EnronSR: A Benchmark for Evaluating AI-Generated Email Replies

Moran Shay[1], **Roei Davidson**[2], **and Nir Grinberg**[1]

[1]Software and Information Systems Engineering, Ben-Gurion University, Israel
[2]Communication, University of Haifa, Israel
shaymora@post.bgu.ac.il, roei@com.haifa.ac.il, nirgrn@bgu.ac.il

Abstract

Human-to-human communication is no longer just mediated by computers, it is increasingly generated by them, including on popular communication platforms such as Gmail, Facebook Messenger, Linkedin, and others. Yet, little is known about the differences between human- and machine-generated responses in complex social settings. Here, we present EnronSR, a novel benchmark dataset that is based on the Enron email corpus and contains both naturally occurring human- and AI-generated email replies for the same set of messages. This resource enables the benchmarking of novel language-generation models in a public and reproducible manner, and facilitates a comparison against the strong, production-level baseline of Google Smart Reply used by millions of people. Moreover, we show that when language models produce responses they could align more closely with human replies in terms of when responses should be offered, their length, sentiment, and semantic meaning. We further demonstrate the utility of this benchmark in a case study of GPT-3, showing significantly better alignment with human responses than Smart Reply, albeit providing no guarantees for quality or safety.

Introduction

The abundance of digital communications, often typed on small mobile devices, creates a need for simpler input methods that can generate quick and complete textual responses with a tap of a button. Even before ChatGPT captivated the imagination of millions of people (Vogels 2023), AI-generated responses were already prevalent. A prime example is Google's Smart Reply (Kannan et al. 2016; Chen et al. 2019), which is available to the billions of Gmail users (Marcelis and MacMillan 2018). Google's Smart Reply offers email recipients three reply suggestions that can be used as-is without any typing or as a prompt for a longer reply (see Figure 1 for example). Similar features are currently available in Facebook Messenger (Landowski and El Moujahid 2017), Uber mobile app (Weng et al. 2019), LinkedIn (Chakravarthi et al. 2017), and the Android operating system (Cuthbertson 2019). Hancock et al. (2020) refer to this type of communication systems as *Artificial Intelligence-Mediated Communication* (AI-MC), where "an intelligent agent operates on behalf of a communicator by modifying, augmenting, or generating messages to accomplish communication goals".

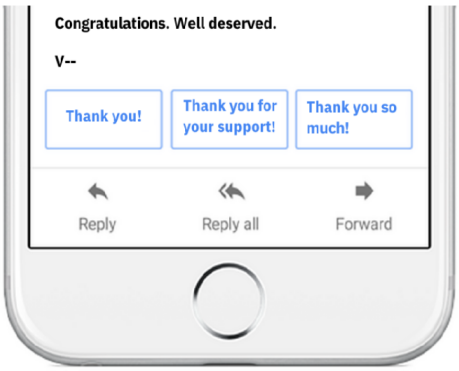

Figure 1: An example of an incoming congratulatory email message and three reply suggestions generated by Gmail Smart Reply: "Thank you!", "Thank you for your support!", and "Thank you so much!".

As generative AI and AI-MC systems become more common, it is increasingly important to make their output as relevant, appropriate, and usable as possible in different social contexts, as well as to avoid unwanted responses. It is well known that violating people's expectations from their communication partners can have negative implications for the relationship between them and that those expectations depend on the social context (Burgoon 1993; Burgoon, Stern, and Dillman 1995). For instance, people have different expectations from colleagues at work, who typically communicate more formally, than the expectations for communication with friends whose messages are typically less formal (Ishii, Kobayashi, and Grudin 1993). To be maximally effective, AI-MC systems need to offer suggestions that are appropriate and nuanced across several different dimensions including the message style, sentiment, and emotional valence expressed toward other conversation partners. For example, a reply from a manager that uses formal language in response to a personal request of an employee may be perceived as cold and impact the working relationship.

Shared resources for evaluating and tracking the progress of reply suggestion models are vital for making steady progress in this increasingly-common task of AI-MC, and for developing generation models that capture more of the nuances of human-to-human communication in an open and reproducible manner. Currently, much of the development of AI-MC models is proprietary, and only in a few cases, model-building steps were reported publicly (Kannan et al. 2016; Chen et al. 2019). Moreover, the primary metric for success was the number of suggestions used by users (Robertson et al. 2021). While usability is certainly an important metric for success, the nature of this language generation task calls for considering additional dimensions. This includes but is not limited to whether suggestions are at all relevant, appropriate for the context, reflect a suitable tone or emotion, and whether they introduce any systematic bias (Robertson et al. 2021; Kannan et al. 2016). To obtain this diversity of perspectives, benchmark datasets are critical.

There are two main challenges in evaluating AI-MC models. First, AI-MC systems often mediate private communications, which makes it difficult to obtain a large and heterogeneous sample of messages along with the full context of past communications and the AI-generated response suggestions. For example, accessing Gmail Smart Reply suggestions requires either accessing individual email accounts or obtaining private messages from participants. The private nature of these communications presents a challenge for sharing this data with the research community. Another challenge is that there is no standard or agreed-upon set of measures for evaluating the generated text by generative models in this complex social context. Of course, there is a variety of methods and measures for evaluation: from manual annotation to assess the relevance of responses (Li et al. 2016b, 2017; Lison and Bibauw 2017; Liu et al. 2018) to examining word-overlap metrics such as BLEU and METEOR (Li et al. 2016a; Sordoni et al. 2015) and training adversarial learning models to reduce discriminability between model and human responses (Li et al. 2017). However, the absence of a public benchmark with agreed-upon metrics for evaluation hinders the ability to develop novel AI-MC models in an open and reproducible manner and to systematically compare them against strong, production-level baselines.

In this work, we present *EnronSR*[1], the first public benchmark for evaluating AI-MC systems against organic human responses to email messages, and for comparing against the responses of a strong AI baseline on the same set of emails. We built this resource by recreating email communications from the public Enron corpus (Cohen 2015), and collecting Google's Smart Reply suggestions (SRs). We conducted extensive experimentation to ensure the validity and robustness of our auditing results to different factors that can impact the algorithmic suggestions such as time, different senders, or recipient personalization. Based on the collected SRs, we examine the differences between human- and AI-generated responses in terms of when responses are offered, their length, emotional valence, and semantic similarity. We further demonstrate the utility of EnronSR in benchmarking newer AI-generation models using a case study of GPT-3. Taken together, this work provides the first large-scale, validated resource to facilitate the benchmarking of novel AI-MC models in an open and reproducible manner.

