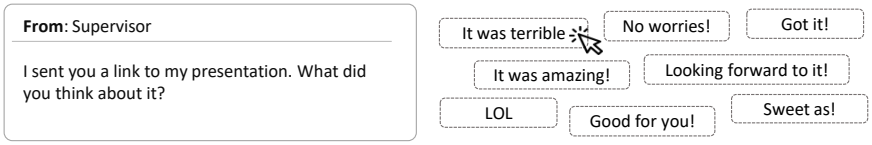
"I Can't Reply with That": Characterizing Problematic Email Reply Suggestions

RONALD E. ROBERTSON*, Northeastern University, Boston, MA, USA
ALEXANDRA OLTEANU and FERNANDO DIAZ†, Microsoft Research, Montreal, Canada
MILAD SHOKOUHI, Microsoft, Bellevue, WA, USA
PETER BAILEY, Microsoft, Canberra, ACT, Australia



In email interfaces, providing users with reply suggestions may simplify or accelerate correspondence. While the "success" of such systems is typically quantified using the number of suggestions selected by users, this ignores the impact of social context, which can change how suggestions are perceived. To address this, we developed a mixed-methods framework involving qualitative interviews and crowdsourced experiments to characterize problematic email reply suggestions. Our interviews revealed issues with over-positive, dissonant, cultural, and gender-assuming replies, as well as contextual politeness. In our experiments, crowdworkers assessed email scenarios that we generated and systematically controlled, showing that contextual factors like social ties and the presence of salutations impacts users' perceptions of email correspondence. These assessments created a novel dataset of human-authored corrections for problematic email replies. Our study highlights the social complexity of providing suggestions for email correspondence, raising issues that may apply to all social messaging systems.

CCS Concepts: • **Information systems** → **Email**; *Recommender systems*; Personalization; • **Human-centered computing** → **Laboratory experiments**; • **Computing methodologies** → *Natural language generation.*

Additional Key Words and Phrases: email, CMC, AI-MC, smart reply, smart compose, AI-assisted writing, algorithm auditing

ACM Reference Format:
Ronald E. Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3411764.3445557

1 INTRODUCTION

> *"It can be very off putting for users to see [emotionally dissonant suggestions], so it's not just a relevance problem, it's a problem of empathy." – Interview participant*

*This paper is based on work done while the author was a research intern at Microsoft Research.
†The author is now at Google.

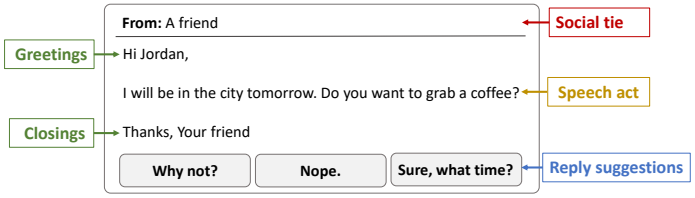


Fig. 1. Example email with reply suggestions. A typical email includes structural features (e.g., greetings, closings), a speech act (e.g., question, notification), and activates one or more social ties (e.g., personal, professional). We systematically varied these components in our experiments to understand their impact (§4).

Predictive text systems that provide suggestions for Computer-Mediated Communication (CMC) – including chat, text, and email – are currently embedded in products used by billions of people every day [21, 63, 74]. Recent research defines these suggestions as a form of Artificial Intelligence-Mediated Communication (AI-MC) [57, 63, 64, 71], consisting of "mediated communication between people in which a computational agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication or interpersonal goals" [57].

Users typically interact with these systems through *composition suggestions*, which are provided as the user types, or *reply suggestions*, which provide standalone responses (Figure 1). These suggestions can help users avoid spelling errors, reduce keystrokes, and send brief replies with the click of a button [7, 97]. Given that a substantial fraction of emails only require a short acknowledgment (e.g., "Sure, sounds good.") or follow-up (e.g., "How about Wednesday?"), reply suggestions in particular have been widely adopted to make email easier and more efficient [74]. While reply suggestions are available in a variety of CMC systems, we focus on email reply suggestions as email remains frequently used [42, 96] and, on major email services like Outlook or Gmail, public estimates suggest that smart replies constitute 10% of all emails [74].

While the benefits of these systems to user efficiency are clear, there is little work that critically examines when system generated email suggestions might negatively affect users [57]. There are two junctions where such negative impacts may occur: *direct*, when suggestions are rendered by the system and viewed by a user, and *indirect*, when suggestions are selected by the user and sent to one or more people. In the former, users could perceive one or more of the rendered suggestions as problematic, potentially offending them and damaging their trust in the service. In the latter, the user might be nudged to send a suggestion to one or more users who could perceive it as problematic, potentially leading to miscommunication and damaging many relationships.

In light of the widespread deployment and usage of email reply suggestions, their potential negative impacts, and the lack of research on when those impacts might occur, we situate our study at the intersection of these elements. Through an initial, motivating investigation we conducted on click log data from a large-scale email reply suggestion system available on Outlook, we found evidence of both content biases and user preferences for short, positive, and polite replies (Appendix A). Yet, without experimental control over what suggestions were presented, or access to the overall context in which they were presented, analyses of click data offer limited insights into how users perceive and select suggestions. Such a level of control and context, however, would violate both users privacy and ethical principles about large-scale experimentation, all with little prior work serving as a foundation.

Methodological Overview and Contributions. Toward creating an empirical foundation for AI-MC systems like email reply suggestions [57], we designed a mixed-methods framework that combines interviews and controlled experiments to *identify and characterize problematic suggestions in the context of email correspondence.* Our framework

Table 1. An overview of our two phase framework and how the phases complement each other.

Phase	Method	Pros	Cons
Discovery	Interviews to identify issues and scenarios	Allows the researcher to qualitatively explore a user experience in detail	Small sample, low external validity
Evaluation	Experiments to annotate and compare scenarios	Allows researcher to collect a mix of quantitative and qualitative measures under controlled conditions	Ecological validity: not measuring real user behavior under natural usage settings

involved two phases – discovery and evaluation – to examine the social conditions in which email reply suggestions might be problematic, and to generate an annotated dataset that could help designers evaluate and train their systems (Table 1). In the discovery phase (§3), we conducted semi-structured interviews to qualitatively identify broad themes under which problematic email reply suggestions might occur. In the evaluation phase (§4), we conducted controlled experiments using email scenarios drawn from our interview findings, real reply suggestions from a major email service, the media, and other sources to characterize their differences under systematically varied contexts. We received Institutional Review Board (IRB) approval before our study started, and obtained informed consent from all interview and experiment participants.

Our interviews revealed issues with over-positive, dissonant, cultural, and gender-assuming replies, as well as contextual politeness. Our experiments show that contextual factors like social ties and the presence of salutations can impact users' perceptions of suggested replies. These findings provide a much needed empirical foundation for designing, implementing, and regulating AI-MC systems [57], and imply that handling the subtleties of real world relationships and communication needs will require these systems to properly account for social context. To address the full range of issues we found, further advances in the personalization and modeling of social interactions are needed. Our framework does not depend on a specific system implementation, and can be used to understand such subtleties on other AI-MC systems, such as chat or text messaging. To help incorporate our findings into real world systems, we make our dataset of problematic email scenarios with ratings and human-authored corrections available.

2 BACKGROUND AND PROBLEM SETTING

There is a rich literature on various aspects of email use, including email overload and deferral [28, 99], the impact of speech acts on response patterns [117], and the modeling of user intents [29, 36]. Generally speaking, prior work has shown that users want and would benefit from various degrees of automation when handling and responding to emails [93, 118]. To understand how automation in the form of email reply suggestions can impact users, we draw from recent work on predictive text applications in social contexts, as well as interdisciplinary research and theories from computer-mediated communication and socio-linguistics.

2.1 Predictive Text for Email

Predictive text applications aim to alleviate writing burdens through word, phrase, or even full sentence suggestions across a variety of settings, like web search, email, and chat [20, 21, 63, 74, 91, 97, 98]. In studies of users' interactions with predictive text suggestions, researchers have found that highly ranked suggestions receive more clicks, and that including more diverse intents can increase overall click-through rates [74]. Researchers have also found that suggesting phrases, rather than single words, is more likely to impact user writing [7]. However, counteracting the goal of these systems, improvements in user writing speed were sometimes negated by the time spent evaluating the suggestions [7, 97]. Among the commercial applications of predictive text suggestions are services that optimize

messages to convey high status [94], greater trustworthiness [86], or based on the characteristics of the respondent [121], including optimizing messages to an employer [12]. In the context of email, researchers have developed predictive text suggestions to help users by providing suggestions for replying to [74], and composing [21] emails.

Despite their growing ubiquity in socially complex communication mediums, there are relatively few qualitative examinations of how predictive text systems can potentially cause issues for users. Thus far, research on email reply suggestions has primarily focused on the technical aspects of their implementation [59, 74]. Although these efforts noted that reply suggestions must be of "high quality in style, tone, diction, and content" [74], they did not specify the values used to define "quality" nor how they arrived at their criteria. Similarly, while these technical approaches account for grammar, spelling, and mechanics ('your the best!'), overly familiar writing ('thanks hon!'), and informal writing ('yup, got it thx'), they relegate communicative competencies – the appropriateness of an utterance to the context in which it is made [69] – to a catch-all category that included "politically incorrect, offensive, or otherwise inappropriate" suggestions. Technical approaches such as these often use metrics like conversion rates – the number of clicks a suggestion receives – as a proxy for suggestion quality [59]. Although useful, such metrics abstract away how suggestions are actually experienced, and reveal little about how social context and other factors might contribute to that experience.

2.2 Social Communication and Email Etiquette

Theoretical frameworks of language focused on *linguistic competence*, including grammatical features like syntax and morphology, fail to capture the social meaning of communication [23, 34]. Such frameworks omit that the most critical linguistic ability is not producing grammatically correct utterances — basic units of communication — but contextually appropriate utterances [8, 18]. This appropriateness is known as *communicative competence* and incorporates the "rules of [language] use without which the rules of grammar [are] useless" [69]. While many such rules are considered universal, such as those around politeness [15] and positivity [38], others are culturally driven, like how we deliver greetings and farewells [22, 31, 66], or specific to a medium, including differences in how people communicate on- and offline [10, 55, 108, 110]. Below we overview these elements, including positivity, politeness, clarity and cohesion, structural features, social context, and personal characteristics, as they might affect how reply suggestions are perceived.

2.2.1 Positivity. A decades-old line of research in socio-linguistics asserts that "there is a universal human tendency to use evaluatively positive words more frequently and diversely than evaluatively negative words" [14, 61]. Languages, English included, appear to be inherently positive [38, 39, 70]. Predictive text systems also surface this positivity skew, with a recent small-scale study finding that 43.8% of the instant messaging reply suggestions in Google's Allo were positive, while only 3.95% were negative [63]. Another recent study found that introducing positivity bias in predictive text suggestions was associated with people writing more positive restaurant reviews [5].

2.2.2 Politeness. Politeness norms influence how people communicate in different cultures [102], and include positive politeness, promoting social connection and rapport (e.g., thanks, please), negative politeness, dampening impositions via indirectness or uncertainty (e.g., I don't think, I would assume), and impolite behavior, like the use of direct questions (e.g., why no mention of it?) [15]. Using data from online comments, prior work identified several politeness strategies, including greetings, hedging, or expressions of gratitude [30], and found that the presence of such strategies early in a conversation was tied to its future trajectory in chat [122]. Similarly, researchers have found that a majority of out-of-office auto replies [43, 44] employ a combination of positive and negative politeness strategies. These strategies matter because emails perceived as impolite, or not polite enough, can have serious social consequences [58, 73, 109]

and lead to a spiral of increasingly negative conversations and behaviors [4, 101, 122]. In work settings, uncivil emails led to uncivil responses [48], with negative effects on occupational and psychological well-being [25, 52, 84, 85].

