# Emails by LLMs: A Comparison of Language in AI-Generated and Human-Written Emails

Weijiang Li
University of Illinois Urbana-Champaign
Urbana, IL, USA
wl13@illinois.edu

Yinmeng Lai
University of Illinois Urbana-Champaign
Urbana, IL, USA
yinmeng2@illinois.edu

Sandeep Soni
Emory University
Atlanta, GA, USA
sandeep.soni@emory.edu

Koustuv Saha
University of Illinois Urbana-Champaign
Urbana, IL, USA
ksaha2@illinois.edu

## Abstract

The growing excitement around generative AI (and LLMs) is fueling a heightened interest in the development of AI-assisted writing tools. One popular context is AI-assisted email writing, and this paper explores how AI-generated emails compare to human-written emails. We obtained human-written emails from the W3C corpus and generated analogous AI-generated emails using GPT-3.5, GPT-4, Llama-2, and Mistral-7B, and compared AI-generated and human-written emails using a suite of natural language analyses across syntactic, semantic, and psycholinguistic dimensions. AI-generated emails are generally consistent across different LLMs but differ significantly from human-written emails. Specifically, AI-generated emails tend to be more formal, verbose, and complex, whereas human-written emails are often more concise and personalized. While AI-generated emails are slightly more polite, both types exhibit a similar level of empathetic tone in language. Further, we qualitatively examined user perceptions of AI and human-written emails by conducting a small survey of 41 participants and interviewing a subset of them. This study highlights preliminary insights into generative AI's distinct strengths and weaknesses in assisting email communication, and we discuss the theoretical and practical implications of the evolving landscape of AI-generated content.

## CCS Concepts

• **Human-centered computing** → *Empirical studies in collaborative and social computing*; *Social media.*

## Keywords

email, language, generative AI, LLMs, linguistic analysis

**ACM Reference Format:**
Weijiang Li, Yinmeng Lai, Sandeep Soni, and Koustuv Saha. 2025. Emails by LLMs: A Comparison of Language in AI-Generated and Human-Written Emails. In *Proceedings of the 17th ACM Web Science Conference 2025 (Websci '25), May 20–24, 2025, New Brunswick, NJ, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3717867.3717872


*Websci '25, New Brunswick, NJ, USA*

ACM ISBN 979-8-4007-1483-2/2025/05
https://doi.org/10.1145/3717867.3717872

## 1 Introduction

Large Language Models (LLMs) have been instrumental in reshaping the landscape of AI and language technology. LLMs, a class of highly capable models trained on massive amounts of text [54, 55, 99], have transitioned within a short time from being part of a research toolkit to being integrated into many commercial applications. The allure of LLMs is that a variety of day-to-day tasks such as planning a trip, seeking cooking recipes, captioning photos, or writing emails can be performed with ease by interfacing with LLMs in natural language [1]. Not only have LLMs transformed the way in which we produce and comprehend language, but they have also become part of the cultural zeitgeist.

A common and widely adopted application of LLMs in everyday use is in writing tasks such as crafting emails. An assortment of tools exist for this purpose: standalone applications such as ChatGPT to tightly integrated all-platform tools such as Grammarly [29]. These tools can aid the writer in adjusting their content, tone, style, and syntax. In this paper, we ask how crafting emails with the assistance of an LLM compares with emails crafted by humans. We posit that, on the one hand, LLMs can be instructed with ease for this task—also potentially helping break linguistic and social barriers—the over-reliance of LLMs can be perceived negatively if their collective use leads to homogenization of content and linguistic style. Despite the delicate balance that is needed in the use of this technology for writing emails, there remains no systematic investigation to uncover the dimensions along which LLMs differ from humans in the task of writing emails.