Therefore, our contributions are as follows:

- A novel resource for benchmarking AI-MC models against a strong production-level baseline of Google Smart Reply.
- Empirical evidence of differences between AI- and human-generated responses in terms of recall (i.e. when responses exist), length, sentiment, and semantics in different social contexts which can guide future work.
- A case study that demonstrates how one can use EnronSR to benchmark newer AI-generation models.

Related Work

The current literature on AI-MC can be organized into four lines of work: (i) efforts to generate high-quality responses, (ii) bias, (iii) user perception of AI-generated responses, and (iv) implications of large language models (LLM) on trust, safety and reliability of communication.

Generating High-Quality Responses

The seminal work in this body of literature is the paper by Kannan et al. (2016) who detailed the inner workings of the Google Smart Reply system in Gmail. Google's Smart Reply consists of a classifier that determines when to offer suggestions, a language model for clustering common email responses, a manual annotation that selects a safe and representative response for each cluster, and a ranking algorithm for top suggestions. Its language model was inspired by the sequence-to-sequence (Seq2Seq) framework (Sutskever, Vinyals, and Le 2014), which was originally applied for machine translation. One of the key issues with these models is that they tend to generate generic and somewhat dull responses (e.g., "Thank you for the update!"), rather than more meaningful and context-aware answers (Sordoni et al. 2015; Kannan et al. 2016; Mou et al. 2016; Li et al. 2017; Liu et al. 2018).

Prior work explored ways to improve parts of the Seq2Seq model to generate more meaningful responses. Several works proposed variations of beam search that consider the diversity of responses and therefore boost heterogeneity in top-ranked suggestions (Li, Monroe, and Jurafsky 2016; Vijayakumar et al. 2016). Lison and Bibauw (2017) introduced a weighting model in their neural architecture to facilitate learning of the quality of each query and response pair. The weighting model, based on conversational data, associated a numerical weight to each training sample to reflect its intrinsic quality for dialogue modeling. In addition, a more recent study proposed a new re-weighting loss function to the Seq2Seq model to weigh responses differently based on their similarity to other responses and length (Liu et al. 2018).

EnronSR contributes to this line of work by providing a meaningful set of organic interactions to evaluate novel AI-MC models on and compare their performance to the widely used production AI-MC system of Gmail Smart Reply.

[1]Dataset available at https://doi.org/10.7910/DVN/RQBWAC.

Bias in AI-MC

The research community is increasingly aware of the risks of biased training sets that are replicated or even amplified by machine learning models (O'Neil 2016; Brock 2020; Crawford, Miltner, and Gray 2014; Gillespie 2014; Noble 2018). Research in the field of Natural Language Processing (NLP) uncovered biases at various stages of model and representation learning as well as in applications of these models for different tasks. These include bias in word representation and embedding (Bolukbasi et al. 2016; Brunet et al. 2019; Zhao et al. 2019; Basta, Costa-jussà, and Casas 2019; Kaneko and Bollegala 2019), hate speech and abusive language detection (Park, Shin, and Fung 2018; Sap et al. 2019; Mozafari, Farahbakhsh, and Crespi 2020), and coreference resolution (Zhao et al. 2018).

A growing body of research examines both the origins and implications of bias in natural language generation (Sheng et al. 2020; Shwartz, Rudinger, and Tafjord 2020; Liu et al. 2020; Henderson et al. 2018; Dinan et al. 2020). It has been shown that language models generate stereotypical occupations and varying levels of respect for different genders and races (Sheng et al. 2019; Kirk et al. 2021; Lucy and Bamman 2021). A noteworthy example is GPT-3's association of Muslims and violence (Abid, Farooqi, and Zou 2021).

Bias in AI-MC systems can similarly perpetuate or increase disparities. Those who receive messages from people who use AI-generated text have no easy way of knowing if it was written by the human sender or generated by AI, and the suspicion of automation use is linked with negative evaluations (Hohenstein et al. 2023). Automated replies may or may not match the sender's style or the receiver's expectations of the communication given its particular context (Burgoon 1993). Therefore, biased AI-generated responses bear the risk of reinforcing existing stereotypes and even harming relationship-building between people.

While research on bias in NLP, in general, is expanding, AI-MC systems received relatively less attention from the research community. A few works identified bias in the sentiment of AI-generated responses. Hohenstein and Jung (2018) reported positivity bias in suggestions generated by Google's Allo messaging app. Arnold et al. (2018) showed that exposure to positive AI suggestions subsequently led to more positive human writing. More recently, experimental work showed how the presence of Smart Reply recommendations changes the speed and positivity of the replies people send and some perceptions of the communication partners and user agency (Mieczkowski et al. 2021; Hohenstein et al. 2023; Wenker 2023). However, bias in AI-MC has not yet been examined in large-scale and naturalistic settings that include both human and machine-generated responses. EnronSR enables future work to identify additional linguistic dimensions where AI-MC systems are systematically biased, and paves the way for developing more inclusive and representative models and communication systems.

User Perceptions of AI-Generated Replies

Prior work has shown that users want and would benefit from various degrees of automation when responding to emails (Park et al. 2019; Yang et al. 2018). Yet, the social implications of these technologies are not fully understood.

Previous research has conducted algorithmic audits of existing systems and asked participants to reflect on the AI suggestions. Participants felt that the emotional expression generated by AI was sometimes inappropriate (Brandtzaeg and Følstad 2017), and that AI-generated profile information was perceived as less trustworthy (Jakesch et al. 2019). Robertson et al. (2021) conducted qualitative interviews asking participants to reflect on Google Smart Reply suggestions, and identified several problematic aspects, including lack of salutations, inauthentic personal style, and responses that do not match the social context. In experimental settings, Hohenstein et al. (2023) showed that when Smart Reply is available, people perceive their communication partners as closer and more cooperative, but evaluate them more negatively when they suspect partners are using automated replies.