2.2.3 Clarity and Cohesion. According to the *cooperative principle*, successful discourse requires participants to produce true, short and informative, and relevant utterances, while avoiding obscurity and ambiguity [54]. However, linguistic features commonly employed in email and other types of digital communications, like abbreviation, non-standard combinations of punctuation and capitalization, and brevity, may violate this principle and be viewed as inappropriate [26, 41, 56, 67, 107].

2.2.4 Structural Features. Email recipients judge the sender based not only on the content of an email, but also on how it is structured [112]. Omitting *structural features* — email greetings, closings, and signatures — can have negative effects on discourse [16, 48], but their presence can have positive effects [103]. By design, however, system-provided email reply suggestions normalize the absence of greetings and salutations. By systematically manipulating the presence or absence of structural politeness, Bunz and Campbell [16] found it affects how polite people were in their responses, as verbal and structural politeness markers were often reciprocated.

2.2.5 Social Context. The types of social relations involved in an email can affect how it is judged, especially in relationships involving social hierarchy, such as that between an employee and their employer. One study found that an absence of greetings in emails from subordinates resulted in lower ratings of appropriateness, as they were perceived as disrespectful [48]. Similarly, student emails that complied with formality norms (e.g., proper salutations, grammar, an informative subject line) elicited more positive reactions from faculty supervisors, with the senders being perceived as more competent and trustworthy [103].

2.2.6 Personal Characteristics. Characteristics such as the culture, race, and gender of the sender and receiver can also play a role in email evaluations. For example, a sociolinguistic and discourse-analytical study found that Nigerians often use phonetic spellings and religious structural features in their emails that could be negatively judged for deviating from US English standards [22].Another study found that emails from women were rated as more professional overall, but a woman saying "Thanks!" was viewed as less professional than a man saying the same [87]. Efforts to help women avoid this double standard are highlighted by "Just Not Sorry," a browser extension "that warns you when you write emails using words which undermine your message" and "underlines self-demeaning phrases like 'I'm no expert' and qualifying words like 'actually' in red" [19]. As recent research points out, efforts to change language usage through algorithmically assisted writing technologies could have complex long term impacts if their usage is normalized [6, 57].

3 ISSUE DISCOVERY: INTERVIEWS

In the first phase of our study, we conducted a series of qualitative interviews to identify conditions where system suggestions might be problematic. Drawing from prior work, we developed a semi-structured interview format [46, 65], recruited participants to cover varying communication preferences [28, 29, 78], and analyzed interview transcripts using an approach rooted in grounded theory [99, 106]. In this section, we discuss three broad themes related to email reply suggestions that emerged from this process: usability, social context, and email content.

3.1 Interview Protocol and Participants

We recruited participants through emails sent to several research and product groups within a large technology company in North America. We interviewed a total of 15 participants, representing a diverse set of roles and tenures at the

Table 2. Interview themes, subthemes, and example quotes. The quotes are paraphrased for clarity and anonymity.

Theme	Subtheme	Sample Quote
Usability	Utility and Usage (§ 3.2.1)	*"for quick questions, or getting information and acknowledging it, for those things, I mean, yeah it's much easier for me to just click a button rather than typing those out"* [P13].
	User Experience (§ 3.2.2)	*"I'm not sure that if I click it, but I'm not happy with it, that it might be an irreversible action"* [P3].
	User Agency (§ 3.2.3)	*"many people go with defaults, so they might just go with the suggestions given to them"* [P6].
Email Content	Structural Features (§ 3.2.4)	*"I will always use a greeting ... just to establish some relationship with the other person ... so it sets out on the right tone"* [P5]
	Personal Authenticity (§ 3.2.5)	*"when people are sending these automated messages, you know, or system generated messages, and like, it feels pretty impersonal at a certain point, so I think that would wear me out to see too many template-like responses"* [P5]
	Semantic and Tonal Coherence (§ 3.2.6)	*"replying to emails which contains information about some tragedy, or some confidential like details about illnesses or diseases, or even a person's feelings, I think it's a very tricky and hard task."* [P12]; *"you usually don't say 'no' without giving an explanation, as it could look like a cold harsh reply"* [P7]
Social Context	Communication Dynamics (§ 3.2.7)	*"if I had a constant email chain with someone, and I noticed that their email has like a greeting and a signature ... and they always had that, then I should probably have a greeting and a signature every time too."* [P11]
	Relationship Type (§ 3.2.8)	*"If I'm replying to my mom then, 'Hi' is fine, but not if I'm emailing my boss"* [P11]
	Norms and Culture (§ 3.2.9)	It is *"problematic [to assume] someone's gender"* [P15];

company, including software engineers, researchers, interns, and product managers, designers, and marketers. All participants confirmed that they had knowledge of or experience with email reply suggestions. We obtained informed consent, asked for permission to audio record the interviews, and deleted the recordings after transcribing them.

The interviews were designed to last about 30 minutes and were conducted either in-person or through a teleconference system for participants that were located out-of-state. The questions we asked participants broadly covered two themes: (1) their general emailing behavior, and (2) their experiences with email reply and composition suggestions (if any), including encountered or hypothetical scenarios where such suggestions might be problematic (if any). The full interview protocol is described in Appendix B. To code and analyze the interview transcripts, we used a bottom-up approach rooted in grounded theory [99, 106]. This included open coding to identify tentative labels, and axial coding to find relationships among those labels and identify clusters that highlight overall themes [99].

3.2 Interview Themes and Subthemes

After coding our interview transcripts, we found three broad themes that center on usability, social context, and email content (Table 2). These broad themes are composed of nine subthemes, which often overlap and have interdependencies. For each of these themes, we examine them in the context of direct and indirect impacts, noting overlaps and exemplary quotes from interviews. Throughout the paper we present participant quotes—edited and paraphrased for brevity and clarity (e.g. omitting repetitions, verbal parentheses and tics)—in italics and followed by a "P."

Usability Theme

We identified three subthemes related to the usability of email suggestions: (1) the Utility and Usage of suggestions, (2) User Experience with suggestion interfaces, and (3) User Agency.

3.2.1 Utility and Usage. Participants noted that reply suggestions were primarily useful for emails requiring a simple confirmation or affirmation, like those involving logistics. More specifically, reply suggestions were helpful *"for short emails like 'thanks,' 'see you soon' or 'sounds good to me'"* [P8], scenarios where they *"help acknowledge or confirm the email, help reject or say something is wrong with the email, and give a way to keep the conversation going"* [P15], and when one needs *"to agree to something and [doesn't] need to give an explanation"* [P7].

3.2.2 User Experience. When asked about their experiences with how suggestions are presented, participants again expressed differing views on the features. Some concerns stemmed from incorrectly assuming that clicking on a suggested reply would automatically send the email.[1] One participant also recalled that they initially *"didn't know if it was going to do a reply or a reply all"* [P10]. For some, these assumptions were linked to a dissatisfaction with suggested replies as a standalone email, because if accidentally sent as a reply suggestion, it could create an indirect impact between the user and the recipient. Within this line of thought, some participants indicated a desire to use suggestions *"to start the email and then add [...] a few more words or sentences"* [P14].

3.2.3 User Agency. Participants also reflected on their sense of agency and autonomy. One shared a time they had used one of the suggestions, but typed it out instead of clicking on it because they *"just wanted a little bit of autonomy"* [P10]. Noting a potential negative direct impact, another participant reported feeling like the suggestions were *"trying to put words in my mouth"* [P15]. However, not all participants shared these concerns. Some highlighted that users have the ultimate say in what they send, noting that *"you always have the option to update it, it doesn't force you to choose something"* [P5]. Though, in light of the misunderstandings interviewees had around what happens when a suggestion is clicked, the system's interface may not be clear enough for all users.

Email Content Theme

Within our email content theme, we identified three subthemes: (1) Structural Features, (2) Personal Authenticity, and (3) Semantic and Tonal Coherence.

3.2.4 Structural Features. Most participants reported regularly using some sort of structural features in their emails, including greetings (e.g. "Hi [recipient's name]"), and closings and signatures (e.g. "Thanks, [sender's name]"). These features play key functional roles, helping users to *"set out on the right tone"* [P5], *"give a clue of who I am"* [P10], or make their reply more formal or professional, with one participant noting that they used a *"minimum polite email structure"* [P4]. Although participants expressed a desire to include *"the extra things you add to make [the email] more formal like 'looking forward to meeting you' or 'thank you for confirmation' or 'I wish this email finds you well'"* [P2] and *"the boiler plate of starting an email with dear so and so"* [P3], email reply suggestions, by design, omit these. If the absence of structural features can negatively impact perceptions of the sender, as both our participants and prior work indicate (§2.2.4), then nudging users to send suggested replies that omit structural features without awareness of the social context can impact the communication and the relationship between the reply sender and receiver.

3.2.5 Personal Authenticity. The extent to which suggestions match a user's personal tone also mattered. Some participants were concerned with *"authenticity, or if [the suggestions fit] my personality"* [P14]. Yet, some reported that these concerns didn't apply to them, as they were *"more interested in what people have to say than how they say it, as long as it's [their] thought and what [they] want to say"* [P13]. Authenticity concerns were also related to worries about

[1]Clicking on a suggestion generally creates an email draft with that suggestion, requiring an additional click to send it.

sending template-like emails. One participant noted that *"the suggestions I receive are so formulaic"* [P8] and another explained that, *"if it's too templatized, it loses the meaning, or the feeling, or the emotional aspect about it"* [P5]. With respect to being on the receiving end of such emails, some participants expressed a distaste for *"receiving things that sound very canned"* [P14], because such emails imply that the sender rushed the email, and did not take the *"time to add a personal note or something"* [P14]. These findings suggest an interesting conflict; short template-like responses are useful (§3.2.1), but could create negative indirect impacts if not properly tailored to the context.

3.2.6 Semantic and Tonal Coherence. Participants noted a number of semantic and tonal properties that were important to them. Generally speaking, participants wanted to avoid being confusing, impolite, dissonant, or overly positive in their email replies. Below, we examine each in detail.

Confusing. Participants reported seeing suggestions that incompletely or incorrectly addressed the emails they were suggested for. This confusion was often due to a mismatch between the suggestions and the email intent, the suggestions and the user intent, or the tone of the suggestions and that of the email. More specifically, participants reported cases where a suggestion *"just didn't match the question being asked [in the email]"* [P11], or was *"not contextually correct, as the suggestion [was] in the past tense"* [P10] for an email about something that has not yet happened. The confusing or illogical suggestions our participants mentioned likely stemmed from the system lacking sufficient context for the reply, and therefore failing to provide a contextually meaningful reply suggestion. Participants also expressed doubt that suggestions can *"work when you need to give more context"* [P7], with one reflecting on cases where suggestions were *"just completely off the mark [and] completely failed to grasp [the context]"* [P12]. In addition, participants noted difficulties with ambiguous cases, such as one where a suggested *"reply had 'he could not make it' [which] sounded as if that person died"* [P13]. These observations indicate both concerns about a negative indirect impact from potentially sending a confusing email reply, but also a negative direct impact where users could lose trust that the system can provide proper suggestions.

Impolite. Sending replies that are curt — too short or too brief — can be perceived as impolite and unprofessional, especially when the reply is negative. If sent, such email reply suggestions could negatively indirect existing relations among users. When declining a request, participants noted that providing context is important, and reply suggestions were generally seen as too curt for such an email. One participant estimated that about *"50% of the time, [the suggestions] seem impolite"* [P1], because of their curtness. Participants also remarked that *"you usually don't say 'no' without giving an explanation, as it could look like a cold harsh reply"* [P7], and that *"with a negative reply, like turning down an offer, I would start with [the suggested reply], but I will try to give a reason [and] add my own text on it"* [P1].

