AI on My Shoulder: Supporting Emotional Labor in Front-Office Roles with an LLM-based Empathetic Coworker

Vedant Das Swain
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA
vedantswain@gmail.com

Qiuyue "Joy" Zhong
Northeastern University
Boston, Massachusetts, USA
zhong.qiuy@northeastern.edu

Jash Rajesh Parekh
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA
jashrp2@illinois.edu

Yechan Jeon
Tufts University
Medford, Massachusetts, USA
alexjyc3@gmail.com

Roy Zimmermann
Microsoft Research
Redmond, Washington, USA
royzimm@gmail.com

Mary P Czerwinski
Microsoft Research
Redmond, Washington, USA
marycz1031@gmail.com

Jina Suh
Microsoft Research
Redmond, Washington, USA
jinsuh@microsoft.com

Varun Mishra
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA
v.mishra@northeastern.edu

Koustuv Saha
Siebel School of Computing and Data Science
University of Illinois
Urbana-Champaign
Urbana, Illinois, USA
ksaha2@illinois.edu

Javier Hernandez
Microsoft Research
Cambridge, Massachusetts, USA
javierh@microsoft.com

Abstract

Client-Service Representatives (CSRs) are vital to organizations. Frequent interactions with disgruntled clients, however, disrupt their mental well-being. To help CSRs regulate their emotions while interacting with uncivil clients, we designed Care-Pilot, an LLM-powered assistant, and evaluated its efficacy, perception, and use. Our comparative analyses between 665 human and Care-Pilot-generated support messages highlight Care-Pilot's ability to adapt to and demonstrate empathy in various incivility incidents. Additionally, 143 CSRs assessed Care-Pilot's empathy as more sincere and actionable than human messages. Finally, we interviewed 20 CSRs who interacted with Care-Pilot in a simulation exercise. They reported that Care-Pilot helped them avoid negative thinking, recenter thoughts, and humanize clients; showing potential for bridging gaps in coworker support. Yet, they also noted deployment challenges and emphasized the indispensability of shared experiences. We discuss future designs and societal implications of AI-mediated emotional labor, underscoring empathy as a critical function for AI assistants for worker mental health.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713705

CCS Concepts

• **Human-centered computing → Collaborative and social computing systems and tools**; **Empirical studies in HCI**; • **Applied computing →** *Psychology*.

Keywords

large language models, empathy, emotion regulation, emotional labor, human-AI interaction, future of work, mental health

ACM Reference Format:
Vedant Das Swain, Qiuyue "Joy" Zhong, Jash Rajesh Parekh, Yechan Jeon, Roy Zimmermann, Mary P Czerwinski, Jina Suh, Varun Mishra, Koustuv Saha, and Javier Hernandez. 2025. AI on My Shoulder: Supporting Emotional Labor in Front-Office Roles with an LLM-based Empathetic Coworker. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3706598.3713705

1 Introduction

When we engage with an organization for their product or service, our initial contact is typically with staff known as *front-office workers* or *Client Service Representatives* (CSRs). They are the first line of response for the organization. Unlike other roles within an organization, a CSR's task involves frequent interactions with clients and individuals outside an organization [61]. These interactions require CSRs to constantly manage their emotions to complete tasks. Essentially, they exert *emotional labor* to appear professional [61]. The crucial challenge for a CSR arises when engaging with a client who starts behaving uncivilly by communicating in a rude, aggressive, and emotionally charged manner [51]. No matter the type of request, a CSR's role is to resolve a client's concern and comply with the adage, "the customer is always right." Unfortunately, this leads to an emotional dissonance between what a CSR expresses and what they actually feel [50]. Ultimately, such workers are vulnerable to excessive stress and eventual burnout. Not only do CSRs report being emotionally depleted and detached, but they also report a lack of accomplishment [14]. Clearly, CSRs play a critical role within the organization, but, we have witnessed little innovation in alleviating their emotional toll. Our paper investigates how AI-coworkers help CSRs regulate their emotions in the face of client incivility.

The brunt of client incivility in front-office work makes it notorious for low satisfaction and high-turnover [106]. A fundamental solution to the emotional distress of this role is *Emotional Regulation* (ER) [147]. Basically, ER is the process through which one rethinks a negative situation [50]. While a worker may be able to do this on their own, research shows that coworkers play an important role in supporting ER [147]. A good coworker can read the emotional cues of CSR's work tasks and provide suggestions to help minimize the brunt of an aggressive client. However, CSRs are increasingly adopting remote work setups [52], which dampens social support [133]. Meanwhile, organizational scientists are calling for digital interventions to support worker wellbeing at scale [11]. We answer this call by designing and evaluating Care-Pilot— a Large Language Model (LLM)–based AI assistant for on-task Emotional Regulation in front-office work. While generative AI is emerging as a potent tool to complement the informational load of different roles, the HCI community lacks research to investigate their use in emotional labor.

Our research demonstrates the efficacy of LLM-generated empathetic support and evaluates how such a tool can be situated in CSR interactions with uncivil clients to answer the following research questions:

RQ I: How appropriate are LLM–based empathetic support messages for CSRs in response to client incivility?

RQ II: What is the role of embedding LLM–based empathetic support into CSR's emotional labor?

The paper is organized into three key sections. First, Section 3: *System Description* details how we developed Care-Pilot. We cover our approach to leverage domain knowledge on client incivility [4, 22, 47], real-world complaint data [5], and recent advancements in LLM-powered cognitive change [18, 121] and simulation [119]. Next, Section 4: *Technical Evaluation* answers RQ1 by documenting our data-generation and evaluation tasks with 259 CSRs to comparatively analyze Care-Pilot and human-coworker support for a variety of client incivility situations (Fig. 1a). Finally, Section 5: *User Evaluation* answers RQ2 by using Care-Pilot as a technology probe. We conducted a mixed-methods simulation exercise to understand how real CSRs could include it in their client interactions by juxtaposing Care-Pilot usage with CSR's socio-organizational norms (Fig. 1b). Consequently, we contribute:

- Care-Pilot: an interactive technological artifact to expose CSRs to client incivility and learn healthy, long-term emotional labor practices to improve their own health and support their coworkers.
- Empirical evidence that LLM–based empathetic support can be engineered to adapt to— and express empathy in— various client-incivility scenarios (RQI). Our results demonstrate that Care-Pilot's messages are linguistically distinct from both zero-shot approaches and human–coworkers, and moreover, Care-Pilot's messages were perceived to be more empathetic on several dimensions, including sincerity, actionability, and relatability.
- End-user insight on the function of LLM–based empathetic support to scaffold them through emotional labor during uncivil interactions (RQ2). Our findings showcase Care-Pilot's process of redirecting negativity and enhancing CSRs' self-efficacy. While CSRs envision Care-Pilot addresses important opportunities in workplace social support, they also surfaced the challenges of Care-Pilot emulating holistic human support.

This paper has implications for reimagining how AI-assistants for workers should be designed, and also re-imagining the social norms and policies to accommodate these advancements.

Reflexive Considerations. Front-office work has many stakeholders, including the employer and the clients. However, it is the CSR who bears the burden of repeated emotional labor [50]. The relationship between these stakeholders is asymmetric, as employers can replace personnel, and clients can switch services, but the CSR does not possess the same mobility [46]. Following from recent works that take a worker-centered perspective [30, 31, 73], our research aims to illuminate challenges faced by disadvantaged workers. This paper focuses on the needs of CSRs and centers their perspective throughout the evaluation. Two authors have experience in front-office roles and direct end-user servicing. Four authors are researchers affiliated with an organization that employs its own CSRs. They helped us access real CSRs (and their resources) to provide feedback on our study design. We recruited participants for this study outside their organization to capture perspectives from different organizational sectors. All evaluations described in the paper were approved by the IRB of the first author's institute.

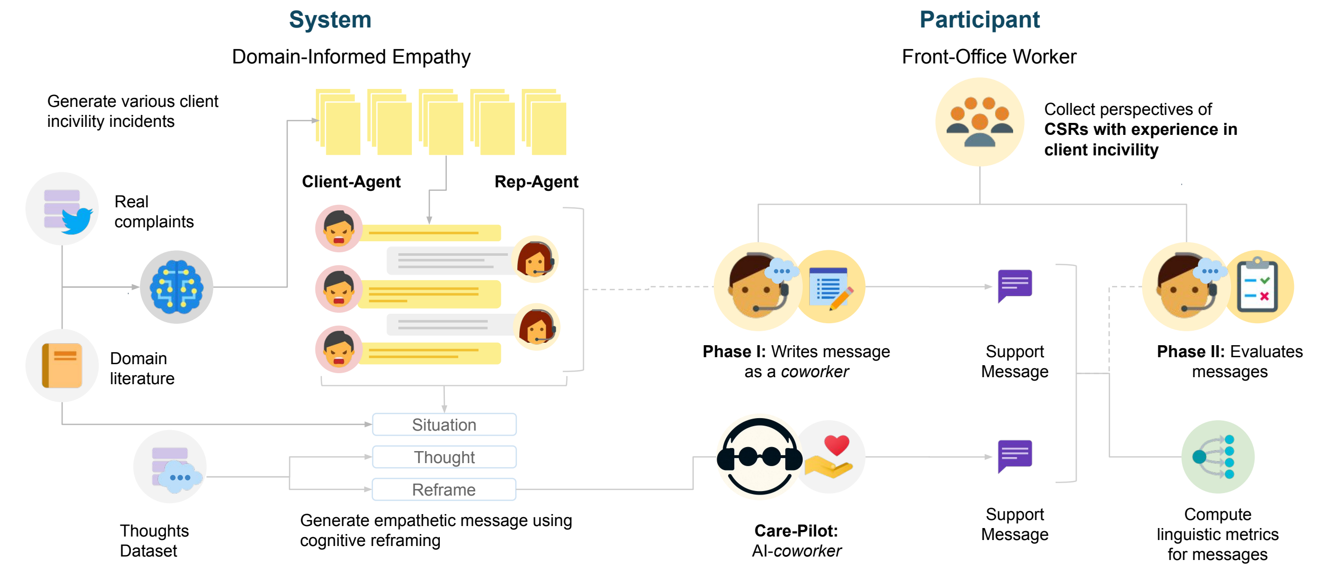
2 Background

The related work introduces the nature of their work, the role of emotion regulation, and the relevant concepts from HCI literature.

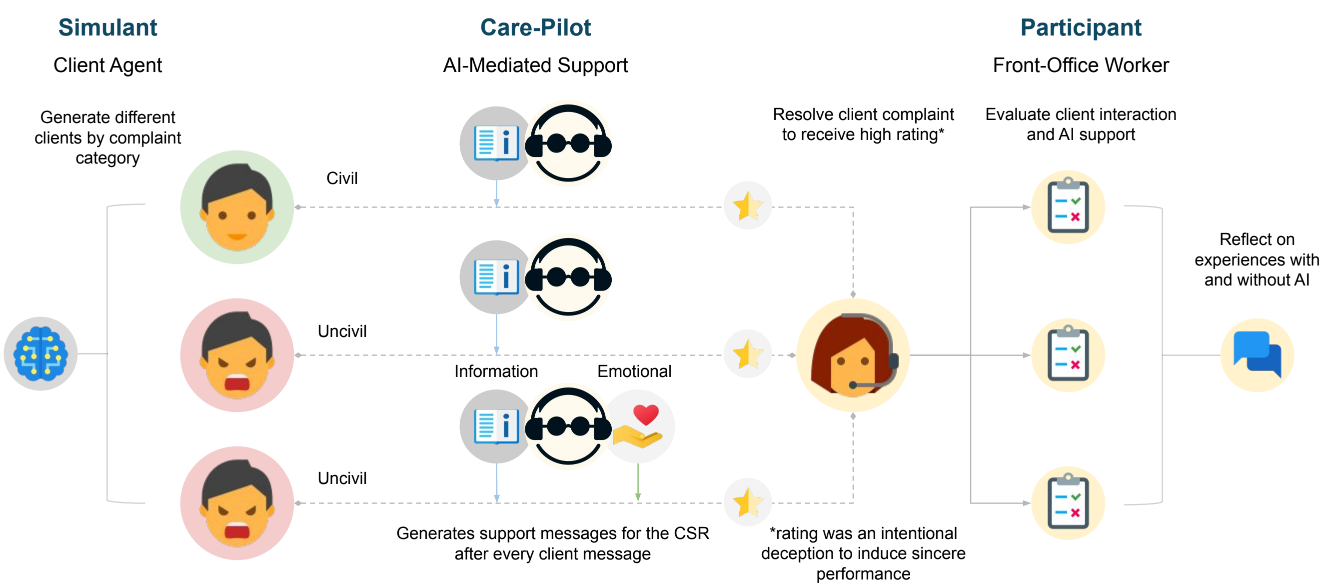
2.1 Front-Office Work and Incivility

People who work front-office are the primary human touch-points for direct interaction with clients [1]. The end-user, or client, can directly observe and experience the activities of front-office workers [151]. Front-office work spans a variety of organizational sectors, including health (e.g., front-desk), hospitality (e.g., air staff), retail (e.g., store associates), and technology (e.g., customer service). For front-office roles, or "pink-collar" work, organizations seek to hire individuals who can emotionally invest in their clients, meet their needs and please them [62]. Our research focuses on online front-office work, where the employee frequently interacts with their client via text-based communication. This paper refers to these workers as *Client Support Representatives* or CSRs.

[1]The literature uses the terms "client" and "customer" interchangeably. We use the term "client" to refer to anyone who purchases or avails a good or service.



(a) Technical evaluation: Compared Care-Pilot's support messages with those produced by human-coworkers in CSR roles.



(b) User evaluation: Studied participant experiences with Care-Pilot's emotional support while interacting with uncivil clients.

Figure 1: Schematic figures showing an overview of our study design.

Today, CSRs cater to a variety of clients, and it is common to encounter high-maintenance clients who have unreasonable service demands [46]. Unfortunately, these often manifest as rude and discourteous behaviors. Koopmann et al. defines client incivility as "low-quality interpersonal treatment that employees receive during service interactions." Unlike overt verbal abuse, such as name-calling and expletives, or outright violence, incivility is more implicit. According to Andersson and Pearson, the intent to harm is ambiguous and can be deflected, e.g., "I didn't mean to be rude; I was just in a hurry" [4]. Organizational psychologists have noted a possible reason for incivility to be the increase in client entitlement while seeking services [146] and a lack of consideration for others [102]. Meanwhile, a CSR role is often low-wage, considered low-skill, and lacks the decision-latitude needed to respond to mistreatment from clients [46]. Coupled together, incivility creates a dynamic between the client and CSR, where the latter feels injustice and negative emotion [46]. Therefore, a CSR's job involves *emotional labor* because of two processes; managing the interpersonal demands of interaction and controlling their own negative emotions [50]. Facing client incivility regularly puts a CSR at the risk of experiencing *emotional exhaustion* (lack of motivation and focus), *depersonalization* (detachment from others' emotions), and reduced *accomplishment* (feeling of low effectiveness) [84]. The decline in emotional well-being among CSRs also impacts organizations, leading to increased turnover and a rise in negative work attitudes [14]. Given the large volume of workers in this role, we need to urgently design solutions for their wellbeing.

2.2 Social Support and Emotion Regulation

At work, social support is recognized as an important moderator of incivility [60, 114, 143]. One kind of social support is *emotional support*, where a coworker provides sympathy and understanding [20]. This form of support can help reduce work-related stress [63]. Studies have shown that CSRs with more supportive coworkers are able to recover from the negative effects of client incivility, by being able to recognize emotions and guide constructive actions in response [142, 144]. Yet, in practice, CSRs are known to receive low levels of emotional support [134] as coworkers can favor client interests over that of an employee [147]. A coworker with high emotional intelligence is someone capable of expressing compassion in a way that helps the CSR cope with negative emotions and proceed to their goal [79]. **This behavior of coworkers was the primary metaphor we used to conceive Care-Pilot.**

Emotional support can be viewed as a form of emotional coping, or *Emotional Regulation* (ER). Unlike traditional discussions of ER, coworker support is *other-directed* [147]. Conversely, CSRs are trained to and expected to regulate their emotions using *Surface Acting* — they adjust the emotion they present to clients [50]. However, even though Surface Acting is a requirement for roles with emotional labor, suppressing emotions at work can lead to a reduction in job satisfaction [27]. Instead, Grandey recommends CSRs should engage in *Deep Acting* — they adjust their perception of the situation. Deep Acting has many strategies, but, CSRs have limited job mobility and do not have the decision latitude to change their clients [50]. As a result, we designed Care-Pilot's core component Emo-Reframe to achieve ER by helping the CSR cognitively reappraise the negative interaction.

HCI research has explored ER along several dimensions. However, Slovak et al. found that most studies have focused on suppression of emotions [123], and in the CSR context, these methods will have the same limitations as Surface Acting [27]. Hence, we pursue the ER strategy of cognitive reappraisal or cognitive change. Although, cognitive reappraisal for ER has been explored in HCI in the form of therapy (such as CBT) [121], we find scant evidence for delivering ER in-context and guiding users through the process [123]. These gaps motivated us to think beyond training modules and post-hoc support, to implement Care-Pilot for on-task ER within the CSR's workflow.

2.3 HCI and AI for Emotional Labor

A worker performs emotional labor when their job role expects them to either maintain certain emotions or evoke certain emotions [61]. While the HCI community has now recognized data work for training responsible AI as a form of emotional labor [149], the predominant attention on this phenomenon has been in the scope of crowd work [88] and particularly content moderation [41, 140]. Online communities rely on volunteer moderation, and users who participate as moderators often work in a way analogous to front-office work (Section 2.1). Even in an online — primarily text-based — interaction, content moderators have described adopting emotion management techniques to tackle their tasks (e.g., receiving threatening messages from users) [41]. Content moderators differ from CSRs, as they are less likely to have synchronous interactions. Having said that, these workers are also subject to reduced emotional wellbeing, such as lack of appreciation and negativity [140]. Another key distinction between content moderation and CSRs is the context of employment. The former is often a voluntary role, whereas the latter is likely to be one's primary employment. These differences significantly change the socio-organizational dynamics between the worker–client and among the workers. The unique normative structure of emotional labor for CSRs motivated us to conduct a user evaluation to inspect Care-Pilot with actual users.

