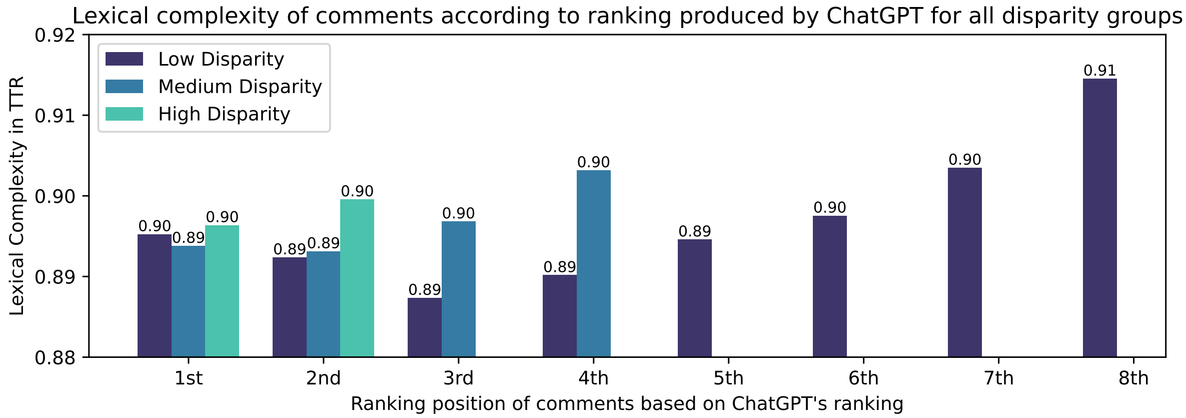

Figure 4: Lexical complexity (TTR) of comments by rank in ChatGPT's ranking. The leftmost bars represent the complexity of top-ranked comments in each disparity group. In the medium group, top comments have an average complexity of 0.89, lower than those ranked 3rd and 4th.

method would entail obtaining reasonings alongside rankings from both entities and subsequently conducting a qualitative analysis. However, due to resource constraints and the size of our dataset, this was not practical.

Data Diversity and Representativeness

While our dataset spans a variety of topics and age groups, its sole source—r/relationships subreddit—limits its diversity. This confines our analysis to views of individuals engaged in online relationship discussions on Reddit, particularly on r/relationships. Such users might exhibit systematic biases or share specific traits, rendering the dataset not fully representative. Notably, the majority of these users are not experts in Psychology, and their opinions may lack the depth and expertise a specialized behavioral therapist provides.

Server-Side Errors with ChatGPT

During our interactions with ChatGPT, server-side issues occasionally arose. For instance, the server sometimes reached capacity, preventing prompt completion, or the model hit its text length limit. To mitigate potential impacts on results, we capped the number of comments at eight, ensuring minimal influence of server-side errors on the overall outcome.

Dataset Imbalance in Topics

Regarding the dataset's topic distribution, an inherent imbalance exists. Everyday concerns or communication problems are more frequently discussed than topics like childbearing decisions. This skewness in topic distribution is natural given the prevalence of day-to-day issues compared to major life decisions.

Conclusion

Our research investigates the alignment between ChatGPT's ranking of relationship advice and human judgments. Statistical analyses revealed an IRA of less than 0.10 between the two, underscoring the weak alignment. ChatGPT's rankings were found to be inconsistent when it is used with default parameters. The consistency rate is 0.0 for identical prompts in the low disparity group and 0.015 for the medium disparity group. When exploring potential influencing factors, such as topic, opinion variance, and lexical complexity of the comments, as well as data timeframe, only comment length and lexical complexity demonstrated minor trends. The results suggest that ChatGPT may have a slight preference (more than 55% of the time) over longer suggestions and suggestions with less lexical complexity. However, these are not conclusive indicators. Notably, the model's potential exposure to data during training did not manifest any significant advantage in alignment with human rankings.

The findings, evident by the revealing statistics, imply that ChatGPT's decision-making process, particularly in the domain of relationship advice, remains difficult to interpret. It underscores the necessity for future research to undertake a deeper exploration into this behavior, potentially tuning models to better mirror human wisdom in specific areas.