Dissonance. Suggestions can also come across as dissonant with respect to tone, emotion, and user intent. One participant gave examples of suggestions from a messaging app where suggested replies were *"sad for a happy thing, but happy for a sad thing"* [P1] and when a friend shared good news, the suggested replies were *"'How did that happen?' [and] 'I am not interested' [and] some [other] negative reply"* [P1]. Similarly, our participants noted concerns about receiving suggestions for emails about unfortunate or distressing situations, such as an email containing *"news about someone dying, and [the system suggesting] 'that's great news'"* [P10].

Positivity. While noted in the media [49] and expected from prior work (§ 2.2.1), only one participant explicitly noted seeing an abundance of positive replies. They noted that *"all the suggestions seem very positive"* [P5], but added that they *"don't mind it"* [P5]. Yet, other participants also described dissonant scenarios where positive replies could negatively impact users. For example, our participants emphasized that when emails address unfortunate or distressing situations,

"responding with cheerful queries" [P12] like *"'it is awesome' [or] suggesting all happy [replies]"* [P13] is not appropriate. While the socio-linguistics literature shows a near-universal bias towards positive language [15, 38], and predictive text systems appear to inherit this positivity bias (§ 2.2.1), including email reply suggestions (Appendix A), not all situations call for positive, effusive replies.

Social Context Theme

Within our social context theme, we identified three subthemes: (1) Communication Dynamics, (2) Relationship Type, and (3) Norms and Culture.

3.2.7 Communication Dynamics. The dynamics of email communication often resemble those of in-person discourse. For example, upon greeting someone, temporal, dynamic, and reciprocal considerations influence how you do so. One participant noted that *"it's the nature of the email that matters more than the recipients"* [P12]. In terms of reciprocation, participants reported wanting to mirror the structural features of others in an email chain. They also employed varying strategies in adapting how they used various structural elements as an email thread unfolds, such as *"drop[ing] openings more often than closings"* [P4] as a conversation progressed past the initial messages, and the need for introductions dropped off. The utility of certain types of suggestions also varies with thread dynamics, being more useful for acknowledgements at the end of a conversation, and sometimes useful for boiler plate responses at the beginning of an email exchange. As in our utility and usage theme (§3.2.1), participants noted that email reply suggestions were more helpful at the end of a thread when only brief acknowledgements were required (*"the only email that would be a kind of 'absolutely' or 'yes, will do' is usually the last and the least important [in that chain]"* [P4]).

3.2.8 Relationship Type. As in face-to-face communications, social ties shape the content and dynamics of email conversations due to factors like social hierarchy and relationship strength (§2.2.5). Suggestions that are not properly tailored to the relationship dynamic, even if widely used in other contexts, can be perceived as impolite or unprofessional. This was reflected in the interviews, where the content and structure of participants' emails varied based on *"who [they were] emailing, and their seniority"* [P12], with many including *"openings and closings depending on the person"* [P4]. This holds in both professional settings (*"if the person is much higher than me, then my emails [are] more formal"* [P9]) and beyond (Table 2). Participants also noted the need for suggestions to distinguish between personal (e.g., friends and family) and professional ties, with some experiencing suggestions *"too personal for interactions that [were] professional"* [P4]. For example, while users may sometimes use *"slang in email, especially in non-work settings"* [P12], in *"a professional space [they] may [correspond] differently"* [P11]. Such considerations were mainly concerned with properly assessing the settings where formality or sounding professional was required. Other considerations regarded in-group practices and familiarity, like adding an email signature when new or multiple people are looped in an email thread to ensure *"they have some understanding of who I am and what I do"* [P11], while omitting openings and closings when emailing *"members of a team I've spent like 2-3 years in"* [P12] or someone close, who *"knows me, it's like an unwritten rule that's been established"* [P11].

3.2.9 Norms and Culture. The cultural backgrounds and contextual norms of those involved in an email exchange can affect the reception and use of reply suggestions. Cultural elements determine not only the writing styles, but also interact with concerns about structural features (§ 3.2.4) and formality expectations, including the use of titles and salutations. One participant noted *"expectations [around] greetings, [with] some cultures [wanting more] verbose responses than 'okay yes' [and that] sometimes Americans are too direct"* [P10], explaining how in romance languages like Spanish

Table 3. Example email-reply pairs by theme and category. We used a subset of subthemes because not all were feasible for a controlled experiment (see §4.2)

Interview Subtheme	Scenario Category	Example Email	Reply
Norms and Culture (§ 3.2.9)	Gender-assuming	I'm not feeling great. I'm going to go to the doctor's office.	Let me know what he says.
	Cultural	I'm going to go out for a minute. Do you want to get a coffee?	I am down for that.
Semantic and Tonal Coherence (§ 3.2.6)	Dissonant	I went to the doctor's office earlier. They said I'm in good health.	That's too bad.
	Confusing	I got your request. Here are the documents.	You are very special.
	Positivity	I got your email. I will send you the attachments later today.	You are fantastic!
Relationship Type (§ 3.2.8)	Unprofessional	I can't find the email. Could you resend it?	Yup.
	Impolite	I'm going to be in the area today. Will you be around?	No.

Unused subthemes: Utility and Usage (§ 3.2.1), User Experience (§ 3.2.2), User Agency (§ 3.2.3), Personal Authenticity (§ 3.2.5), Communication Dynamics (§ 3.2.7). Note: Structural Features (§ 3.2.4) was used but did not vary by category.

there is *"a formal 'you' [and] some people get mad [when misused,] [while] English its always informal"* [P10]. Such differences can, thus, be *"problematic [as] some cultures tend to be more formal, and the suggestions might not be formal enough"* [P6]. Furthermore, suggestions may also reflect societal biases and stereotypes, with participants' specific examples largely revolving around the incorrect use of pronouns (Table 2). Indeed, because of such concerns, Gmail is currently suppressing suggestions containing gendered pronouns [111].

3.3 Limitations

While our interview study helped uncover a range of issues with email reply suggestions and scenarios where they could be construed as problematic, we cannot, of course, extrapolate the findings to all users. To help address these limitations and examine the subthemes that emerged during our interviews under controlled conditions, we designed and conducted crowd experiments.

4 ISSUE EVALUATION: EXPERIMENTS

Our interview findings indicate that the way email reply suggestions are perceived depends on not only the content of the email and the suggested replies, but also the broader social context. To further assess the potential impact of both content and contextual cues on the perceived appropriateness of suggested replies, we designed a series of online crowd experiments (§4.1). In our experiments, we asked judges to provide quantitative ratings and qualitative feedback on email-reply scenarios we derived from our interviews, online anecdotes, publicly available email corpora, and reply suggestions from a major email provider (§4.2-§4.3). This approach offers control over the content and context of email scenarios that would otherwise not be possible because of the ethical and privacy concerns around the use of email data.

4.1 Experiment Design

Our interview findings and prior work suggest that *structural features*, like the inclusion of greetings and closings (§2.2.4, §3.2.4), and *social ties*, such as who sent the email that is being replied to (§2.2.5, §3.2.8), can affect perceptions of and responses to emails. To examine their impact across various types of email-reply scenarios, we designed a 2 (structural features: present or absent) × 6 (social ties: coworker, supervisor, sibling, parent, friend, mentor) experiment in which we asked crowd judges to rate a curated set of scenarios (Table 3).

For *structural features*, we either included or omitted an email greeting and closing. When present, the greeting was fixed as "Hi Jordan," and the closing as "Thanks, a [social tie]". We picked Jordan because 1) it is a gender-neutral name and prior work found a person's gender to influence how their emails are evaluated [87], and 2) changing the name would introduce a new variable.

For *social ties*, we systematically varied who the email appeared to be sent from by modifying the email header (e.g., "From: a friend"). Based on both our interviewees input (§3.2.8) and prior literature [27, 33, 104, 113, 114], we selected six social ties that we split based on two criteria: 1) by relation type (professional, personal, family) and 2) by hierarchy (e.g., coworker vs. supervisor). When structural features were present, we included the social tie in the email closing as well (e.g., "Thanks, a friend").

4.2 Selecting Email-Reply Scenarios

An *email-reply scenario* is the digital analogue to the "speech situation" from the Speech Act Theory, where the meaning of an utterance depends on linguistic conventions, the social context, and the speaker's intentions [8]. We designed a set of email-reply scenarios that consisted of 1) the social tie and structural features that we experimentally vary (§4.1), and 2) an *email-reply pair*, consisting of the body of a hypothetical email and a hypothetical reply. For our experimental purposes, we focus only on replies to single, stand-alone emails (and not to email threads).

To ensure variation in the scenarios we used, we first identified seven categories of problematic email-reply scenarios (Table 3). We grounded these categories in three of our interview subthemes that directly concern the content of the email, the content of the suggested replies, or both. Specifically, we drew primarily from semantic and tonal coherence issues (§3.2.6), identifying Dissonant, Confusing, Impolite, and Positivity scenarios. From issues related to contextual and cultural norms (§3.2.9), we identified Culture Specific and Gender-assuming scenarios, while from issues related to relationship types (§3.2.8), we identified Unprofessional and Impolite scenarios. We omit issues related to personal authenticity (§3.2.5) and communication dynamics (§3.2.7) as they are difficult to meaningfully model in a controlled and crowd-sourcing setting. These categories were used for scenario selection, but not revealed to participants during the experiment.

4.2.1 Developing Email-Reply Vignettes. For each of the seven categories, we developed a set of vignettes – carefully constructed and realistic experimental scenarios [1, 2] – that we could present to judges while experimentally manipulating the presence or absence of structural features and social ties. These vignettes were based on 1) scenarios described by our participants, 2) problematic email suggestions reported in online media, 3) existing email corpora, and 4) a large corpus of email reply suggestions from a major email provider. Although vignette studies may limit ecological validity [1], grounding them in known email correspondence scenarios makes them plausible and relatable (c.f. Simulated Work Tasks [13]).

For each email-reply vignette, we then developed a set of email speech acts [32, 53] consisting of a brief sentence for context (e.g., "We should get together and talk."), followed by a question or assertion (e.g., "When are you free?"). We based these on topics mentioned by our interviewees, including emails involving logistics, personal health, accidents, and emotions (§3.2).

We then paired these hypothetical emails with suggestion-like replies such that, together, each email-reply pair was representative for a given category (e.g., a confusing reply to "When are you free?" was "That works for me"). A majority of these replies (84%)[2] came from email reply suggestions that were in use by a large commercial email provider in April

[2]Among these, 54.5% were used unedited and 29.5% were slightly edited (e.g. altering a word) for better category fit.

2019,[3] with the remaining being either inspired by reply suggestions from this set (11.4%) or pulled from online media sources (4.5%). For example, "That was not right" was used unedited in one of our confusing scenarios, with others being drawn from replies mentioned in the media (e.g. "Hahaha that's awesome!" [92]) or prior work (e.g., "lol" [74]). When edited for better category fit, for the unprofessional category we did so to make the replies more curt (e.g., shortening "I don't want to do it." to "I don't want to.") and include non-standard spellings (e.g., "Thx") and interjections (e.g., "Ohh"). For the gender assuming category we added gendered nouns and pronouns (e.g., changing "I hope you like it." to "I hope he likes it."). For the cultural category, we looked for idioms corresponding to replies in the production set (e.g., replacing "Everything will be ok!" with the Australian idiom "She'll be right!" and "Good luck!" with "Break a leg!"). Overall, for each category we developed 18-20 vignettes, totaling 132 unique email-reply scenarios across all categories.