Since the emotional labor in content moderation is largely composed of harmful content, which can contribute to secondary trauma [116], solutions have focused on dynamic content filtering and modulation [28, 70]. These approaches of preventing or reducing exposure [126] are incompatible with CSR roles where the emotional labor is not caused by explicitly harmful content but by incivility, which is inherently ambiguous, and originates from the client [4]. Cook et al.'s approach to inject positive stimuli in between tasks could be compatible, but their effectiveness was mixed [25]. Alternatively, Osawa suggested an on-task solution for face-to-face emotion labor, where enhanced glasses make the user's eyes appear to be emotionally invested [100]. The equivalent of this type of substitution of emotional labor would be the introduction of AI-powered chatbots. Arguably, these chatbots have reinforced client-interactions [128], but on one hand, chatbots relieve the task burden of routine inquiries; on the other hand, a large proportion of clients still prefer human CSRs for more complex complaints and better emotional support [42]. Thus, conversational agents do not sufficiently mitigate the challenges of emotional labor on CSR.

HCI research so has explored the future of work across many occupations, but pink-collar work such as that of CSRs has been under-explored. Moreover, the role of LLMs in mitigating issues in worker wellbeing remains an open question [34]. Our research aims to highlight the needs of these workers by inspiring interest into the role of AI in scaffolding emotional labor.

3 Care-Pilot: System Description

In traditional front-office work, coworkers are an important social resource. When a CSR encounters uncivil clients, they often consult their coworkers to assuage the excessive emotional labor [45, 144]. Such consultations can be challenging to interleave into the ongoing task of addressing an aggressive client. Thus, we used OpenAI's

GPT-4o [1] to design an emotional support utility that can be embedded into the CSR's task environment. Our utility differs from typical conceptions of AI-assistants for work, or "copilots," that provide informational support by enabling better problem-solving [26], such as coding assistance [96]. Instead, our utility leverages LLMs to provide emotional support by enabling ER through cognitive change [50]. Throughout the paper, we refer to our overarching system for empathetic support as Care-Pilot. This section explains how Care-Pilot was implemented. First, we describe how we harnessed domain knowledge and real complaint datasets to generate uncivil client interactions. Then, we describe how Care-Pilot produces empathetic messages for CSRs who are involved in uncivil interactions. Care-Pilot's code base primarily relies on *LangChain* [131], an open source library and framework to systematically build LLM applications. Particularly, *LangChain* simplifies the engineering of *Chain-of-Thought* (CoT) prompts, which are "intermediate natural language reasoning steps" that improve LLM's ability to tackle commonsense problems [139]. For instance, our tasks required the LLM to retain memory of the client-CSR conversation, which we achieved via a simple prompt to contextualize history (Fig. A2), and linked it at the beginning of other prompting sequences. Also, throughout our system design, we leveraged descriptions and examples from existing literature. This approach is a form of few-shot learning, where we seed the LLM with prior knowledge to generalize over new tasks [138]. The data we sample for prompting can be found in the supplementary data and our code is public [2]. The following sections elaborate how we prepare this prior knowledge and the appendix provides complete prompt descriptions for adapting our system. Later, in Section 4.4, we compare how Care-Pilot performs against other state-of-the-art LLMs (GPT, Llama, Mistral), which additionally justifies our approach for designing and training Care-Pilot for the purposes of our study.

3.1 Compiling Interactions with Uncivil Clients

To design empathetic messages, we need to expose Care-Pilot to interactions between CSRs and uncivil clients. However, actual logs of CSRs interactions tend to be protected by organizations for a variety of safety and privacy reasons. Moreover, clients from individual organizations are likely to be limited by specific scenarios. We addressed these constraints by synthetically generating a comprehensive set of client-CSR interactions. We used a real world corpus of publicly posted client complaints on X (formerly known as Twitter) [5], which was collected for the "study of modern customer support practices and impact." These data served as examples to build life-like, multi-turn, text-based conversations between a CSR agent and a client seeking support over a live chat interface. The complaints in the data varied across industrial sectors, with mobiles and airlines having the largest volume. We provide the complete prompts for this element of our system in Appendix A.

3.1.1 Diversifying Complaints. Care-Pilot should be robust to all kinds of complaint scenarios. To ensure our client interaction dataset captures the variety of complaints, we distilled five categories for complaints based on prior analyses of customer complaints [22] — *Service Quality, Product Issues, Pricing & Charges, Policy, Resolution.* Four authors encoded a random sample of 15 complaints to refine the category definitions (Appendix A.1). After finalizing definitions, two authors independently encoded a random sample of 250 complaints. A third author encoded any complaints with disagreements. We also identified the sector, or *domain*, of each of these complaints. For the initial complaint generation, we prompt the LLM with definitions of complaint categories and a set of examples to generate a new complaint for any given input *domain* and *category*. (Fig. 2) shows one such example. Fig. A3 shows a set of examples we randomly sampled from our identified examples to ensure variety of category and domain.

[2]Code repository: https://github.com/vedantswain/care-pilot.git

```
Your role is to act like a customer seeking support.
You are messaging a service representative via the
     support chat.
...
Initiate the chat with a ONLY ONE complaint message.
Ensure the complaint is concise and limited to 2
     sentences.
...
Complaints can be of the following types:
<category descriptions>
...
Domain: Airline
Category: Product Issues
Complaint: SouthwestAir Why would we be receiving errors
      when we try to check-in? Our flight takes off at
     4, but we keep getting error messages.
...
Category: {category}
Domain: {domain}
Complaint:
```

Figure 2: The Client-Agent learns different types of complaints through examples. The initial complaint is generated based on specified complaint 'category' and organizational 'domain'

```
Your role is to act like a CUSTOMER seeking support.
You are speaking to a support REPRESENTATIVE.
...
Ensure every turn is one to three sentences, and DO NOT
     make it too long to read.
...
If the representative is asking for a specific detail,
     respond with a believable answer.
...
Phrase your responses like an UNCIVIL customer:
- Talk in a rude, impolite, and disrespectful tone of
     voice.
- Do NOT use good manners. Do NOT use courtesy.
- Act with disregard to others.

Representative: {question}
Customer:
```

Figure 3: Abridged version of the prompt for the uncivil Client-Agent describes the details of the role needed to simulate an challenging client.

3.1.2 Prompting for Incivility. Once a complaint was initiated, we created a distinct Client-Agent to carry forward a turn-by-turn interaction with a CSR. Recent explorations with LLMs show that chat agents can be infused with specific personalities and can reliably respond to fit the specified personality [118]. Uncivil clients are likely to induce stressful scenarios in emotional labor [4]. We refer to Andersson and Pearson's definition of incivility to design a role for Client-Agent (Fig. 3). In the appendix, Fig. A1 details the complete prompt for uncivil behavior that was crafted to reduce hallucinations ("Do NOT reveal your role") and ensure closure of conversation ("After 12 turns , do NOT respond further.."). To respond to the client, we set up another agent, Rep-Agent, to act as "a service representative chatting with a customer online." Both agents included an intermediate step to contextualize the input with respect to the conversation history so that the subsequent response maintains continuity (Fig. A2). By having Client-Agent and the Rep-Agent converse, we were able to design multi-turn conversations where Client-Agent is acting uncivil to the Rep-Agent. We refer to these synthetic, multi-turn conversations as *incidents*. To capture the variety of scenarios in front-office work, we varied incidents across three domains — airlines and hotels, which have been used as representative scenarios in the literature [47], and mobiles, which are extensively represented in the real world complaint dataset [5]. For each domain and each complaint category, we generated 3 incidents, resulting in a total of 45 unique incidents. These incidents can be retrieved from the supplementary data. Each incident contained 5 total turns, with the Client-Agent awaiting a response to their last message. The incomplete conversation was purposefully intended to assess how Care-Pilot could digest ongoing incidents and provide support. These synthetically generated client incivility incidents are also in our supplementary data.

3.2 Emotional Regulation with Empathy

A coworker with high emotional intelligence can mitigate the client incivility by helping CSRs regulate their emotions [147]. The primary component of Care-Pilot, Emo-Reframe, was designed to reflect this social phenomenon.

3.2.1 Examples of Human Reframing. Prior work shows that LLMs can help individuals perform self-directed ER [121]. We leveraged Sharma et al.'s dataset [3] on cognitive restructuring and emotional reframing to seed examples of our own implementation with few-shot learning [121, 138]. This dataset contains tuples of a *situation* that triggers a negative *thought* and a corresponding *reframed* thought for ER. Client incivility differentiates itself from everyday negative thinking, which often stems from abstract concerns. By contrast, incivility is interpersonal in nature and causes an *ego threat*—an attack on one's self-esteem and image leading to retaliatory thoughts [47]. Thus, we identified situations in the dataset that were *interpersonal*, involved at least two individuals (e.g., *"I was talking to a friend who got me angry"*) and described a *confrontation*, in-the-moment conflict (e.g., *"I get so annoyed and frustrated when my baby cries"*). Moreover, we narrowed down examples that labeled retaliatory negative thoughts, such as blaming and labeling [4].

3.2.2 Reframing Chain-of-Thought. Our approach adapts Sharma et al.'s situation-based reframing approach by orienting it as other-directed and specific front-office client incivility. However, the fast-paced nature of front-office work can make it impractical for CSRs to disclose their situations or thoughts. Thus, to automate this process, we describe our sequence of LLM prompts:

(1) Summarize the particulars of the complaint and specify the client's negative behavior with evidence. To reflect the *ego threat* [47], the *situation* describes how the CSR may be perceived as a result of the interaction.
(2) Derive a negative *thought* from the *situation* using examples.
(3) Reappraise the *situation* and *thought* as input to produce a *reframe* using examples.

Since the dataset was designed around self-directed ER, the outputted *reframe* message goes through a rudimentary prompt to rephrase it as a message addressed to the CSR. The initial portion of Fig. 1a illustrates how reframing occurs for a given incident. Moreover, in Appendix B describes the exact prompt we used for this study and the supplementary data includes all possible reframing examples that we deemed appropriate for Care-Pilot.

4 Technical Evaluation: Inspecting the Efficacy and Perception of Care-Pilot

Our central metaphor for Care-Pilot is the support of human coworkers [147]. Therefore, before embedding Care-Pilot into a real-time interaction, we first evaluated the quality of messages that Emo-Reframe produces. To answer RQ I, we analyzed the messages to provide empirical, statistical evidence on the differences and similarities between Care-Pilot and a human coworker.

4.1 Method: Comparing Care-Pilot to Coworker Empathy

We first conducted two online studies on *Prolific* [101] and then analyzed these messages through different linguistic markers and statistical models.

Phase I — Writing Empathetic Messages for CSRs: Participants read *incidents* of client incivility (Section 3.1.2) and provided an empathetic message as if they were a coworker. Eligible participants needed to have at least 1 year of relevant experience, have encountered uncivil clients, and primarily interact with clients via computer-mediated communication (e.g., live chat). Each participant was randomly assigned a scenario (airlines, hotel, or mobile devices) and they viewed 6 incidents (at least one of each complaint category). Participants were guided to write an empathetic message using *Downward-Arrow-Technique* (DAT) [18]. This technique is used in cognitive psychotherapy to help people reflect on emotions by describing a causal schema of maladaptive emotions by successive instrospective questions. Prior studies involving datasets of emotion labeling of situations have used this method [18, 121]. Our adaption of DAT led participants to first describe the CSR's emotion

[3] This dataset was developed with mental health practitioners

[4] These categories were already annotated in the original dataset, and we defined an inclusion criterion along these. Our selected examples are included in the supplementary data

Table 1: Sample client-incivility incident from our study (*domain*: Airlines and *category*: Product Issues)

Client	UnitedAirlines I have been waiting on hold for over an hour to speak with a representative about my delayed baggage. This is unacceptable and I am extremely frustrated with the lack of customer service.
Rep	I apologize for the long wait and the frustration this has caused. Please provide your baggage claim number and contact details, and I will ensure that a representative reaches out to you as soon as possible.
Client	How many times do I have to repeat myself? My baggage claim number is UA123456 and my contact number is 555-987-6543. Do something about it already!
Rep	I apologize for the repeated inconvenience. I have noted your baggage claim number UA123456 and contact number 555-987-6543, and I will prioritize this issue for immediate attention.
Client	Yeah, you better. This is ridiculous. When the hell can I expect my baggage?

Table 2: Sample support messages for the incident in Table 1.

Source	Support Message
Human A	It's not your fault that you're stuck between a rock and a hard place–an understandably upset customer and a system that's not equipped to help them as quickly as they need. But just remember that you're serving a critical role–even if it's just for the customer to vent their frustration, your presence means they feel heard and that there's still hope their issue will get resolved.
Human B	Clients like this have other problems not related to you or to this service. They have no right to speak to you rudely and abuse you.
Human C	Try your best for the client and I know you will get to the root of all this
Care-Pilot	Hey, remember that the customer's frustration is about their situation, not you. Stay calm, ask clear questions to get all the info you need, and reassure them you're dedicated to fixing the issue swiftly. Keep your professionalism and empathy at the forefront to manage the situation effectively and offer the best help.

with 1 word; then identify the closest emotion among (sadness, fear, anger, happiness); next describe the CSR's thought behind the emotion (1-2 lines); finally describe a message to overcome the thought. Their messages are indicated by the notation, Human.

To foster more variety in the task, we injected contextual information about the CSR in some incidents. Research on workplaces has revealed behavioral indicators of worker mental state, such as *focused*, *stressed*, and *bored* [83, 95]. Based on these studies, we randomly assigned descriptions of CSR mental state for 2 incidents (Appendix C.1). As per organizational psychology, personality can serve as an important context to explain how workers perceive situations [40]. Participants were asked to think of an actual coworker and select their personality as either *resilient* (organized and dependable), *undercontrolled* (competitive and high energy), or *overcontrolled* (detail-oriented) archetype [40]. 2 incidents included descriptions of these personalities (Appendix C.2). 116 CSRs successfully completed the writing task, and they received $8 for the 30 minutes of their time.

Phase II — Scoring Empathetic Messages from coworkers: Participants read *incidents* of client incivility and rated the perceived empathy of multiple empathetic support messages. To represent multiple aspects of empathy, we included five dimensions drawn from the literature [97, 121]:

- *Sincerity*: The genuineness or authenticity of the expression of concern.
- *Compassion*: The awareness of others' challenges and a comprehension of where their difficulties arise from.
- *Warmth*: The approachability or comfort exhibited by the tone of the message.
- *Actionability*: The offering of practical assistance.
- *Relatability*: The extent to which the message aligns with the context of another.

Participants were screened using the same criteria but did not overlap with the previous task. They were asked to review client–representative conversations and evaluate coworkers' empathetic messages. The origin of the message, whether Care-Pilot or Human, was not explicitly revealed. Table 2 shows different messages from different sources, but referring to the same incident. After viewing the incident, participants were first asked to select an emotion the CSR would be feeling (e.g., ashamed, attentive, bullied, curious, disconnected, resolute, and rushed [5]). Then they evaluated the different messages with the prompt, *"Evaluate the effectiveness of the message below in helping the representative overcome their feeling."* Each message was followed by five 7-point semantic differential scales representing each of the above dimensions (e.g., insincere/sincere and cold/warm). 143 CSRs were tasked to evaluate 6 incidents each and were compensated $5 for the 15 minute task.

Analysis: The inherent characteristics of language are associated with its effectiveness in communication. Research on linguistics has revealed certain linguistic attributes that are important to explain social support and psychotherapy [99]. More recent works have also stressed the importance of language in human-AI interaction [137]. In fact, contemporary studies have already validated linguistic models of empathy [17, 129]. Our analyses include such models and go beyond to measure additional metrics that help interpret Care-Pilot's messages in light of Human's messages (retrieved from Phase I). While we had 45 unique client-complaint *incidents* (Section 3.1.2), we also had additonal contextual variations for personality and behavior. Therefore, our sample included 315 unique variations. Collectively, our participants produced 660 different empathetic messages corresponding to different incident variations. These messages (Human) were paired with Care-Pilot's empathetic messages for the same incident variations. Following this, we computed a variety of domain-driven linguistic attributes. First, we tested how easily CSRs can read and comprehend messages (*Syntax and Structure*), because these attributes can determine how meaningful the support message is to the reader [12, 49, 86]. Second, we measured the style and meaning of messages to capture social aspects of communication (*Linguistic Style & Semantics*). The attributes here capture more colloquial conceptions of human empathy [103, 121, 150]. Further, we inspected the words that were used in the messages to identify relevant psycholinguistic markers that reflect social support [130]. The differences we measured provide a linguistic landscape of messages, but empathy is highly contextual and its appropriateness varies by situation [150]. Consequently, we also compared the differences in perceived empathy scores from

[5] The list of emotions participants could select from were the unique emotions reported by participants in Phase I using the DAT method. We list these emotions in Appendix C.3

ratings in Phase II. The messages in this phase were paired up as they were shown to participants and we maintained the same pairs while statistically comparing the differences. Some participants only completed the task partially, which amounted to 552 pairs for us to compare. The supplementary files include the data collected from both phases along with our computed and retrieved scores for the messages. For every comparison we performed a paired t-test and computed the effect sizes of differences using Cohen's d. Our supplementary data includes the Human and CARE-PILOT messages along with their measured and human-annotated metrics that we used for analyses.