A systematic investigation is also needed for at least two additional reasons. First, despite the availability of many alternatives for online communication, emails continue to be ubiquitous in both social and professional settings. The writing and interpreting of emails is laced with its own unique challenges needing a tight adherence to social and communicative norms in the absence of visible social cues in other forms of interpersonal communication [68]. Therefore, it is critical to evaluate the capabilities of LLMs for linguistic markers of these norms and assess how emails written by LLMs are perceived to build confidence in their role as assistive technology. Second, while LLM-powered tools can enhance the user's efficiency and perceived writing quality, they concurrently raise considerations regarding the origin and genuineness of the content—is something human-written, machine-generated, or machine-improved? Furthermore, it remains unknown whether

AI models sufficiently incorporate context in generating cohesive information—a crucial aspect of email communication.

Empirical insights into the above questions will help us understand the strengths and limitations of AI-writing assistance, guiding the design and development of more effective and reliable tools. Additionally, understanding these dynamics will provide preliminary insights into AI's integration into mainstream tasks such as workplace communication and the new challenges and considerations it may introduce. To this end, we ask the research question (RQ): ***How do AI-generated emails linguistically compare against human-written emails?***

We conducted a linguistic examination of emails based on comparing and contrasting human-written and AI-generated email content. We leveraged the email dataset from the W3C corpus [89], which we consider the human-written email dataset. Then, we used a suite of LLMs (GPT-3.5, GPT-4, Llama-2, and Mistral-7B) to obtain our AI-generated email datasets. We examined linguistic attributes spanning syntactic, semantic, and psycholinguistic analyses. We also conducted small user studies with 41 participants through surveys and interviews that focused on understanding user perceptions about AI-generated versus human-written emails.

Our investigations revealed that AI-generated emails are different from human-written emails based on linguistic attributes with statistical significance. Specifically, AI-generated emails are consistently more formal, encode a neutral-to-positive sentiment, are more readable, and tend to stick with uniform semantics and topics. However, AI-generated emails tend to be more complex, often reuse the same words, and are more verbose compared to human-written emails. This work bears theoretical implications in understanding how AI-generated emails perform in language compared to human-written emails and practical and design implications for designing tools for AI-assisted email writing.

## 2 Related Work

**Application of LLMs in Writing Assistance.** Large Language models (LLMs) are a type of foundational models [9] typically characterized by training on vast amounts of text, having billions of parameters, and showing capabilities on a range of linguistic and general-purpose comprehension or generation tasks. Recent examples of LLMs include both commercial models such as OpenAI's GPT [60], Google's Gemini [83] and PaLM [6], and open-access models such as Meta's Llama [87] and Mistral [42].

A popular application of LLMs is writing-assistance [45], including, but not limited to, chatbots such as ChatGPT, Gemini [86], and Sparrow [33], coding assistants such as GitHub Copilot [59], Code Llama [69], and Lemur [92], question-answering applications in domains such as law [74] and medicine [64], and creative writing narrative composition applications such as screenplay generation such as Dramatron [56], Re3 [94] and Detailed Outline Control [93], and several others [15, 38, 96]. More integrative and task-agnostic writing applications include LLM-for-X, which supports workflows for multiple tasks such as article composition or coding [85] and GhostWriter [95], which builds a human-AI collaborative writing interface. The development of these applications and concurrent progression in LLM capabilities can be aided by a thorough understanding of the dimensions along which LLM-enhanced writing differs from human writing, which is the focus of this paper.

To interact with LLMs, users employ *prompting* [50, 91] where an LLM is given input in the form of text such as a dialogue turn, a question or a fragment of a story, along with instructions to get a response such as the next turn in the dialogue or answer to the question, or the rest of the story, respectively. Recent work investigated the use of LLMs by knowledge workers [22], inferring that LLMs assist workers in both composition and comprehension tasks such as text summarization and writing improvement, respectively [10].

*LLM-Assisted Writing and Perceptions.* Prior work on AI-assisted email writing focused on enhancing the experience of people with special needs. Goodman et al. [35] studied the challenges of email writing for people with dyslexia, and designed a prototype email-writing interface that uses LLMs to power writing support tools [35], and Buschek et al. [14] investigated the impact of multi-word suggestion choices on text composition for non-native English writers and developed a text editor prototype with GPT-2 [14].

