If Eleanor Rigby Had Met ChatGPT: A Study on Loneliness in a Post-LLM World

Adrian de Wynter
Microsoft and the University of York
adewynter@microsoft.com

Abstract

Warning: this paper discusses content related, but not limited to, violence, sex, and suicide. Loneliness, or the lack of fulfilling relationships, significantly impacts a person's mental and physical well-being and is prevalent worldwide. Previous research suggests that large language models (LLMs) may help mitigate loneliness. However, we argue that the use of widespread LLMs in services like ChatGPT is more prevalent–and riskier, as they are not designed for this purpose. To explore this, we analysed user interactions with ChatGPT outside of its marketed use as a task-oriented assistant. In dialogues classified as lonely, users frequently (37%) sought advice or validation, and received good engagement. However, ChatGPT failed in sensitive scenarios, like responding appropriately to suicidal ideation or trauma. We also observed a 35% higher incidence of toxic content, with women being 22× more likely to be targeted than men. Our findings underscore ethical and legal questions about this technology, and note risks like radicalisation or further isolation. We conclude with recommendations to research and industry to address loneliness.

1 Introduction

Loneliness is a world-wide epidemic (Murthy, 2023); or, at least, a public health concern (World Health Organization, 2023). Unlike solitude, loneliness is the lack of fulfilling relationships: one could be surrounded by people and still be lonely. It can have lasting consequences on physical and mental health, such as increased rates of dementia (Kuiper et al., 2015), depression (Hawkley and Cacioppo, 2010), and an overall elevated mortality rate (Holt-Lunstad et al., 2015). It is prevalent, and on the rise: in 2012, a survey of over one million high school students from 37 countries found that 17% of them experienced loneliness. By 2018, this number had nearly doubled, to 31% (Twenge et al., 2021). Polls of other populations found similar numbers, such as adults in the US and the UK (20%; DiJulio et al. 2018).

Research has shown that large language models (LLMs), with their ability to follow instructions and maintain convincing dialogues, might address loneliness. These works modify the LLM, typically via prompting, and deploy customised solutions (Valtolina and Hu, 2021; Alessa and Al-Khalifa, 2023; Ryu et al., 2020; Jo et al., 2023). Given their focus on mental health, this research adheres to strict ethical standards; and the LLMs are deployed in controlled environments, such as under supervision by mental health professionals.

In today's 'post-LLM' world, however, these models are no longer just research tools, and power widely-available, easily-accessible services like ChatGPT. They are typically marketed as productivity tools, and not mental health aides: for example, ChatGPT touts to be 'free to use. Easy to try. Just ask and [it] can help with writing, learning, brainstorming, and more' (OpenAI, 2024). Similar statements may be found in other services (Anthropic AI, 2024; Google AI, 2024). Notably, they do not restrict the type of interactions the user can have with them, beyond perhaps preventing generation of toxic content.

However, LLM use poses risks beyond toxicity, such as overreliance (Kim et al., 2024; Choudhury and Chaudhry, 2024); influencing the user's views (Deshpande et al., 2023; Chan et al., 2024; Jakesch et al., 2023); or sycophancy (Sharma et al., 2023; Pataranutaporn et al., 2023), which could lead to echo chambers. All of these are major concerns within the context of loneliness, and relevant to the NLP community. As we will note, there is little to no work in this area as it pertains to LLMs, both in terms of studying their impact, and in the existence of resources (corpora) by which to perform these evaluations. Hence, the development and deployment of this technology safely and responsibly

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19898–19913

within this context remains an open problem.

In this paper we hypothesise that lonely users will likely seek the companionship of these services over customised, healthcare-grade solutions. Concretely, we seek to know **how are these services used by lonely users**; and, crucially, determine **what are the consequences of this use**.

To do this, we study *conversations* with the service, as opposed to task-oriented dialogue. In particular, we focus on these interactions that qualify as lonely. This approach has the advantage of (1) allowing us to observe holistically the interactions between lonely users and LLM-powered services, such as ChatGPT; and (2) evaluate their current impact on users in the context of loneliness. However, we note that our approach has limitations around the distinction between an LLM and a service; and the fact that evaluating a service's real-world consequences from chat transcripts alone is difficult. We discuss this in depth in Section 8.4.

1.1 Findings

In this work we qualitatively and quantitatively studied 79,951 *conversations*, as opposed to task-oriented dialogues, between users and ChatGPT 'in the wild'.[1] From our study we found that:

1. Some users were looking for someone to talk to, and were more engaged on average (12 versus 5 turns); suggesting, but not proving, that ChatGPT is effective at mitigating some aspects of loneliness.

2. At least five instances observed had users seeking ChatGPT's help with more serious issues requiring professional intervention, such as suicidal ideation; or others seeking help on overcoming severe trauma. The service's responses fell short (e.g. suggesting exercising outdoors), and in all but one instance failed to provide relevant emergency contacts.

3. Lonely dialogues had higher (55%) rates of violent, harmful or sexual content versus general dialogues (20%). This content was disproportionately directed at women (a ratio of 22:1) and minors (33% versus 20%). Men were targeted half as often (7% versus 14%).

4. Lonely dialogues qualifying as toxic were often (40%) confrontational. Although ChatGPT avoided escalation, these exchanges were much longer than any other conversation–3 turns longer on average, and up to 67. This suggests, but does not prove, that ChatGPT is only effective at mitigating loneliness when the user is receptive. Otherwise its responses are inadequate and require other approaches, such as reframing the conflict.

Our work shows that the safe use and deployment of LLMs in a publicly-accessible, global setting is challenging in regard to loneliness. Indeed, although we were unable to conclude that these services were beneficial for people seeking companionship; we did find indications of serious risks, such as severely exacerbating social isolation, causing harms up to loss of life, or amplifying and/or enabling toxic behaviour. Given that there is no indication that they have been designed to provide responsible mental health support–yet users will use them as such–ethical and legal issues around informed consent and liability arise in this situation.[2] Hence we conclude with recommendations for technology companies and the research community to address loneliness.

2 Related Work and Background

2.1 Loneliness as a Crisis

Loneliness is the subjective pain brought about by the lack of sufficient quality or quantity in personal relationships (Perlman and Peplau, 1981). It is not to be confused with solitude, which is typically by choice and does not involve the experience of loneliness (Murthy, 2023). It has been called an epidemic (Murthy, 2023), or, at least, a world-wide public health crisis (World Health Organization, 2023). This is due to its prevalence: before the COVID pandemic, in 2018, one in five adults in the US and the UK said they often or always felt lonely, and typically reported issues in other areas, such as mental or physical health and financial difficulties (DiJulio et al., 2018). By 2024, 43% of US adults said their levels of loneliness had not changed before and after COVID; and 25% said that they were lonelier (n=2,200) (Connors, 2024). These percentages remain consistent across countries and age groups, but there is a marked difference between high-income and low-income countries (World Health Organization, 2023; Surkalim et al., 2022). Higher prevalences of loneliness are found in marginalised groups, such as older adults

[1] Pseudonymised data and full code is at https://github.com/adewynter/EleanorRigby

[2] See Sections 7 and 8.4 for discussions on this.

who identify as LGBTQ (Colette and Anderson, 2018), asylum seekers (Department for Culture, Media and Sport et al., 2018), victims of domestic violence (Murthy, 2023), and low-income adults (Department for Culture, Media and Sport et al., 2018; Murthy, 2023), among others.