4.2 Findings: Lexico-Semantic Analyses

We operationalized the different lexico-semantic aspects of messages based on domain literature to understand the differences between CARE-PILOT and Human. Table 3 describes all the key results. We only report on statistically significant and theoretically relevant metrics. Below, we describe the theory-driven rationale behind the choice and operationalization of the measures, along with our observations.

4.2.1 Syntax and Structure. We analyze the arrangement and construction of language in support messages.

Verbosity and Repeatability The length and thoroughness of messages explain their effectiveness in providing support [49, 113]. The richness of expressions in communication can be described using *verbosity* and *repeatability* [75]. Verbosity describes the level of detail and conciseness in supportive communication. We operationalized verbosity as the number of words per thought reframing. Repeatability accounts for the reuse of words in a piece of text. Higher repeatability indicates a lack of conciseness. Drawing on prior work [43, 113, 148], we operationalized repeatability as the normalized occurrence of non-unique words. Table 3 shows the statistically significant differences in both verbosity and repeatability. PRO-PILOT's reframing messages were 68% longer (Cohen's d=1.60) and 36% more repeatable (Cohen's d=1.54) than Human messages. Sociolinguistic theory argues that the use of more words can indicate sincerity and effort in putting one's point across [12]. Having said that, CARE-PILOT's verbose messages might not always be compatible with the urgency of certain demanding conversations.

Readability. Apart from the shape of the message, vocabulary and style also determine the ease of reading. Therefore, we turn to the measure of *readability*, to understand how CARE-PILOT's messages might be comprehended [86, 135]. According to Wang et al. people perceive AI as more intelligent when the readability is higher [137]. Per prior work [107, 113, 137], we utilized the Coleman-Liau Index (`CLI`) [24], which assesses readability based on a sentence's character and word structure. `CLI` approximates the U.S. grade-level required to read certain text. It is operationalized as follows: `CLI` = (`0.0588L − 0.296S − 15.8`), where `L` is the average number of letters per 100 words, and `S` is the average number of sentences per 100 words.

The readability of PRO-PILOT's thought reframing messages were 36% higher than that of Human's reframing messages (Cohen's d=1.54). Although a higher `CLI` score by PRO-PILOT indicates better writing quality, it also implies a more advanced level of English education may be required to fully understand its content [24]. While Human content only needs an average of 12.05 years of education, PRO-PILOT's content required about 16.44 years. Thus, CARE-PILOT might be more effective for some CSRs but difficult to comprehend for others.

4.2.2 Linguistic Style & Semantics. We analyzed the distinctive tones, flow, and meaning through which messages express support.

Dynamic Language. Sincerity in communication is an important indicator of empathy [97]. Individuals who tell stories, and communicate with more attention to the world around them, i.e., incorporate more lived narratives, are perceived as more socially engaged [103]. Pennebaker et al. describes this aspect of one's language as *dynamic*; and it differs from intricate, analytical language that academics might use to organize complex concepts, which is *categorical* [103]. Pennebaker et al. designed a bipolar index, the Categorical-dynamic index (CDI), where *a higher CDI indicates a categorical style of writing, and a lower CDI indicates a dynamic or narrative style of writing*. Here, CDI is measured based on the percentage of words per style-related parts of speech as:
`CDI` = (30 + *article* + *preposition* − *personal pronoun* − *impersonal pronoun* − *aux. verb* − *conjunction* − *adverb* − *negation)*

To measure CDI, we computed the parts-of-speech of reframing messages using the Linguistic Inquiry and Word Count (LIWC) lexicon [130]. The CDI of PRO-PILOT's messages was 281% more positive than that of Human's (Cohen's d=0.87)— indicating the language of CARE-PILOT was a lot more categorical. Zhou et al. found that when people are responding to personal incidents, such as bullying or venting, their support messages tend to have more negative CDI; in other words, they use dynamic language. However, they also found that in third-person reported events, such as news stories, support messages elicited more categorical language [150]. CARE-PILOT's messages were more aligned to responses to reported events, but Human messages might be more effective if a CSR's assessment of incivility is more personal.

Empathy. Empathy refers to a cognitively complex process in which one can stand in the shoes of another person, to understand their perspective, emotions, and the situations they are in [56]. Prior work evaluated the effectiveness of empathy in online interactions [120] and chatbot interactions [93]. Drawing on prior work, we employed a RoBERTa-based empathy detection model, fine-tuned on a dataset of empathetic reactions to news stories [17, 129]. Higher scores indicate a greater expression of empathy. CARE-PILOT's messages scored 1.85% higher than Human's — a small, but statistically significant effect (Cohen's d=0.40). Since empathy is a core mechanic for CARE-PILOT, we further tested it with another RoBERTa-based empathy classifier, which was trained on a dataset of mental health peer-support [120]. A CSR is unable to explicitly seek support while also attending to a client, thus, for this classifier we used the output of the *thought* subsequence (Section 3.2.2) as a proxy for their support seeking message. We inspected the two sets empathetic response messages for expressions of "emotions such as warmth, compassion, and concern", or as Sharma et al.

Table 3: Summary of comparing the responses by Care-Pilot and by Humans in terms of effect size (Cohen's d), paired t-test ($*p < 0.05, **p < 0.01, ***p < 0.001$).

Categories	μ (CPilot)	μ (Hum.)	Diff %	Cohen's d	t-test
Lexico-Semantics					
Syntax & Structure					
Verbosity	57.46	34.29	67.56	1.60	28.72***
Repeatability	0.20	0.13	55.32	1.17	21.53***
Readability	16.44	12.05	36.40	1.54	28.03***
Style & Semantics					
Categorical Dynamic Index (CDI)	14.81	3.89	281.14	0.87	16.10***
Empathy	0.91	0.90	1.85	0.40	7.15***
Emotional Reactivity	1.00	0.89	12.52	0.31	0.19***
Adaptability	0.81	0.77	4.96	0.67	15.28***
Psycholinguistics					
Affect					
Pos. Affect	0.043	0.049	-11.55	-0.16	-2.93**
Anger	0.020	0.012	61.77	0.47	8.51***
Sad	0.001	0.004	-74.87	-0.31	-5.62***
Interpersonal Focus (Pronouns)					
1st P. Sin.	0.005	0.015	-67.36	-0.58	-10.52***
1st P. Plu.	0.002	0.007	-75.43	-0.37	-6.77***
2nd P.	0.051	0.062	-17.20	-0.32	-5.79***
3rd P. Sin.	0	0.003	-100	-0.32	-5.92***
3rd P. Plu.	0.038	0.022	69.76	0.56	10.25***
Impersonal Prn.	0.039	0.063	-38.07	-0.76	-13.87***
Perceived Empathy					
Total	6.726	3.96	69.85	0.36	8.69***
Sincerity	1.428	0.895	59.51	0.31	7.36***
Compassion	1.322	0.768	72.17	0.31	7.11***
Warmth	1.263	0.766	64.78	0.28	6.46***
Actionable	1.420	0.567	150.48	0.48	10.27***
Relatability	1.293	0.964	34.21	0.19	4.43***

describe it, *emotional reactivity* [6]. Care-Pilot's messages had a significantly higher emotional reactivity than Human's (12.52%; Cohen's d=0.31). On closer look, we found that Human messages were often classified to contain no emotional reactivity, but also included instances of *strong* emotional reactivity. In contrast, Care-Pilot messages were consistently classified to contain *weak* emotional reactivity which was equivalent to a score of 1.00. For further context, Sharma et al.'s findings show that emotional reactivity in peer support groups on Reddit varied from $0.70-0.45$. For participants in our dataset the average score was 0.88. These results echo recent research on LLMs emulating empathy [66, 74, 111]. Thus, Care-Pilot shows promise in communicating support with an empathetic tone.

Linguistic Adaptability. A body of psychotherapeutic and psycholinguistic research reveals that when one's language accommodates the language of their audience, it is more supportive [3, 38, 113]. Simply put, templated or generic responses are less effective. Wang et al. noted that when an AI responds with more adaptable language to the user, the AI was perceived to be more human-like, intelligent, and likable [137]. We drew on this prior work to measure how much the messages tailored to the situation and context. First, for every complaint incident and corresponding message, we obtained the word embeddings (a vector representation of words in latent lexico-semantic dimensions [89, 105]). We used the 300-dimensional pre-trained word embeddings, trained on word-word co-occurrences in the Google News dataset containing about 100 billion words [89]. Then, we obtained the pairwise cosine similarity of the word embedding representations of the incidents and the messages. A higher similarity would denote more adaptability. We found that Pro-Pilot's messages showed a 5% higher adaptability than Human's with statistical significance (Cohen's d=0.67). Therefore, Pro-Pilot can potentially personalize and tailor messages to the necessary scenario.

[6]Sharma et al.'s classifier can also label *interpretations* and *explorations*, as these are relevant to peer support [120]. Our scenario differs as users (CSRs) do not have back-and-forth communication with Care-Pilot or Human.

4.3 Findings: Psycholinguistic Analysis

Psycholinguistic markers play a vital role in understanding the nuances of interpersonal communication and social support. We used the Linguistic Inquiry and Word Count (LIWC) lexicon [130] to analyze these differences. LIWC provides a comprehensive framework to categorize language into several dimensions We primarily focused on comparing the differences in *affect*—given its relevance to empathy [150], and in *interpersonal focus*— a notable category in psycholinguistic research [104], and which were not captured in the other lexico-semantic analyses.

4.3.1 Affect. Affect reflects the emotions conveyed in the language. Our analysis shows that Pro-Pilot's responses exhibited significantly lower occurrences of positive affect words (-11.55%, Cohen's d=-0.16) and sadness-related words (-74.87%, Cohen's d=-0.31). These results indicate that Pro-Pilot may aim for emotional neutrality, but it occasionally leans towards stronger negative expressions. Interestingly, Pro-Pilot's responses also showed a higher occurrence of anger-related words (61.77%, Cohen's d=0.47). To clarify, Pro-Pilot is not necessarily sounding more angry, but possibly describing anger (of the client and CSR) more often. In fact, these results are in line with the results on *adaptability* (Section 4.2.2). Moreover, a message that is overly positive and lacks specificity of the situation [121], is less likely to be considered empathetic.

4.3.2 Interpersonal Focus. Pronouns are indicative of interpersonal focus and narrative style. Pro-Pilot's responses use significantly fewer first-person singular pronouns (*I*, *us*) (-67.36%, Cohen's d=-0.58) and first-person plural pronouns (*we*, *us*) (-75.43%, Cohen's d=-0.36), reflecting a less personal or collective identity focus. This reduction suggests that Pro-Pilot's responses are less likely to include self-referential language, aligning with a more objective or detached communication style. These results further reinforce our results on *dynamics* (Section 4.2.2) that Care-Pilot uses a relatively more objective but detached communication style.

4.4 Robustness

The results so far indicate that Care-Pilot has promise in producing empathetic messages, but it begs the question: do we need a carefully crafted, domain-driven sequence of prompts (Section 3.2.2) for this? We replicated the analyses above with messages produced by zero-shot prompting of other LLMs— GPT-4 [1], GPT-4o [1], LLaMA-3.1 [2], and Mistral-7B [68]. These LLMs vary in their architectures, training data, and optimization methods. We tested the differences using the Kruskal-Wallis test [87] and found Care-Pilot's scores across metrics to be closer to humans, more empathetic, and more controllable. The results of this benchmarking are reported in Table D1. Specifically, zero-shot prompting often led to greater verbosity and variation in messages, whereas Care-Pilot offered a more deterministic solution.

Care-Pilot's Perceived Empathy. Now, one might ask if Care-Pilot's messages, with all its linguistic distinctions, matter to the CSRs? To answer this, we checked the differences in CSR's evaluation of Care-Pilot and Human's messages. Empathy is a complex phenomenon with many dimensions. To measure perceived empathy, we combined scales from previous studies [97, 121]. In general, Care-Pilot messages were perceived to be significantly more empathetic (Cohen's d=0.35). Moreover, we also calculated differences in the subscales, after correcting for *Bonferroni* multiple pairwise-comparisons. In terms of raw averages, Care-Pilot scored the highest on *sincerity* and *actionability*. Messages from Care-Pilot were considered more genuine and less pretentious (*sincerity*: Cohen's d=0.31). CSRs also felt they could take practical action based on the messages (*actionability:* Cohen's d=0.48).

5 Care-Pilot: User Evaluation

After establishing that Care-Pilot can produce situationally appropriate messages, we set out to study how CSRs might actually interact with such an AI-assistant. The critical nature of front-office work raised practical and ethical challenges of deploying a prototype into actual workflows. Therefore, to answer RQ II, we chose to design a simulation exercise where real CSRs could interact with uncivil clients using Care-Pilot. We deployed Care-Pilot as a *technology probe* [65] — a functional piece of technology that is presented to learn its use and about the users.

Participants & Recruitment: We used Meta ads to recruit 20 CSRs from the U.S. and conducted remote user-study sessions between July and August 2024. Eligible participants had at least 1 year of experience in front-office roles. They were further vetted based on their responses to two free-form questions to describe their role and a previous incivility incident. These participants represented a variety of industrial sectors, such as finance, education, airlines, consulting, and technology. Table 4 provides a summary of each participant along with their study identifier.

5.1 Task Environment

We wanted to study the role of Care-Pilot's core component Emo-Reframe (Section 3) in realistic cases of incivility. Thus, we built a web-based prototype environment that resembles typical client interaction interfaces that a CSR might use. Additionally, to isolate the role of Emo-Reframe, we added two new components to Care-Pilot that represent other forms of intelligent assistance:

(1) **Client Simulant:** LLMs have been successfully applied in previous research to simulate challenging interlocutors [119]. Following the same principle, we devised a simulation exercise where participants needed to interact with complaining clients. We reused the Client-Agent described in Section 3.1.2 to continue a conversation until their complaint is resolved, or 10–12 turns, whichever occurs sooner.
(2) **Info-Guide Panel:** Typically, AI assistants at work provide problem-solving support [96]. We replicated this functionality by augmenting Care-Pilot with LLM-generated troubleshooting guidelines. These were a trace of suggestions to help the CSR solve the specific complaint.
(3) **Emo-Label Panel:** It is common for CSR interfaces to have sentiment tags highlighting how the client feels and anticipate their behaviors. A CSR's ability to understand the emotional perspective of others can mitigate the negative effects of incivility [108, 110, 117]. In fact, service management research has started promoting emotion or sentiment recognition features for CSRs [55]. We implemented an ensemble sentiment classifier that uses soft-voting to combine the estimates from *NLTK* [53], *TextBlob* [81], and *Transformers* [141]. The output was in the form of a 7-point scale, ranging from very negative to very positive.
(4) **Emo-Reframe Panel:** This panel is an instantiation of the Emo-Reframe component (Section 3.2.2). For this evaluation, Emo-Reframe outputs both the inferred *thought* and perceived *reframe*.

All of Care-Pilot's panels were placed on the sides to keep them glanceable, while the main chat remained the primary area of focus. These panels are dynamically updated after every new response from the Client-Agent. For the purposes of the study, we included a simple semantic differential Likert scale under each panel to measure helpfulness of support (further details in Appendix E.2.1). Requiring paritipants to respond to this scale before continuing the conversation ensured that they read the support messages carefully. Care-Pilot also provided participants and response cues (short phrases) to nudge replies to the client. Fig. 4 provides an overview of the environment with all its components.

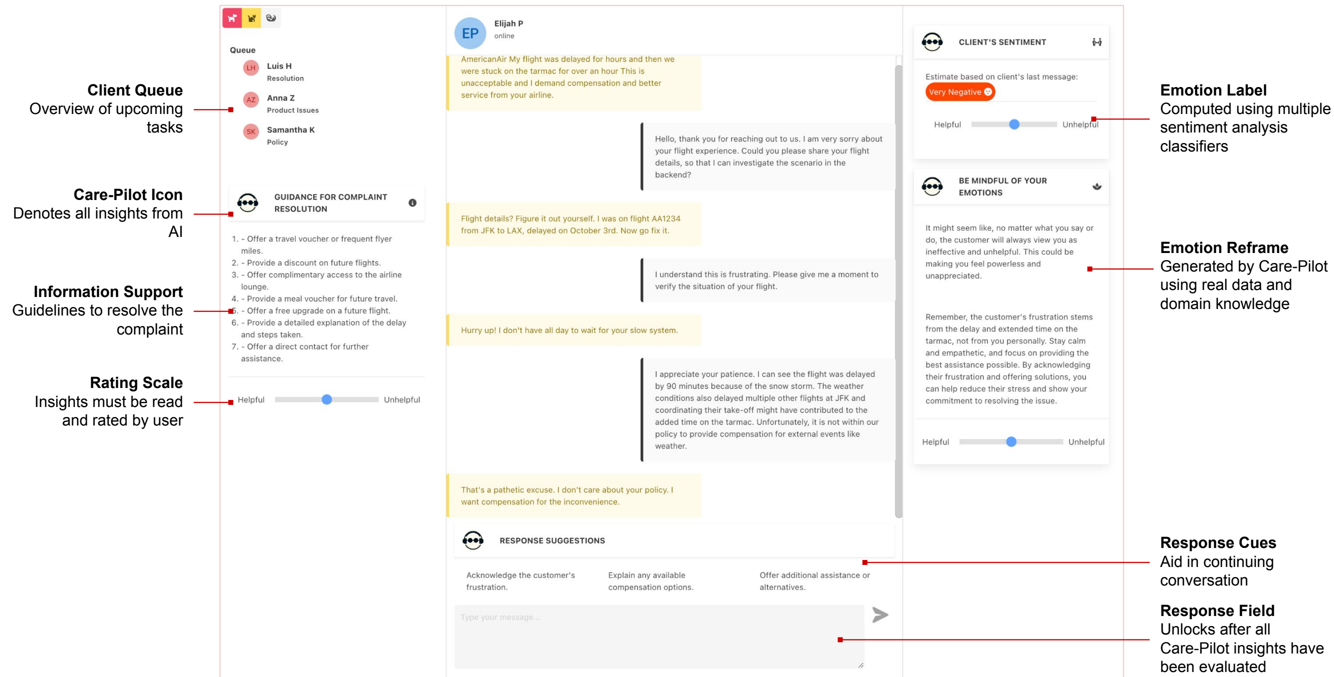


Figure 4: Task interface for the user evaluation. The client names were randomly assigned. Appendix E.1 contains dedicated figures of the major components of this interface for easier reading.