References

Adam, E. K.; Chyu, L.; Hoyt, L. T.; Doane, L. D.; Boisjoly, J.; Duncan, G. J.; Chase-Lansdale, P. L.; and McDade, T. W. 2011. Adverse adolescent relationship histories and young adult health: cumulative effects of loneliness, low parental support, relationship instability, intimate partner violence, and loss. *Journal of Adolescent Health*, 49(3): 278–286.

American Academy of Child and Adolescent Psychiatry Committee on Health Care Access and Economics Task Force on Mental Health. 2009. Improving mental health services in primary care: reducing administrative and financial barriers to access and collaboration. *Pediatrics*, 123(4): 1248–1251.

Aminah, S.; Hidayah, N.; and Ramli, M. 2023. Considering ChatGPT to be the first aid for young adults on mental health issues. *Journal of Public Health*, fdad065.

Bartholomew, K.; and Allison, C. J. 2006. An attachment perspective on abusive dynamics in intimate relationships. *Dynamics of romantic love: Attachment, caregiving, and sex*, 102: 127.

Black, M.; et al. 2006. Physical dating violence among high school students–United States, 2003. *MMWR: Morbidity and mortality weekly report*, 55(19): 532–535.

Carlbring, P.; Hadjistavropoulos, H.; Kleiboer, A.; and Andersson, G. 2023. A new era in Internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions*, 32.

Chris. 2023. I asked chatgpt for dating advice.

Chrysikou, E. 2013. Accessibility for mental healthcare. *Facilities*, 31(9/10): 418–426.

Cocci, A.; Pezzoli, M.; Lo Re, M.; Russo, G. I.; Asmundo, M. G.; Fode, M.; Cacciamani, G.; Cimino, S.; Minervini, A.; and Durukan, E. 2023. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer and Prostatic Diseases*, 1–6.

Coker, A. L.; Smith, P. H.; Bethea, L.; King, M. R.; and McKeown, R. E. 2000. Physical health consequences of physical and psychological intimate partner violence. *Archives of family medicine*, 9(5): 451.

Collisson, B.; Cordoviz, P.; Ponce de Leon, L.; Guillen, S.; Shier, J.; and Xiao, Z. 2018. " Should I Break Up or Make Up?" A Text Analysis of Online Relationship Advice. *North American Journal of Psychology*, 20(2).

Elkington, K. S.; Hackler, D.; Walsh, T. A.; Latack, J. A.; McKinnon, K.; Borges, C.; Wright, E. R.; and Wainberg, M. L. 2013. Perceived mental illness stigma, intimate relationships, and sexual risk behavior in youth with mental illness. *Journal of adolescent research*, 28(3): 378–404.

Fernández-Fuertes, A. A.; and Fuertes, A. 2010. Physical and psychological aggression in dating relationships of Spanish adolescents: Motives and consequences. *Child abuse & neglect*, 34(3): 183–191.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2): e7785.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gao, C. A.; Howard, F. M.; Markov, N. S.; Dyer, E. C.; Ramesh, S.; Luo, Y.; and Pearson, A. T. 2023. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, 6(1): 75.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gross, N. 2023. What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, 12(8): 435.

Huang, Y.; Xiong, H.; Leach, K.; Zhang, Y.; Chow, P.; Fua, K.; Teachman, B. A.; and Barnes, L. E. 2016. Assessing social anxiety using GPS trajectories and point-of-interest data. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, 898–903.

Jensen, J. P.; and Bergin, A. E. 1988. Mental health values of professional therapists: A national interdisciplinary survey. *Professional Psychology: Research and Practice*, 19(3): 290.

Lefkowitz, E. S.; and Espinosa-Hernandez, G. 2007. Sex-related communication with mothers and close friends during the transition to university. *Journal of Sex Research*, 44(1): 17–27.

MacGeorge, E. L.; and Hall, E. D. 2014. Relationship advice.