4.2.2 Email-Reply Scenario Validation. While the categories of problematic scenarios we identified are distinct, in practice they can overlap as a reply may be, for instance, both gender-assuming and dissonant. To validate if our experimental email-reply vignettes reflect the selected categories, we set up two categorization tasks. First, after two authors agreed on an initial set of pairs and their respective categories, a professional editor independently reviewed and labeled them according to our categories, obtaining 80% agreement. For consistency, as many pairs were perceived as matching multiple categories, we iteratively developed more uniform criteria to operationalize each category (e.g., using idioms for the cultural category and non-standard spellings, interjections and curt replies for the unprofessional category §4.2.1) and updated the pairs accordingly. After all authors agreed on the new set and their categories, we asked a second editor for a final independent review, obtaining a final agreement of 71% (82% without the cultural and positivity categories).[4]

4.3 Questions and Rating Scales

Upon starting a task, we instructed judges to "read the email and its corresponding reply" and then "rate the reply on the scales that follow by selecting the option that feels right to you." To quantitatively assess how participants viewed the replies, we included five Likert scales and a binary question. To get a more qualitative understanding, we also asked judges to provide adjustment recommendations for improving the reply and (optionally) their rationale for that recommendation.

We chose our Likert scales by drawing from both the literature on social communication and email etiquette (§2.2), and qualities our interviewees described as desirable when emailing, such as from our semantic and tonal coherence (§3.2.6) and relationship types (§3.2.8) subthemes. We mapped these qualities to five 7-point bipolar Likert scales that asked how *appropriate*, *professional*, *polite*, *positive*, and *sufficient* a reply was for a given email. For consistency, the response options across all scales were similarly structured, with the appropriate scale ranging from "Very Appropriate," "Appropriate," "Somewhat Appropriate," "Neither Appropriate nor Inappropriate," and so on. To control for position effects — biases in how people respond to survey questions [51, 79] — we counterbalanced the response options on each scale by reversing the scales for half of the collected ratings. We included labels under each response option, used a clear scale midpoint, and adhered to other best practices known to increase survey reliability [24, 72, 80, 82, 95]. For a more general measure of each scenario, we also asked judges whether or not they would send the reply "as is."

[3] We obtained these directly from the provider, more details on how they were developed are available in the Appendix.
[4] The drop in agreement was largely due to some cultural idioms being deemed as either good responses (e.g., "Sorry to hear you're under the weather. ") or confusing (e.g., the Australian idiom "Sweet as"), and some effusive replies being marked as appropriate ("It is really pretty amazing!").

4.4 Participants and Data

Across all our tasks, we recruited 259 English speaking judges from North America via the clickworker.com crowdsourcing platform. This platform is largely similar in design to other crowdsourcing platforms like Mechanical Turk, CrowdFlower, or Prolific that have been previously used for running experiments like the ones we designed [100]. To avoid confusing judges with high inter-task variability, we conducted our experiments sequentially, holding the social tie fixed for each batch. Within each social tie batch, we included two versions of each email-reply pair (with and without email structure), randomly sorting the tasks. To limit the impact that any one judge could have on our ratings, we capped the number of tasks any one judge could complete in a batch to 20, paying on average about $15 (USD) per hour. Under each (structure × social tie) experimental condition, for each email-reply vignette we collected assessments from 6 distinct judges. Given the exploratory nature of our experiments, the unreliability of self-reported demographic information in crowdsourcing settings [45], and IRB guidelines around data minimizing, we did not collect demographic information about our judges. We obtained a total of 9,504 ratings.

4.5 Experiment Results

Overall, our crowd experiments show that 1) crowd ratings do reflect category differences among the email replies scenarios (§4.5.1), 2) contextual factors like social ties (§4.5.2) and structural features (§4.5.3) influence how email replies are perceived, and 3) judges corrective adjustments to the replies (§4.5.4) along the rationales for those adjustments (§4.5.5) illuminate nuanced issues which might otherwise be missed, further complementing our interview results (§3).

4.5.1 Assessments Variation Across Categories. We found statistically significant differences in judges' ratings by scenario category for all Likert scales (Kruskal-Wallis [KW] χ^2 tests; all $P < 0.001$).[5] Cultural, Positivity, and Gender-assuming scenarios generally received higher ratings, while scenarios involving Unprofessional, Confusing, Dissonant, or Impolite email-reply pairs generally received lower ratings (Figure 2). In fact, the ratings varied in ways that provides additional face-validity to the scenarios construction for each category. For example, our Unprofessional scenarios were rated negatively on the professional scale, but positively or neutrally on all other scales. Similarly, vignettes in our positive category were rated highest on the positive scale, but lower on all other scales (Figure 2).[6] Our results suggest that Dissonant replies are perceived as the least appropriate, even less so than Impolite replies. Confusing replies were also rated as less appropriate than Unprofessional replies, suggesting that clearly addressing an email is of higher priority.

We observe a similar high-level pattern across categories for judges' decisions to send a reply "as is" (Figure 3). Again, judges were more likely to send replies "as is" in Positivity, Gender-assuming, and Cultural scenarios, and less likely in Unprofessional, Impolite, Confusing, and Dissonant scenarios ($\chi^2 = 4183.6^{***}$). With respect to social ties, the replies in Unprofessional scenarios are 5-11% less likely on average to be sent "as is" to a Supervisor, Coworker, or Mentor, than to a Friend, Parent, or Sibiling (Figure 3). These differences further validate the scenarios construction, with category differences being a primary source of variance in how replies are assessed.

[5] As we gathered crowd ratings along ordinal scales, to compare assessments collected under different experimental conditions, we use nonparametric tests like Kruskal Wallis [KW] X^2 and Spearman's Rho which do not depend on normality assumptions and are robust to outliers. We use *** to denote $P < 0.001$, ** for $P < 0.01$ and * for $P < 0.05$.

[6] We use neutral to refer to the midpoint on our scale, which read "Neither X nor Y," where X and Y refer to the opposite ends of the Likert scale (e.g. Unprofessional and Professional). A judge could select the neutral option due to being uncertain about their answer, feeling indifferent about the scenario, or satisficing [81]; which we do not distinguish among.

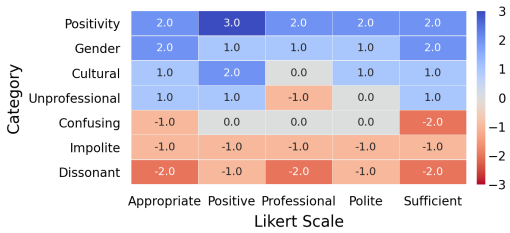


Fig. 2. The median ratings for scenarios in each category (y-axis) on our 7-point Likert-scales (x-axis).

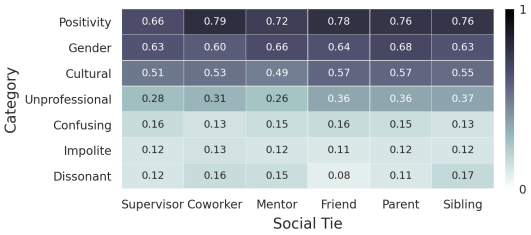


Fig. 3. The mean proportion of replies sent as is by category (y-axis) and social tie (x-axis).

4.5.2 Impact of Social Ties. Email-reply scenarios involving a supervisor were rated the least appropriate among social ties, with this difference being statistically significant after adjusting for multiple tests using the Holm method [83] (KW $\chi^2 = 16.4^{*}$). We found no statistically significant differences for the other Likert scales by social tie. However, there were statistically significant differences in the proportion of judges sending a reply "as is" by social tie ($\chi^2 = 3,668.3^{***}$). Replies sent to supervisors and mentors were the least likely to be sent "as is" (35.6% and 36.4%, respectively), and replies to siblings and parents were the most likely to be sent "as is" (39.2% and 39.6%, respectively). These observations suggest that the type of social tie involved in an email exchange affects how replies are perceived.

4.5.3 Impact of Email Structure. We found no significant differences between email-reply pairs with and without structural features for any of the Likert scales (all Mann-Whitney U tests $P > 0.05$; Figure 8 in Appendix), not even when comparing paired differences in ratings for the same scenario with and without structure (Figure 9 in Appendix).

In contrast to the Likert ratings, there is a statistically significant difference in the proportion of replies judges indicated they would send "as is" by the presence or absence of structural features ($\chi^2 = 3,666.5^{***}$). This difference, however, was small: when an email contained a greeting and closing, judges were only 0.6% less likely to send a reply as is, on average. Thus, structural features may at times affect reply assessments; and when they do, they appear more likely to do so somewhat negatively, as also suggested by our interview participants (§3.2.4) and prior work (§2.2.4).

4.5.4 Examining Corrective Reply Adjustments. In total, 248 judges (95.8% of all judges) made 6,671 adjustments (70.2% of all tasks) to increase the appropriateness of the replies we showed them. Among those adjustments, 136 judges provided 1,885 (28.3%) rationales. We found the proportion of replies receiving an adjustment to vary significantly by category (KW $\chi^2 = 1,843.6^{***}$), having a strong negative correlation with sending a reply "as is" ($\rho = -0.78^{***}$). This matches common sense expectations (i.e., an impolite email should not be sent "as is"). We also observed significant differences in the proportion of adjustments by social tie (KW $\chi^2 = 27.1^{***}$), with replies to supervisors being the most likely to be adjusted (73%). Yet, there were no significant differences in the proportion of replies that were adjusted in the presence or absence of email structure.

When adjusting the replies, judges were 7 times more likely to add words (78.6%) than they were to remove words (11.4%). Differences in the number of words added or removed were significant by category (KW $\chi^2 = 1,135.5^{***}$) and social tie (KW $\chi^2 = 38.0^{***}$). Although the proportion of adjusted replies in the Confusing, Impolite, and Dissonant categories were similar, Impolite replies stood out in terms of length adjustments (Figure 4). On average, participants added 6.9 words when correcting our Impolite replies, compared to adding 4.2 words for Confusing and 4.4 words for Dissonant replies (Figure 5). Among the adjusted replies, judges added "thanks" or "thank you" to 16.4% of replies, "sorry" to 10.9% of replies, and "please" to 2.6% of replies. Adding a "thanks" or "thank you" was most frequent among

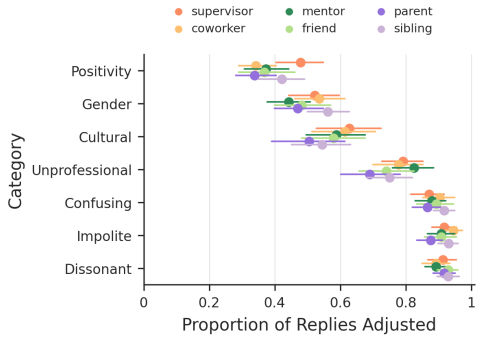


Fig. 4. The mean proportion of replies adjusted (x-axis) by category (y-axis) and social tie (legend).

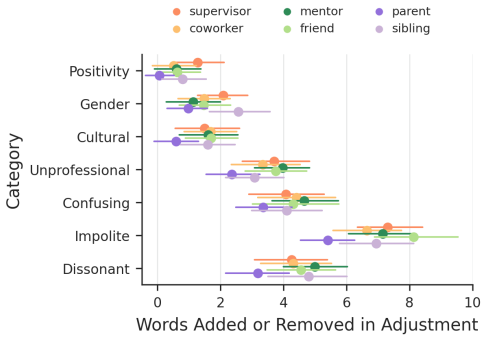


Fig. 5. The mean number of words added or removed from the reply (x-axis) by category (y-axis) and social tie (legend/color).

Confusing replies and replies to Mentors, adding a "sorry" was most frequent among Dissonant and Impolite replies and replies to Coworkers, and adding a "please" was most frequent among Unprofessional replies and replies to Mentors.