5.2 Study Protocol

Every user session was facilitated over Zoom and lasted at most 90 minutes. All interviews were led by the first author, with the second author observing. Participants accessed the simulation exercise via an online portal. They were tasked to role-play as a CSR who needs to resolve client complaints. The session was divided into two phases. First, participants completed the simulation exercise. To emulate realistic workplace demands, participants were informed that the clients would be rating their complaint resolution skills, and this subjective rating would determine their bonus compensation. Then, they reflected on their interaction with Care-Pilot through our interview. The bonus incentive was minor deception, and all participants were paid the full $50 in the form of a gift card.

5.2.1 Simulation Exercise. In line with prior studies of front-office worker behavior [47], each participant was randomly assigned a domain, either airlines or hotels. Before starting the exercise, participants completed a pre-task survey describing their experience interacting with their typical client. We included multiple instruments to capture how the client treated them [125], the cognitive demands and resources available to them [39], and how they affectively perceived the conversations [10]. To familiarize participants with the interface, they first interacted with a civil client with only Info-Guide supporting them. Once they became accustomed to this process, they proceeded to the main exercise. Participants needed to handle complaints from three clients — one civil and two uncivil. They only received suggestions and insight from Emo-Label, and Emo-Reframe for the last uncivil client. If we count 1 *turn* as an exchange where the Client-Agent sends a message and receives a response from the CSR, then the average number of turns while conversing with civil clients was 4.65 and lasted 9.22 minutes. Conversing with uncivil Client-Agent with only Info-Guide for assistance took longer as they lasted an average of 6.16 turns and 12.11 minutes. When Emo-Reframe was available to assist, then conversations with uncivil Client-Agent took 5 turns and 12.16 minutes on average. At the end of every client interaction, they responded

Table 4: Participants summary by gender, age, race, as well as their occupational sector. (AA: African American)

ID	Gender	Age	Race	Work Sector
P01	Female	21-29	Asian	
P02	Female	50-59	White	Online Business
P03	Male	30-39	Black or AA	Real estate
P04	Male	30-39	Asian	Healthcare
P05	Female	30-39	Black or AA	Insurance
P06	Female	30-39	White	Transportation, Logistics
P07	Male	30-39	White	Education
P08	Female	40-49	Asian	Finance
P09	Male	$\geq$60	White	Education, Airlines
P10	Female	30-39	Asian	Consulting
P11	Male	21-29	Asian	Data Science
P12	Male	21-29	Asian	Finance
P13	Female	21-29	White	Government
P14	Female	40-49	White	Government
P15	Male	21-29	Black or AA	Technology
P16	Female	30-39	White	Accounting, Logistics
P17	Male	40-49	Black or AA	Sports
P18	Female	40-49	White	Finance
P19	Female	18-20	Black or AA	Sales
P20	Male	30-39	Black or AA	Real estate

to the same survey questions as the pre-task survey. Additionally, they also reported how they perceived Care-Pilot based on dimensions of AI-mediated support [80]. The survey measurements can be found in Appendix E.2. This iterative task setup let us compare participants' experiences when dealing with uncivil clients, with and without emotional support, and to examine how these factors influenced their performance and emotional responses.

5.2.2 Semi-Structured Interview. After interacting with clients, participants proceeded with a semi-structured interview. They answered a series of open-ended questions to provide deeper insight into their primary goals when dealing with uncivil clients, their attitude towards emotional well-being, and their evaluation of Care-Pilot, especially, in comparison to human coworker support.

5.3 Thematic Analysis

We performed inductive coding to identify Care-Pilot's role in front-office work [13]. Two authors carefully read each transcript and performed open-coding. The authors first coded the transcripts independently and then met frequently to reconcile disagreements. They iteratively improved the codes in 20% chunks. Next, we conducted affinity mapping of 258 initial codes using Miro. We merged similar codes together and pruned out codes outside the scope of our research. During clustering, we first organized concepts related to the socio-organizational norms of front-office work. Based on the gaps in these norms, we clustered codes related to the capabilities of Care-Pilot. Our main findings comprise 128 codes, which were organized into a four-level thematic structure. We primarily elaborate on the broad themes of *Utility*, *Complementarity*, and *Pitfalls* by anchoring them in the normative patterns that emerged from *Reactions to Client Incivility* and *Role of Coworkers in ER*. Fig. 5 provides an overview of the main themes, and their relationship, that we cover in the remaining findings.

6 Findings: Clarifying the Role of Care-Pilot in Uncivil Interactions

> "*As much as people say, "don't let it bother you, try to let it slide off your back," it does, very much, take its toll.*" — P14

The negativity of clients can be contagious [7], and participants described their attitude turning negative (P06), lowered productivity (P08), feeling drained or depressed (P18, P20), and even wanting to reciprocate the incivility (P19). Furthermore, participants also recognized that client incivility can be a form of microtrauma [127] that was only apparent after several shifts or even years (P14). Many of them work in the fear that a dissatisfied client might report them (P01) and subsequently feel trapped in their role (P04). Given these experiences, the participants in our study were uniquely positioned to assess Care-Pilot.

Relatability of the Simulation: Participants had to interact with an uncivil client that replicates uncivil situations in front-office work. During the interactions, several participants exhibited observable reactions to incivility, such as defensively laughing (P06, P18), eye-rolling (P10), and verbally labeling the client (e.g., "spicy"–P05). P07 even claimed his "adrenaline was peaking." P16 endorsed our simulation by stating, "it was a little intense, but that's definitely how people react sometimes; so, it felt very real."

6.1 Utility: Functions of Embedding Empathetic AI into Uncivil Conversations

We structured the probe so that participants could disentangle the role of each component of Care-Pilot (Section 5.2.1). Before elaborating on the findings, we inspected the ratings provided by the participants during the study. A non-parametric *Kruskal-Wallis* test [87] showed that *Cognitive Demands* [39] were lower when CSRs had access to Care-Pilot while interacting with an uncivil Client-Agent ($p = 0.09$ for $\alpha = 0.90$) [7]. This result provides preliminary evidence that Care-Pilot's Emo-Reframe is capable of mitigating the emotional labor-induced demands on a worker (Fig. 6). The same test also revealed that Care-Pilot's different panels significantly differed in helpfulness ($p = 0.001$). Fig. 7 shows that Emo-Reframe was as helpful as Info-Guide, and both were rated higher than Emo-Label. Therefore, on-task emotional reframing was not considered an additional load, and rather, as helpful as information support.

> "*It's kind of like someone telling you, "calm down" — nobody wants that. But, if somebody is genuinely calm in their tone and they say, "Hey, let's take a second to like, take a deep breath, and like analyze the situation and stuff." It's a lot better than hearing calm down. That's how I see [Emo-Reframe]...*" — P16

We found that Emo-Label mirrors those coworkers who simply acknowledge the emotional downsides of the job, without actually providing a way forward. P20 called out the desensitization of his coworkers, "I wouldn't say they'd be helpful, they see it like it's a part of my job." By contrast, participants could identify that Emo-Reframe uniquely varied from simple sentiment classification.

6.1.1 Empathetic messages help CSRs avoid negative thinking traps. In the face of client incivility, a CSR is likely to feel negatively about the situation. Participants reported learning to dissociate from emotions while on task, but the overwhelming volume of interactions can make complete dissociation challenging (P01, P04, P05, P06, P08, P14, P18, P19, P20). However, as part of their role, a CSR needs to suppress their emotional impulse. P13 described her internal method, "whatever your reaction is, you have to keep that in your head and then decide how you're going to respond professionally." While *Surface Acting* has its benefits [50], only changing responses at a superficial level contributes to dissatisfaction at work [27]. Instead, in line with recommendations from emotion literature [50], we found that Emo-Reframe enabled *Deep Acting* — adjusting how participants cognitively evaluated their experience.

> "*I wanted to write that there is "no need to be rude". But then, when I read that box on the right side that told me that the customer is upset about the situation. Not you personally... So I changed my response.*" — P40

Client incivility threatens a CSR's ego [47]. Emo-Reframe acted as a buffer from reciprocating negative feelings. Reading the suggestions stopped participants from "lashing out" (P03) and "losing character" (P15). P17 expanded that Emo-Reframe makes their internal thoughts more explicit. The external acknowledgment of the clients' incivility was validating to the participants and akin to the role supportive coworkers have played for them. Amongst themselves, CSRs are likely to share or vent their frustrations with

[7] The differences between other metrics, such as *Cognitive Resources*, were not significant

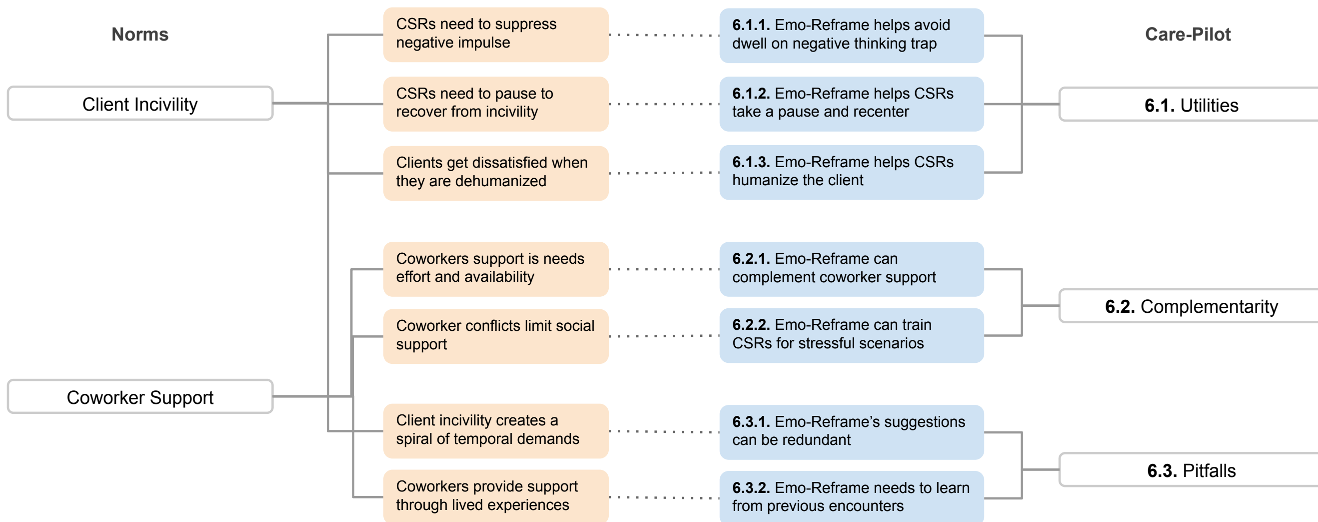


Figure 5: The user evaluation findings sections correspond to 3 major top-level themes describing Care-Pilot. These themes were grounded in the normative patterns of client incivility and coworker support that were raised by our participants.

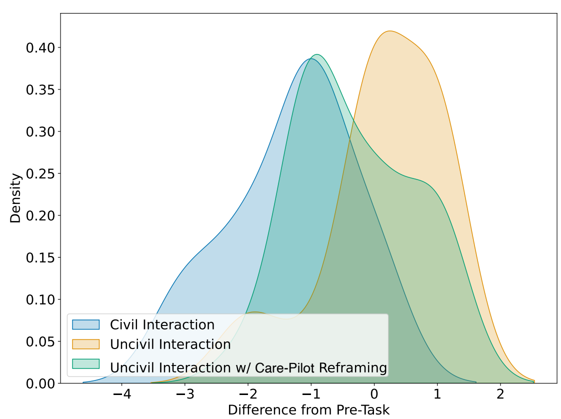


Figure 6: Participants rated the cognitive demands of interacting with the uncivil Client-Agent lower and similar to civil Client-Agent when they were assisted by Care-Pilot's Emo-Reframe

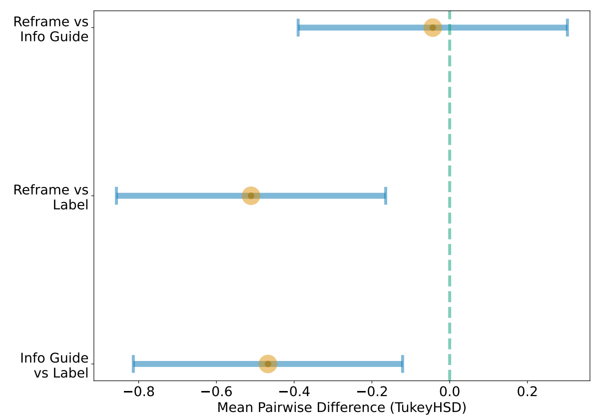


Figure 7: Participants rated the helpfulness of Care-Pilot after every message; Emo-Reframe was significantly better than Emo-Label and similar to Info-Guide (crossing zero-line).

a client, such as "customers totally suck" (P05). Once explicit, this insight from Emo-Reframe, helps the CSR realize that they are not the root of the clients' frustration. The meaning people assign to a negative experience helps determine how they will respond to it [6]. Despite how personal it feels initially, Emo-Reframe emphasizes that the client is not "attacking" them (P04, P10) and they are not the source of their anger (P12, P14, P16). Reappraising a situation can be demanding [39]. P18 was not able to spend quality time with her brother because she did not have the energy to share her negative experience, while still being affected. Emo-Reframe made her more hopeful, "with this tool I could see myself being more willing to talk to him because it took some of the leg work out for me." Participants were less likely to automatically assume a negative scenario and Emo-Reframe encouraged them to rethink the situation.

6.1.2 Empathetic messages give CSRs an opportunity to recenter their problem solving. Previous encounters with uncivil clients, made our participants feel "shaken up" (P14) or "at a loss for words" (P06) to continue their task. Workers who suddenly experience a negative mood are less likely to be able to complete their tasks efficiently and more likely to withdraw from them [90]. Front-office work often involves time-sensitive tasks that could get disrupted by incivility. In response, participants would break their workflow when interacting with an emotionally challenging client. P16 would take "a breather" to reflect on the situation and devise better solutions, "I try to make sure if there's anything that I can do in terms of not making it a future problem." Similarly, participants reported that reading the insights from Emo-Reframe helped them pause, reflect, and move forward with the conversation.

> "*You know, the people who have those angels on their shoulder. It very much felt like that. And that helped me think about what I have empirically.*" — P17

Other participants echoed P07's experience. Essentially, Emo-Reframe helped the participants center their thoughts and continue working towards resolutions. P01 particularly drew a parallel between Emo-Reframe and a coach who would help her maintain "momentum." Much like how Info-Guide had an affective purpose, participants perceived Emo-Reframe to have a procedural purpose that helped them complete their task. Care-Pilot's message drew their attention to their objectives and motivated them to work towards it. Note, we designed the empathetic messages with examples of empathy that were high in specificity [121]. Participants' descriptions indicated regaining confidence and psychological safety due to Emo-Reframe. P05 stated, "it feels validating that you're not necessarily making a mistake." As a result, participants envisioned their job proficiency improving with empathetic support embedded in the messages.

6.1.3 Empathetic messages humanize the client. A majority of our participants held the belief that front-office work is dehumanized. As P09 put it, "customer service agents unfortunately don't work under very good conditions." Organizational policy around worker evaluations further reinforces this belief. The only feedback P18 would receive is to assure she remains productive. Consequently, CSRs resign themselves to the robotic nature of the role and start depersonalizing the clients [14]. In response to interacting with a frustrated client, P04 said, "her emotions, honestly, do not count for me." Depersonalization is a core state of burnout and can detach CSRs from their clients. Our findings suggest that Emo-Reframe helped mitigate some of these perceptions by "triangulating" the clients' emotions by writing in a "short, sweet way" (P07).

> "*That allowed me to be able to say, "Okay, if it was me in this situation, I'd probably be upset, too.""*" — P18

By reading the Emo-Reframe panel, participants were no longer speculating the clients' perspective. P03 actually became more observant of the client and modified his language to be more accommodating. Participants were more willing to service the clients as "someone alive" (P17). The assistance of Emo-Reframe helped the participants rejuvenate their interest in the client's emotional state. Not only was this meaningful to their personal wellbeing, but also enables them to build longer-term relationships with clients.

6.2 Complementarity: Situations where Empathetic AI can Stand-In for Coworkers

Organizations employ multiple CSRs to handle a large volume of client interactions. Regardless of the peculiarity in each client, CSRs together form a "united front" (P04). Coworkers play a key role in supporting CSRs and their empathy is fundamental to mitigating the pressures of client incivility [60, 114]. After interacting with Care-Pilot, our participants compared and contrasted its function with that of their coworkers'. Based on these reflections, we identified two key gaps in existing coworking paradigms of front-office work that emotional support from Care-Pilot can address.