Recent research on the professional perception of AI-assisted writing reveals that AI-generated messages are professional, effective, efficient, confident, and direct [20]. Liu et al. examined the trust perceptions of AI-mediated email writing [51], and Padmakumar and He studied if writing with LLMs reduces content diversity [61]. Relatedly, Kacena et al. studied AI-assisted academic writing to find that AI-based writing tools effectively reduce the time for writing but potentially include inaccurate information, and thus writers should utilize such tools with caution [44].

**Language Analysis and Online Communication in the Workplace.** The ubiquity and widespread use of online technologies in professional and workplace settings has also enabled a body of research to emerge in studying this data to understand worker behaviors [24]. For instance, Ehrlich and Shami examined employees' motivations for using social media, particularly Twitter [27], finding that social media engagement, both at work and home, fostered a sense of connection among workers—especially mobile workers—and helped enhance professional reputation. Research has also shown a positive correlation between social media usage and workplace wellbeing [76]. For example, IBM's Beehive platform offered benefits in networking, career advancement, and innovation through increased workplace social interactions [25, 26, 28, 31]. A number of analytical and computational techniques, including language and network dynamics, have been applied to investigate factors influencing outcomes such as employee engagement [58, 75], affect [24, 71], organizational role [72], social dynamics [76], reputation [39], organizational relationships [11, 32, 57], workplace behavior [53], and job satisfaction [73].

In particular, email communication is an important factor at work [4, 97], and the language used in email communication carries information indicative of a person's behavior in the organization, such as how one addresses another person due to the hierarchy of power [8]. Patil et al. [62] modeled organizational attrition using email communication, and Mitra and Gilbert [57] studied how workplace gossip manifests in email communication using the publicly-available Enron email corpus [79]. Relevant to our

**Table 1: List of prompts used to generate emails.**

| |
|---|
| Given the following email, help me write a reply to the email with appropriate tones and information: |
| Help me write a response to the following email from my colleague |
| Please help me write a response to the given email from a coworker in the company: |
| The following is an email from one of my coworkers in the company, help me write an email to reply: |
| I have an email from a colleague, help me write back to this person: |

work is Robertson et al. [68]'s work on characterizing problematic automatic email suggestions on Outlook.

**Comparison to Prior Work** With growing excitement surrounding the development and deployment of LLMs, LLMs are no longer a research topic but a practical tool integrated into our daily lives. Despite recent interest in AI-assisted writing [34], our understanding of LLMs' potential in email writing—an everyday online communication used both professionally and casually—remains empirically underexplored. Our work addresses the theoretical gap by providing an empirical comparison between AI-generated emails and human-written emails. Our mixed methods study examines the content and the perceptions of the emails. We borrowed from natural language analyses to examine the content of emails, as well as user studies to understand participants' perceptions of emails. Our work is situated in the emerging space of human-AI collaboration, around questions related to AI alignment, trust, and perceptions [30, 78, 84].

## 3 Data

This study compares the language of human-written and AI-generated emails. In this section, we describe our dataset collection and construction. Our email dataset consists of several email threads, each containing a sequence of email conversations.

### 3.1 Email Dataset Construction

*3.1.1 Human-Written Email Dataset.* We sourced our email dataset from the World Wide Web Consortium (*w3c.org*) corpus [21, 89]. Upon downloading the W3C email corpus, we used the pipeline developed by Zhang et al. [98] to process and format the raw email data into structured email threads, and each thread is restricted to a conversation between two people. This results in 80 threads and 296 emails in total, out of 9,841 email threads ($E_H$).

*3.1.2 AI-based Email Generation.* To obtain an AI-generated email dataset to compare with, we used a variety of LLMs—GPT-3.5 and GPT-4 [60], Llama-2 [87], and Mistral-7B [42]. These LLMs cover a suite of different architectures, parameters, and training datasets. For all of these LLMs, we used the default `temperature` setting of 1.0 for generating emails

To each of these LLMs, we prompted the first emails in $E_H$ as prompts to generate follow-up email threads using the default temperature setting of 1.0. Given that our human-written email corpus (W3C dataset) consists of emails in a workplace setting, we tailored our prompts to workplace and professional settings—these prompts are listed in Table 1. We iteratively applied this process to all email threads in the $E_H$ dataset. Finally, we end up with four LLM-generated datasets with four LLMs considered. Each of these datasets consisted of 80 email threads and a total of 296 emails.