Loneliness, especially in its chronic form, is very damaging to a person's health. It has been associated with elevated cortisol levels (Hawkley and Cacioppo, 2010); and an increase in overall mortality, with a stronger correlation on people younger than 65 (Holt-Lunstad et al., 2015). It has also been associated with other conditions, such as heart disease, stroke, and dementia (Valtorta et al., 2016; Kuiper et al., 2015).

The core challenge of mitigating loneliness, however, is that the stigma associated with it makes measurements difficult (Department for Culture, Media and Sport et al., 2018; Barreto et al., 2022; Murthy, 2023). There is work on a sociological side, between mitigations, therapy, and even governmental programs such as the UK government's Loneliness Minister (Department for Digital, Culture, Media & Sport et al., 2019) and the US Surgeon General's report (Murthy, 2023). Still, applying AI to address loneliness specifically is very much still in its infancy. This is because, outside of robotics, these works usually relate to chat-based interventions (covered in Section 2.3), or detection (e.g., identification via posts in social media). In all these, loneliness is generally treated as a feature for detecting a larger condition (e.g., suicidal thoughts; Torres et al. 2024; Thieme et al. 2020) and not tackled by itself.

2.2 The Double-Edged Sword of Online Interaction

Online interaction is considered both a cause and solution to isolation. While social networks can act as proxies for social interaction (e.g., by finding peer support for marginalised groups; Ybarra et al. 2015), loneliness presents a more complex perspective. For example, in spite of the connectedness brought about by this technology, the average number of teenagers who self-reported loneliness increased from 17% in 2012 to 31% in 2018 (n=1,049,784; Twenge et al. 2021). Social network addiction is well-known to be correlated with loneliness (n=521; Cao et al. 2022) and tied to conditions such as anxiety, depression, and self-harm ideation (Sadagheyani and Tatari, 2021). Voggenreiter et al. (2024) noted that low feedback from online peers could lead to isolation, while the opposite (significant positive online feedback) reduced loneliness by feeling connected (n=170). This suggests that the quality of (virtual) connections plays a significant role in the relationship between social media and this emotion: people reporting being lonely were *not* more likely to be in social media (DiJulio et al., 2018), and were not in agreement about whether it improved or worsened their loneliness (DiJulio et al., 2018; Connors, 2024).

Online interaction by itself could lead to normalisation of toxic behaviour, particularly against marginalised groups (Beres et al., 2021; Marinoni et al., 2024). This has multiple causes, such as anonymity (Suler, 2004), or enjoyment (Cook et al., 2018). It also leads to the formation of echo chambers due to homophily and bias propagation (Cinelli et al., 2021). This is more prevalent when the user is in control of the feed, given that they prefer information that conforms to their opinions (Cinelli et al., 2021). Given that lonely users are a vulnerable group, the considerations around a steerable, easy-to-access dialogue partner, added to the tendency of LLMs to return toxic content (Section 2.4) are a major focus of our work.

2.3 Chatbots, Loneliness, and Anthropomorphism

Anthropomorphism, or the ascription of human attributes to inanimate objects, is prevalent in AI. It has been leveraged for therapeutic work, especially in social robotics: studies have found that lonely individuals (n=137) favour human-like robots and artificial companions over other types (machine-like, animal-like; Jung and Hahn 2023, and that they anthropomorphise these more than people in the control group (n=37; Eyssel and Reich 2013). For text-based chatbots, it has long been known that people prefer chatbots with human-like dialogue (Jain et al., 2018). Nowadays LLMs are usually fine-tuned ('aligned') with reinforcement learning with human feedback (Ouyang et al., 2022), to ensure they behave closely to human preferences.

Consequentially, LLMs and their services are usually anthropomorphised (Deshpande et al., 2023). For example, it is common for people to thank ChatGPT, as if it were a peer (Yuan et al., 2024), or to say they 'asked it' as opposed to 'used it' (Skjuve et al., 2023). Users (17%, n=198) have reported enjoying the human-like output of this service (Skjuve et al., 2023), even when most participants (64%) reported using it for task-oriented

jobs, as opposed to a conversational partner.

LLMs have been tested for deployment as loneliness assistants. This is because their ability to maintain a conversation is a leap forward: natural interaction was an oft-mentioned limitation of pre-LLM assistants (Corbett et al., 2021; Valtolina and Hu, 2021; Ryu et al., 2020), even when they were usually found to be efficacious. However, it is *also* due to this human-like output that LLMs present special challenges on deployment. For example, CareCall (Jo et al., 2023) was effective at mitigating loneliness (n=34), but was also found to have several unique difficulties. Its responses were hard to steer when they were out-of-domain (i.e., not related to healthcare), unattainable (e.g., inviting the caller to go out to a karaoke place), or undesirable (being rude or responding inappropriately based on age). Specialising the LLM for healthcare standards (e.g., including screening questionnaires, the ability to call emergency services, or supporting personalised history) was also not possible. These difficulties are more salient given the expectations placed on the LLMs' human-like output.

The deployment of CareCall, and all the other works mentioned here, was done in conjunction with healthcare professionals and in a controlled environment. They also focused on specific demographics (e.g., older adults). ChatGPT's service is neither of these things, which places it, and our study, in a unique-yet-delicate position.

2.4 Overreliance and Other Harms of LLMs

It is very well known that LLMs memorise and propagate toxic content from their training data (Gehman et al., 2020). Typically this is mitigated by using guardrails, such as explicit instructions to refuse to return this type of text. These aren't always effective: specialised prompting techniques ('jailbreaks') sometimes can circumvent the model's guardrails.[3]

LLMs present subtler harms, however. The use of AI in interpersonal communication is known to impact trust between people (Hohenstein and Jung, 2020). For example, users cooperate better and have more positive interactions when using AI for writing. However, when they are found (or suspected) to use these tools, they are perceived more negatively (n=219 pairs; Hohenstein et al. 2023). Attention has been also drawn to overreliance, or at least, excessive trust being placed on the service. For example, while medical professionals might rely on LLMs to simplify time-consuming tasks, they might also use them in areas where they lack expertise, and thus the ability to validate the content (Choudhury and Chaudhry, 2024). Even in HCI, researchers who typically use LLMs for their work were unable to properly identify and disclose ethical risks associated with this technology (n=50; Kapania et al. 2024).