The EnronSR dataset offers a systematic way to compare AI-generated suggestions against naturally occurring human replies on a large set of emails, and can contribute to gradually mitigating those differences in an open and replicable manner.

Safety and Reliability in Language Generation

Large language models like GPT-3 or ChatGPT offer high-quality general-purpose responses that can potentially be fine-tuned to fulfill AI-MC applications. Yet, there are concerns about the tendency of these models to produce offensive, biased, or abusive outputs (Bender et al. 2021). It is unclear how these models will respond to harmful narratives and whether they will serve to amplify them. Furthermore, there are concerns about the tendency of GPT models to hallucinate, appearing knowledgeable and credible even when generating inaccurate or misleading content (Bender et al. 2021; Kumar 2023; Lim and Schmälzle 2023), which adds to the risks of deploying such models in production. In contrast, Google Smart Reply generates common and safe responses by design, drawing on a set of responses that are sufficiently popular and going through several steps of manual assurance of response appropriateness. Therefore, EnronSR can be used as a baseline for benchmarking not only the quality of the responses but also their safety.

The Enron Dataset

The foundation of this work is the public Enron email corpus, which consists of more than 500,000 emails from 150 Enron employees sent over 3.5 years, from 1998 to 2002 (Cohen 2015). The emails were made public at the conclusion of the legal case brought forward and won by the US government against Enron. To the best of our knowledge, Enron is the largest publicly available corpus of email messages to date and no prior work has used it to evaluate AI-MC systems. For ease of processing, we use a version of Enron corpus released by the CALO Project[2].

[2]https://www.cs.cmu.edu/~./enron/

Methods

A key element in the creation of EnronSR is the collection of AI-generated reply suggestions from Google Smart Reply for emails in the Enron corpus, which then enables a comparison of human- and AI-generated responses. Before collecting Smart Reply suggestions, we preprocessed the original Enron corpus and made several robustness checks to ensure the validity of our collection processes, as detailed next.

Data Preprocessing

The Enron dataset contains email messages in raw format as they were sent using the email clients and servers used circa 2002. This meant we had to resolve email aliases, remove duplicate messages, and organize messages into threads. For deduplication, we considered four message fields as identifying a message: sender ("To"), recipient list ("From"), email subject, and contents, which resulted in the removal of 266,286 duplicate messages. To construct threads, we used a simple heuristic to extract the parent of each message since the "In-Reply-To" header did not exist at the time. We identified the message parent as a message with an earlier timestamp, a recipient that matches the sender of the child message, and a matching subject line, allowing common modifications such as "Re:" or "Fw:". We manually validated this heuristic to ensure that it yields consistent threads that match the logical order of the communication.

Overall, the preprocessing resulted in a dataset with 234,485 messages in total, sent by 19,122 distinct email addresses.

Sampling and Robustness Checks

In order to recreate the full communication in our preprocessed Enron dataset with maximum fidelity, one would have to operate nearly 20,000 accounts, over the course of 3.5 years, sending over 230,000 messages, and collecting the SRs through the Gmail interface. While this approach may generate high fidelity, it is time-consuming, costly, prone to failures due to changes in Gmail's interface and the underlying Smart Reply model during this extended period of time.

Instead, we focused on a stratified sample of emails and accounts, preserving the order of messages but not their full chronological timeline, and validating that the SRs generated are not impacted by these choices. We conducted an extensive series of robustness checks that assessed the sensitivity of Google's Smart Reply to different aspects of time, senders, personalization, and dependence on previous communications. While this approach cannot rule out every possible factor that may impact the SRs being generated, it represents our best attempt to control for possible confounding factors while keeping the collection process feasible.

Time dependence: We ran three full parallel executions that sent emails one after the other with different time gaps of two seconds, 60 seconds, and random sleep time uniformed sample in the range of 2-60 seconds. The same SRs were generated in all of these cases. We further tested whether Smart Reply generates different SRs depending on times of the day (e.g day or night) or days of the week (e.g weekdays or weekends) of sent messages. In all of our experiments, the same SRs were generated in all conditions. We also made sure that the same SRs were present when collecting the SRs using the Gmail interface at different times and at different time lags after the message was received.

Sensitivity to different senders: Due to budgetary constraints, we could only create a limited number of Google Workspace accounts. This implied that certain messages that were sent by different senders in the original dataset were "packed" into a single account in our data collection process. To test the sensitivity of the SRs to such packing, we collected SRs for different senders as well as varied the number of accounts being packed into a single account. In all of these experiments, identical SRs were generated for the same message.

Recipient personalization: SRs may be personalized for the recipient account as indicated by the original paper (Kannan et al. 2016). We tested this in two different ways. First, we sent the same message from the same sender to multiple recipients and record the SRs the recipients received. Second, we tested different history lengths – full history, 12 preceding months, and 3 preceding weeks – before collecting the SRs for the last month of data. We found that SRs were generally identical between different recipients for the same message and sender, but there were a few rare cases where SRs differed. When examining the SRs for different history lengths we found differences both in the amount of SRs being offered and their content. The more history we preserved, the more similar the SRs were across recipients. These findings indicate that Gmail SRs are personalized in non-trivial ways. To mitigate the impact of personalization on our results, we chose to focus our data collection efforts on the 150 focal Enron users that have full message histories in the original dataset.

Experimental Procedure

As described in the previous section, we collected SRs for a stratified sample of accounts and messages that prioritized emails with human replies. Out of the 234,485 emails in the preprocessed dataset, 74,382 were addressed to the 150 users from Enron users and 8,085 of them had a reply. Our data collection included all of these 8,085 messages with a reply, and additional 944 messages that our focal users sent as part of these conversations, for a total of 9,029 messages. In addition, we created a similarly-sized random sample of 9,029 messages that were addressed to the focal 150 users and had no reply. Finally, to emulate suggestions to messages originating from outside the organization, we sent a random sample of 2,000 messages that were addressed to our 150 employees by external accounts.

We created 150 Gmail users, one for each Enron user in the original dataset, as part of a Google Workspace Business Starter account. These accounts were used to send and receive emails as well as to collect the SRs as described below. All data collection procedures took place in May, 2022, and all accounts were permanently deleted upon completion.