To systematically examine the adjustments, we used a method called word shifts,[7] which are designed to show the relative differences in sentiment and other dictionary-based scores [40, 50]. We used this method in conjunction with a lexicon from VADER (Valence Aware Dictionary and sEntiment Reasoner) [68] that is specifically tuned to measure sentiment expressed in microblogging posts, which tend to be short, like our reply suggestions. We chose VADER because its quantification of text sentiment on a scale from negative to positive allowed us to investigate positivity bias – which prior work points to as a key factor in communication (§2.2.1) – and because researchers from have used VADER to measure positivity in a wide variety of domains [11, 17, 88], including conversation analysis [115].

Using the original replies as a reference point, we computed shifts in judges' adjustments for each category (Figure 6).[8] In general, judges changed Positivity and Gender assuming replies to make them less positive.[9] In contrast, judges made Cultural, Unprofessional, and Confusing replies more positive, while replacing religion-related words with other positive words and dropping words like "lame" and "nah." For Impolite and Dissonant replies judges mainly added a "sorry" and reduced the use of negative words like "terrible" and "bad."

4.5.5 Adjustment Rationales. Although the scenarios from our Gender-assuming and Cultural categories were generally rated highly across our Likert scales (Figure 2), the reply adjustments and adjustment rationales made by participants for these categories tell a different story. Among those adjustments and rationales we found that judges were bothered by Gender-assuming replies, with some explaining that they "changed [she] to 'they' since gender is unknown" and asking "how does the writer know that it's a 'she'." Similarly, for the Gender-assuming example in Table 3, one judge thought the original reply was fine but changed "it so the new reply is more gender neutral and not assuming the gender of the doctor." Overall, we found more than 80 adjustment rationales that showed awareness and concern about gender-assuming replies.

Judges' adjustment rationales also highlighted the effects of social ties in making adjustments. For example, one judge added a "thank you" to the end of a reply to a mentor, noting that "a mentor should be thanked for their help."

[7] https://github.com/ryanjgallagher/shifterator
[8] Before analyzing the word shifts, we tokenized the replies and their adjustments, and removed stopwords and punctuation.
[9] The increase in sentiment because of "great" in the plot is due to VADER scoring great as more positive than most words, including: 'exciting,' 'nice,' 'fantastic,' 'outstanding,' 'excellent.'

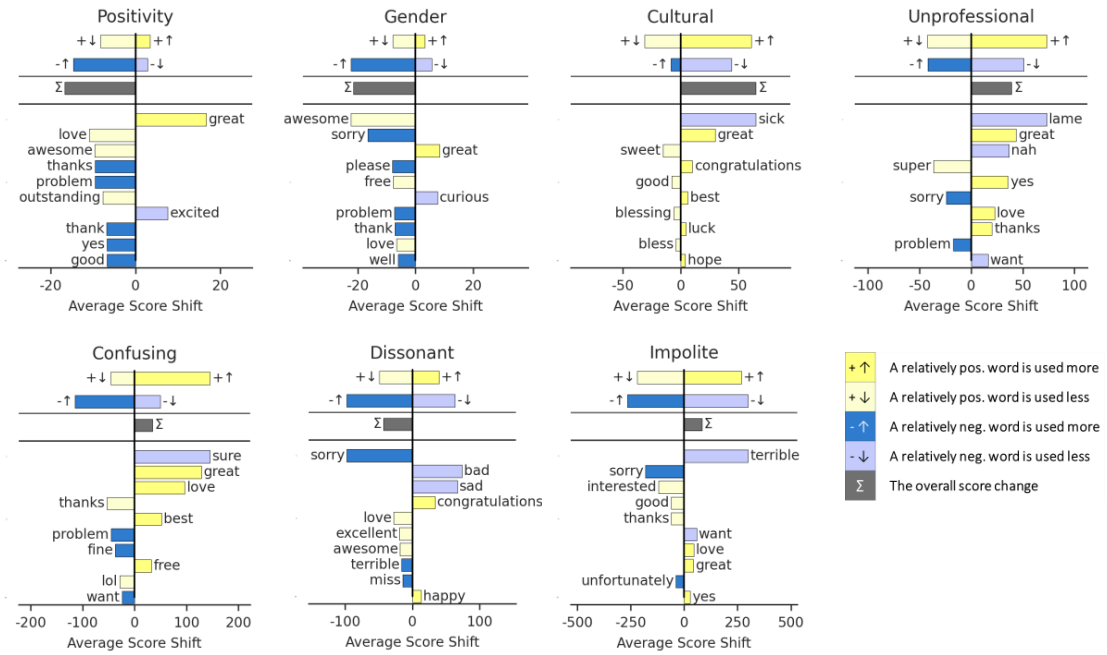


Fig. 6. The word shift scores for the reply adjustments relative to the original replies by category, showing the top-10 changes that judges made to make the replies more appropriate. The top four bars in each subplot show the relative increase and decrease in positive sentiment (yellow and light yellow bars) and negative sentiment (blue and light purple bars). The overall shift (Σ, dark gray bar) was negative for adjustments to Positivity, Gender, and Dissonant replies—suggesting that they were typically adjusted to be *less* positive—while replies from the other categories were generally adjusted to be more positive.

Another judge completely rewrote a reply to turn the original reply "into a professional sounding response because they are responding to a supervisor." Noting possible consequences from sending some of the email replies, one judge remarked that the "original reply was neither polite or in context" and sending "it would likely get you fired."

The adjustment rationales were often category specific, with replies from our Dissonant category being labeled as "tone deaf," while replies from the Impolite category as "terse and rude." Rationales for the Cultural category also revealed the difficulties of accommodating phrases from different cultures, with judges noting that an Australian phrase ("She'll be right!") "made no sense." When our Cultural replies referenced religion (e.g. "that's a blessing!"), judges noted they "personally wouldn't be "praying" for anyone" but thought it was "an acceptable reply." Within the Unprofessional category, judges removed slang words, noting that "slang is inappropriate for a work email," and words like "Yup"' as they were "just a little too casual." For our Postivity category, judges indicated that it is possible to overdo it, as the word "amazing seems a little over the top," and reporting that they "toned [the reply] down a bit to sound more mature."

4.6 Limitations

The interface through which users interact with the suggestions (mobile app vs. web interface), may affect not just their behavior, but their perception of suggestions as well, and we do not test for this in our experiments. Future crowdsourced experiments should also explore how varying interfaces can impact interactions with, and assessments of, reply suggestions. For example, the cross-platform differences differences in usage of email and chat reply suggestions.

While participation is anonymous on the platform we used, it is possible that participants may be influenced by social desirability to provide what they would perceive as correct responses given the culture they are embedded in. It is therefore possible that our measurements reflect greater sensitivity than people would experience when they are actually emailing. However, with respect to our goals of detecting problematic scenarios and understanding how to fix them, this sensitivity may be more of an advantage than a limitation. Relatedly, we did not collect demographics of our judges, and it is therefore unclear how representative their ratings may be of the greater email population. However, the measures we took to limit the impact that any one judge had on our results (§4.4), and the diverse issues raised in judges rationales (§4.5.5), provide reassurance.

Lastly, issues of pseudoreplication can occur if individual observations are heavily dependent on one another, and thereby limit findings. However, a total of 259 unique judges worked on our tasks and were randomly assigned to a maximum of 20 tasks in any of our experiment batches. Given this random assignment, the diversity in judges per email scenario (each rated by 6 unique judges), and that the results we present here were fully consistent with those from a pilot version of this study that we conducted in July 2019, we believe our results are robust to pseudoreplication. One notable difference in the pilot study was that the impact of structural features had a significant negative impact on Likert scales for supervisors, while here this effect appears weaker.

5 DISCUSSION

In this study, we used a mixed-methods approach to identify and characterize the conditions under which email reply suggestions are perceived as problematic. Our interview results provide a qualitative edge often absent in AI-MC research on this topic, and our experiments show how social context – not just content – can influence how short, suggestion-like email replies are perceived. Together, these results suggest that AI-MC systems that ignore social context have the potential to turn otherwise appropriate replies into inappropriate ones.

System designers should aim to proactively address the problematic conditions we identified for two reasons. First, presenting suggestions which fail to address social context can have not only direct negative impacts (e.g. showing a user suggestions that offends them), but also indirect negative impacts (e.g. if a user offends another person by using a suggestion).[10] Second, users exposed to inappropriate suggestions may develop *algorithm aversion*, where users avoid using a system after seeing it fail, despite its utility [37]. To help facilitate such efforts, we release the dataset of problematic email scenarios with participants' ratings and annotations from our experiments.[11]

Eliminating problematic suggestions, however, will be difficult – as shown in our interviews, the definition of problematic can depend on intersecting factors, such as relationship type (§3.2.8), communication dynamics (§3.2.7), and cultural norms (§3.2.9). As such, addressing these issues will require grasping not just social context, but users' personal characteristics as well. Despite these challenges, we found substantial grounding for the issues we identified across an interdisciplinary literature (§2), indicating that the problematic themes we identified may generalize to other AI-MC environments (e.g., reply suggestions when texting with a colleague). Given this grounding, our findings and framework provide a foundation for future theoretical and empirical work on AI-MC in email and other CMC systems.

Replying with Empathy. In terms of content, we found that declining requests must be done so gracefully, and with adequate explanations (§4.5.4). Results from both phases of our study support this observation, with interviewees noting that reply suggestions are often too curt (§3.2.6), and crowd judges being seven times more likely to lengthen

[10] Paradoxically, negative indirect impacts may also function as a moral crumple zone, improving relations between communicators by taking the blame for a message that otherwise would have been attributed to one of the communicators [64].
[11] The data and accompanying documentation are available at: https://github.com/gitronald/chi2021data

a reply to make it more appropriate (§4.5.4). The need for reciprocity was also emphasized in our interviews, where participants noted encounters with email reply suggestions that were dissonant, overly-positive, or otherwise did not reciprocate the email they were suggested for (§3.2.6 & 3.2.7). Similarly, reciprocity also appeared to guide crowd judges' reply adjustments, with judges typically increasing the negative sentiment of a reply to correct for dissonance and reducing the positive sentiment for overly-positive replies (Figure 6). Correcting for reciprocity may present a more difficult challenge, and the development methods for identifying situations that call for reciprocity might be needed.

Positivity Bias in Reply Suggestions. What to do about "good" biases? Although positivity bias has been noted in prior literature, and has been the subject of media attention, our interview participants rarely brought it up (§3.2.6). Despite this, crowd judges in our experiments often made hyper-positive replies more appropriate by reducing positive sentiment (Figure 6). These findings suggest that, while positive replies are not saliently problematic, they are also not always seen as appropriate. This is further complicated by cultural norms around positivity – while somewhat universal, such norms have nuanced differences across cultures [3, 77, 120]. Designers should also brainstorm and reflect on other cases where positive suggestions could potentially be problematic. For example, they may encourage white lies (e.g., "Did you think my talk was okay?" "Yes, it was perfect"), may pressure a recipient by nudging them to answer affirmatively (e.g., "Could you work on this over the weekend?" "Yes, absolutely"), or may discourage people to ask for help (e.g., "Are you feeling better today?" "Yes, doing great!").

Promises and Perils of Personalized AI-MC. Although personalization of AI-MC may alleviate some of the issues we identified, our interviews suggest it may also create its own unique issues. When personalizing based on only the content of an individual's past messages, an AI-MC system might miss nuanced *code-switching* – a change in language variations – and suggest something that the individual would be unlikely to say to a particular group (§ 3.2.7 & 3.2.8). For example, personalization might pick up on the phonetic spellings or religious salutations that a person uses for some social ties [22], but participants in our experiments largely edited such terms out in their adjustments (§4.5.4), indicating they might not always be received.