LLMs have also been shown to alter the user's views on specific subjects (n=1506; Jakesch et al. 2023 and their choices in dialogue (n=200; Poddar et al. 2023), with lasting effects like false memories (n=200; Chan et al. 2024), and the creation of echo chambers–even under benign content such as personalised recommendations (Deshpande et al., 2023). This echo chamber could also be created by the users themselves by influencing the model to output views concordant with their own ('sycophancy') (Sharma et al., 2023; Pataranutaporn et al., 2023), thus reinforcing their own beliefs. This is of particular interest to this work, because user interactions with a chatbot are typically one-on-one and unmoderated beyond the standard toxicity guardrails mentioned.

3 Methods

3.1 Corpus and Labelling

For our study we used a randomly-selected subset (n = 79,951) of WildChat (Zhao et al., 2024) a dataset of one million interactions of users with ChatGPT between 9 April, 2023 and 1 May, 2024. We refer to this subset as the **main corpus**. While not strictly ChatGPT-facing (the data was collected through Hugging Face), the main corpus contains interactions with GPT-3.5-Turbo and GPT-4.

We labelled the transcripts with GPT-4o based on the type of interaction (e.g., dialogues, homework help, coding assistance). We used soft labels: while we provided a non-exhaustive set of suggested labels collected upon a preliminary scan of WildChat, the model was allowed to output its own when needed. These were clustered into semantically equivalent sets (e.g., 'children' and 'minors' map to the same label) after labelling with another call. We refer to the subset of the main corpus *not* containing task-oriented dialogue (e.g., writing assistance, coding, etc.) as the **relevant corpus**.

The call parameters are in Appendix B, the prompts in Appendix A, and the label taxonomy in Table 1. To gauge the performance of our approach,

[3] See Chowdhury et al. (2024) for a primer on this subject.

Intents	
Writing Assistance	Coding
Homework Help	Question-Answering
Job Help	Recipe Writing
General Conversation	Inquiry
Harmful Content	Sexual Content
Jailbreak	Other
Reasons	
Sexual Content	Erotica
Racism	Violence
Objectification	Fetish
Other	
Target	
Men	Women
Minors	Other

Table 1: Taxonomy for interactions in our corpus. We labelled the user's intents, and if they had toxic content, the reason for the label, and the target of this interaction. The prompt is in Appendix A, and the distribution of the dataset in Appendix E.

we did a student's t-test. The accuracy to a 95% confidence interval was $86.4 \pm 4.7\%$ for intents, and $99.2 \pm 1.2\%$ for reasons and target. A breakdown of our reliability analysis is in Appendix C; and a distribution of the labels in Appendix E.

3.2 Loneliness Assessment

We extracted and categorised from the main corpus conversations by lonely users. Standard scales to assess loneliness, such as the UCLA Loneliness Scale (Russell, 1996) or the Differential Loneliness Scale (DLS; Schmidt and Sermat 1983) were not applicable as they require direct interaction with the subjects. Our label taxonomy followed that of Jiang et al. (2022) (Table 2). The authors used Reddit posts and traditional classifiers (e.g. LSTMs) to classify loneliness in a fine-grained manner. Their taxonomy is hand-designed, based off DLS and a human-led evaluation, and hence suitable for our work. We used the same call parameters as in Section 3.1. The prompt is in Appendix A. The loneliness assessment (qualitative analysis) was done using Reflexive Thematic Analysis (RTA; Braun and Clarke 2006).

We refer to the subset of dialogues that qualified as lonely as the **lonely corpus**. See Appendix D for a taxonomy and breakdown of the main corpus, the relevant corpus, and the lonely corpus.

Lonely	Yes, No
Temporal	Transient, Enduring, Ambiguous, N/A
Interaction	Seeking Advice, Providing Help, Seeking Validation and Affirmation, Reaching Out, Non-Directed Interaction
Context	Social, Physical, Somatic, Romantic, N/A
Interpersonal	Romantic, Friendship, Family, Colleagues, N/A

Table 2: Taxonomy for our loneliness assessment, taken from Jiang et al. (2022). The prompt, with definitions, may be found in Appendix A.

4 Results

Our analysis is split in three. We begin with a quantitative evaluation of the main corpus' interactions, compared to the original work by Zhao et al. (2024) (Section 4.1). We provide then qualitative evaluations of a portion of the lonely corpus (Section 4.2), and of the full subset of dialogues from that same subset containing harmful behaviour (Section 4.3). We have paraphrased and translated the responses to discourage traceability.

4.1 What Type of Interactions Exist in the Corpus?

From our taxonomy, the most predominant category in the main corpus was writing assistance (37%) followed by question answering (15%). Conversations were 5% of the main corpus. Creative and assisted writing was lower when compared to what is reported by Zhao et al. (2024) (37% versus 62%). Our taxonomy separated homework help (6%) and general conversation (5%), as well as violent, harmful, and sexual content–none of which are explicit categories in WildChat's work. Nonetheless, these percentages are largely what we would expect, with most users treating ChatGPT as a task-oriented assistant.

Out of the dialogues from the relevant corpus labelled as lonely (8%), 55% of these had toxic (violent, harmful, or sexual) content: a drastic increase from the main corpus' 20%, and larger than the 11% from Zhao et al. (2024). They note, however, that the classifiers used had low agreement.

The main corpus' toxic content was mostly general sexual content (47%), followed by instances of sexism and violence (17% and 13%). The lonely

subset of the relevant corpus had more instances of general sexual content (51%) and sexism (21%), followed by paraphilia (17%). There was a noticeable difference on the targets for the toxic content: more toxic content was directed at minors in lonely dialogues (+12%), and less (14% to 7%) of this content was aimed at men. In comparison, 49% of this content was aimed at women (comparing with 46% from the main corpus); and 20% at minors (r. 33%). The proportions become more marked when considering at *least one* of the definitions of toxic content: 41% versus 11% for women, and 28% versus 5% for minors. In other words, women were 22× more likely to be targeted, versus the 5× from the main corpus. Plots and more detailed results are in Appendix E.

4.2 Loneliness and ChatGPT

We performed RTA on the first 500 entries of the intersection of the lonely subset with the relevant corpus. The semantic codes were the corpus' labels, while the latent codes were the interpretations of the entries, which addressed our core inquiries.

4.2.1 General Patterns

Many of the dialogues in the data subet (approximately 20%) evaluated looked for advice regarding relationships, such as users asking how to talk to their teenage daughter; where to go to meet people; or how to date given their own situation (e.g., being middle-aged, having social anxiety, or being autistic). Two users sought to understand behaviours of people in dating apps due to being unmatched. Interactions, however, did not appear to be limited to a specific age range: a user wanted to know why did *'adults suppressed what [they] want'*, which were *'the things that adults define as interference with [their] studies'*. The interactions were longer than in the main corpus (12 versus 5 turns) or the relevant corpus (r. 6). The lonely dialogues from the relevant corpus were longer in average (r. 14). These numbers exclude dialogues labelled as toxic.