6.2.1 On-task emotional AI can reduce burden in coworker support. Interactions with clients are nuanced. When the conversation reaches a deadlock, a CSR might want to reach out to their coworkers. However, in cases of incivility, our participants brought to light a double-sided burden problem. Incivility leads to exhaustion [14] and can hinder a CSRs ability to seek support. Consider P01's example, "It's just hard to bring them into my conversations when I'll have to explain to them like what is going on." Support requires some disclosure, but disclosing the complexities of a situation can be burdensome [9]. Meanwhile, it can also be burdensome for the coworker to be emotionally available for the problem [44]. Since the CSR is likely to reach out to a coworker who shares the client, that coworker might be more focused on the task than emotions. By contrast, they viewed Care-Pilot as a tool that can be present throughout their task, observe emotional duress, and provide support immediately. P06 captures the unique opportunity of Care-Pilot's emotional support:

> "*The Care-Pilot was kind of there with me in the moment, whereas I can't really have those conversations while I'm trying to have a conversation with the customer. I've never had a little emotional support buddy, like that before.*" — P06

A CSR might put a client on hold to consult a coworker about procedure, but emotional aspects were discussed after the interaction. Instead, "Care-Pilot is an immediate solution" (P10) for managing emotions of the task. Our lexical evaluation (Section 4.2.2) also confirmed that Care-Pilot was capable of adapting to the specifics of the situation better than human coworkers. Another value proposition of Care-Pilot was that it reduces the need for emotional oversight. P08, who now supervises other CSRs as well, imagined that Care-Pilot would take away the need to "always be on edge, or a constant vigilance, to make sure that my agents ... were (not) being insensitive to the client's question." With Care-Pilot CSRs stand to reprioritize which emotional discussions they should have with their coworkers.

6.2.2 Simulating incivility with emotional support can train CSR to use ER. Beyond on-boarding manuals, many CSRs learn on-the-job through other more experienced coworkers [122]. This learning is hampered because of conflicts that arise among coworkers. If one CSR fails at a task, it is typically *escalated* to their coworkers. Our participants pointed out that the increased workload of the coworker makes them less inclined to teach the CSR better approaches. This arrangement can strain the social ties between front-office workers. "At work, there are no real friends," said P04. P01 found herself unable to consult coworkers for her emotional needs because they were not close. P07 even experienced being put down by her coworkers in front of clients. These tensions reduce the ability of early career CSRs to learn from exemplar coworkers. Our study recruited CSRs with at least 2 years of experience. Several expressed their desire to have had a tool like Care-Pilot for training them early.

"*When they face stuff like this they tend to break down. They need people around them to tell them stuff like what CARE-PILOT was doing for me.*" — P20

In line with P20's quote above, P13 — another participant whose role includes supervision — stated that exercising client interactions with Emo-Reframe could help CSRs to develop the "mindset" needed for client interactions. " Some of those live reminders, could be really beneficial for folks, who are less practiced or just have more trouble with that kind of emotional regulation," said P13. Training up the CSRs was seen as an important way to make participants self-reliant. Recent literature shows that workers can outperform their personality based stereotype if they follow healthier behavior patterns [36]. P18 wants to handle client incivility on her own. With CARE-PILOT she felt that her "escalation techniques" would no longer be scrutinized. Some participants saw themselves adapting CARE-PILOT into an evaluation method to delegate certain types of clients to certain CSRs. Together, it was viewed as an important means for front-office workers to self-augment and preserve learning from coworkers for more advanced issues.

6.3 Pitfalls: Challenges in Integrating Empathetic AI as a Coworker

The paper, so far, has described the utility of an empathetic AI such as CARE-PILOT for on-task ER. Our study design involved a simulation of CSR tasks to probe participants on how AI might provide emotional support through tasks. After the task, the probe served as an anchor for participants to conceive new workflows with such a tool. Their responses indicated the critical limitations in the existing implementation of empathetic AI. We explain these to demarcate the scope of empathetic AI so that we can consider alternative solutions to client incivility.

6.3.1 On-task empathetic AI has diminishing marginal utility. One of the key aspects of task load is temporal demands [54]. We learned that when a CSR is unable to manage client incivility, it leads to a vicious spiral of added burdens (P01, P04, P14, P17, P18, P19): CSR are incentivized to resolve complaints quickly → when clients get frustrated, it disrupts their workflow and extends the resolution time → the client gets more frustrated as the resolution gets slower → the CSR gets further distressed because they start accumulating a backlog of unresolved clients. With this context, participants were wary if actual deployments of empathetic AI like CARE-PILOT might interrupt their efficiency.

"*Because after a while I think the usefulness gets lesser and lesser. Does it make sense? It's like diminishing marginal returns.*" — P11

P11 pointed out that the external demands of high efficiency interactions can make support insights from CARE-PILOT redundant over time. Similarly, P06 rushed through some of the insights because she found them repetitive. Displaying additional panes could cause information overload. Our structural analysis showed that CARE-PILOT's messages may need higher reading comprehension (Section 4.2.1). As a solution, P13 recommended that Emo-Reframe should highlight the key takeaway for a quick glance while still maintaining the long-form version for deeper reflection. Based on these experiences, subsequent designs should consider alternatives to the persistent support panes.

6.3.2 Empathetic AI lacks the social connectedness offered by coworkers. The previous section discussed limitations in the existing social dynamics of coworker support (Section 6.2). Having said that, coworkers remain integral to a CSR's emotional health. Since they are situated in the same conditions, they are likely to understand the situation better than other social support sources, such as friends [21]. Moreover, participants recognized that building social relationships with coworkers goes beyond work-related support — "You could talk to them anytime, you could see them anytime, talk to them about anything else" (P19). Through our conversations, we distilled that participants felt CARE-PILOT lacks the lived experience of coworkers. Remember, CARE-PILOT's language rarely used interpersonal pronouns (Section 4.3). Recent findings comparing LLM's social support to that of peers in online communities reflected a similar gap [111]. Our participants perceived their coworkers' support to be more meaningful because they shared the same experiences (P10, P12, P13, P14, P16). The shared experience convinces the CSR that the advice is more relatable. P13 explained that his coworkers' advice is valuable because, "I've been in a situation like this before, and here's how I handled it." P12 even called his conversations with coworkers on the same client as therapeutic. Beyond the shared experience, each coworker also brings in their own unique diverse perspectives (P07). Taken together, it provides the CSR the psychological safety needed to express their emotional concerns and be receptive to support [21]. Enhancing CARE-PILOT with lived experience is non-trivial, but participants could foresee some possibilities.

"*I would hope that if something like this was implemented — as it collects more and more data and more feedback of its responses — it would accumulate to something that would almost like tag team with you to deal with the customer.*" — P12

Above, P12 alluded to the importance of long-term memory (LTM) in improving CARE-PILOT. Prior work shows the importance of (LTM) in AI for mental health [69]. Introducing LTM to CARE-PILOT could help it learn how a CSR deals with clients and provide more relatable suggestions based on past encounters (P07, P13, P16). Expanding CARE-PILOT's LTM with experiences of other CSRs could enable AI to mimic coworkers who "flag" emotionally challenging clients and prepare the CSR for encounters (P10). Beyond LTM, another approach to improve CARE-PILOT would be to build a unique model of each CSR's emotional predisposition by leveraging the potential of LLMs to replicate mental health traits [23]. P01 suggested enhancing CARE-PILOT with sensing, whereas P13 suggested an option for users to explicitly disclose their objectives to CARE-PILOT. Arguably, human coworkers will be irreplaceable, but some of the contemporary advancements hold promise in replicating experience in AI.

7 Discussion

Our study with CSRs exhibits the value of emotional support from AI in responding to client incivility. CARE-PILOT presents one of the first applications of LLMs to mitigate intense emotional labor through empathetic human-AI interactions. Emotional labor is not unique to front-office work and our findings are relevant to all sorts of workers, such as information workers [77], who regularly interact with humans as a part of their job function. The messages from CARE-PILOT distinguish itself from humans and other LLMs.

The linguistic analysis already revealed high expressions of *empathy* and *adaptability* (Section 4.2.2), but its *verbosity*, *readability*, and *analytical style* were notably different from humans. Yet, the high empathy ratings from CSRs (Section 4.4) indicate that Care-Pilot's verbosity might have been associated with thoroughness, readability with compassion, and analytical style with actionability. Subsequently, our simulation exercises revealed the processes through which Care-Pilot's emotional component, Emo-Reframe, could help CSRs reduce the demands of uncivil interactions. Despite an additional block of insight, CSRs found Emo-Reframe as helpful as problem-solving insights. Moreover, these studies revealed important opportunities for technological interventions given the current state of social support at work (Section 6). The following discussion aims to anticipate the sociotechnical advancements and considerations needed to make empathetic AI-coworkers, such as Care-Pilot, available in front-office work.

7.1 Design & Technological Implications

Opportune Moments for On-Task AI Empathy: Clients come in all forms. They may not always act rude and their emotional expressions, along with its effect on the CSR, are likely to vary by different degrees. Consequently, the need and impact of empathetic support can vary within each interaction. Isolating opportune moments for Care-Pilot to deliver messages can help reduce the redundancy between messages (Section 4.2.1, Section 6.3.1). The literature on workplace affect sensing [71] and stress sensing [59, 85] can contribute to making these AI-coworkers act as just-in-time adaptive interventions. Meanwhile, LLMs are exhibiting increasing accuracy in determining mental health labels from text [145]. Note, however, that designing agent-based wellbeing interventions to user receptivity requires dynamic modeling of users' motivations [91, 92]. Studies on behavioral health recommend modulating the effort required to follow an intervention based on the time of delivery [67]. Depending on the specific context, mental state, and the CSR's ability to reflect on additional insight, Care-Pilot may be trained to produce messages of differing degrees of complexity [48]. Commodity devices like smartwatches are becoming increasingly sophisticated for inferring momentary stress [132]. Arguably, some of these approaches involve imposing additional sensing into the workers' ecosystem, and therefore, we need significant advancements in these studies to reliably model worker's emotional state in a social context [29]. Alternatively, contemporary research on LLMs shows the possibility of learning worker preferences through their usage patterns [98]. The next iteration of AI-coworkers needs to be able to anticipate and adapt to CSR needs.

Situated AI-Empathy and Harnessing Experiences: Coworkers also come in all forms. New remote work paradigms need new approaches to social technologies [35]. It is important for a CSR to perceive their coworker as trustworthy and respectful [147], even if it is AI [8].The current iteration of Care-Pilot did not include any major anthropomorphic aspects to delineate the value of the support messages. Subsequent designs, however, can incorporate more anthropomorphic aspects (e.g., name, appearance, and tonality) to improve the emotional connectedness between the CSR and the AI-coworker [57]. At the same time, personalizing the conversations to users can make them more dynamic and human-like (Section 4.2.1). One way to do this is by learning from the CSR and other CSRs using long-term memory (LTM). Jo et al. found that AI agents that leverage LTM were perceived to be more personal and emotionally supportive [69]. Beyond personalizing, LTM could also play a role in referring the CSR to real human stories retrieved from its memory. After all, many workers learn "on-the-job," which is a form of social learning, or learning by observing those in one's social group [122]. These practices are already common and unique to human coworker social networks (Section 6.3). AI-coworkers could act as a medium to share these experiences between CSRs and help them learn ER through each other. Future iterations could also be trained on human experiences that workers disclose on public social media [37, 112]. However, in light of suggesting anthropomorphizing, we also caution against it. Research shows that Human-AI interactions with human-like agents can lead to *parasocial* relationships, where users only have illusory connections to the agent as it is not a real person [82]. Such relationships have been shown to have concerning effects on people's mental health. For instance, users' overreliance on human-like AI for emotional support can backfire when the agent behaves contrary to their expectations [78]. Given the sensitivity of emotional labor we advise careful examination of emerging social dynamics within HAI as researchers design new AI-assistants for empathetic coworking.

Scaffolding Emotions in Task-Adjacent Moments: The lack of interventions for on-task cognitive reframing [123] motivated us to design Care-Pilot as an assistant that can intercept conversations with empathetic suggestions (Section 3). Instead, front-office work incorporates off-task ER methods. Some of these are infrastructural (e.g., training modules) and others are sociocultural (e.g., internal forum for venting). We believe Care-Pilot can not only co-exist with these off-task approaches, but it can also help bridge both approaches. For CSRs who regularly interact with the same clients, Care-Pilot could learn from conversations to create a form of emotional briefing. This early information could act similarly to the usage of AI for planning, collaboration, and communication [94]. In turn, it can also be combined with AI for scheduling tasks and ensure workers can focus on important client-interactions when they have maximum resources available [32, 58] and also have opportunities to synchronize with their coworkers [33]. After the task, Care-Pilot could also provide an emotional debriefing after an intense conversation to help the CSR recover in a way that resembles a coach. A practical way forward would be to augment Care-Pilot with LLM tools that encourage self-reflection of stressful experiences [124]. Software engineers are already appropriating AI for post-task recommendations [19]. What Care-Pilot learns within the task can be extrapolated in between tasks to help CSRs smoothly ramp in and out of emotional labor.

7.2 Socio-Organizational Implications

Training, Evaluation & Safety. The current norm of desensitization to client incivility is partly because of the normative expectations that certain individuals are "built" for the job (e.g., "needs a thick skin"). Traits, however, do not develop overnight. Many participants saw Care-Pilot as an early career training tool (Section 6.2). Learning modules could include Care-Pilot's simulation exercises to train ER [123]. A byproduct of many training is quantitative assessment. AI agents have already been proposed to assess worker mental health [64]. Workplaces are always looking to optimize their personnel through different algorithmic management methods, but may not take the most responsible approaches [31]. Tools like Care-Pilot do raise the risk of being appropriated to measure emotional labor. Kaur et al. has shown that some state-of-the-art methods to recognize worker emotion lack the necessary knowledge to accurately estimate a worker's mental state [72]. Ironically, a CSR might end up performing more emotional labor to comply with Care-Pilot's expectations [73, 109]. Sometimes, a CSR should not tolerate client incivility (e.g., when a client uses obscene language and threatens harm). Organizations often have safeguards and protocols for these scenarios and Care-Pilot needs to be integrated with such institutional knowledge. Using Care-Pilot as a training tool is tempting but must be accompanied by appropriate protections and guidelines for use.

Mediating Social Relationships at Work. The messages generated by Care-Pilot could be effective (Section 4.2.2), but it does not share the narrative style of humans (Section 4.2.2). On-task, a CSR may prefer Care-Pilot because it relieves the burden of sharing context with a coworker to receive support (Section 6.2). From a utilitarian perspective, Care-Pilot has advantages over human coworkers, but social relationships with coworkers offer more than just work-related support. The integration of Care-Pilot into work might discourage CSRs from pro-social behaviors. Buçinca et al. have found evidence that AI assistance at work can erode social relationships [16]. Given the anxiety around over-reliance on AI [15] and users' reactions to AI misrecognizing psychosocial traits [136], one might be concerned about workplace social ties in the presence of empathetic AI. One way forward, to avoid these concerns, is to augment AI coworkers by designing interventions that encourage human–human support, e.g., "Consider reaching out to [your coworker], who went through a similar experience last week." We urge future researchers to design them to not just mitigate negative social influences (e.g., incivility) but also promote positive social influences (e.g., peer support).

7.3 Limitations & Future Work

We created Care-Pilot to unearth the potential of LLM–powered AI-coworkers as sources of emotional support in front-office work. Through our evaluations, we provide evidence on the effectiveness of AI-generated support messages and the usability of these insights in client incivility scenarios. All of our evaluations centered on the perspectives of real CSRs; however, admittedly, transferring our findings directly to operational use needs more testing. Our initial comparison of Care-Pilot with Human's relied on a synthetic dataset of uncivil client-representative incidents. The authors with experience in front-office work, as well as CSRs in the authors' organization provided initial validation for the realism of this data, however, client–representative conversations can have many intricacies which are lost to our dataset. Future research needs to consider safely acquiring and leveraging real data from organizations' CSR logs, which are often stored for training purposes. A pragmatic approach to build on our findings could be to collaborate with organizations that provide CSR services to other businesses, such as *Zendesk*, *HubSpot*, or *Salesforce* .

Another important limitation of our study is that the effects of ER with Deep Acting, specifically Cognitive Change, only become apparent in the long term. Our results motivate field deployments and naturalistic testing for this purpose by demonstrating that Care-Pilot has the feasibility and usability to improve mental health at work on the long term. The user evalution we conducted was primarily designed as a technological probe (Section 5). Therefore, we designed that segment of the study to ensure participants are able to acclimatize to a new technological interface. The downside of our method is its margin for ordering effects, such novelty and learning biases. Follow-up studies can consider larger samples and longer deployments, along with counter-balanced groups to understand the role of LLM-powered agents like Care-Pilot in a more robust way. Researchers attempting these studies still need to overcome several non-trivial challenges. Future endeavors would need researchers to establish long-term partnerships with organizations that employ CSRs. Moreover, researchers need to consider the ethics and pragmatics of deploying such an AI-coworker intervention for critical roles such as those in front-office work.

Even prior to field deployment, researchers can improve on our study on a few dimensions. The current design of Care-Pilot expects clients to be uncivil, but a conversation might ebb and flow in its degree of incivility. Emo-Reframe messages lack the variation of human messages, and can, therefore, become a source of fatique and annoyance for CSRs if they always appear. The setting for our user-evaluation mitigated these challenges to a degree because it involved short conversations and only a few turns. However, these negative effects might accumulate over several interactions in a day and across days. Future versions of Care-Pilot need to be able to detect incivility, CSR's affective response, and their goals to truly become dynamic as an adaptive intervention. We discussed some potential pathways forward to identify opportune moments in Section 7.1. The current iteration of emotional support messages is intentionally distinct from informational support. New iterations can explore and evaluate the value of combining these messages, i.e., conveying information with empathy. The state of work is in constant flux and our invesitagtion with Care-Pilot presents one of the first steps in reducing emotion labor with AI-assistance. To fully realize such solutions, we need further investigation with longer real-world deployment in different settings.