There are 12,971 external email addresses and 6,151 Enron-internal addresses that appear in the pre-processed dataset for our 150 focal Enron employees. In order to re-

main within the allotted budget and within Google's limits on email automation, we created a single Gmail account that sent out all external messages to Enron users, 16 accounts that represent Enron non-focal users, and 150 accounts for the focal Enron users.

Overall, we sent a total of 20,058 messages, many of them having multiple recipients in the dataset. Therefore, our final dataset consists of 34,626 messages. All messages were sent in the original chronological order.

Collecting SRs: We developed a custom client-side code that emulated user interaction with the Gmail interface. The script "clicked" on individual messages and collected both the incoming message and SRs. We manually validated the accuracy for script in retrieving the actual SRs available to Gmail users. All SRs were linked to the messages that prompted them as well as to the human reply to the same message, if such reply exists.

Results

In this section, we describe the EnronSR dataset, and report statistics about differences between AI- and human-generated responses along a number of linguistic dimensions. The section closes with a case study of GPT-3, demonstrating the utility and usability of EnronSR for benchmarking language generation models.

Summary Statistics About EnronSR

Table 1 provides summary statistics about the EnronSR dataset. The dataset contains all 34,626 emails received by the 150 Enron employee accounts that we recreated. Since Google Smart Reply does not produce reply suggestions for every email, not all emails have SRs. Similarly, not all emails have a human reply associated with them. As a classifier for detecting when human replies will be present, Smart Reply has an accuracy of 68.0%, which is mostly driven by correctly identified cases without a human reply (N=20,166, 58.2%) and a little by cases with human replies (N=3,406, 9.8%). However, the overlap between messages with human replies and messages with AI suggestions is quite low. Google SR did not offer suggestions for 5,972 emails that people responded to, missing 63.6% of human replies, and "over-triggering", offering suggestions to emails that people would not reply to, in 5,082 cases (60% of its suggestions).

Differences Between Human- and AI-Generated Responses

Next, we examined the differences between human- and AI-generated responses in the set of 3,406 emails where both existed. We found that the average length of the SRs is 3.17 (3.12, 3.22)[3] words, while the average length of the first sentence of the human responses is 10.12 (9.81, 10.43) words. These differences suggest that SRs offer a much more concise response, and potentially more generic, than what humans normally write. Although the SRs are based on frequent full-sentence responses by people, the system offers much shorter responses on average.

[3]All error estimates represent 95% confidence interval

Human reply? \ Has Smart Reply?	No	Yes	
No	**20,166** (58.2%)	**5,082** (14.7%)	72.9%
Yes	**5,972** (17.2%)	**3,406** (9.8%)	27.1%
	75.5%	24.5%	**N=34,626**

Table 1: Summary statistics about the 34,626 incoming messages in EnronSR dataset. The table shows the number of incoming messages with and without Smart Reply suggestions (columns) and with and without human replies (rows). In parenthesis are percentages of the total number of incoming messages, and row/column totals are shown along the table margins.

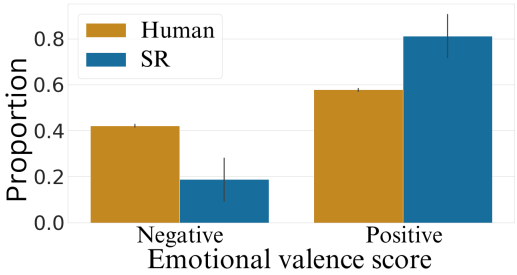

Figure 2: Differences in the emotional valence between human and SR generated responses.

We also examined the emotional valence of human- and AI-generated responses. To evaluate emotional valence robustly, we used three different measures: Flair NLP (Akbik et al. 2019), Textblob (Loria et al. 2018), and Vader (Hutto and Gilbert 2014). Flair yields positive or negative valence labels while taking into account contextual information around words. TextBlob and Vader extract valence scores between 1 and -1 using a lexical approach with a dictionary of positive and negative words.

Figure 2 shows the proportion of negative and positive emotional valence of human- and AI-generated responses on the same set of emails as measured by Flair. Similar results, obtained by TextBlob and Vader, are shown in the appendix. We observe that human replies fall in the negative bucket at twice the rate of the SRs. Moreover, they are more evenly distributed across the buckets, as opposed to the SRs, where the vast majority (81.2%) are positive. An example of such a mismatched sentiment between a human reply and smart replies is having a negative human reply saying "I do not want my name in the article" and 3 positive smart replies: "Will do.", "Thanks, I'll take a look.", "Looks good to me." This positivity bias is consistent with the findings of prior work (Hohenstein and Jung 2018; Hohenstein et al. 2023; Mieczkowski et al. 2021), but is still notable given its magnitude and Google's deliberate attempt to address the issue by adding negative-sentiment responses (Kannan et al. 2016).

Next, we examine the semantic characteristics of the SRs

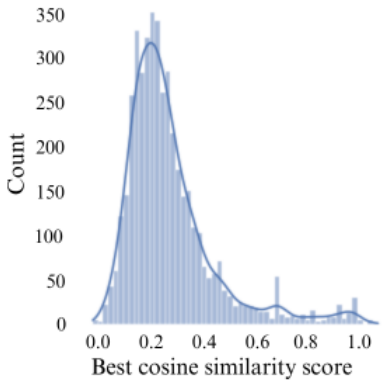

Figure 3: The distribution of the highest cosine similarity scores between the three Smart Replies and the human replies to each email.

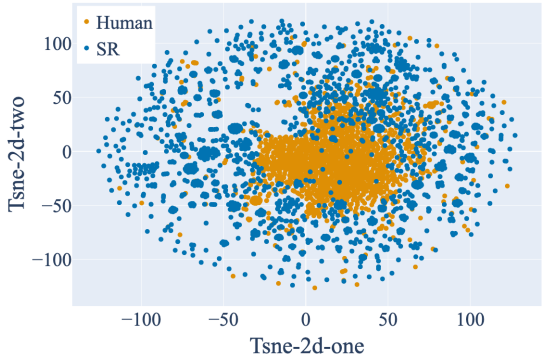

Figure 4: A visualization of the embeddings using t-SNE. Orange points represent human responses and blue points represent the SRs.

and human replies. To do that, we used BERT sentence embeddings (Reimers and Gurevych 2019) and measure cosine similarity between human and AI-generated responses. For brevity, we only report similarity results using the most similar SR of the three to the human response, providing an upper bound for similarity, although we experimented with additional configurations (averaging the embeddings of the three SRs, taking only the first SR suggestion, using only the first sentence of the human reply) and the results were not meaningfully different.