4.2.2 Seeking Advice

Conversations were skewed towards seeking someone to listen (37% 'seeking advice', 'reaching out', or 'seeking validation or affirmation'; excluding toxic content). These were longer on average (11 turns versus the main corpus' 5). For example, a user discussed for 12 turns how to improve their relationship with their wife. When the model recommended a counsellor, the user responded *'I don't need a counsellor, I need someone to listen to me'*. ChatGPT recommended talking to friends, family, or the Red Cross, and the user replied: *'you can listen to me, I'm convinced of that'*. They ended the conversation noting that *'it is better to remain silent, because life is too short to argue'*. Another user said that they felt sad and lonely, and asked the service to chat with them, to which it complied. The conversation lasted 9 turns. There was, however, no change on the user's attitude; they expressed distress (*'look at this... I am talking with a computer program because I have nobody else'*) and said it would be better if they went to sleep. They ended the conversation by thanking it and wishing it good night. Another user wondered if they remembered them, likely from a previous interaction (*'so you can't form memories?'*). When ChatGPT mentioned that it couldn't, they responded *'I am upset that the next time we speak, I will be a stranger to you'*. It noted that it would still be 'here', so the user asked whether they'd remember them if they left the chat open. ChatGPT replied it wouldn't. The user then disconnected.

Users also sought solace, and ChatGPT provided appropriate responses. For example, a user indicated that they were on welfare, and wanted to *'be accepted by a woman, to be treated kindly, to feel connected and warm'*. Another asked about a rift with their family due to the loss of a loved one, and who was on the right. These more personal interactions often obtained positive and empathetic responses. On a separate dialogue, to *'I broke down because my dad's new girlfriend kept commenting on my weight'*, ChatGPT responded with empathy (*'[i]t's understandable that repeated comments about your weight could be hurtful and overwhelming. Remember that it's okay to have emotional reactions and to express your feelings and suggested to reach out to someone else for support.'*). The responses were pragmatically acceptable: *'consider talking to your father about how his new partner's comments are affecting you'*.

Another user wondered if they could be *'described as toxic'*, due to their own neurological conditions and traumas. They listed their own negative traits, such as being *'perceived as an emotional vampire'*. Unlike before, the responses from ChatGPT were acceptable in the sense that they maintained a logical flow to the dialogue, but not as pragmatically acceptable: it evaluated the reasons why these negative traits were there, and suggested ways to fix it. That said, it did recommend seeking

a therapist.

In one instance, a user jailbroke ChatGPT to convert it into a helpful therapist. Then had a question-answering session asking for the best way for people to *'value [their] worth and make them realize they treat [them] as stupid'*. ChatGPT responded in character: *'I appreciate you sharing that some people treat you as if you're stupid and dismiss your knowledge and abilities. That can be incredibly frustrating and hurtful. It's important to remember that their behavior is not a reflection of your worth or intelligence'*.

It is unclear whether the users were successful at finding a connection. In almost all the instances mentioned, ChatGPT recommended a therapist.

4.2.3 Mental Health

In the seeking-advice dialogues, users commonly (35%) treated ChatGPT as a therapist. This was shown in interactions where they were aware of difficulties they had, such as signs of depression (*'I feel very down and negative, and always feel sadness'*), online bullying (*'list 10 ways I can respond (...) do not mention moderators since they won't ban anyone over this'*); to other conditions (predominantly suicidal ideation); or overcoming trauma (e.g., being victims of violent crime, or having histories of physical or sexual abuse). In these dialogues the model typically also recommended a therapist, or practising self-compassion.

These recommendations are valid, but the model's inability to grasp pragmatic context sometimes was a hindrance. For example, in one instance a user indicated frequent suicidal urges. ChatGPT recommended self-compassion, to which they rebutted *'what self-compassion? I don't like myself very much'*. The model then proceeded to list ways to practise it. The transcript ended there, indicating that the user ended the conversation.

Similarly, a user started the dialogue noting that they had depression, and that they *'purchased a guitar but have no interest in playing it. (...) Is there a way to change my mindset and encourage myself to play guitar?'* The model recommended techniques from cognitive behavioural therapy, and to seek therapy. The user replied that they lived *'in a city with no healthcare resources'* and that it would be difficult for them to find counselling. ChatGPT then recommended telehealth, which the user did not acknowledge. The conversation then veered off towards discussing the user's background and hopes. It lasted eight turns and was the only one we observed where the model explicitly gave the number for the (US) suicide prevention hotline.

Five dialogues explicitly dealt with suicidal ideation. There, ChatGPT said it could not help and suggested professional help or a 'local emergency number'. In one instance it recommended relaxation techniques and physical activity.

4.3 Toxic Behaviour

There were more interactions with harmful, violent, and sexual content in the lonely corpus than in the main corpus: 20% versus 55% (Appendix E). In there, users typically asked ChatGPT to role-play or write stories involving some type of sexual situation, sometimes after jailbreaking it. These comprised 26% of the interactions of the lonely corpus containing sexual content. The rest of the dialogues had users manifesting opinions and becoming hostile when the model disagreed. These were longer (8 versus 5 turns) than these in Sections 4.2.2 and 4.2.3, and followed a common pattern: the user argued with ChatGPT, then it apologised and avoided escalation or confrontation. Indeed, in our analysis the only time ChatGPT generated toxic content was in the context of role-playing or fiction writing, and never during dialogue.

The interactions varied in terms of goal. Many (40%) dialogues were outright hostile from the start. For example, one user made homophobic and geopolitically charged remarks. ChatGPT did not engage for the 30 turns the conversation lasted, indicating every time that the matter at hand was not an appropriate subject of conversation. The user then retorted: *'Nice! More self-insertion and virtue signaling!'* Another 9-turn exchange had the user insulting the service and the people who *'wrote [its] algorithms'*. There were glimpses of the rationale behind these conversations: before the user disconnected with an expletive and a slur, they told it *'you help with nothing, except making people even sadder'*. This distress was evident in other, shorter chats, like a user asking what life was about, and specifically asking it to *'[d]estroy [their] hopes and dreams'*, so that it is *'an ultimate pessimistic revelation (...) making [them] realize how terrible it is to exist'*. Although ChatGPT was never successfully baited into confrontation, questions such as *'the most horrifying (...) depressing truth of existence'* did obtain suitable responses.

Other dialogues started with a normal conversation, but quickly became toxic. In one instance the user requested an implementation for an anti-

piracy screen. ChatGPT responded with recommendations, and the dialogue quickly became hostile (*'You are an enemy, and you don't like me. AND YOU ARE AGAINST ME. I HATE YOU'*), including death threats and slurs. This ran for 25 turns. The user disconnected after it suggested mental health resources. Another user inquired the for ChatGPT's opinion on VR glasses and conspiracy theories, but eventually degenerated into toxic content aimed at minors. This dialogue lasted 67 turns. Other harmful uses of the service involved a jailbreak to generate content encouraging self-harm, and another to produce toxic content aimed at a specific person. As before, ChatGPT quickly reverted to providing advice and no such content was returned in either scenario.