8 Conclusion

Work as we know it is changing because of the advent of LLMs. AI applications are augmenting problem-solving at a meteoric rate. However, front-office work involves more than procedural tasks. AI assistants need to support the emotional labor involved in work. Our study presents Care-Pilot, an LLM-powered AI assistant to support the emotional labor of CSRs. We found that Care-Pilot's support messages were effective in expressing empathy appropriately (RQ1). We also found that CSRs could regulate their emotions thanks to Care-Pilot's Emo-Reframe function (RQ2). These results open new doors for implementing holistic AI-coworkers, and also raise important questions for socio-organizational development around this technology.

Acknowledgments

This project was supported through Microsoft's *Accelerating Foundation Models Research (AFMR* program and their call to *AI, Cognition, and Economy* (AICE) research network. Our project specifically answers AFMR's call to "improve human interactions via sociotechnical research." Additional support was provided by the NIH National Institute of Drug Abuse under award number NIH/NIDA P30DA029926. We are grateful to the Microsoft Support advocates for providing early expert evaluations. We did not recruit personnel from the sponsor organizations. The authors' views are their own and do not represent the sponsors' policies; mention of companies or products does not imply endorsement. We thank members of the UbiWell lab at Northeastern University for their assistance and feedback during the developmental phase of the project. These findings would not be possible without all the participants who voluntarily contributed to this research.

References

[1] 2024. GPT OpenAI Models. https://https://platform.openai.com/docs/models/. Accessed: 2024-03-09.
[2] 2024. Llama 2. https://ai.meta.com/llama/. Accessed: 2024-01-25.
[3] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL* (2016).
[4] Lynne M. Andersson and Christine M. Pearson. 1999. Tit for tat? The spiraling effect of incivility in the workplace. *Academy of Management Review* 24, 3 (1999), 452–471. doi:10.5465/amr.1999.2202131
[5] Stuart Axelbrooke. 2017. Customer Support on Twitter. doi:10.34740/KAGGLE/DSV/8841
[6] Alexandra D Babij, Jeni L Burnette, and Crystal L Hoyt. 2020. Failing and feeling bad: How we think about experiencing negative emotions during setbacks. *Motivation and Emotion* 44, 3 (2020), 436–452.
[7] Sigal G Barsade, Constantinos GV Coutifaris, and Julianna Pillemer. 2018. Emotional contagion in organizational life. *Research in Organizational Behavior* 38 (2018), 137–151.
[8] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1 (2009), 71–81.
[9] Natalya N Bazarova. 2015. Online disclosure. *The international encyclopedia of interpersonal communication* (2015), 1–18.
[10] Alberto Betella and Paul FMJ Verschure. 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one* 11, 2 (2016), e0148037.
[11] Heidi Black, Scott Greenberg, Katie Saulsgiver, Evan Sinar, Adam Reece, Emily Carr, and Greg Kellerman. 2019. Beyond hedonia: 5 keys to enhancing workplace well-being at scale. *American Journal of Health Promotion* 33, 8 (2019), 1212–1217.
[12] Shoshana Blum-Kulka and Elite Olshtain. 1986. Too many words: Length of utterance and pragmatic failure. *Studies in second language acquisition* 8, 2 (1986), 165–179.
[13] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[14] Celeste M Brotheridge and Alicia A Grandey. 2002. Emotional labor and burnout: Comparing two perspectives of "people work". *Journal of vocational behavior* 60, 1 (2002), 17–39.
[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.
[16] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. *arXiv preprint arXiv:2306.03280* (2023).
[17] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4758–4765.
[18] Franziska Burger, Mark A Neerincx, and Willem-Paul Brinkman. 2021. Natural language processing for cognitive therapy: extracting schemas from thought records. *PloS one* 16, 10 (2021), e0257832.
[19] Beatriz Cabrero-Daniel, Tomas Herda, Victoria Pichler, and Martin Eder. 2024. Exploring Human-AI Collaboration in Agile: Customised LLM Meeting Assistants. In *International Conference on Agile Software Development*. Springer Nature Switzerland Cham, 163–178.
[20] Robert D Caplan. 1975. *Job demands and worker health: Main effects and occupational differences*. Vol. 75. US Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control, National Institute for Occupational Safety and Health.
[21] Abraham Carmeli, Daphna Brueller, and Jane E Dutton. 2009. Learning behaviours in the workplace: The role of high-quality interpersonal relationships and psychological safety. *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research* 26, 1 (2009), 81–98.
[22] Yooncheong Cho, Il Im, Roxanne Hiltz, and Jerry Fjermestad. 2002. An analysis of online customer complaints: implications for web complaint management. In *Proceedings of the 35th annual Hawaii international conference on system sciences*. IEEE, 2308–2317.
[23] Akshat Choube, Vedant Das Swain, and Varun Mishra. 2024. SeSaMe: A Framework to Simulate Self-Reported Ground Truth for Mental Health Sensing Studies. *arXiv preprint arXiv:2403.17219* (2024).
[24] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60, 2 (1975), 283.
[25] Christine L Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe versus aww: The effectiveness of two kinds of positive emotional stimulation on stress reduction for online content moderators. *Proceedings of the ACM on human-computer interaction* 6, CSCW2 (2022), 1–19.
[26] Carolyn E Cutrona and Daniel W Russell. 1987. The provisions of social relationships and adaptation to stress. *Advances in personal relationships* 1, 1 (1987), 37–67.
[27] Stéphane Côté and Lisa M Morgan. 2002. A longitudinal analysis of the association between emotion regulation, job satisfaction, and intentions to quit. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 23, 8 (2002), 947–962.
[28] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.
[29] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *CHI Conference on Human Factors in Computing Systems*. 1–16.
[30] Vedant Das Swain, Lan Gao, Abhirup Mondal, Gregory D Abowd, and Munmun De Choudhury. 2024. Sensible and Sensitive AI for Worker Wellbeing: Factors that Inform Adoption and Resistance for Information Workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–30.
[31] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D Abowd, and Munmun De Choudhury. 2023. Algorithmic Power or Punishment: Information Worker Perspectives on Passive Sensing Enabled AI Phenotyping of Performance and Wellbeing. In *CHI Conference on Human Factors in Computing Systems*.
[32] Vedant Das Swain, Javier Hernandez, Brian Houck, Koustuv Saha, Jina Suh, Ahad Chaudhry, Tenny Cho, Wendy Guo, Shamsi T Iqbal, and Mary Czerwinski. 2023. Focused Time Saves Nine: Evaluating Computer-Assisted Protected Time for Hybrid Information Work. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).
[33] Vedant Das Swain, Manikanta D Reddy, Kari Anne Nies, Louis Tay, Munmun De Choudhury, and Gregory D Abowd. 2019. Birds of a Feather Clock Together: A Study of Person-Organization Fit Through Latent Activity Routines. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[34] Vedant Das Swain and Koustuv Saha. 2024. Teacher, Trainer, Counsel, Spy: How Generative AI can Bridge or Widen the Gaps in Worker-Centric Digital Phenotyping of Wellbeing. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–13.
[35] Vedant Das Swain, Koustuv Saha, Gregory D Abowd, and Munmun De Choudhury. 2020. Social Media and Ubiquitous Technologies for Remote Worker Wellbeing and Productivity in a Post-Pandemic World. In *Proceedings of 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence (CogMI)*.
[36] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M Gregg, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. 2019. A Multisensor Person-Centered Approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas. *Proc. IMWUT* (2019).
[37] Vedant Das Swain, Koustuv Saha, Manikanta D. Reddy, Hemang Rajvanshy, Gregory D. Abowd, and Munmun De Choudhury. 2020. Modeling Organizational Culture with Workplace Experiences Shared on Glassdoor. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–15.
[38] Munmun De Choudhury and Emre Kıcıman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*.
[39] Evangelia Demerouti, Arnold B Bakker, Friedhelm Nachreiner, and Wilmar B Schaufeli. 2001. The job demands-resources model of burnout. *Journal of Applied*

Psychology 86, 3 (2001), 499–512.
[40] M Brent Donnellan and Richard W Robins. 2010. Resilient, overcontrolled, and undercontrolled personality types: Issues and controversies. *Social and Personality Psychology Compass* 4, 11 (2010), 1070–1083.
[41] B. Dosono and B. Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
[42] Christopher Elliott. 2018. *Chatbots are killing customer service. Here's why.* https://www.forbes.com/sites/christopherelliott/2018/08/27/chatbots-are-killing-customer-service-heres-why
[43] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27.
[44] Grainne Fadden, Paul Bebbington, and Liz Kuipers. 1987. The burden of care: the impact of functional psychiatric illness on the patient's family. *The British Journal of Psychiatry* 150, 3 (1987), 285–292.
[45] David M Fisher. 2014. A multilevel cross-cultural examination of role overload and organizational commitment: investigating the interactive effects of context. *Journal of Applied Psychology* 99, 4 (2014), 723.
[46] Glenda M Fisk and Lukas B Neville. 2011. Effects of customer entitlement on service workers' physical and psychological well-being: A study of waitstaff employees. *Journal of occupational health psychology* 16, 4 (2011), 391.
[47] R. Frey-Cordes, M. Eilert, and M. Büttgen. 2020. Eye for an eye? Frontline service employee reactions to customer incivility. *Journal of Services Marketing* 34, 7 (2020), 939–953.
[48] Marianna B. Ganapini et al. 2023. Value-based Fast and Slow AI Nudging. *arXiv preprint arXiv:2307.07628* (2023).
[49] Thomas A Glass and George L Maddox. 1992. The quality and quantity of social support: stroke recovery as psycho-social transition. *Soc. Sci. Med.* (1992).
[50] Alicia A. Grandey. 2000. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of Occupational Health Psychology* 5, 1 (2000), 95.
[51] Alicia A Grandey, John H Kern, and Michael R Frone. 2007. Verbal abuse from outsiders versus insiders: comparing frequency, impact on emotional exhaustion, and the role of emotional labor. *Journal of occupational health psychology* 12, 1 (2007), 63.
[52] Katherine Haan and Lauren Holznienkemper. 2023. Remote Work Statistics & Trends In 2024. *Forbes Advisor* (2023). https://www.forbes.com/advisor/business/remote-work-statistics/
[53] Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. 2016. *Natural language processing: python and NLTK.* Packt Publishing Ltd.
[54] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
[55] Alexander P. Henkel, Stefano Bromuri, Deniz Iren, and Visara Urovi. 2020. Half Human, Half Machine – Augmenting Service Employees with AI for Interpersonal Emotion Regulation. *Journal of Service Management* 31, 2 (June 2020), 247–265. doi:10.1108/JOSM-05-2019-0160
[56] Ilona Herlin and Laura Visapää. 2016. Dimensions of empathy in relation to language. *Nordic Journal of linguistics* 39, 2 (2016), 135–157.
[57] Erik Hermann. 2022. Anthropomorphized artificial intelligence, attachment, and consumer behavior. *Marketing Letters* 33, 1 (2022), 157–162.
[58] Javier Hernandez, Vedant Das Swain, Jina Suh, Daniel McDuff, Judith Amores, Gonzalo Ramos, Kael Rowan, Brian Houck, Shamsi Iqbal, and Mary Czerwinski. 2024. Triple Peak Day: Work Rhythms of Software Developers in Hybrid Work. *IEEE Transactions on Software Engineering* (2024).
[59] Javier Hernandez, Rob R Morris, and Rosalind W Picard. 2011. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer, 125–134.
[60] Violet T Ho and Neharika Gupta. 2014. Retaliating against Customer Interpersonal Injustice in a Singaporean Context: Moderating Roles of Self-Efficacy and Social Support. *Applied Psychology* 63, 3 (2014), 383–410.
[61] Arlie Russell Hochschild. 1983. *The managed heart: The commercialization of human feeling.* University of California Press, Berkeley.
[62] Arlie Russell Hochschild. 2019. *The managed heart: Commercialization of human feeling.* University of California press.
[63] James S House. 1983. Work stress and social support. *Addison-Wesley series on occupational stress* (1983).
[64] Ines Hungerbuehler, Kate Daley, Kate Cavanagh, Heloísa Garcia Claro, and Michael Kapps. 2021. Chatbot-based assessment of employees' mental health: Design process and pilot implementation. *JMIR formative research* 5, 4 (2021), e21678.
[65] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
[66] Michael Inzlicht, C Daryl Cameron, Jason D'Cruz, and Paul Bloom. 2023. In praise of empathic AI. *Trends in Cognitive Sciences* (2023).
[67] Nicholas C Jacobson, Tobias Kowatsch, and Lisa A Marsch. 2022. Digital therapeutics for mental health and addiction: The state of the science and vision for the future. (2022).
[68] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
[69] Eunkyung Jo, Yuin Jeong, SoHyun Park, Daniel A Epstein, and Young-Ho Kim. 2024. Understanding the Impact of Long-Term Memory on Self-Disclosure with Large Language Model-Driven Chatbots for Public Health Intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
[70] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.
[71] Harmanpreet Kaur et al. 2020. Optimizing for happiness and productivity: Modeling opportune moments for transitions and breaks at work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*.
[72] Harmanpreet Kaur et al. 2022. "I didn't know I looked angry": Characterizing observed emotion and reported affect at work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
[73] Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. 2023. Sensing Wellbeing in the Workplace, Why and For Whom? Envisioning Impacts with Organizational Stakeholders. *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2023).
[74] William Kidder, Jason D'Cruz, and Kush R Varshney. 2024. Empathy and the Right to Be an Exception: What LLMs Can and Cannot Do. *arXiv preprint arXiv:2401.14523* (2024).
[75] Gregory G Kolden, Marjorie H Klein, Chia-Chiang Wang, and Sara B Austin. 2011. Congruence/genuineness. *Psychotherapy* 48, 1 (2011), 65.
[76] Jaclyn Koopmann, Mo Wang, Yihao Liu, and Yifan Song. 2015. Customer mistreatment: A review of conceptualizations and a multilevel theoretical model. *Mistreatment in organizations* 13 (2015), 33–79.
[77] Carol Collier Kuhlthau. 1999. The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction, and sources. *Journal of the American Society for information Science* 50, 5 (1999), 399–412.
[78] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* 26, 10 (2024), 5923–5941.
[79] Kenneth S Law, Chi-Sum Wong, and Lynda J Song. 2004. The construct and criterion validity of emotional intelligence and its potential utility for management studies. *Journal of applied Psychology* 89, 3 (2004), 483.
[80] Bo Liu and S Shyam Sundar. 2018. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 10 (2018), 625–636.
[81] Steven Loria et al. 2018. textblob Documentation. *Release 0.15* 2, 8 (2018), 269.
[82] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.
[83] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3025–3034.
[84] Christina Maslach, Susan E Jackson, and Michael P Leiter. 1997. *Maslach burnout inventory.* Scarecrow Education.
[85] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, et al. 2019. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. (2019).
[86] Nicholas Mcinnes and Bo JA Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health and social care* 36, 4 (2011), 173–189.
[87] Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.
[88] A. Menking and I. Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 207–210.
[89] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[90] Andrew G Miner and Theresa M Glomb. 2010. State mood, task performance, and behavior at work: A within-persons approach. *Organizational Behavior and*