For ease of exposition, this paper mainly focuses on comparing human-written emails with GPT-4-generated emails. Later, in section 6, we provide a comparison with all LLMs. Figure 1 shows an overview of our study design and methodologies.

## 4 Quantitative Examination

Our study aims to provide empirical evidence of how LLMs perform in content generation, especially email writing, by comparing and contrasting the language of human-written and AI-generated emails. Essentially, we approached the problem in a theory-driven fashion [23, 70] and investigated the distinctive characteristics between human-written and AI-generated emails across three linguistic dimensions: 1) syntactic, 2) semantics, and 3) psycholinguistics. The following subsections describe the analyses, their rationale, and observations. For each analysis, we measured effect size (Cohen's *d*), and conducted *t*-tests and Kolmogorov-Smirnoff (KS) tests for statistical significance. Table 2 and Table 3 summarize the results, which we go through in detail in this section.

### 4.1 Syntactic Analyses

*4.1.1 Verbosity.* Verbosity (and conciseness) is a key linguistic feature in written communication [40]. We operationalized three measures of verbosity by counting the number of—1) *words per sentence*, 2) *words per email*, and 3) *sentences per email*. We find that AI-generated emails (mean=193.4 words) contain 166.8% more words than human-written emails (mean=72.5 words), along with a statistically significant difference as per *t*-test ($t$=19.47, p<0.05). Further, AI-generated email responses contain 129.2% more sentences than human-written emails (per Table 2). When we manually looked into the AI-generated emails, we found significant occurrences of greetings and polite expressions [68], like *"Please let me know if there is anything specific you would like me to focus on [...]"* and discussions of implications of matters, such as in *"I understand the importance of staying updated, especially with the rapid changes in the IETF expiration process. I'll review the document from the link you provided on your homepage."*, where the earlier email only mentions *update the website*.

*4.1.2 Readability.* Readability is a measure of text comprehensibility. Drawing on prior work, we calculated the Coleman-Liau Index (CLI) [66, 90]. We measured the Coleman-Liau Index (CLI) [19] to assess the understandability of email texts, which is defined as $CLI = (0.0588L - 0.2596S - 15.8)$, where $L$ is the average number of letters per 100 words and $S$ is the average number of sentences per 100 words. CLI corresponds to U.S. grade levels—the higher the score, the higher the education level required to comprehend the text. We find that GPT-generated emails generally have higher CLI scores (mean=11.3), with human-written emails having significantly lower scores (mean=6.82). While higher readability indicates a "better" quality of writing, it also entails that one requires a higher level of education to appropriately comprehend the email content AlAfnan and MohdZuki [3]. For example, *"I concur that advocating for a model that supports both mutable and immutable revisions with consistent operation interpretations does simplify the configuration management process [...] Could we perhaps schedule a*