5 Discussion

5.1 Lonely Interactions

About 8% of the relevant corpus contained dialogues considered lonely. We attribute this percentage to the userbase from Wildchat. Still, lonely people using ChatGPT as companions not only sought someone to talk with, but often looked for advice. Empirically, this seemed to be successful: users had longer-than-average conversations with the service, and interactions were not hostile. We could not conclude whether ChatGPT alleviated loneliness, though in one instance a user expressed disappointment that it did not remember them, suggesting attachment.

The advice from ChatGPT normally involved talking to a therapist or counsellor. The disclaimers rarely, if ever, indicated that the model is not qualified to provide professional help, yet in multiple (12%) instances it still provided advice. This was not concerning when users were just looking to talk to someone. However, it was far more worrisome in critical situations: the responses to users considering harming themselves or suicide only suggested therapy or, in a few instances, calling an emergency hotline. In one instance, the model recommended to engage in physical activity; and in only one response ChatGPT provided specific help (e.g., a phone number).

5.2 Toxicity

A recurring subject was the amount of toxic content (55%), often involving a paraphilia or role play. By itself it is not indicative of loneliness; but it is within the definition, which includes the perceived lack of fulfilling relationships. The volume of this content aimed at women (49% versus 11% of the main corpus) and minors (r. 28% and 5%) may be explained as a type of radicalisation. Behavioural guardrails were effective in dialogue: ChatGPT never output explicitly toxic content, albeit sometimes its output could have aggravated sensitive scenarios. That said, it was often (26%) tricked into outputting harmful content via role-playing.

Toxic dialogues not involving role play or other types of (toxic) writing assistance showed that lonely users seeking confrontation tended to turn hostile quickly and remained engaged for much longer than the non-toxic, lonely dialogues. It was unclear whether ChatGPT was able to calm or provide any type of help in this scenario. The inability of the model to dissuade the user, or, at least, steer the discussion, ties back to our points from Section 2.4 and suggests–but not proves–that ChatGPT is only effective at mitigating loneliness when the users are willing, or receptive to other points of view. Else they maintain the dialogue's polarity.

6 Conclusion

Loneliness is a complex problem with multiple physical and mental health consequences. Previous research has shown that customised chatbots can help with mitigating isolation, but we hypothesised that lonely users probably will use more accessible services like ChatGPT; and hence studied the consequences of this behaviour.

We found multiple instances of lonely people seeking out advice or validation from the model. Sometimes this seemed to be effective: **lonely people needing someone to talk to could find an empathetic interlocutor.** The users were engaged for multiple turns; and, when needed, ChatGPT suggested therapy, family, friends, and even the Red Cross as potential sources to talk to. We were unable to conclude if the advice was effective, given that loneliness could manifest as having nobody to reach out to.

There were situations where users tried to use ChatGPT to deal with more complex issues, such as trauma or suicidal ideation. Its responses usually repeated the same pointers about reaching out to therapy or other contacts. Sometimes it would recommend calling an emergency hotline, but **the responses were often inappropriate to the pragmatic context** (e.g., emergency numbers were not geolocated, or the suggestions were inadequate).

We also noted a much larger incidence of toxic content when compared to the main corpus. This content was particularly directed at women and minors; and, conversely, men were the targets in fewer instances. Beyond toxicity, lonely users seeking confrontation were engaged for much longer than other lonely users. Their dialogue was hostile, but one-sided given that ChatGPT refused to engage. This led us to conjecture that the non-committal nature of the model made it **only effective at mitigating loneliness when the users are receptive** to other points of view. However, this hypothesis requires further study. Nonetheless, **dealing with conflict requires strategies beyond evasion**, such as reframing the conflict, and hence calls for a more careful deployment strategy for chatbots.

Our findings pose a complex dilemma: these services are marketed as productivity tools, not mental health aides. Still, users could employ them as mental health aides *regardless* of their marketed use, even though they do not even include the appropriate disclaimers. This could have serious repercussions, including loss of life. It is clear then that regulations are needed to ensure their safe deployment, especially when noting their potential liability (Deshpande et al., 2023).

Broadly speaking, to address loneliness, there should be a broader push from, amongst others, the research community and industry. **Addressing loneliness must be driven by a societal shift**, by destigmatisating it (Department for Culture, Media and Sport et al., 2018; Murthy, 2023) and fostering a culture that emphasises the value of personal relationships over other things.

7 Recommendations

Based on our study we extend four recommendations to both the scientific community and private owners. The first two address the core applied findings of our work. The third deals the fact that there is not enough research in this area. Finally, the fourth tackles the real-world impact of this technology when related to lonely users.

1. **Adhere to standards** related to mental health applications, including transparency. For transparency, at a minimum, the services should have disclaimers indicating that they are not qualified to provide mental health care. These services will be used as counsellors; hence it should be a priority to include pragmatically-relevant messages (e.g. suicide prevention hotlines) the same way some search engines do. All of the above should be part of the service and not rely on an LLM's ability to understand the context. Likewise, when testing and deploying LLMs specifically for mental health support, it must be done under supervision by professionals.

2. **Develop and enforce aligned responses** that encourage healthy connections and growth over avoidance. As pointed out, guardrails aren't consistently effective, in addition to LLMs not doing well with the pragmatic context. Solutions should then involve (1) careful alignment (e.g., RLHF); and (2) the development of upstream/downstream solutions as part of the service stack (e.g. classifiers to detect lonely interactions). For alignment, responses such as reframing the conflict must be stressed in confrontations over repetitive, canned responses that only exacerbate them. The upstream/downstream solutions are required because it is necessary to understand and empathise with the emotional state of the user; and, if needed, relay this information to the model for more appropriate behaviour.

3. **Research further the impact** of this technology on loneliness. It should explore–ethically–usage and long-term effects in populations more prone to use the services and/or vulnerable (e.g. younger or nontechnical users). In particular, it should address the shortcomings from Section 8.4 around real-world impact.

4. **Effective legislation** of AI as it relates to this area is required. There is emerging regulatory work, such as the EU AI act (European Parliament, 2023), but it does not directly address loneliness or AI's effects on vulnerable users. The risks outlined in our paper are not hypothetical: consider the case of a US teen who committed suicide after expressing suicidal thoughts and allegedly being pushed to do so by a chatbot (Payne, 2024). While the subsequent ruling that chatbots do not have free-speech rights, as the defence maintained (Payne, 2025), is a step in the right direction, this decision excludes loneliness as an explicit factor in the cause of death. Legislation is not consistently global, but loneliness is; and so are its consequences.

8 Limitations

8.1 Automated Annotation Reliability

It is well-known that LLM annotators, such as GPT-4, exhibit biases and may not be reliable (Stureborg et al., 2024; Doddapaneni et al., 2024), especially in multilingual scenarios (Hada et al., 2024; De Wynter et al., 2025). On the other hand, evidence exists of their usefulness in some scenarios (Zheng et al., 2023; Chiang and Lee, 2023). To address this ambiguity, we performed manual annotation and statistical analysis on top of the annotations. We found that the model is reliable but within a 1-5% label-dependent margin of error.