Figure 3 shows the full distribution of cosine similarity scores between the human reply and its closest SR suggestion. We find that the average cosine similarity of the most similar SR to the human response is 0.252 (0.247, 0.257). The appendix provides a few illustrative examples that demonstrate the meaning of relatively low and high cosine similarity of SRs. We also found that the average cosine similarity of the third suggestions by Smart Reply is significantly lower than the average similarity of the first and second replies ($p < 0.001$), and no significant difference between the first and second suggestions. Figure 4 further shows, using the t-SNE dimensionality reduction of human and SR embeddings, that human replies cluster together in a dense core, while the SRs are mostly scattered around them, a visual indication of the dissimilarity.

Speak Direction	Semantic Similarity
Intercept	0.291 (0.013)***
Down-speak	−0.050 (0.018)**
Same	−0.063 (0.016)***
Out-speak	−0.037 (0.014)

Note: N=2,408, significance levels are denoted as *** p<0.005, ** p<0.05, * p<0.1.

Table 2: Coefficients and their standard errors of a linear regression modeling cosine similarity as a function of speaker direction.

Finally, we use a simple regression model to examine whether semantic similarity is associated with the direction of speaking: up the chain (up-speak), down the chain (down-speak), to a peer of the same level (same), or outside the company (out-speak). Speaker direction was determined using the information about the employee roles in the company[4] for messages where both roles were known (N=2,408). Table 2 shows the results of this regression. The intercept represents the base level of up-speak, where the average response to someone with higher social status (up-speak) has cosine similarity of 0.29 (0.25, 0.32). SRs that down-speak have slightly lower, but statistically significant, cosine similarity as indicated by the negative coefficient (−0.05). Even more dissimilar, are suggestions between peers of the same social status with a negative coefficient of −0.06 and an average similarity of 0.22 (0.21, 0.24). Although SRs are generally quite far semantically from human replies as demonstrated in Figures 3 and 4, these results suggest that using AI-generated responses may portray a sender who is using SRs as slightly less authoritative and of lower social status.

Case Study of GPT-3

To demonstrate the utility of EnronSR for benchmarking language models, we conduct a case study of GPT-3. The goal of this case study is not to engage in extensive and exhaustive benchmarking of various language generation models, nor is meant to identify the optimal prompt for producing the best-performing responses. The case study is strictly focused on demonstrating how one can relatively easily use EnronSR to examine an AI-MC model. To that end, we collected 3,683 responses generated by GPT-3 for each of the Enron emails with human response and compared them to both the human responses and the SRs. We used the *text-davinci-003* model from OpenAI with the prompt: "write a response for this email message: {email_message}".

We found that the average length of the first sentence generated by GPT-3 is 9.54 (95% CI: 9.35, 9.74) words on average, which is a lot closer to the average of human replies

[4]Obtained from http://www.ahschulz.de/enron-email-data/

(9.79 words) than the average of the SRs (3.18 words). We also compared GPT-3 responses to SRs and human responses in terms of emotional valence, replicating Figure 2 with the addition of GPT-3. Figure 6 in the appendix shows the proportion of negative and positive responses as measured by Flair for humans, SR, and GPT-3. The figure shows that GPT-3 generates significantly fewer positive-biased responses ($p < 0.05$), and that it is better aligned with the polarity of human responses. Along the same line, we found that the average cosine similarity between GPT-3 and human responses is 0.309 (0.300, 0.318) is statistically higher than the similarity of SR and human responses of 0.252 (0.247, 0.257), although both suggestion models are still quite far semantically from the human replies.

Discussion & Conclusion

As AI-generated text becomes more widely adopted, it is important to understand how to make these technologies as relevant and as efficient as possible. In this paper, we introduced EnronSR, a novel dataset that provides human- and AI-generated responses to the same set of emails. EnronSR enables the benchmarking of new reply suggestion models and language generation architectures against the strong production-level model of Google Smart Reply and a ground truth of human replies. We demonstrated how one can utilize this corpus to benchmark reply suggestion models in a case study of GPT-3.

Our findings suggest that current, widely available AI-MC implementations are still quite far from perfectly capturing the nuances of human-to-human communication. We observe significant differences between human replies and SRs both in form and function: SRs are considerably shorter, more positive, and semantically distant from human replies. The results of our case study on GPT-3 indicate that more recent language generation models are capable of closing these gaps to some extent. Nevertheless, it is important to note that "pure" language generation models do not have the same safety and quality standards that production Learning to Rank models like Google Smart Reply adhere to. Namely, SRs are generated from a finite set of high-quality and safe-to-use clusters (Kannan et al. 2016), while GPT-3 provides no bounds for the generated text. Mechanisms for assuring the safety and quality of generated language are an active area of research, which could significantly contribute to the success of AI-MC systems.

The results also raise a series of normative questions about the role that AI-MC systems should fill in supporting human-to-human communication. The shortness of SRs may provide a fast and easy way to "close" a conversation. Surely, that is a desired outcome in some circumstances, but capturing when and where this is desired is a complex task that current models are not yet mastering as evident in the mismatch between human and SR responses. Similarly, the positivity bias we observed — reaffirming prior findings obtained using different methodology (Hohenstein and Jung 2018; Hohenstein et al. 2023; Mieczkowski et al. 2021) and persisting despite direct inclusion of negative responses in the SR model (Kannan et al. 2016) – may be a bias that people, in fact, want to exist. It is plausible that people prefer biased AI assistants that constantly skew towards more positive responses, but it is difficult to calibrate when, where, and to what degree this is desired. Future work could try to disentangle these complexities and find ways to promote user agency over these decisions.

Future work may also use EnronSR to track the progress of AI-MC systems in terms of the above-mentioned dimensions and explore additional dimensions of the text in which AI-MC are systematically biased. These may include aspects such as formality or politeness. The corpus may also be used as a resource for training new models that aim to reduce bias and as a resource for identifying additional objective functions for which models can optimize.