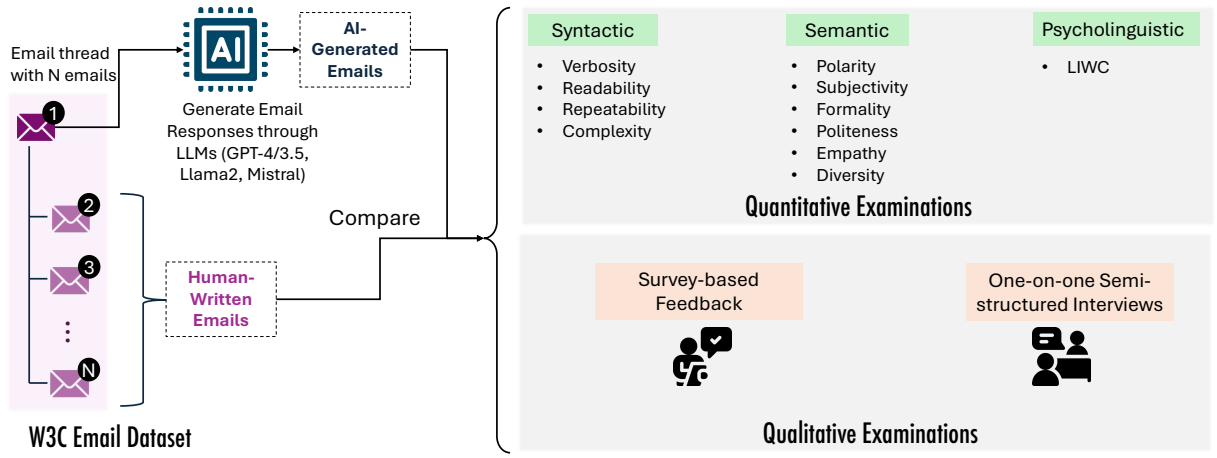



**Figure 1: A schematic overview of our study design.**

*time to discuss this further [...]"* is a response generated by GPT-4 with a CLI score of 17.56 which indicates that it is likely that only a professional would be able to understand this. On the other hand, *"I guess I don't really get what it means to "checkout" an old revision that will be replaced on checkin and not branched. To mean they seem like dissimilar things that need to interoperate."* is a human-written response, with a CLI score of 8.73. The key difference is that LLM-generated texts tend to use more complicated words, longer sentences, or fewer sentence breaks; whereas humans prefer daily vocabulary. This shows that humans are more likely to *naturally write emails* in a more casual style, whereas LLMs are trained to convey the message more professionally.

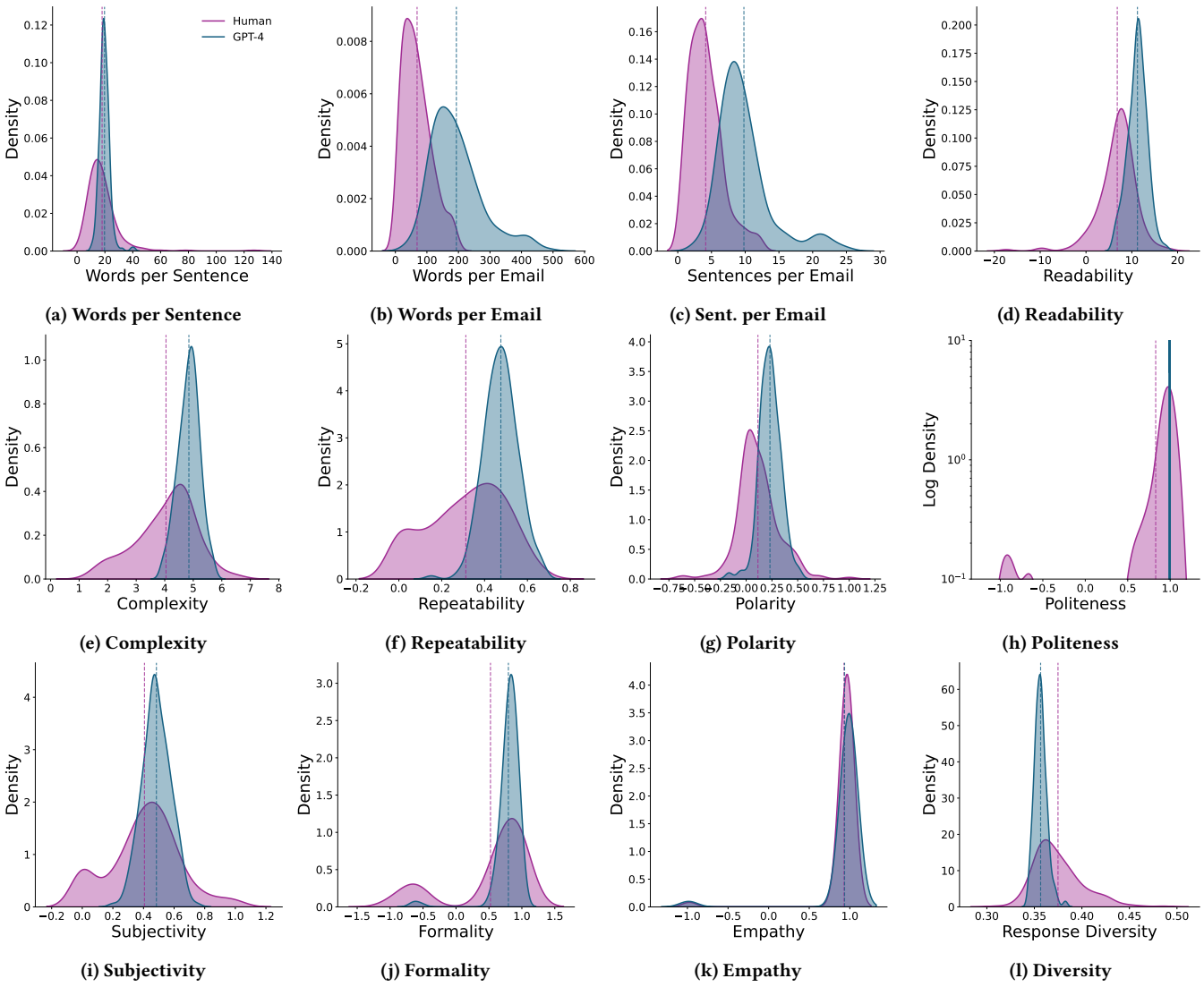
*4.1.3 Complexity.* Complexity is a metric that presents the average length of words per sentence across the entire text. Prior work noted that complexity is negatively associated with the perceived likelihood of an AI agent [90]. Different from readability, which measures character-to-word ratio, complexity considers word lengths within sentences. A higher score indicates longer words are used, suggesting a more sophisticated vocabulary. For example, in one of the emails generated by an LLM: *"Thank you for sharing the new URL scheme proposal. It definitely seems like something that's moving forward quickly, especially with implementations already in place. I can take a look at the draft and will provide feedback or sign off by [...]"* whereas, for the same email, the human-written email response was *"Ok, educate me. What is a VEMMI???"*. From this example, we can observe that LLMs use longer, multi-clause sentences with sophisticated vocabulary while humans prefer short, direct ones with simpler language. On average, GPT-generated emails (mean=4.85) scored 19.8% higher than human-written emails (mean=4.05) with statistical significance (Cohen's $d$=1.02, $t$=10.6, $p$<0.01).