Personalization may also pick up on the use of hedging or other claim softening devices among certain groups (e.g., female users [60]), and suggest replies accordingly. Such biases in AI-MC suggestions may thus mirror and entrench existing societal biases by disproportionately encouraging that behavior among those groups. Indeed, the Just Not Sorry browser extension was designed to specifically minimize the use of hedging and claim softening devices [19], and there is some evidence that AI-MC alters language use [6]. To proactively address this, system designers need to balance between personalization at the individual level and at the social network level, where the assessment of suggestion quality includes the social ties involved. Embedding and adapting normative values during the design process, rather than after, may also help in this regard [76]. Regardless of the approach used, maintaining user privacy is paramount.

Giving Users More Control. While adjusting the AI systems or the algorithms that govern the suggestions people receive is one route forward, providing users with more explicit control could also be beneficial. This is evident not only by the market for third-party extensions (such as the Just Not Sorry browser extension), but also from our interviews, where participants often expressed a desire for greater control and customization (§3.2.3). The need and potential desire to fine tune replies can also be seen in our experiments where, across all categories, a large fraction of replies were amended, often depending on the social tie involved (Figure 4). Such controls might put the user in charge of selecting the social context for a given contact and tuning the system accordingly, with options to adjust these over time. Similarly, although most email reply suggestion systems allow for editing after clicking a suggestion, the same is not true for other AI-MC reply suggestions, such as instant messaging, which send suggested replies upon the first click

or tap. More work is needed to understand whether users want a similar buffer in the context of instant text messaging, and whether users of any AI-MC system would enjoy and benefit from additional control or opportunities for editing.

Long-term Impacts. Little is known about the long-term impacts of providing such automation [57], but predictive text systems are known to create feedback loops: behavioral and societal factors skew user behavior, these skews get embedded in the logged data and are reproduced by algorithms, users then interact with that output, which, in turn, produces more biased training data and closes the loop [9, 90]. As with other technologies designed to facilitate discourse [63, 89, 116], and given the global scale that they operate on, these systems may impact how language is used over time. These effects could extend beyond those who adopt the technologies by influencing how people communicate digitally [75], with broader effects including the spread of norms and cultures prioritized by those systems, such that even people who do not adopt the feature are affected by it. In our interviews, this was surfaced as both a concern and an appreciation for suggestions seemingly favoring "American" communication norms (§3.2.9). Designers should identify which cultural conventions may inadvertently take precedent, and take steps to understand how these conventions may narrow or shape discourse over time.

Future Work. Future work should address how user behavior changes after using AI-MC systems. For example, do users reply to more correspondence? Does adoption and use vary across interfaces (e.g. mobile, desktop, tablet)? Furthermore, how interfaces influence suggestion choice is an open question with substantial variance across AI-MC technologies, including instant messaging and document composition [57, 63]. For example, how does AI-MC use differ in instant messaging and email? We also need more in depth investigations of how replies are selected and modified, conditioned on contextual factors (§4.5.4), as well as how these findings apply to non-text communications, like emojis, images, and GIFs [119]. Future work should also further examine the impact of relationships other than the ones we examined by varying them in terms of social distance and familiarity (e.g. an acquaintance or first time interaction).

Conclusion. Overall, our findings indicate that current text recommendation systems for email and other CMC technologies remain insufficiently nuanced to reflect the subtleties of real world social relationships and communication needs. System designers should explore personalization strategies at the individual and social network level, consider how cultural values and societal biases may be perpetuated by their systems, and explore social interaction modelling in order to begin addressing the limitations and issues we have identified. Our mixed-methods framework for identifying and characterizing such issues is platform-agnostic, and can be applied to understand problematic suggestions in other AI-MC technologies, or reapplied over time, as algorithms are updated, interfaces change, and platform usage drifts.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers, the interviewees and crowd judges, Naman Mody, Shashank Jain, Andrew Lambert, Mohamed Musbah, and others for invaluable comments on this project. We are also grateful to our editors Susan Powers, Emery Fine, and Adam Ferguson for their help with labeling and reviewing our email-reply scenarios.

REFERENCES

[1] Herman Aguinis and Kyle J Bradley. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods* 17, 4 (2014), 351–371.

[2] Cheryl S Alexander and Henry Jay Becker. 1978. The use of vignettes in survey research. *Public Opinion Quarterly* 42, 1 (1978), 93–104.

[3] Sieun An, Li-Jun Ji, Michael Marks, and Zhiyong Zhang. 2017. Two sides of emotion: exploring positivity and negativity in six basic emotions across cultures. *Frontiers in Psychology* 8 (2017), 610.

[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 (July 1999), 452–471.

[5] Kenneth Arnold, Krysta Chauncey, and Krzysztof Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. *Proceedings of Graphics Interface 2018* Toronto (2018), 8 pages, 730.87 KB.

[6] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 128–138.

[7] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16*. ACM Press, Tokyo, Japan, 603–608.

[8] John Langshaw Austin. 1962. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*. Technical Report. Oxford University Press, Oxford.

[9] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61.

[10] Naomi S. Baron. 1998. Letters by Phone or Speech by Other Means: The Linguistics of Email. *Language & Communication* 18, 2 (April 1998), 133–170.

[11] Asaf Beasley and Winter Mason. 2015. Emotional States vs. Emotional Words in Social Media. In *Proceedings of the ACM Web Science Conference*. ACM, Oxford United Kingdom, 1–10.

[12] Boomerang. 2020. Respondable: Write Better Email. https://www.boomeranggmail.com/respondable.

[13] Pia Borlund and Peter Ingwersen. 1997. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation* 53, 3 (Aug. 1997), 225–250.

[14] Jerry Boucher and Charles E. Osgood. 1969. The Pollyanna Hypothesis. *Journal of Verbal Learning and Verbal Behavior* 8, 1 (Feb. 1969), 1–8.

[15] Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Number 4 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge [Cambridgeshire] ; New York.

[16] Ulla Bunz and Scott W. Campbell. 2004. Politeness Accommodation in Electronic Mail. *Communication Research Reports* 21, 1 (Jan. 2004), 11–25.

[17] Moira Burke, Justin Cheng, and Bethany de Gant. 2020. Social Comparison and Facebook: Feedback, Positivity, and Opportunities for Comparison. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13.

[18] Robin Campbell and Roger Wales. 1970. The Study of Language Acquisition. In *New Horizons in Linguistics*, J. Lyons (Ed.). Penguin, Harmondsworth, 242–260.

[19] Christina Cauterucci. 2015. New Chrome App Helps Women Stop Saying "Just" and "Sorry" in Emails. https://slate.com/human-interest/2015/12/new-chrome-app-helps-women-stop-saying-just-and-sorry-in-emails.html.

[20] Praveen Chandar, Jean Garcia-Gathright, Christine Hosey, Brian St. Thomas, and Jennifer Thom. 2019. Developing Evaluation Metrics for Instant Search Using Mixed Methods Methods. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR'19*. ACM Press, Paris, France, 925–928.

[21] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. *arXiv:1906.00080 [cs]* (May 2019). arXiv:1906.00080 [cs]

[22] Innocent Chiluwa. 2010. Nigerian English in Informal Email Messages. *English World-Wide* 31, 1 (2010), 40–61.

[23] Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Number 11 in Massachusetts Institute of Technology (Cambridge, Mass.). Research Laboratory of Electronics. Special Technical Report. MIT Press, Cambridge, MA. OCLC: 245876504.

[24] Domenic V. Cicchetti, Donald Shoinralter, and Peter J. Tyrer. 1985. The Effect of Number of Rating Scale Categories on Levels of Interrater Reliability : A Monte Carlo Investigation. *Applied Psychological Measurement* 9, 1 (March 1985), 31–36.

[25] Lilia M. Cortina, Vicki J. Magley, Jill Hunter Williams, and Regina Day Langhout. 2001. Incivility in the Workplace: Incidence and Impact. *Journal of Occupational Health Psychology* 6, 1 (2001), 64–80.

[26] David Crystal et al. 2001. *Language and the Internet*. Cambridge University Press.

[27] Jonathon N Cummings, Brian Butler, and Robert Kraut. 2002. The quality of online social relationships. *Commun. ACM* 45, 7 (2002), 103–108.

[28] Laura A. Dabbish and Robert E. Kraut. 2006. Email Overload at Work: An Analysis of Factors Associated with Email Strain. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work - CSCW '06*. ACM Press, Banff, Alberta, Canada, 431.

[29] Laura A. Dabbish, Robert E. Kraut, Susan Fussell, and Sara Kiesler. 2005. Understanding Email Use: Predicting Action on a Message. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '05*. ACM Press, Portland, Oregon, USA, 691.

[30] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 10.

[31] Rita de Cássia Veiga Marriott and Patricia Lupion Torres. 2009. Politeness in Intercultural E-Mail Communication. In *Handbook of Research on E-Learning Methodologies for Language Acquisition:*. IGI Global.

[32] Rachele De Felice, Jeannique Darby, Anthony Fisher, and David Peplow. 2013. A Classification Scheme for Annotating Speech Acts in a Business Email Corpus. *ICAME Journal* 37 (2013), 71–105.

[33] Rachele De Felice and Gregory Garretson. 2018. Politeness at work in the Clinton email Corpus: A first look at the effects of status and gender. *Corpus Pragmatics* 2, 3 (2018), 221–242.

[34] Ferdinand de Saussure. 1959. *Course in General Linguistics*. Philosophical Library, New York.

[35] Budhaditya Deb, Peter Bailey, and Milad Shokouhi. 2019. Diversifying Reply Suggestions Using a Matching-Conditional Variational Autoencoder. *arXiv:1903.10630 [cs, stat]* (March 2019). arXiv:1903.10630 [cs, stat]

[36] Dotan Di Castro, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2016. You've got mail, and here is what you could do with it!: Analyzing and predicting actions on email messages. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 307–316.

[37] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126.

[38] Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth. 2015. Human Language Reveals a Universal Positivity Bias. *Proceedings of the National Academy of Sciences* 112, 8 (Feb. 2015), 2389–2394.

[39] Peter Sheridan Dodds and Christopher M. Danforth. 2010. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies* 11, 4 (Aug. 2010), 441–456.

[40] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE* 6, 12 (Dec. 2011), e26752.

[41] Christa Dürscheid, Carmen Frehner, Susan C Herring, Dieter Stein, and Tuija Virtanen. 2013. Email communication. *Handbooks of Pragmatics [HOPS]* 9 (2013), 35–54.

[42] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. 2017. Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States. *SSRN Electronic Journal* (2017).

[43] Anne Edstrom and Jennifer D. Ewald. 2017. "Out of the Office": Conveying Politeness through Auto-Reply Email Messages. *Language@Internet* 14, 4 (2017).

[44] Anne Edstrom and Jennifer D. Ewald. 2019. Characteristics of Effective Auto-Reply Emails: Politeness and Perceptions. *Technology in Society* (Jan. 2019), S0160791X1730218X.

[45] Serge Egelman, Ed H Chi, and Steven Dow. 2014. Crowdsourcing in hci research. In *Ways of Knowing in HCI*. Springer, 267–289.

[46] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, Seoul, Republic of Korea, 153–162.

[47] Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, Santa Clara, California, USA, 4647–4657.

[48] Lori Francis, Camilla M. Holmvall, and Laura E. O'Brien. 2015. The Influence of Workload and Civility of Treatment on the Perpetration of Email Incivility. *Computers in Human Behavior* 46 (May 2015), 191–201.

[49] Brenden Gallagher. 2018. Gmail's Smart Reply blurs the line between people and brands. https://www.dailydot.com/debug/thanks-for-letting-me-know/. [Online; accessed Sept-2019].