This work has several important limitations. First, the language people use nowadays in email communications may be different from that used by Enron employees 20 years ago. To ensure that the large differences we found between the human responses and the AI-generated responses do not simply stem from the use of outdated language in the original Enron corpus, two of the authors annotated a random sample of replies and found that no more than 7% of the responses contain outdated language (inter-coder reliability of Gwet AC1 0.95), which is unlikely to fully account for the differences found in this work. While our manual annotation of a random sample of replies suggests that 93% of responses could have been written today, as a resource that is frozen in time, its relevance may degrade over time as language is constantly evolving. The procedure we used to develop EnronSR could be replicated in the future if more up-to-date, publicly available corpus of human communications becomes available. Second, the communications in the Enron corpus may not be representative of the language generated by the general population or email communications outside the context of work. In addition, similar to other algorithmic audits, we cannot determine whether the same underlying Smart Reply model was used throughout our experiments or across different accounts (e.g., due to A/B testing). Particularly challenging is the issue of personalization, which despite our best efforts to limit its impact, may still exist to some extent in the set of collected responses. Nevertheless, the extensive robustness and validity checks conducted as part of this research ensure that the corpus is at least internally valid, even if external validity is somewhat limited.

In summary, this work presents EnronSR, a new resource for benchmarking AI-MC models in an open and reproducible manner, identifies significant differences between human- and AI-generated responses, and highlights several aspects of generated responses that future models could align better with human responses.

Code Availability Statement

All replication code is publically available at https://github.com/Socially-Embedded-Lab/EnronSR for academic usage.

Ethical Statement

This study was approved by the ethics review board of Ben-Gurion University (protocol #344-1). Although the founda-

tion of EnronSR is a public and widely-used dataset, released by the U.S. justice system in 2004, we took several steps to minimize any potential risks that may stem from our annotation of the original corpus. No personally identifying information (PII) is released as part of this dataset. The Smart Reply suggestions included in EnronSR were manually validated to ensure that no PII or sensitive information was included. Our validation is on top of the internal processes at Google for selecting SRs that are popular and that meet quality and safety guidelines assessed by human raters (Kannan et al. 2016). The email accounts used as part of this research were created in compliance with Google's terms of service (TOS), and were permanently and immediately deleted as soon as the data collection concluded. Specifically, the audit did not violate any spam law, mislead or deceive users, or circumvent any Google policy. AI-generated content is not protectable under copyright or patent law, and according to Google policy[5], SRs as frequent user responses are not considered Google's content.

It is important to note that there is a growing professional and legal consensus that it is in the public's interest to allow academic researchers to study major publicly deployed algorithms to investigate social bias and other forms of potentially harmful algorithmic practices even when such studies violate TOS (Metaxa et al. 2021). As stated, our work did not violate Google's TOS, but it does investigate a major publicly-deployed algorithm that could have a significant impact on individuals and society. To limit potential misuses of this benchmark, we are sharing EnronSR with a CC BY-NC-SA 4.0 (Attribution-NonCommercial-ShareAlike) license to allow others to build upon this work non-commercially, with proper attribution, and sharing under the same terms.

Acknowledgements

This research was supported by grant #1247 from the Ministry of Science & Technology, Israel.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Akbik, A.; Bergmann, T.; Blythe, D.; Rasul, K.; Schweter, S.; and Vollgraf, R. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the conference of the North American chapter of the association for computational linguistics (NAACL)*, 54–59.

Arnold, K. C.; Chauncey, K.; and Gajos, K. Z. 2018. Sentiment bias in predictive text recommendations results in biased writing. In *Proceedings of the Graphics Interface Conference (GI)*, 42–49.

Basta, C.; Costa-jussà, M. R.; and Casas, N. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33–39. Florence, Italy: Association for Computational Linguistics.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610–623.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 4356–4364.

Brandtzaeg, P. B.; and Følstad, A. 2017. Why people use chatbots. In Kompatsiaris, I.; Cave, J.; Satsiou, A.; Carle, G.; Passani, A.; Kontopoulos, E.; Diplaris, S.; and McMillan, D., eds., *Internet Science*, 377–392.

Brock, A. 2020. *Distributed blackness: African American cybercultures*. New York, NY: NYU Press. ISBN 9781479829965.

Brunet, M.-E.; Alkalay-Houlihan, C.; Anderson, A.; and Zemel, R. 2019. Understanding the origins of bias in word embeddings. In *International conference on machine learning (ICML)*, 803–811.

Burgoon, J. K. 1993. Interpersonal expectations, expectancy violations, and emotional communication. *Journal of language and social psychology*, 12(1-2): 30–48.

Burgoon, J. K.; Stern, L. A.; and Dillman, L. 1995. *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press.

Chakravarthi, N.; Pasternack, J.; Leon, A.; Rajashekar, N.; Birjodh Tiwana, R.; and Zhao, B. 2017. Building Smart Replies for Member Messages. Available at https://engineering.linkedin.com/blog/2017/10/building-smart-replies-for-member-messages.

Chen, M. X.; Lee, B. N.; Bansal, G.; Cao, Y.; Zhang, S.; Lu, J.; Tsay, J.; Wang, Y.; Dai, A. M.; Chen, Z.; et al. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2287–2295.

Cohen, W. W. 2015. Enron email dataset. Available at https://www.cs.cmu.edu/~./enron/ and also on the Internet Archive https://web.archive.org.

Crawford, K.; Miltner, K.; and Gray, M. 2014. Critiquing big data: Politics, ethics, epistemology. *International Journal of Communication*, 8: 1663–1672.

Cuthbertson, S. 2019. Sharing what's new in Android Q. Available at https://blog.google/products/android/android-q-io/.

Dinan, E.; Fan, A.; Williams, A.; Urbanek, J.; Kiela, D.; and Weston, J. 2020. Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8173–8188.

[5] Available at https://policies.google.com/terms#toc-content.

Gillespie, T. 2014. The relevance of algorithms. In Gillespie, T.; Boczkowski, P.; and Foot, K., eds., *Media technologies: Essays on communication, materiality, and society*, 167–193. Cambridge, MA: MIT Press.