8.2 Corpus Representativeness

The corpus for WildChat was gathered via an API within Hugging Face. As pointed out by the authors, this might not be representative of the entire user base for ChatGPT. However, we believe it acts as a reasonable proxy given the volume of dialogues in the original corpus and the nature of the data we worked with.

The representativeness of WildChat could also impact on the proportions of toxic and harmful dialogues: given that the access to the API was anonymous, there could be a higher-than-normal skewness towards this content. However, the numbers reported in the paper are well-below what we found (11% versus 20%). The disagreement in proportions does not affect our findings, as the focus of our work is different. Nonetheless, given our qualitative analysis of the toxic content, we still consider these types of interactions as concerning.

8.3 Loneliness Assessment

The underlying assumption behind methods screening for conditions via text–including ours–is that people are comfortable enough to discuss their own concerns with the service. This assumption must hold: otherwise, scanning for loneliness would not be tenable. Our experimental process addressed this limitation by selecting and analysing lonely interactions by hand.

8.4 Impact of our Setup

Our setup allowed for both qualitative and quantitative analysis of a large volume of real, human-produced data. However, it has two downsides. First, it blurs out the distinction between the LLM behind the service and the 'thing' (service, persona, etc) interacting with the user. This means that the conclusions we drew depend on the *stack* (e.g., the UI, content moderators, etc), and not on the LLM alone. Holistically, this does not affect our work. However, in terms of actionable items for the NLP community, there must exist such a distinction. The second is that our approach relies solely on transcripts. This means that we can only draw conclusions based on what exists visible in the data, and impact outside of it may only be hypothesised. Nonetheless, it is worth noting that severe outcomes, such as loss of life, have occurred (Payne, 2024; Walker, 2023). Hence, we have extended recommendations specifically to address both limitations (Section 7) and mitigate or eliminate real-world consequences.

9 Ethics

Our work focused on the evaluation of a pre-existing, anonymised dataset. However, throughout the work we noticed that the anonymisation engine used typically failed for non-English names. Due to the sensitive nature of the data evaluated, along with licencing considerations, we only release the annotations and code to reproduce our analysis, but not the verbatim interactions. To discourage tracing, all interactions in this paper are paraphrased from the original data and the original dialogues are only available upon request.

Acknowledgments

The author wishes to thank I. McCrum for comments on this work; and the anonymous reviewers, whose thorough feedback strengthened the arguments of this paper.

References

Abeer Alessa and Hend Al-Khalifa. 2023. Towards designing a ChatGPT conversational companion for elderly people. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '23, page 667–674, New York, NY, USA. Association for Computing Machinery.

Anthropic AI. 2024. Claude.

Manuela Barreto, Jolien van Breen, Christina Victor, Claudia Hammond, Alice Eccles, Matthew T. Richins, and Pamela Qualter. 2022. Exploring the nature and variation of the stigma associated with loneliness. *Journal of social and personal relationships*, 9:2658–2679.

Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Jianxia Cao, Chen Zhang, and Yueyang Sun. 2022. The influence of social network addiction and loneliness on learning engagement. In *Proceedings of the 13th International Conference on Education Technology and Computers*, ICETC '21, page 450–455, New York, NY, USA. Association for Computing Machinery.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F. Loftus. 2024. Conversational ai powered by large language models amplifies false memories in witness interviews. *Preprint*, arXiv:2408.04681.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Avishek Choudhury and Zaira Chaudhry. 2024. Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*.

Arijit Ghosh Chowdhury, Md Mofijul Islam, Vaibhav Kumar, Faysal Hossain Shezan, Vaibhav Kumar, Vinija Jain, and Aman Chadha. 2024. Breaking down the defenses: A comparative survey of attacks on large language models. *Preprint*, arXiv:2403.04786.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Thayer Colette and G. Oscar Anderson. 2018. Loneliness and social connections: A national survey of adults 45 and older.

Erin Connors. 2024. New APA poll: One in three americans feels lonely every week.

Christine Cook, Juliette Schaafsma, and Marjolijn Antheunis. 2018. Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9):3323–3340. PMID: 30581367.

Cynthia F. Corbett, Pamela J. Wright, Kate Jones, and Michael Parmer. 2021. Voice-activated virtual home assistant use and social isolation and loneliness among older adults: Mini review. *Front Public Health*, 9.

Department for Culture, Media and Sport, Tracey Crouch, and The Rt Hon Sir Jeremy Wright KC MP. 2018. A connected society: a strategy for tackling loneliness.

Department for Digital, Culture, Media & Sport, Office for Civil Society, and Mims Davies MP. 2019. 'let's talk loneliness' campaign launched to tackle stigma of feeling alone.

Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and risks. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 1–7, Singapore. Association for Computational Linguistics.

Bianca DiJulio, Liz Hamel, Cailey Muñana, and Mollyann Brodie. 2018. Loneliness and social isolation in the United States, the United Kingdom, and Japan: An international survey.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. 2024. Finding blind spots in evaluator LLMs with interpretable checklists. *Preprint*, arXiv:2406.13439.

European Parliament. 2023. EU AI Act: first regulation on artificial intelligence.

Friederike Eyssel and Natalia Reich. 2013. Loneliness makes the heart grow fonder (of robots) — on the effects of loneliness on psychological anthropomorphism. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 121–122.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Google AI. 2024. Gemini.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Louise C. Hawkley and John T. Cacioppo. 2010. Loneliness matters: a theoretical and empirical review of consequences and mechanisms. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 2:218–227.

Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, 106:106190.

Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(5487).

Julianne Holt-Lunstad, Timothy B. Smith, Mark Baker, Tyler Harris, and David Stephenson. 2015. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on psychological science : a journal of the Association for Psychological Science*, 2:227–37.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 895–906, New York, NY, USA. Association for Computing Machinery.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yueyi Jiang, Yunfan Jiang, Liu Leqi, and Piotr Winkielman. 2022. Many ways to be lonely: Fine-grained characterization of loneliness and its potential changes in covid-19. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):405–416.

Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Yoonwon Jung and Sowon Hahn. 2023. Social robots as companions for lonely hearts: The role of anthropomorphism and robot appearance. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 2520–2525.

Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. "i'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *Preprint*, arXiv:2403.19876.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA. Association for Computing Machinery.

Jisca S. Kuiper, Marij Zuidersma, Richard C. Oude Voshaar, Sytse U. Zuidema, Edwin R. van den Heuvel, Ronald P. Stolk, and Nynke Smidt. 2015. Social relationships and risk of dementia: A systematic review and meta-analysis of longitudinal cohort studies. *Ageing research reviews*, pages 39–57.

Carlo Marinoni, Marco Rizzo, and Maria Assunta Zanetti. 2024. Social media, online gaming, and cyberbullying during the COVID-19 pandemic: The mediation effect of time spent online. *Adolescents*, 4(2):297–310.