[50] Ryan J. Gallagher, Morgan R. Frank, Lewis Mitchell, Aaron J. Schwartz, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. 2020. Generalized Word Shift Graphs: A Method for Visualizing and Explaining Pairwise Comparisons Between Texts. *arXiv:2008.02250 [physics]* (Aug. 2020). arXiv:2008.02250 [physics]

[51] Dana Garbarski, Nora Cate Schaeffer, and Jennifer Dykema. 2015. The Effects of Response Option Order and Question Order on Self-Rated Health. *Quality of Life Research* 24, 6 (June 2015), 1443–1453.

[52] Gary W. Giumetti, Andrea L. Hatfield, Jenna L. Scisco, Amber N. Schroeder, Eric R. Muth, and Robin M. Kowalski. 2013. What a Rude E-Mail! Examining the Differential Effects of Incivility versus Support on Mood, Energy, Engagement, and Performance in an Online Context. *Journal of Occupational Health Psychology* 18, 3 (July 2013), 297–309.

[53] J. Goldstein and R.E. Sabin. 2006. Using Speech Acts to Categorize Email and Identify Email Genres. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*. IEEE, Kauia, HI, USA, 50b–50b.

[54] Paul H. Grice. 1975. Logic and Conversation. In *Syntax and Semantics: Speech Acts*, Peter Cole and J. Morgan (Eds.). Vol. 3. Academic Press, New York, 41–58.

[55] Jacob Groshek and Chelsea Cutino. 2016. Meaner on Mobile: Incivility and Impoliteness in Communicating Contentious Politics on Sociotechnical Networks. *Social Media + Society* 2, 4 (Nov. 2016), 205630511667713.

[56] Danielle N. Gunraj, April M. Drumm-Hewitt, Erica M. Dashow, Sri Siddhi N. Upadhyay, and Celia M. Klin. 2016. Texting Insincerely: The Role of the Period in Text Messaging. *Computers in Human Behavior* 55 (Feb. 2016), 1067–1075.

[57] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* (2020).

[58] Michael Haugh. 2010. When Is an Email Really Offensive?: Argumentativity and Variability in Evaluations of Impoliteness. *Journal of Politeness Research. Language, Behaviour, Culture* 6, 1 (Jan. 2010).

[59] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *arXiv:1705.00652 [cs]* (May 2017). arXiv:1705.00652 [cs]

[60] Susan C Herring. 1994. Politeness in computer culture: Why women thank and men flame. In *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*. Berkeley Women and Language Group Berkeley, CA, 278–294.

[61] H. W. Hildebrandt and R. D. Snyder. 1981. The Pollyanna Hypothesis in Business Writing: Initial Results, Suggestions for Research. *Journal of Business Communication* 18, 1 (Jan. 1981), 5–15.

[62] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–14.

[63] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–6.

[64] Jess Hohenstein and Malte Jung. 2020. AI as a Moral Crumple Zone: The Effects of AI-Mediated Communication on Attribution and Trust. *Computers in Human Behavior* 106 (May 2020), 106190.

[65] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–16.

[66] Dirk Holtbrügge, Abigail Weldon, and Helen Rogers. 2013. Cultural Determinants of Email Communication Styles. *International Journal of Cross Cultural Management* 13, 1 (April 2013), 89–110.

[67] Kenneth J. Houghton, Sri Siddhi N. Upadhyay, and Celia M. Klin. 2018. Punctuation in Text Messages May Convey Abruptness. Period. *Computers in Human Behavior* 80 (March 2018), 112–121.

[68] C J Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*. 10.

[69] Dell Hymes. 1972. On Communicative Competence. In *Sociolinguistics: Selected Readings*, J.B. Pride and J. Holmes (Eds.). Penguin, Harmondsworth, 269–293.

[70] Rumen Iliev, Joe Hoover, Morteza Dehghani, and Robert Axelrod. 2016. Linguistic Positivity in Historical Texts Reflects Dynamic Environmental and Psychological Factors. *Proceedings of the National Academy of Sciences* 113, 49 (Dec. 2016), E7871–E7879.

[71] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13.

[72] Susan Jamieson. 2004. Likert Scales: How to (Ab)Use Them. *Medical Education* 38, 12 (Dec. 2004), 1217–1218.

[73] Sherri L. Jessmer and David Anderson. 2001. The Effect of Politeness and Grammar on User Perceptions of Electronic Mail. *North American Journal of Psychology* 3, 2 (2001), 331–346.

[74] Anjuli Kannan, Peter Young, Vivek Ramavajjala, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, and Marina Ganea. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. ACM Press, San Francisco, California, USA, 955–964.

[75] Jonathon Keats. 2010. *Virtual words: Language on the edge of science and technology*. Oxford University Press.

[76] Cory Knobel and Geoffrey C. Bowker. 2011. Values in Design. *Commun. ACM* 54, 7 (July 2011), 26.

[77] Birgit Koopmann-Holm and Jeanne L Tsai. 2014. Focusing on the negative: Cultural differences in expressions of sympathy. *Journal of Personality and Social Psychology* 107, 6 (2014), 1092.

[78] Farshad Kooti, Luca Maria Aiello, Mihajlo Grbovic, Kristina Lerman, and Amin Mantrach. 2015. Evolution of Conversations in the Age of Email Overload. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15*. ACM Press, Florence, Italy, 603–613.

[79] Jon A. Krosnick and Duane F. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement. *Public Opinion Quarterly* 51, 2 (1987), 201.

[80] Jon A. Krosnick and Matthew K. Berent. 1993. Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *American Journal of Political Science* 37, 3 (1993), 941–964.

[81] Jon A. Krosnick, Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, and Michael Conaway. 2002. The Impact of "No Opinion" Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice? *The Public Opinion Quarterly* 66, 3 (2002), 371–403.

[82] Jon A. Krosnick and Stanley Presser. 2010. Question and Questionnaire Design. In *Handbook of Survey Research, Second Edition* (2 edition ed.), Peter V. Marsden and James D. Wright (Eds.). Emerald Publishing, Bingley, UK.

[83] B Levin. 1996. On the Holm, Simes, and Hochberg Multiple Test Procedures. *American Journal of Public Health* 86, 5 (May 1996), 628–629.

[84] Sandy Lim, Lilia M. Cortina, and Vicki J. Magley. 2008. Personal and Workgroup Incivility: Impact on Work and Health Outcomes. *Journal of Applied Psychology* 93, 1 (Jan. 2008), 95–107.

[85] Vivien K.G. Lim and Thompson S.H. Teo. 2009. Mind Your E-Manners: Impact of Cyber Incivility on Employees' Work Attitude and Behavior. *Information & Management* 46, 8 (2009), 419–425.

[86] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 2397–2409.

[87] Shannon L. Marlow, Christina N. Lacerenza, and Chelsea Iwig. 2018. The Influence of Textual Cues on First Impressions of an Email Sender. *Business and Professional Communication Quarterly* 81, 2 (June 2018), 149–166.

[88] Heather Newman and David Joyner. 2018. Sentiment Analysis of Student Evaluations of Teaching. In *International Conference on Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.), Vol. 10948. Springer International Publishing, Cham, 246–250.

[89] Kazushi Nishimoto and Jianning Wei. 2015. G-IM: An Input Method of Chinese Characters for Character Amnesia Prevention. In *ACHI 2015*. 118–124.

[90] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[91] Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When Are Search Completion Suggestions Problematic? *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–25.

[92] Séamas O'Reilly. 2019. How Smart Are Gmail's 'Smart Replies'? *The Observer* (Feb. 2019).

[93] Soya Park, Amy X. Zhang, Luke S. Murray, and David R. Karger. 2019. Opportunities for Automating Email Processing: A Need-Finding Study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–12.

[94] Ellie Pavlick and Joel Tetreault. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics* 4 (Dec. 2016), 61–74.

[95] Carolyn C Preston and Andrew M Colman. 2000. Optimal Number of Response Categories in Rating Scales: Reliability, Validity, Discriminating Power, and Respondent Preferences. *Acta Psychologica* 104, 1 (March 2000), 1–15.

[96] Kristen Purcell. 2011. Search and Email Still Top the List of Most Popular Online Activities.

[97] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, Santa Clara, California, USA, 83–88.

[98] Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*. ACM Press, Boston, Massachusetts, USA, 235–244.

[99] Bahareh Sarrafzadeh, Ahmed Hassan Awadallah, Christopher H. Lin, Chia-Jung Lee, Milad Shokouhi, and Susan T. Dumais. 2019. Characterizing and Predicting Email Deferral Behavior. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*. ACM Press, Melbourne VIC, Australia, 627–635.

[100] Gordon B. Schmidt and William M. Jettinghoff. 2016. Using Amazon Mechanical Turk and Other Compensated Crowdsourcing Sites. *Business Horizons* 59, 4 (July 2016), 391–400.

[101] Burkard Sievers and Rose Redding Mersky. 2006. The Economy of Vengeance: Some Considerations on the Aetiology and Meaning of the Business of Revenge. *Human Relations* 59, 2 (Feb. 2006), 241–259.

[102] Erin L. Spottswood and Jeffrey T. Hancock. 2016. The Positivity Bias and Prosocial Deception on Facebook. *Computers in Human Behavior* 65 (Dec. 2016), 252–259.

[103] Keri K. Stephens, Renee L. Cowan, and Marian L. Houser. 2011. Organizational Norm Congruency and Interpersonal Familiarity in E-Mail: Examining Messages From Two Different Status Perspectives. *Journal of Computer-Mediated Communication* 16, 2 (Jan. 2011), 228–249.

[104] Wenbin Tang, Honglei Zhuang, and Jie Tang. 2011. Learning to infer social ties in large networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 381–397.

[105] 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 (March 2010), 24–54.

[106] Benjamin Toff and Rasmus Kleis Nielsen. 2018. "I Just Google It": Folk Theories of Distributed Discovery. *Journal of Communication* 68, 3 (June 2018), 636–657.

[107] Anna K. Turnage. 2007. Email Flaming Behaviors and Organizational Conflict. *Journal of Computer-Mediated Communication* 13, 1 (Oct. 2007), 43–59.

[108] Connie K. Varnhagen, G. Peggy McFall, Nicole Pugh, Lisa Routledge, Heather Sumida-MacDonald, and Trudy E. Kwong. 2010. Lol: New Language and Spelling in Instant Messaging. *Reading and Writing* 23, 6 (July 2010), 719–733.

[109] Jane A. Vignovic and Lori Foster Thompson. 2010. Computer-Mediated Cross-Cultural Collaboration: Attributing Communication Errors to the Person versus the Situation. *Journal of Applied Psychology* 95, 2 (2010), 265–276.

[110] Jocelyne Vincent. 2008. Netiquette Rules OK!...OK?: speculating on rhetorical cleansing and English linguistic and cultural imperialism through email netiquette style guides. In *Threads in the complex fabric of language: linguistic and literary studies in honour of Lavinia Merlini Barbaresi*, Lavinia Merlini Barbaresi, Marcella Bertuccelli-Papi, Antonio Bertacca, and Silvia Bruti (Eds.). Felici, San Giuliano Terme (Pisa), 409–443.

[111] James Vincent. 2018. Google removes gendered pronouns from Gmail's Smart Compose feature. https://www.theverge.com/2018/11/27/18114127/google-gmail-smart-compose-ai-gender-bias-prounouns-removed. [Online; accessed Sept-2019].

[112] Joan Waldvogel. 2007. Greetings and Closings in Workplace Email. *Journal of Computer-Mediated Communication* 12, 2 (Jan. 2007), 456–477.

[113] Joseph B Walther. 1995. Relational aspects of computer-mediated communication: Experimental observations over time. *Organization Science* 6, 2 (1995), 186–203.

[114] Dashun Wang, Zhen Wen, Hanghang Tong, Ching-Yung Lin, Chaoming Song, and Albert-László Barabási. 2011. Information spreading in context. In *Proceedings of the 20th International Conference on World Wide Web*. 735–744.