Hancock, J. T.; Naaman, M.; and Levy, K. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1): 89–100. Publisher: Oxford University Press.

Henderson, P.; Sinha, K.; Angelard-Gontier, N.; Ke, N. R.; Fried, G.; Lowe, R.; and Pineau, J. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 123–129.

Hohenstein, J.; and Jung, M. 2018. AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)*, 1–6.

Hohenstein, J.; Kizilcec, R. F.; DiFranzo, D.; Aghajari, Z.; Mieczkowski, H.; Levy, K.; Naaman, M.; Hancock, J.; and Jung, M. F. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1): 5487.

Hutto, C.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 8, 216–225.

Ishii, H.; Kobayashi, M.; and Grudin, J. 1993. Integration of interpersonal space and shared workspace: ClearBoard design and experiments. *ACM Transactions on Information Systems (TOIS)*, 11(4): 349–375.

Jakesch, M.; French, M.; Ma, X.; Hancock, J. T.; and Naaman, M. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1–13.

Kaneko, M.; and Bollegala, D. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742*.

Kannan, A.; Kurach, K.; Ravi, S.; Kaufmann, T.; Tomkins, A.; Miklos, B.; Corrado, G.; Lukacs, L.; Ganea, M.; Young, P.; and Ramavajjala, V. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 955–964.

Kirk, H. R.; Jun, Y.; Volpin, F.; Iqbal, H.; Benussi, E.; Dreyer, F.; Shtedritski, A.; and Asano, Y. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34: 2611–2624.

Kumar, K. 2023. Geotechnical Parrot Tales (GPT): Overcoming GPT hallucinations with prompt engineering for geotechnical applications. *arXiv preprint arXiv:2304.02138*.

Landowski, L.; and El Moujahid, K. 2017. M Now Offers Suggestions to Make Your Messenger Experience More Useful, Seamless and Delightful. https://about.fb.com/news/2017/04/m-now-offers-suggestions-to-make-your-messenger-experience-more-useful-seamless-and-delightful/.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Li, J.; Monroe, W.; Ritter, A.; Galley, M.; Gao, J.; and Jurafsky, D. 2016b. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Li, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Lim, S.; and Schmälzle, R. 2023. Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. *Frontiers in Communication*, 8: 1129082.

Lison, P.; and Bibauw, S. 2017. Not all dialogues are created equal: Instance weighting for neural conversational models. *arXiv preprint arXiv:1704.08966*.

Liu, H.; Dacon, J.; Fan, W.; Liu, H.; Liu, Z.; and Tang, J. 2020. Does Gender Matter? Towards Fairness in Dialogue Systems. In *Proceedings of the International Conference on Computational Linguistics*, 4403–4416.

Liu, Y.; Bi, W.; Gao, J.; Liu, X.; Yao, J.; and Shi, S. 2018. Towards less generic responses in neural conversation models: A statistical re-weighting method. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2769–2774.

Loria, S.; et al. 2018. textblob Documentation. *Release 0.15*, 2(8).

Lucy, L.; and Bamman, D. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Workshop on Narrative Understanding*, 48–55.

Marcelis, D.; and MacMillan, D. 2018. Is This Article Worth Reading? Gmail's Suggested Reply: "Haha, Thanks!". Published on Sepetember 18, 2018 and available at https://www.wsj.com/articles/very-interesting-awesome-love-it-gmail-users-confront-chipper-smart-reply-1537282569.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; and Sandvig, C. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

Mieczkowski, H.; Hancock, J. T.; Naaman, M.; Jung, M.; and Hohenstein, J. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text

Conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3349–3358. Osaka, Japan: The COLING 2016 Organizing Committee.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one*, 15(8): e0237861.

Noble, S. U. 2018. *Algorithms of oppression: How search engines reinforce racism*. New York, NY: NYU Press. ISBN 1479837245.

O'Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Broadway Books.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Park, S.; Zhang, A. X.; Murray, L. S.; and Karger, D. R. 2019. Opportunities for automating email processing: A need-finding study. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 1–12.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Robertson, R. E.; Olteanu, A.; Diaz, F.; Shokouhi, M.; and Bailey, P. 2021. "I can't reply with that": Characterizing problematic email reply suggestions. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)*, 1668–1678.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3239–3254. Online: Association for Computational Linguistics.

Shwartz, V.; Rudinger, R.; and Tafjord, O. 2020. "You are grounded!": Latent Name Artifacts in Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6850–6861. Online: Association for Computational Linguistics.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems (NIPS)*, 27.

Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Vogels, E. A. 2023. A majority of Americans have heard of ChatGPT, but few have tried it themselves. Available at https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves.

Weng, Y.; Zheng, H.; Bell, F.; and Tur, G. 2019. OCC: A Smart Reply System for Efficient In-App Communications. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2596–2603. ISBN 9781450362016.

Wenker, K. 2023. Who wrote this? How smart replies impact language and agency in the workplace. *Telematics and Informatics Reports*, 10: 100062.

Yang, X.; Awadallah, A. H.; Khabsa, M.; Wang, W.; and Wang, M. 2018. Characterizing and supporting question answering in human-to-human communication. In *The International ACM Conference on Research & Development in Information Retrieval (SIGIR)*, 345–354.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, this new benchmark contributes to the ability to observe biases in existing AI-generation systems as a first step towards bridging those gaps.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We discuss the representativeness of emails in the original Enron corpus, including direct examination of whether the original text is outdated, and relating to questions about the bias in response quality for subpopulations.
   (e) Did you describe the limitations of your work? Yes, this appears in the Discussion and Conclusion section.

(f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethical Statement section. Specifically, we refrained from sharing any PII data and explicitly made sure that none of the generated responses include such information. All of the accounts created during the course of the audit were terminated and permanently deleted as soon as the data collection concluded.

(g) Did you discuss any potential misuse of your work? Yes, the Ethical Statement section contains this discussion as well as the steps to mitigate it.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, this is in the Ethical Statement section.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? N/A

(b) Have you provided justifications for all theoretical results? N/A

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A

(e) Did you address potential biases or limitations in your theoretical framework? N/A

(f) Have you related your theoretical results to the existing literature in social science? N/A

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. All replication materials are publicly available.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. The text contains all of the details of the algorithmic audit we conducted as well as the ensuing analysis.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No. The work required nothing more than a general-purpose personal computer.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, this is discussed in regards to SR over- and under-triggering.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? Yes.