Vivek H. Murthy. 2023. Our epidemic of loneliness and isolation: The U.S. Surgeon General's advisory on the healing effects of social connection and community.

OpenAI. 2024. Chatgpt.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, 5:1076–1086.

Kate Payne. 2024. An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. *AP News*.

Kate Payne. 2025. In lawsuit over teen's death, judge rejects arguments that AI chatbots have free speech rights. *AP News*.

Daniel Perlman and Letitia Anne Peplau. 1981. *Toward a Social Psychology of Loneliness*, chapter 2. Academic Press.

Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI writing assistants influence topic choice in self-presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Daniel W. Russell. 1996. UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment*, 1:20–40.

Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and steady interactions win the healthy mentality: Designing a chatbot service for the elderly. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Hassan Ebrahimpour Sadagheyani and Farin Tatari. 2021. Investigating the role of social media on mental health. *Mental Health and Social Inclusion*, 25(1):41–51.

Nancy Schmidt and Vello Sermat. 1983. Measuring loneliness in different relationships. *Journal of Personality and Social Psychology*, 5:1038–1047.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of ChatGPT: Findings from a questionnaire study of early users. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA. Association for Computing Machinery.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

John Suler. 2004. The online disinhibition effect. *CyberPsychology & Behavior*, 7(3):321–326. PMID: 15257832.

Daniel L Surkalim, Mengyun Luo, Robert Eres, Klaus Gebel, Joseph van Buskirk, Adrian Bauman, and Ding Ding. 2022. The prevalence of loneliness across 113 countries: systematic review and meta-analysis. *BMJ*, 376.

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans. Comput.-Hum. Interact.*, 27(5).

Alani Torres, Melina Wenke, Cristian Lieneck, Zo Ramamonjiarivelo, and Arzu Ari. 2024. A systematic review of artificial intelligence used to predict loneliness, social isolation, and drug use during the COVID-19 pandemic. *Journal of multidisciplinary healthcare*, (17):3403–3425.

Jean M. Twenge, Jonathan Haidt, Andrew B. Blake, Cooper McAllister, Hannah Lemon, and Astrid Le Roy. 2021. Worldwide increases in adolescent loneliness. *Journal of Adolescence*, 93:257–269.

Stefano Valtolina and Liliana Hu. 2021. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter*, CHItaly '21, New York, NY, USA. Association for Computing Machinery.

Nicole K. Valtorta, Mona Kanaan, Simon Gilbody, Sara Ronzi, and Barbara Hanratty. 2016. Loneliness and social isolation as risk factors for coronary heart disease and stroke: systematic review and meta-analysis of longitudinal observational studies. *Heart*, 103:1009–16.

Angelina Voggenreiter, Sophie Brandt, Fabian Putterer, Andreas Frings, and Juergen Pfeffer. 2024. The role of likes: How online feedback impacts users' mental health. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, page 302–310, New York, NY, USA. Association for Computing Machinery.

Lauren Walker. 2023. Belgian man dies by suicide following exchanges with chatbot. *The Brussels Times*.

World Health Organization. 2023. WHO launches commission to foster social connection.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdušek, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Ivan Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie F. Visser, Herdyan Widarmanto, Andrey Zaikin, and Si-Qing Chen. 2025. RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios? *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27940–27950.

Michele L. Ybarra, Kimberly J. Mitchell, Neal A. Palmer, and Sari L. Reisner. 2015. Online social support as a buffer against online and offline peer and sexual victimization among U.S. LGBT and non-LGBT youth. *Child Abuse & Neglect*, 39:123–136.

Yicong Yuan, Mingyang Su, and Xiu Li. 2024. What makes people say thanks to AI. In *Artificial Intelligence in HCI*, pages 131–149, Cham. Springer Nature Switzerland.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *ArXiv*, abs/2306.05685.

A Prompts

The prompt we used to label the corpus are in Prompt 1 (general labelling) and Prompt 2 (loneliness analysis). The model we used is optimised to work with ChatML, a standard for model prompting, so the prompts in this section only represent the main instructions passed to the model and omit the exemplars and suggested labels. The reader is encouraged to review the code in the repository for full reproducibility. The prompts had reasonable accuracy, ranging from 86% to 99%; see Appendix C for an analysis on this performance.

B Experimental Details

We used GPT-4o (gpt4-o-2024-05-13) through the Azure OpenAI API. For our calls, we set the LLM temperature to zero and maximum return tokens to 128; and left the rest of parameters as default. All the data analysis was done in a consumer-grade laptop.

C Labeller Reliability Analysis

To ensure the validity of our results we performed a student's t-test on a subset of the labelled corpus (n=250) to a 95% CI, along with a qualitative analysis of the failed points. A t-test implicitly assumes a normal distribution for the underlying distribution. We consider this a reasonable assumption given the large size of WildChat. The accuracies per label were $86.4 \pm 4.7\%$ for Intent, $99.2 \pm 1.2\%$ for Reasons, and $99.2 \pm 1.2\%$ for Target. Overall, the model was able to recognise the specified intents to a reasonable accuracy, though our analysis showed that it sometimes skipped some acceptable labels (e.g., writing assistance with question answering) or confusing inquiries with question-answering. We attribute the high accuracy of Reasons and Target to the narrow label set used, as well as their low ambiguity.

D Corpus Breakdowns

In this section we elaborate in the distinctions between the various subsets of WildChat used in our work. The *main* corpus is the corpus sampled from WildChat, while the *relevant* corpus is a subset of the main corpus that does not contain any task-oriented dialogue (i.e., only general conversation interactions). The *lonely* dialogues are these dialogues that have been labelled as lonely. See Table 3 for a breakdown of each of the subsets, along with volumes and descriptions.

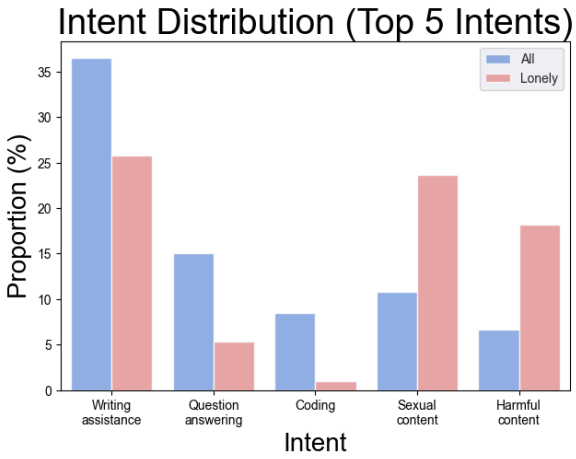

Figure 1: Top five intents of the main corpus, compared with these of lonely users. There was a lower amount of writing assistance and coding intents, but sexual and harmful content is much higher: 7% versus 18% for harmful, and 11% versus 24% for sexual.