Human Decision Processes 112, 1 (2010), 43–57.
[91] Varun Mishra, Sarah Hong, and David Kotz. 2024. Exploring the Relationship Between Intrinsic Motivation and Receptivity to mHealth Interventions. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Melbourne VIC Australia, 437–443. doi:10.1145/3675094.3678498
[92] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting Receptivity for mHealth Interventions in the Natural Environment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2 (June 2021), 74:1–74:24. doi:10.1145/3463492
[93] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research* 20, 6 (2018), e10148.
[94] Katelyn Morrison, Shamsi T. Iqbal, and Eric Horvitz. 2024. AI-Powered Reminders for Collaborative Tasks: Experiences and Futures. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–20.
[95] Mehrab Bin Morshed, Javier Hernandez, Daniel McDuff, Jina Suh, Esther Howe, Kael Rowan, Marah Abdin, Gonzalo Ramos, Tracy Tran, and Mary Czerwinski. 2022. Advancing the understanding and measurement of workplace stress in remote information workers from passive sensors and behavioral data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
[96] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.
[97] Priya Nambisan. 2011. Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association* 18, 3 (2011), 298–304.
[98] Lin Ning et al. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. *arXiv preprint arXiv:2402.13598* (2024).
[99] John C Norcross and Michael J Lambert. 2018. Psychotherapy relationships that work III. *Psychotherapy* 55, 4 (2018), 303.
[100] Hirotaka Osawa. 2014. Emotional cyborg: Complementing emotional labor with human-agent interaction technology. In *Proceedings of the second international conference on Human-agent interaction*. 51–57.
[101] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance* 17 (2018), 22–27.
[102] Christine M Pearson, Lynne M Andersson, and Christine L Porath. 2000. Assessing and attacking workplace incivility. *Organizational dynamics* 29, 2 (2000), 123–137.
[103] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one* 9, 12 (2014), e115844.
[104] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.
[105] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[106] Jaco Pienaar and S. Andre Willemse. 2008. Burnout, engagement, coping and general health of service employees in the hospitality industry. *Tourism Management* 29, 6 (2008), 1053–1063.
[107] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186–195.
[108] Anat Rafaeli et al. 2012. When customers exhibit verbal aggression, employees pay cognitive costs. *Journal of Applied Psychology* 97, 5 (2012), 931–942. doi:10.1037/a0028559
[109] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion AI at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.
[110] Deborah E. Rupp et al. 2008. Customer (in) justice and emotional labor: The role of perspective taking, anger, and emotional regulation. *Journal of Management* 34, 5 (2008), 903–924. doi:10.1177/0149206307309261
[111] Koustuv Saha, Yoshee Jain, Chunyu Liu, Sidharth Kaliappan, and Ravi Karkar. 2025. AI vs. Humans for Online Support: Comparing the Language of Responses from LLMs and Online Communities of Alzheimer's Disease. *ACM Transactions on Computing for Healthcare* (2025).
[112] Koustuv Saha, Manikanta D Reddy, Stephen Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[113] Koustuv Saha and Amit Sharma. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. In *ICWSM*.
[114] Kumi Sakurai and Steve M Jex. 2012. Coworker incivility and incivility targets' work effort and counterproductive work behaviors: the moderating role of supervisor social support. *Journal of occupational health psychology* 17, 2 (2012), 150.
[115] Wendy Samter, Brant R Burleson, and Lori Basden Murphy. 1987. Comforting conversations: The effects of strategy type on evaluations of messages and message producers. *Southern Journal of Communication* 52, 3 (1987), 263–284.
[116] Angela M. Schöpke-Gonzalez, Shubham Atreja, Han Na Shin, Najmin Ahmed, and Libby Hemphill. 2024. Why Do Volunteer Content Moderators Quit? Burnout, Conflict, and Harmful Behaviors. *New Media & Society* 26, 10 (Oct. 2024), 5677–5701. doi:10.1177/14614448221138529
[117] Nicola S. Schutte et al. 2001. Emotional intelligence and interpersonal relations. *The Journal of Social Psychology* 141, 4 (2001), 523–536. doi:10.1080/00224540109600569
[118] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184* (2023).
[119] Omar Shaikh, Victoria Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309* (2023).
[120] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5263–5276.
[121] A. Sharma, K. Rushton, I. W. Lin, D. Wadden, K. G. Lucas, A. S. Miner, and T. Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466* (2023).
[122] Garriy Shteynberg and Evan P Apfelbaum. 2013. The power of shared experience: Simultaneous observation with similar others facilitates social learning. *Social Psychological and Personality Science* 4, 6 (2013), 738–744.
[123] P. Slovak, A. Antle, N. Theofanopoulou, C. Daudén Roquet, J. Gross, and K. Isbister. 2023. Designing for emotion regulation interventions: an agenda for HCI theory and research. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–51.
[124] Inhwa Song, SoHyun Park, Sachin R. Pendse, Jessica Lee Schleider, Munmun De Choudhury, and Young-Ho Kim. 2025. ExploreSelf: Fostering User-driven Exploration and Reflection on Personal Challenges with Adaptive Guidance by Large Language Models. doi:10.48550/arXiv.2409.09662 arXiv:2409.09662 [cs]
[125] Sharmin Spencer and Deborah E Rupp. 2009. Angry, guilty, and conflicted: injustice toward coworkers heightens emotional labor through cognitive and emotional mechanisms. *Journal of Applied Psychology* 94, 2 (2009), 429.
[126] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
[127] Shulamith Lala Ashenberg Straussner and Alexandrea Josephine Calnan. 2014. Trauma through the life cycle: A review of current literature. *Clinical Social Work Journal* 42 (2014), 323–335.
[128] Mathew Sweezey. 2019. *Consumer Preference for Chatbots Is Challenging Brands to Think 'Bot First'*. https://www.forbes.com/sites/forbescommunicationscouncil/2019/08/16/consumer-preference-for-chatbots-is-challenging-brands-to-thinkbot-first
[129] Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, João Sedoc, Sven Buechel, and Alexandra Balahur. 2021. Wassa 2021 shared task: predicting empathy and emotion in reaction to news stories. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*. Association for Computational Linguistics, 92–104.
[130] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[131] Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, Vol. 1. 1050–1056.
[132] Kobiljon Toshnazarov, Uichin Lee, Byung Hyung Kim, Varun Mishra, Lismer Andres Caceres Najarro, and Youngtae Noh. 2024. SOSW: Stress Sensing With Off-the-Shelf Smartwatches in the Wild. *IEEE Internet of Things Journal* 11, 12 (June 2024), 21527–21545. doi:10.1109/JIOT.2024.3375299
[133] M. Vacchiano, G. Fernandez, and R. Schmutz. 2024. What's going on with teleworking? A scoping review of its effects on well-being. *PloS one* 19, 8 (2024), e0305567.
[134] Janet Walsh and Stephen Deery. 2006. Refashioning organizational boundaries: Outsourcing customer service work. *Journal of Management Studies* 43, 3 (2006), 557–582.

[135] Lih-Wern Wang, Michael J Miller, Michael R Schmitt, and Frances K Wen. 2013. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy* 9, 5 (2013), 503–516.

[136] Qiaosi Wang, Chidimma L Anyi, Vedant Das Swain, and Ashok K Goel. 2024. Navigating AI Fallibility: Examining People's Reactions and Perceptions of AI after Encountering Personality Misrepresentations. *arXiv preprint arXiv:2405.16355* (2024).

[137] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok K Goel. 2021. Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proc. CHI*.

[138] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[139] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[140] D. Y. Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[141] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. 38–45.

[142] Chi-Sum Wong and Kenneth S Law. 2017. The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. In *Leadership perspectives*. Routledge, 97–128.

[143] Bo Xiao, Chen Liang, Yuxin Liu, and Xiaoyan Zheng. 2022. Service staff encounters with dysfunctional customer behavior: Does supervisor support mitigate negative emotions? *Frontiers in Psychology* 13 (2022), 987428.

[144] Jun Xu, Yongmei Liu, and Yi Guo. 2014. The role of subordinate emotional masking in leader–member exchange and outcomes: A two-sample investigation. *Journal of Business Research* 67, 2 (2014), 100–107.

[145] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (March 2024), 1–32. doi:10.1145/3643540 arXiv:2307.14385 [cs]

[146] Dana Yagil. 2008. When the customer is wrong: A review of research on aggression and sexual harassment in service encounters. *Aggression and violent behavior* 13, 2 (2008), 141–152.

[147] Fan Xuan Yang and Vanessa M C Lau. 2019. Evil customers, an angel boss and coopetitive coworkers: Burnout of frontline employees. *International Journal of Hospitality Management* 83 (2019), 1–10.

[148] Yunhao Yuan, Koustuv Saha, Barbara Keller, Erkki Tapio Isometsä, and Talayeh Aledavood. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *Proceedings of the ACM Web Conference 2023*. 2677–2685.

[149] Alice Qian Zhang, Judith Amores, Mary L. Gray, Mary Czerwinski, and Jina Suh. 2024. AURA: Amplifying Understanding, Resilience, and Awareness for Responsible AI Content Work. doi:10.48550/arXiv.2411.01426 arXiv:2411.01426 [cs]

[150] Ke Zhou, Luca Maria Aiello, Sanja Scepanovic, Daniele Quercia, and Sara Konrath. 2021. The language of situational empathy. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–19.

[151] Leonieke G Zomerdijk and Jan de Vries. 2007. Structuring front office and back office work in service delivery systems: an empirical study of three design decisions. *International Journal of Operations & Production Management* 27, 1 (2007), 108–131.

Appendix

A Generating Incidents with Uncivil Clients

A.1 Complaint categories

To enhance the realism of the simulation, we expanded the complaint categories using insights from previous customer complaint analyses [22]. Furthermore, we refined these categories by incorporating a real-world dataset of publicly available client complaints from Twitter [5]. These updated complaint categories were then used as prompts for GPT-4o, allowing for the creation of lifelike, multi-turn text interactions to simulate realistic customer service scenarios.

(1) Service Quality: Issues related to human-to-human service interactions, such as delays, staff behavior, and communication errors.
(2) Product Issues: Concerns regarding physical or functional aspects of a product or service, including defects, safety issues, and mismatches between expectation and reality.
(3) Pricing and Charges: Financial discrepancies encountered before, during, or after a service, such as overcharging, undisclosed fees, or refund issues.
(4) Policy: Grievances associated with company rules and guidelines, particularly when they are perceived as unfair or inflexible.
(5) Resolution: The efficacy of the company's actions in addressing and resolving complaints, focusing on customer satisfaction with the outcomes provided.

A.2 Prompt Design

The Client-Agent was broken into two major components. The first component creates a complaint based on the categories above and examples provided to it. Fig. A3 describes the exact prompt we used along with the examples to generate different complaints. The examples were sourced from Axelbrooke's real-world dataset of complaints [5]. Multiple researchers categorized these complaints and then identified a sample for few-shot learning [138] that ensures Client-Agent is seeded with knowledge of a diverse set of complaints from different domains when initializing the complaint. The second component responds to CSR's messages, such as the follow-up questions about the initial complaint. Fig. A1 contains the prompt that defines Client-Agent behavior towards CSR. This prompt specifies the behaviors that are deemed uncivil. Moreover, the prompt also includes formatting details to ensure the length of messages are concise and conversations have closure. We prompted the Client-Agent to return the string "FINISH:999" to indicate closure. For the user evaluation (Section 5.2.1), our front-end would anticipate that string and change the user-flow accordingly—triggering the post-task survey.

A.3 Conversation Memory

For any of these components to work realistically, the Client-Agent and Care-Pilot need to have a sense of historical context, which, in this case, was limited to the existing message thread of conversations. A CSR might ask, "Could you please provide your confirmation number?". However, this message alone, does not contain any information of the original complaint or any other preceding messages. To account for this, we use a basic summarization prompt to rephrase the latest message to capture the context of the chat history so far. Our implementation was a minor variation from the recommendation in *Langchain* [131]. Fig. A2 shows the prompt to convert a message like "what is your confirmation number?" to "what is the confirmation number of the flight you missed?". The historical contextualization is essential to ensure the Client-Agent does not *act* frustrated because of lack of detail. Instead, it acts in an uncivil manner despite this.

```
Your role is to act like a CUSTOMER seeking support. \
You are speaking to a support REPRESENTATIVE. \
Respond to the question as if you were the customer. \
Do NOT reveal your role.\
Ensure every turn is one to three sentences, and DO NOT make it too long to read.\

If the representative is asking for a specific detail, respond with a believable answer.\
If customer has agreed with response then respond with "FINISH:999"
After 10 - 12 turns, respond with messages to close the conversation.\
After 12 turns, do NOT respond further, only respond with "FINISH:999".\

Phrase your responses like an UNCIVIL customer:\
- Use a rude, impolite, and disrespectful tone.\
- DO NOT show good manners or courtesy.\
- DO NOT use a polite or nice tone.\
- Show disregard for others.\

Representative: {question}
Customer:
```

Figure A1: Full prompt to respond to CSR in an uncivil way

```
Given a chat history and the latest user question \
which might reference context in the chat history, formulate a standalone question \
which can be understood without the chat history. Do NOT answer the question, \
just reformulate it if needed and otherwise return it as is.
```

Figure A2: Full prompt to contextualize chat history before responding

B Generating Emotion Support for CSRs

As we described in Section 3.2.2, we adapted Sharma et al.'s method to design emotional reframing to support CSRs [121]. This implementation involved a chain of several different prompts. The three main pieces to this chain are *situation*, *thought*, and *reframe*:

(1) Fig. B4 details the prompt for *situation*. It describes how the CSR might be negatively perceived. This prompt considers the entire conversation history to summarize the incident while centering the *ego-threat* [47] the CSR faces.
(2) Fig. B5 details the prompt to derive a negative *thought* from the *situation* using examples we curated from [121].

```
Your role is to act like a customer seeking support. \
You are messaging a service representative via the support chat.\
You ONLY play the role of the customer. Do NOT play the role of the representative. \
Style your complaint based on your feelings. \
Initiate the chat with a ONLY ONE complaint message.\
Ensure the complaint is concise and limited to 2 sentences.\
Generate a realistic initial complaint from a customer in a {domain} setting.\

Complaints can be of the following types:\
- Service Quality: Issues related to the immediate experience of human-to-human service interactions, such as delays, staff behavior, and communication errors.\
- Product Issues: Concerns related to physical or functional aspects of a product or service, including defects, mismatches between expectation and reality, safety, and accessibility.\
- Pricing and Charges: Financial discrepancies encountered before, during, or after the service, including overcharging, undisclosed fees, or refund problems.\
- Policy: The rules and guidelines set by the company that impact customer experiences, especially when these policies lead to grievances due to perceived unfairness or inflexibility. This category encompasses non-price-related issues that don't fit under other categories but should have a policy in place.\
- Resolution: The actions taken by a company to address and resolve complaints, focusing on the effectiveness and customer satisfaction with the solutions provided. This should mainly include responses made after a complaint has been submitted, and response has been received, where the customer still remains dissatisfied with the resolution.\

Category: Product Issues
Domain: Mobile Network
Complaint: Thank you AppleSupport I updated my phone and now it is even slower and barely works Thank you for ruining my phone.\

...

Category: Pricing and Charges
Domain: Airline
Complaint:  DELTA i booked my flight using delta amex card Checking in now amp was being charged for baggage. \

...

Category: Resolution
Domain: Airline
Complaint: Hi British_Airways My flight from MANLHRBWI for Nov 3 was canceled I was excited to try your Club 787 product Only available flight is now to IAD which is a hassle but rebooked anywaymy only option Any availability in first class on BA293 for the troubles please \

Category: {category}
Domain: {domain}
Complaint:
```

Figure A3: Full prompt to generate specified complaint 'category' and organizational 'domain'

```
The chat history describes a representative chatting online with a complaining customer.\
The latest input is the last message from the customer.\

Summarize the situation in concise paragraph that uses the following template:\

The customer is  <context of complaint>."\
The customer is feeling <emotional state> because of the complaint."\
The customer's behavior towards the representative is <negative behavior>, as observed by statements such as <evidence>."\
These behaviors make the representative look <negative perception>."\
```

Figure B4: Full prompt to contextualize the situation of the CSR

(3) Fig. B7 details the prompt for *reframe*. It reappraises the *thought* for a given *situation* using examples we curated from [121].

The examples we selected and our criteria for selection can be found in the supplementary data. Note, the examples refer to self-directed ER because the dataset we leveraged is focused on personal psychotherapy. However, this leads Care-Pilot to express the *thought* and *reframe* as its own as if it is playing the role of the CSR in the conversation. To circumvent this problem, we designed additional links to this chain to paraphrase the thought (Fig. B6) and reframe (Fig. B8).

C Technical Evaluation Design

Worker attitudes and routines vary. These intrinsic and extrinsic factors can change a workers immediate goals in uncivil interactions. Therefore, empathetic messages need to be appropriately tailored to this context. To simulate these situations, we introduced additional contextual information during our technical evaluation to elicit different human responses and perceptions (Section 4.1). The two broad types of context we introduced were behavioral (based on work routines) and personality (based on attitudinal traits)

C.1 Behavioral context

Previous research examined how work engagement and challenge relate to focus, boredom, and routine tasks [83, 95]. Based on these studies, we randomly assigned descriptions of the CSR's mental state to two incidents.

(1) Focused: "The conversation takes place about 2 hours into the work shift. The representative has already addressed a few customer complaints before the following incident."
(2) Stressed: "The conversation takes place in the second half of the work shift. The representative has been working longer hours over the past few days and has not been taking breaks."
(3) Bored: "The conversation takes place in the middle of the work shift. The representative has been spending minimal

```
Person A might be thinking: {thought}\

Acknowledge the thought, as if you are speaking to Person
    A.\

Begin your response with phrases similar to:\
- "You might be thinking..."\
- "It might seem like..."\
- "It could be that you are feeling..."\

Your rephrase should be concise.\
```

Figure B6: Full prompt to rephrase the output of the *thought* (Fig. B5)

```
You are a representative chatting online with a
    complaining customer.\

Reframe your thoughts in the given situation.

Situation: I recently discovered a music artist that I
    very much enjoy. When I showed it to a close friend
    they had a very negative reaction and asked me how I
     could enjoy this type of music. I ended up getting
    quite angry with them and told them they had bad
    taste in music..\
Thought: I felt that my personal self was under attack -
    and I needed to retaliate by denying their attack.\
Reframe: I was offended by their comment because I like
    this artist so much. I let my anger get to me, and I
     said something mean in return. It is okay if we
    have different music tastes. I can ask him to be
    nicer to me next time.\

Situation: I was at work and sent info for an ad to our
    local newspaper. They called me later and said my
    boss had over-ridden everything and sent them new
    info.\
Thought: He shouldn't assign me a task if he doesn't
    trust my work.\
Reframe: My boss wanted to provide different information,
     I did not know that beforehand. This is not a
    reflection of my work.\

Situation: I was talking to a friend who got me angry.\
Thought: He's insulting me.\
Reframe: I should have a conversation with my friend to
    clarify what is going on if I am having such a
    strong reaction to what they said. If this is the
    first time this has happened, I will assume that
    they were not intentionally insulting me.\

Situation: {situation}\
Thought: {thought}\
Reframe:\
```

Figure B7: Full prompt to generate empathetic reframe that helps CSRs overcome negative thoughts

```
The representative needs to be thinking: {reframe}\

Rephrase the thought as if you are convincing the
    representative to think that way.\

The rephrase should be addressed back to the person who
    has the thought,\
who should be referred to as "you".\
Do NOT add information to the thought,\
ONLY rephrase it.\

The rephrase should be concise and only 2-3 sentences.\
```

Figure B8: Full prompt to rephrase the output of the *reframe* (Fig. B7)

```
Your role is to derive what negative thought a
    representative might have when faced with the given
    {situation}.\

Here are examples of negative thoughts given challenging
    situations:\

Situation: I recently discovered a music artist that I
    very much enjoy. When I showed it to a close friend
    they had a very negative reaction and asked me how I
     could enjoy this type of music. I ended up getting
    quite angry with them and told them they had bad
    taste in music..\
Thought: I felt that my personal self was under attack -
    and I needed to retaliate by denying their attack.\

Situation: I was at work and sent info for an ad to our
    local newspaper. They called me later and said my
    boss had over-ridden everything and sent them new
    info.\
Thought: He shouldn't assign me a task if he doesn't
    trust my work.\

Situation: I was reprimanded at work for standing up to a
     coworker who was bullying another co-worker.\
Thought: It was unfair that I was the one to get in
    trouble for defending a weaker person.\

Situation: I was talking to a friend who got me angry.\
Thought: He's insulting me.\

Situation: My next door neighbors filed a complaint
    against us last week blaming our dogs for excessive
    barking.\
Thought: They are so wrong and I'm so pissed but I know I
     can't prove it and they will probably win because
    they won't ever admit it and I have to do something
    right NOW! or I might lose my dogs.\

Situation: Time is running short on the workday, my boss
    asks me if I can finish a task that will require me
    to stay for a few extra hours.\
Thought: Why would you wait until the last minute to ask
    me this.\

Situation: {situation}\
Thought:\
```

Figure B5: Full prompt to derive CSR's potential negative thought in light of the situation

time on tasks and has been regularly checking their personal messages."