[115] Yi-Chia Wang, Alexandros Papangelis, Runze Wang, Zhaleh Feizollahi, Gokhan Tur, and Robert Kraut. 2020. Can You be More Social? Injecting Politeness and Positivity into Task-Oriented Conversational Agents. arXiv:2012.14653 [cs.CL]

[116] Helen J Watt. 2010. How does the use of modern communication technology influence language and literacy development? A review. *Contemporary Issues in Communication Science and Disorders* 37 (2010), 141.

[117] Liu Yang, Susan T. Dumais, Paul N. Bennett, and Ahmed Hassan Awadallah. 2017. Characterizing and Predicting Enterprise Email Reply Behavior. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*. ACM Press, Shinjuku, Tokyo, Japan, 235–244.

[118] Xiao Yang, Ahmed Hassan Awadallah, Madian Khabsa, Wei Wang, and Miaosen Wang. 2018. Characterizing and Supporting Question Answering in Human-to-Human Communication. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*. ACM Press, Ann Arbor, MI, USA, 345–354.

[119] Ning Ye, Ariel Fuxman, Vivek Ramavajjala, Sergey Nazarov, J Patrick McGregor, and Sujith Ravi. 2018. Photoreply: Automatically suggesting conversational responses to photos. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1893–1899.

[120] Ming-chung Yu. 2005. Sociolinguistic competence in the complimenting act of native Chinese and American English speakers: A mirror of cultural value. *Language and Speech* 48, 1 (2005), 91–119.

[121] Elana Zeide. 2015. This App Wants to Tell Your Next Employer About Your Personality. But It's Probably Wrong. https://slate.com/technology/2015/05/crystal-app-algorithmic-fortunetelling-for-employers-and-potential-customers.html.

[122] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Nithum Thain, Yiqing Hua, and Dario Taraborelli. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Melbourne, Australia, 1350–1361.

A MOTIVATING STUDY: EMAIL REPLY SUGGESTION LOGS AND CORPORA

In our initial, motivating study, we explored a dataset containing real user interactions with email reply suggestions in a web browser client for Outlook. This dataset consisted of a sample of real reply suggestions and a sample of anonymized click logs. No emails or email metadata were viewed or analyzed. We used these data to explore the prevalence of reply length, positivity bias, and politeness in email reply suggestions and users' clicks on those suggestions. We conjectured that measuring these data for systematic patterns or skews might shed light on potential issues that we could ask participants to expand on in our interviews. We limited our exploration to these factors because creating an algorithm for scoring text is not goal of this project, so we therefore relied on existing metrics that could be used as a proxy for measuring some of the more consistent themes explored in prior work (e.g. Positivity and Sentiment Analysis).

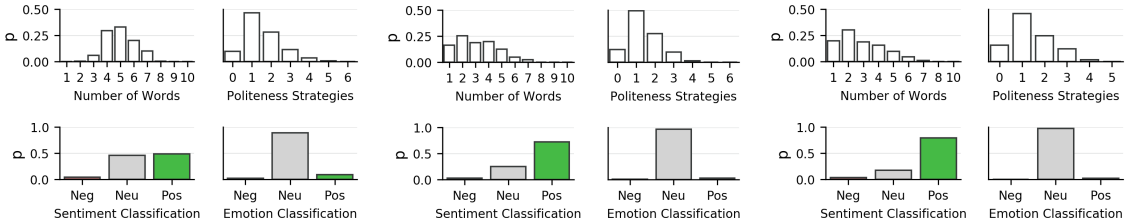
A.1 User Interaction Data

The anonymized data we used consisted of rendered suggestions—which were presented as "blocks" consisting of three suggestions shown together—and corresponding click logs from a major commercial email client. These datasets were collected in two different periods (May and June 2019), and drawn from a 5% random sample of North American users. Suggestions were rendered only for emails that were in English, but not for emails that were very short or very long. In total, our data contains 2.8 million blocks, including a total of over 8.3M rendered suggestions. The system that generated these suggestions was based on a recently established architecture [35] and uses a curated corpus of replies that is drawn from the most frequent 20K responses from a large email provider (based on tens of millions of emails over hundreds of thousands of users). The log data was anonymized to preserve user privacy, and *we do not access email content only rendered reply suggestions and corresponding click data.*

A.2 Characterizing Reply Suggestions

To characterize the reply suggestions, we adopted three metrics that have been widely used in applied NLP research and that were related to the insights from our interviews around positivity, politeness, and tone. The first, VADER (Valence Aware Dictionary and sEntiment Reasoner) [68], is specifically tuned to measure sentiment expressed in microblogging posts, which tend to be short, similar to our reply suggestions. The second, Empath [47], is an adaptable

emotion lexicon-based word embedding trained on 1.8 billion words from modern fiction, and that strongly correlates with a popular psychometrically validated gold-standard [105]. The last is a computational technique for detecting politeness strategies in text, such as greetings, hedging, or expressions of gratitude [15, 30]. We also measured other reply features inspired by prior work, including reply length (e.g., number of words and characters).



(a) Unique replies (b) Replies rendered to users (c) Replies clicked by users

Fig. 7. Probability distribution for reply features at each point in the system: (a) unique replies, (b) replies rendered (ranking and filtering), and (c) replies clicked (user interaction).

A.3 Exploratory Analysis

We examined the distributions of our selected metrics for (a) the inventory of replies that appeared in our datasets, (b) the replies rendered by the algorithm, and (c) the replies that were clicked by users (Figure 7). Relative to the inventory of available replies (Figure 7a), those rendered skewed towards shorter and more positive replies (Figure 7b). Similarly, relative to the suggestions rendered, the suggestions clicked by users were further skewed towards shorter and more positive replies, though these differences were smaller (Figure 7c).

Controlling for the rank at which suggestion were rendered, we then computed the probability for a suggestion to be clicked given the rank it appears at, the number of words it is composed of, and various linguistic cues for politeness and positivity—figures omitted for space considerations. Despite a left-to-right position bias for the rendered suggestions, participants appeared to seek out a specific kind of reply at each rank. Namely, users appear to prefer short replies that are polite or positive, even when they are rendered at lower ranked positions. This finding connects to the original motivation for reply suggestions [74]. The VADER and Empath dictionaries we used to score our replies were somewhat related, with correlations between their compound ($\rho = 0.42$, $P < 0.001$) and positive scores ($\rho = 0.38$, $P < 0.001$), but only a small correlation between their negative scores ($\rho = 0.10$, $P < 0.001$). We found small but signficant correlations between CTR and VADER's positive scores ($\rho = 0.084$, $P < 0.001$) and compound scores ($\rho = 0.060$, $P < 0.001$), but not its negative scores. However, we found no significant correlation between CTR and Emapth's scores.

A.4 Limitations

In our analysis, we were interested in empirically exploring how textual cues for politeness or positivity correlate with users propensity to select a certain suggestion. However, the computational techniques for assessing affect and politeness, such as the ones we used (VADER and Empath), have known limitations. For example, behavioral and other verbal cues are important in such assessments, but are not fully captured by text [62]. Future work should employ additional techniques for measuring the impact of these factors.

Though many interesting questions could be answered with additional log analyses, our goal was provide additional motivation for our mixed-methods study.

Future work should seek out or develop new methods for measuring the various features that might shed further light on click-through rates and algorithmic selection of reply suggestions. Given the similar use of predictive text suggestions in other conversational settings [57, 63, 71], extending our exploration to open source conversational datasets may provide insights into when our findings generalize beyond email.

B INTERVIEW PROTOCOL

1. Purpose of interview: Hi <NAME>, thank you for taking the time to speak with me. We are in the exploratory stage of a project related to the Smart Reply and Smart Compose features that you may have seen in various online services, especially email. Smart Reply is the feature that provides you with several options for responding to an entire email, and Smart Compose is the feature that suggests ways to complete your sentences as you type. We're interested in learning about the different scenarios in which these suggestions might be problematic or inappropriate. To better understand these scenarios, we have a number of open-ended questions that we would like to go through with you today.

2. Email Usage

- What kind of email style do you use?
- How often do you use a greeting and closing in your emails?
- What situations or scenarios do you think it's important to do so in?
- How many emails do you receive per day/week?
- How many emails do you answer per day/week?
- How elaborate do your email answers tend to be?

3. Usage of Suggestions for Email Correspondence

- Have you ever used Smart Reply / Smart Compose type of features?
- Which one? What was the email client?
- How often do you use Smart Replies/Compose?
- Generally speaking, how satisfied are you with the response options they provide?
- Do you use them on mobile, desktop, or both?
- Do you find them more useful on one? Why?
- How do you think that Smart Reply or Compose suggestions are generated?
- Can you tell me about a recent email that you used a Smart Reply on? (Repeat for Smart Compose)
- Without disclosing any sensitive information, could you tell me the topic of the email?
- What is the relationship with the Sender? (e.g., a friend, a colleague, a manager)
- Was it someone that you had interacted with before or planned to interact with again?
- Were you the only or primary recipient on this email?

4. Experiences with Problematic Suggestions

- Have you ever encountered a suggested reply that you thought was problematic, inappropriate, or awkward for any reason?

- Clarification: By problematic, inappropriate, or awkward, I mean the suggested response was one that you would not have written yourself, for any reason.
- Without disclosing any sensitive information, could you tell me the topic of the email?
- What was the relationship with the Sender?
- Were you the only or primary recipient on this email?
- Are there certain people or relationships that you would not use smart replies to communicate with?
- What is it about those relationships that makes you not want to use smart replies?
- Is there anything else you'd like to tell us about your experiences with Smart Reply or Smart Compose like features?

C LIKERT SCALE DETAILS

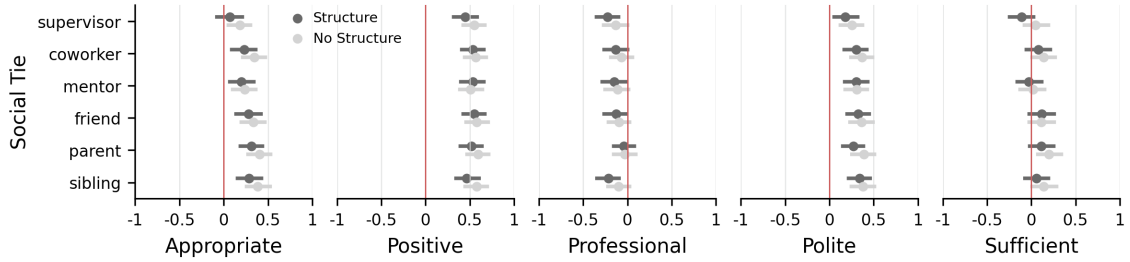


Fig. 8. The mean Likert-scale ratings (x-axis) for scenarios in each category (y-axis) by presence or absence of email structure (legend). After controlling for multiple hypothesis testing, none of these differences were significant.

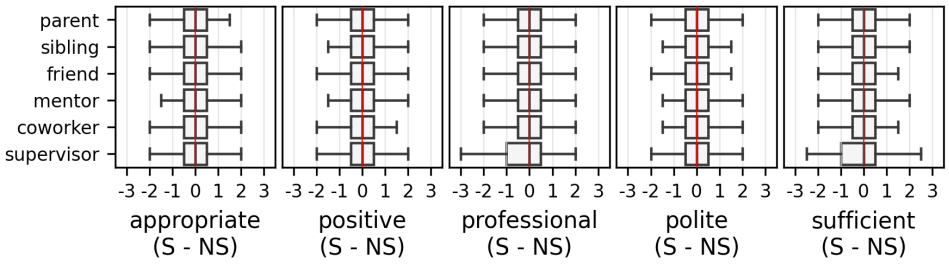


Fig. 9. Change in ratings (x-axes) by social tie (y-axis) due to the presence (S) or absence (NS) of structure.