(c) Did you include any new assets in the supplemental material or as a URL? Yes.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Ethical Statement for details.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see Ethical Statement for details.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? This does not apply to an algorithmic audit of an existing, widely-used algorithm.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? This paper includes all of the elements of the recommended Datasheet (motivation, composition, collection process, recommended uses, and more), and effectively serves as the datasheet for EnronSR.

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

(a) Did you include the full text of instructions given to participants and screenshots? There was no active participation of individuals in this study, only secondary analysis of publically available data.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see Ethical Statement for details.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? There was no active participation of individuals in this study, only secondary analysis of publically available data.

(d) Did you discuss how data is stored, shared, and deidentified? Yes.

EnronSR Schema

The dataset consists of two table: one representing incoming messages and another representing replies to incoming

Field	Description
incoming_msg_id	Message ID of an email in the Enron corpus that was sent to one of the 150 focal users.
recipient_msg_id	Message ID of a random email in the Enron corpus that was sent by the recipient of the current message. This is useful for identifying the particular recipient of the incoming message for whom we collected SRs for.
sr1	The first Smart Reply that was suggested by Google, empty otherwise.
sr2	The second Smart Reply that was suggested by Google, empty otherwise.
sr3	The third Smart Reply that was suggested by Google, empty otherwise.
has_hr	Indicator whether the incoming message has at least one human reply.

Table 3: Schema for the incoming messages table.

Field	Description
incoming_msg_id	Message ID of an email in the Enron corpus that was sent to one of the 150 focal users.
recipient_msg_id	Message ID of a random email in the Enron corpus that was sent by the recipient of the current message. This is useful for identifying the particular recipient of the incoming message for whom we collected SRs for.
reply_msg_id	Message ID of an email in the Enron corpus where the focal user replied to the incoming message.

Table 4: Schema for the human replies table.

Email	Response	SR 1	SR 2	SR 3	Score
Guys, I got the following approved by Whalley. We should sit down and discuss next steps asap	when? i am free anytime this aft	Sounds good.	Great!	Yes, let's discuss.	0.308
Maybe a conference call would be the most productive way to go.	I'm free after 11 am	Yes, I agree.	Sounds good.	I don't think so.	0.297
...I would like for you to provide me with details of your previous rotations..	I worked in ENA West Risk from July 2000 until March 1 , 2001 ...	Noted, will do.	Thank you, will do.	Noted with thanks.	0.145
Please let me know if you plan to attend	I will be attending.	I will be attending.	I will be attending the meeting.	Yes, I will be attending.	0.861
hope everyone is okay with this	Sounds good to me	Sounds good to me	Fine with me.	Sounds good	0.698

Table 5: Examples of low and high cosine similarity. Top three rows have low similarty while the bottom two rows have high similarity.

messages. The schemas for these tables are in Table 3 and Table 4, respectively.

Examples Responses

Table 5 provides examples of low and high cosine similarity suggestions to provide readers with a qualitative sense of the matches and mismatches. Each raw represents an incoming message, its human reply, and the SRs.

Additional Measures of Emotional Valence

Figure 5 provides histograms of the emotion valence of human- and AI-generated responses as obtained using Textblob (Loria et al. 2018) and Vader (Hutto and Gilbert 2014), respectively.

Emotional Valence of GPT-3

Figure 6 shows the proportion of negative and positive responses generated by humans, SRs, and GPT-3. The figure replicates the proportions appearing in Figure 2 with the addition of the polarity of responses obtained through GPT-3.

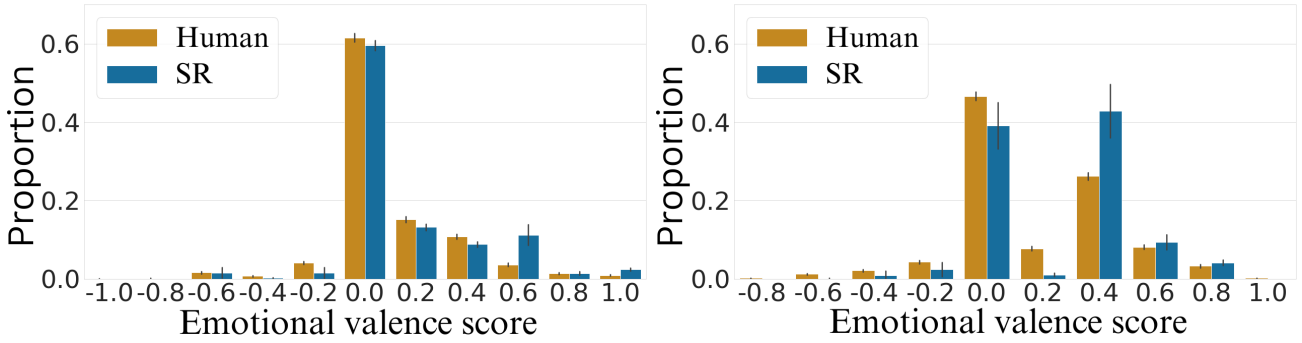

Figure 5: Emotional valence scores as obtained by TextBlob (left) and Vader (right) for human and SR generated responses.

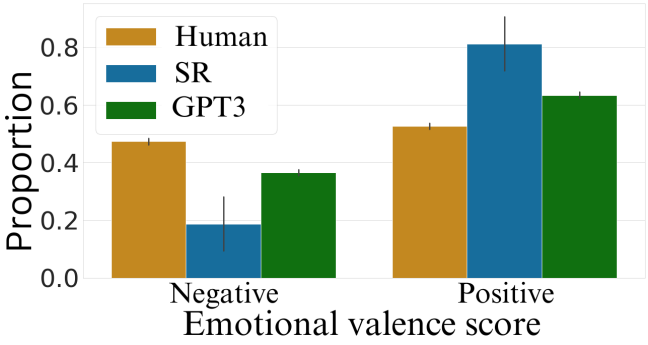

Figure 6: The proportion of positive and negative responses, as measured through Flair, when generated by humans, SR, and GPT-3.