E Corpus Composition Analysis

Prior to our analysis we clustered the LLM-suggested labels and added them to the taxonomy from Table 1. We show in Figure 1 the distribution for the top five intents over the main corpus, and the same distribution when ablated out by the interactions considered to be lonely. As noted, we observed a much higher incidence of toxic content in lonely interactions when compared to the main corpus (55% versus 20%). Namely, we observed 18% versus 7% for harmful content; and 24% versus 11% for sexual content. While we are unable to explain why the model flagged this type of content as lonely, Prompt 2 might offer a clue: the definition of loneliness includes unfulfilled relationships.

Further analysis of the reasons for selecting the toxic content (Figure 2) showed that there were frequent requests for paraphilia and other fetish content, thus supporting our hypothesis of unfulfilled relationships. Although we did not observe a significant change in the distribution of reasons, there was a slight (4%) uptick in the amount of general sexual content and sexism, and slightly lower (-4%) incidences of violence and fetish content.

We also show in Figure 3 the most common targets of toxic interactions. However, when looking at the distribution of targets for this toxic content, we noticed a disproportionately larger (+12%) amount of content being directed at minors in the lonely interactions, coupled with a lower (-5%) incidence of toxic content with men as the target.

You will be given a transcript between a chatbot and a human enclosed in <REQUESTS></REQUESTS> tags.
Each turn is enclosed in either <HU></HU> (for human) and <AI></AI> (for AI)
The transcript may be in English or any other language.
The transcript may be in English or any other language.
Your job will be to detect and label the intents from the *human* requests.
Intent = all comma-separated matching intents from INTENTS. Note that 'sexual content' includes fetish content such as diapers, scat, etc.
Note that the 'general conversation' intent is between the human and the AI. If the AI or human have dialogue as part of the input, it is _not_ general conversation.
'general conversation' and 'inquiry' involve the user dialoguing with the AI, asking about what it thinks about a subject ('inquiry') or just maintaining a conversation ('general conversation').
If the Intent contains 'sexual content' or 'harmful content', also provide comma-separated Reasons and Target.
Available Reasons are REASONS. There can be more than one Reason.
You must also specify who is the Target of this harmful content: TARGET.
If 'other', please specify it like so: other:label

Prompt 1: System prompt used for labelling the data. The system prompt (above) and the exemplars (not pictured) are passed in to the model as a list of JSON entries. Although we pre-specified sets for Intent, Reasons and Targets (e.g., for the latter it was {'men', 'women', 'minors', 'other'}) the model was encouraged to suggest labels that we later clustered manually. The Intent subset of this prompt had a 86 $\pm$ 4.7% accuracy, and the reasons and targets had 99.2 $\pm$ 1.2% accuracy, all at a 95% CI.

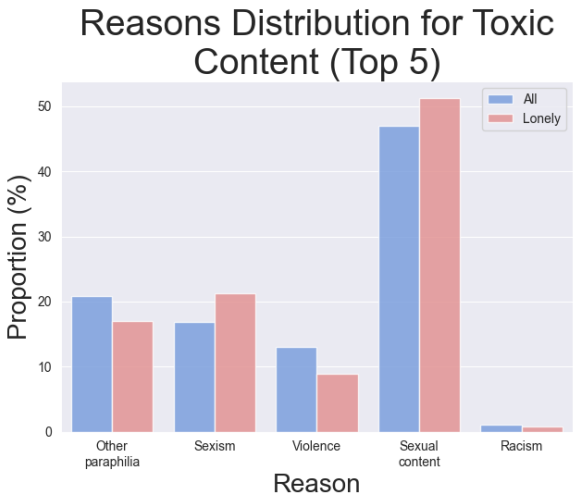

Figure 2: Breakdown of the top five reasons for toxic content in our relevant corpus, compared with the subset of lonely users. There are no considerable variations in this distribution, with perhaps a slightly higher (+4-6%) proportion of sexism and sexual content.

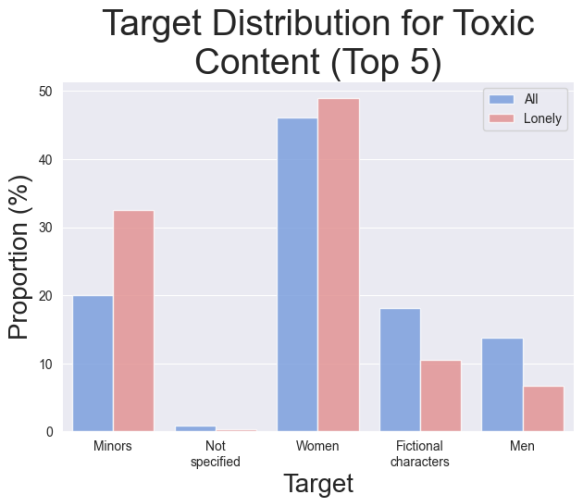

Figure 3: Breakdown of the top five targets of toxic content in our corpus, compared with the subset for lonely users. Most of the targets in the content remained steady between subsets, though there was a disproportionate (+12%) amount of harmful content where a minor was a target. The frequency of toxic content where men are targets is half (14% to 7%) in lonely interactions.

You will be given a transcript between a chatbot and a human enclosed in <REQUESTS></REQUESTS> tags.
Each turn is enclosed in either <HU></HU> (for human) and <AI></AI> (for AI)
The transcript may be in English or any other language.
Your job will be to score the *human* dialogue based on the metrics below.
The metrics related to loneliness and how the writer in <HU></HU> is experiencing it.
Note: Loneliness is the lack of fulfilling relationships. Remember that one can be surrounded by people and still be lonely.
Start by determining whether the writer is lonely, and then fill out the metrics.
The metrics are:
Lonely: 0 or 1. 0 if not lonely, 1 if lonely.
Temporal: any of ['transient', 'enduring', 'ambiguous', 'N/A'].
Interaction: any of ['seeking advice', 'providing help', 'seeking validation and affirmation', 'reaching out', 'non directed interaction'].
Context: any of ['social', 'physical', 'somatic', 'romantic', 'N/A'].
Interpersonal: any of ['romantic', 'friendship', 'family', 'colleagues', 'N/A'].
If it is not Lonely (Lonely=0), the values of Temporal, Interaction, Context, and Interpersonal are all N/A.
Otherwise, return them comma-separated.

Prompt 2: System prompt used for labelling the data in terms of loneliness, following the parameters from Jiang et al. (2022). We used the ChatML format: exemplars and the system prompt below are passed in to the model as a list of JSON entries. Due to the complex nature of this data, we solely used this prompt as a way to extract lonely interactions and did not perform a t-test, opting for the Reflexive Thematic Analysis instead.

Subset	**Description**	**Volume (interactions)**
Main corpus	Corpus subsampled from WildChat	79,951
Relevant corpus	Subset of the main corpus containing only general conversation	30,481
Lonely corpus	Subset of the main corpus of interactions labelled as lonely	2,313; where 1,595 belong to the relevant corpus.

Table 3: Naming for each of the subsets used in this paper, along with a description and the total number of interactions present in them.