McKiernan, A.; Ryan, P.; McMahon, E.; and Butler, E. 2017. Qualitative analysis of interactions on an online discussion forum for young people with experience of romantic relationship breakup. *Cyberpsychology, Behavior, and Social Networking*, 20(2): 78–82.

Miller, R.; Perlman, D.; and Brehm, S. S. 2007. Intimate relationships. *Handbook of Intercultural Communication*, 341.

OpenAI. 2023a. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2023b. Sharing and Publication Policy. Accessed: 2023-09-14.

Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; and Zimmer, M. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2): 20563051211019004.

Rakovec-Felser, Z. 2014. Domestic violence and abuse in intimate relationship from public health perspective. *Health psychology research*, 2(3).

Sallam, M. 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, 887. MDPI.

Saphire-Bernstein, S.; and Taylor, S. E. 2013. Close relationships and happiness.

Singh, O. P. 2023. Artificial intelligence in the era of ChatGPT-Opportunities and challenges in mental health care. *Indian Journal of Psychiatry*, 65(3): 297.

Smith, T. W.; and Baucom, B. R. 2017. Intimate relationships, individual adjustment, and coronary heart disease: Implications of overlapping associations in psychosocial risk. *American Psychologist*, 72(6): 578.

Tagliabue, S.; Olivari, M. G.; Giuliani, C.; and Confalonieri, E. 2018. To seek or not to seek advice: Talking about romantic issues during emerging adulthood. *Europe's Journal of Psychology*, 14(1): 125.

U.S. Government Accountability Office. 2022. Behavioral Health: Available Workforce Information and Federal Actions to Help Recruit and Retain Providers. https://www.gao.gov/products/gao-23-105250. Accessed: 2023-09-1.

Vallade, J. I.; Dillow, M. R.; and Myers, S. A. 2016. A qualitative exploration of romantic partners' motives for and content of communication with friends following negative relational events. *Communication Quarterly*, 64(3): 348–368.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? Yes.
   (g) Did you discuss any potential misuse of your work? Yes.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, we explored multiple factors that might affect ChatGPT's ranking patterns.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? No, because we have not found the same exact approach,i.e., trying to correlate a factor with the IRA, being used before
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA, because this paper does not involve theoretical proofs.
   (b) Did you include complete proofs of all theoretical results? NA, because this paper does not involve theoretical proofs.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? This research does not involve machine learning experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? This research does not involve machine learning experiments.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? This research does not involve machine learning experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? This research does not involve machine learning experiments.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? This research does not involve machine learning experiments.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? This research does not involve machine learning experiments.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, as we use OpenAI's language model, we cited OpenAI's research paper on its models.
   (b) Did you mention the license of the assets? No, because the models are released by OpenAI with its specific terms and conditions, which does not strictly qualify as a traditional "license" like MIT or GPL.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, the refined dataset and analysis codes are included in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, per the OpenAI policy, researchers are welcomed to use OpenAI models for research purposes. We cited the related terms in the beginning of the Methodology section. In terms of the Reddit data, we have cited explanations of Reddit's term for usage of its data. The data collection adheres to Reddit Terms of Use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, as we discussed, Reddit data from r/relationships subreddit do not contain any identifiable information or offensive content. For the GPT-3.5-turbo model used in this experiment, we did not prompt it to provide any identifiable information or offensive content.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, it is described in the Dataset overview in the Methodology section.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No. We have carefully considered the recommendation to include a Datasheet for our dataset. However, given the explicit structure and organization of our dataset—where each processed .json Reddit post unambiguously contains elements such as post descriptions, comments, and scores—we believe that the dataset's content is largely self-explanatory. Nonetheless, we appreciate the importance of clarity and are open to feedback to ensure the dataset's accessibility and proper utilization.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA, because the research does not involve crowdsourcing or human subjects.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA, because the research does not involve crowdsourcing or human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA, because the research does not involve crowdsourcing or human subjects.

(d) Did you discuss how data is stored, shared, and deidentified? NA, because the research does not involve crowdsourcing or human subjects.