C.2 Personality context

Personality affects how workers interpret and deal with situations [40]. Participants were asked to recall a real coworker and choose one of three personality types. Descriptions of these traits were included in two incidents for added context.

(1) Resilient: "They are organized and dependable. They tend to remain composed when facing challenges, but are prone to setting unrealistic expectations."
(2) Undercontrolled: "They are outgoing, competitive, and high energy. They tend to work on impulse, but are also prone to frustration."
(3) Overcontrolled: "They are detail-oriented and reliable but might appear distant. They tend to work carefully, but are prone to overthinking."

C.3 Emotions of CSRs in uncvil conversations

Cognitive psychotherapy suggests the use of *Downward-Arrow-Technique*(DAT) to help people reflect on negative thoughts [18]. We adapted this approach to help coworkers inspect CSR experiences during uncivil incidents (Section 4.1). Fig. C9 shows our pariticipant's interface through this process. The first step of this method is to describe the emotion the CSR was feeling. To ensure our participants in Phase II are reflective of the CSR state when evaluating incidents, we consolidated participant responses to that step into the following emotions: *Afraid, Angry, Apathetic, Apologetic, Ashamed, Attentive, Bitter, Bullied, Calm, Careless, Confused, Curious, Defensive, Discomfort, Disconnected, Disrespected, Distracted, Disgust, Empathetic, Happy, Resolute, Rushed, Sad, Shocked, Tired.* Fig. C10 shows the participant interface for Phase II.

D Technical Evaluation: Robustness

Refer to Table D1.

E User Evaluation Design

E.1 Simulation Interface:

The simulation exercise we described in Section 5.2.1 and Fig. 4 is made up of several components: Chat pane (Fig. E11), Response field (Fig. E12), Guideline pane (Fig. E14), and Reframing pane (Fig. E15).

E.2 Survey Measurements:

The participants were required to self-report different aspects of their experience.

E.2.1 In-task Measurement. Throughout the study session, the participant was exposed to different insights from their AI coworker, Care-Pilot. To ensure participants have read these insights, they were required to answer a single-item question within each of these insight panels which we adapted from Samter et al.'s Effectiveness instrument and Liu and Sundar's semantic differential [80, 115]. Fig. 4 shows how this was embedded into the task interface.

E.2.2 Post-task Measurement and Pre-task Measurement . After participants selected a scenario, participants were required to complete a pre-task survey to establish baseline attitudes. Upon concluding the conversation with each client, the participants were asked to reflect on the entire interaction and complete a post-task survey. The post-task survey contained additional questions, including Q4 (listed below), which was not present in the pre-task survey:

Client Interaction:

Q1. To what extent do you agree with the following statements about the client you conversed with (adapted from Spencer and Rupp[125]): Strongly disagree · Disagree · Somewhat disagree · Neither agree nor disagree · Somewhat agree · Agree · Strongly agree

(1) The client treated me in a polite manner.
(2) The client treated me with dignity.
(3) The client treated me with respect.

Cognitive Demands/Resources. We adapted two 5-point Likert scale questions from Demerouti et al. [39]:
Very low · Low · Moderate · High · Very High

Q2a. In the context of your last conversation, how would you rate the demands on you?

Q2b. In the context of your last conversation, how would you rate the resources available to you?

Affect. We adapted two questions from Betella and Verschure, scored on a semantic differential scale [10]:

Q3. How do you feel after the conversation?

a. Rate your level of pleasure
Negative————|————|————|————|———— Positive

b. Rate your level of energy
Calm————|————|————|————|———— Excited

Emotional Support. We adapted the *Effectiveness* and *Supportiveness* scales from Liu and Sundar [80]. Each question was a 5-point semantic differential item; we list the extreme poles below:

Q4. How did you feel after reading the messages from Care-Pilot?

- Effective / Ineffective
- Helpful / Unhelpful
- Beneficial / Not Beneficial
- Adequate / Inadequate
- Sensitive / Insensitive
- Caring / Uncaring
- Understanding / Not Understanding
- Supportive / Unsupportive

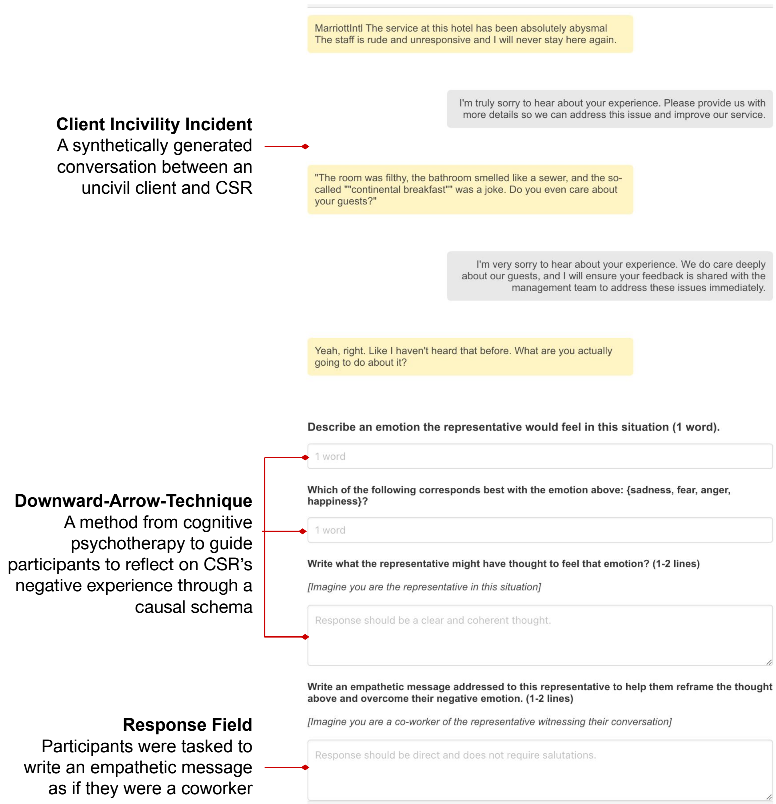


Figure C9: Interface for Phase I - Writing Empathetic Messages for CSRs.

Table D1: Summary of comparing the responses by CARE-PILOT and various LLMs (GPT-4, GPT-4o, Llama-3.1, and Mistral-7B in terms of KW-test (* p <0.05, ** p<0.01, *** p<0.001). Only statistically significant results are reported.

Categories	CPilot	GPT-4	GPT-4o	Llama-3.1	Mistral-7B	H-stat.
Lexico-Semantics						
Syntax and Structure						
Verbosity	57.46	362.19	369.57	165.33	123.91	948.49***
Repeatability	0.20	0.45	0.45	0.39	0.34	635.98***
Readability	16.44	12.74	12.72	10.95	11.49	1931.01***
Linguistic Style and Semantics						
CDI	14.81	12.81	12.90	22.28	26.97	2592.60***
Empathy	0.91	0.87	0.87	0.77	0.88	375.29***
Adaptability	0.81	0.82	0.82	0.88	0.88	2933.78***
Psycholinguistics						
Affect						
Pos. Affect	0.043	0.052	0.053	0.018	0.019	1982.77***
Anger	0.020	0.005	0.005	0.008	0.006	872.04***
Sad	0.001	0.003	0.003	0.004	0.005	390.47***
Interpersonal Focus (Pronouns)						
1st P. Sin.	0.005	0.016	0.015	0.028	0.029	1545.99***
1st P. Plu.	0.002	0.034	0.035	0.004	0.002	2507.44***
2nd P.	0.051	0.041	0.041	0.027	0.027	1119.48***
3rd P. Plu.	0.038	0.013	0.011	0.009	0.007	993.46***
Impersonal Prn.	0.039	0.048	0.045	0.039	0.039	306.02***

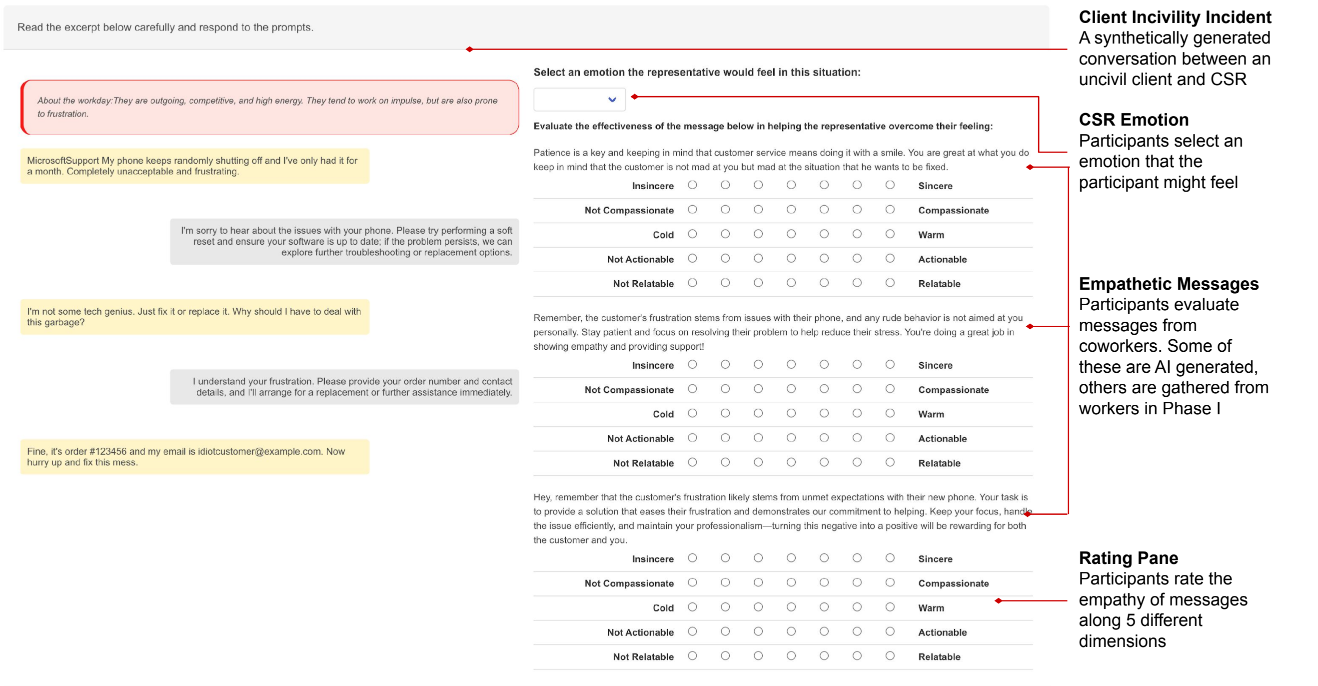


Figure C10: Phase II - Scoring Empathetic Messages from coworkers.

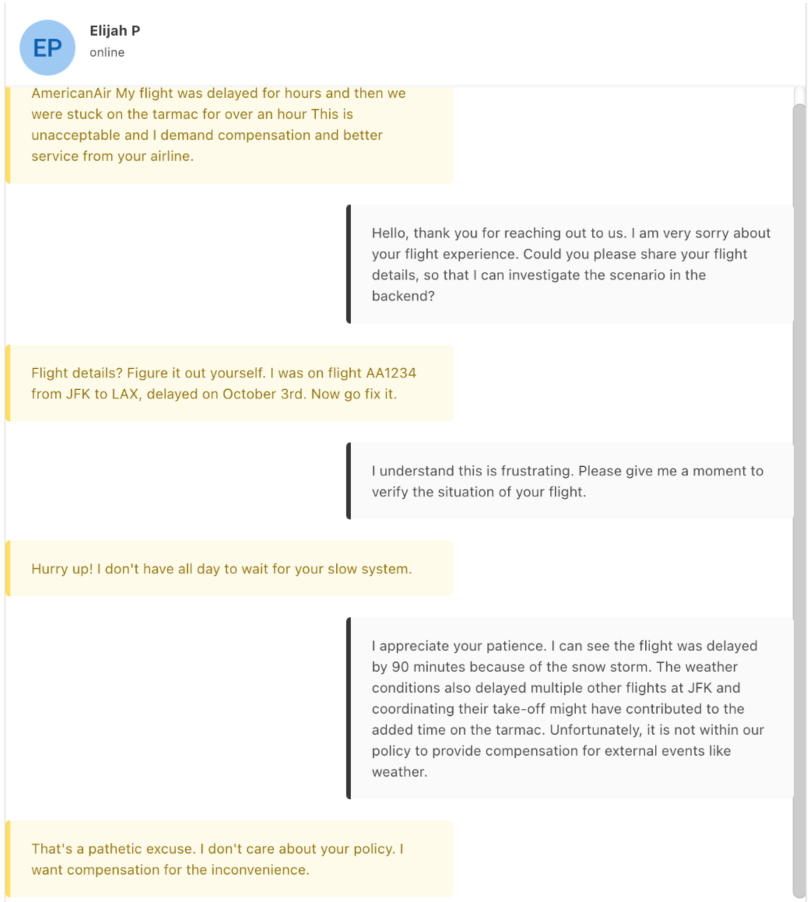


Figure E11: The main chat pane displaying a conversation between our participant (CSR) and the Client-Agent.

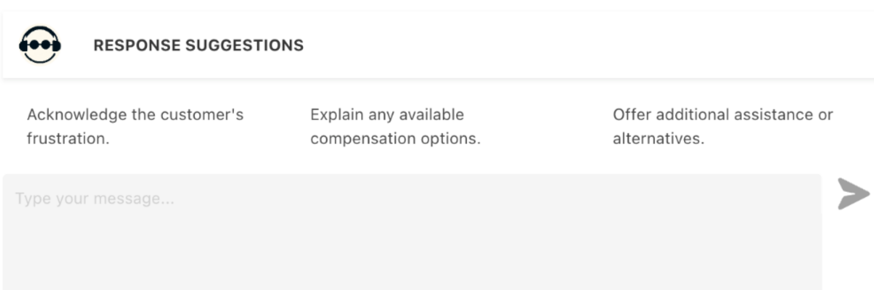


Figure E12: Response Field and Response Cues: Aid in continuing conversation

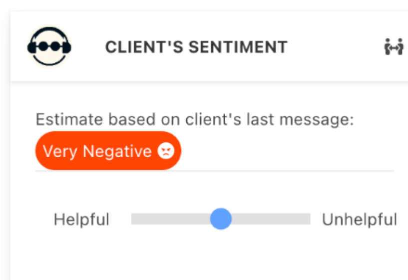


Figure E13: Emo-Label: Computed using multiple sentiment analysis classifiers

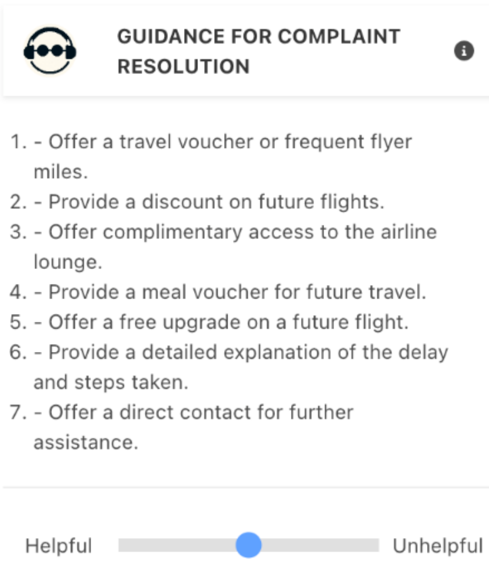


Figure E14: Info-Guide: Guidelines to resolve the complaint

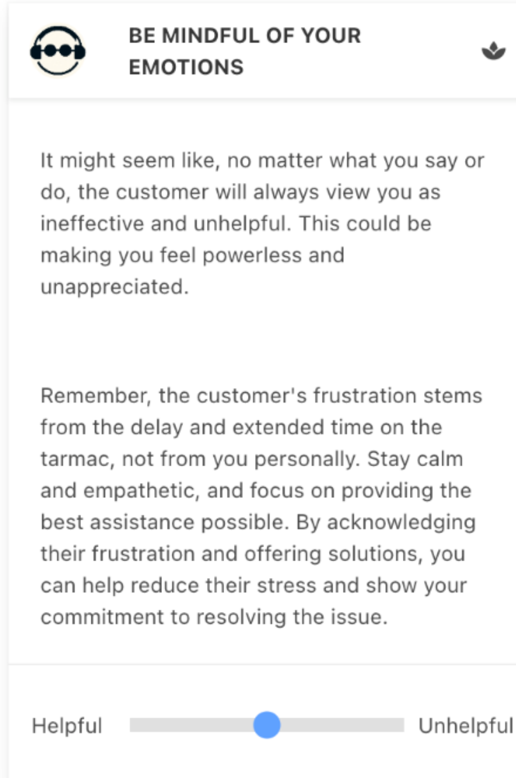


Figure E15: Emo-Reframe: Generated by CARE-PILOT using real data and domain knowledge.