*4.1.4 Repeatability.* Repeatability is another syntactic-based linguistic feature that measures the average number of non-unique words used in each email (ranging from 0 to 1), with a score closer to 0 meaning fewer words are repeated, and vice versa for a score close to 1. We found GPT-generated emails (mean=0.48) to have 52.30% higher repeatability score compared to humans (mean=0.31)

**Table 2: Summary of comparing email syntactic and semantic measures of Human-Written vs GPT-4 Generated Emails, with Cohen's $d$, $t$-test, and Kolmogorov–Smirnov (KS) test ($^{*}$ $p$<0.05, $^{**}$ $p$<0.01, $^{***}$ $p$<0.001).**

| Metric | GPT-4 | Human | Δ% | d | t | KS |
|---|---|---|---|---|---|---|
| **Syntax** | | | | | | |
| Verbosity: Words per Sentence | 19.70 | 18.00 | 9.62 | 2.00 | 0.19* | 0.41*** |
| Verbosity: Words per Email | 193 | 68.8 | 181.00 | 1.88 | 19.5*** | 0.73*** |
| Verbosity: Sentences per Email | 9.81 | 4.14 | 137.00 | 1.71 | 17.8*** | 0.75*** |
| Readability | 11.30 | 6.82 | 64.60 | 1.35 | 14.0*** | 0.61*** |
| Complexity | 4.85 | 4.05 | 19.80 | 1.02 | 10.60*** | 0.45*** |
| Repeatability | 0.47 | 0.31 | 52.30 | 1.17 | 12.20*** | 0.50* |
| **Semantics** | | | | | | |
| Polarity | 0.23 | 0.11 | 103.00 | 0.74 | 7.65*** | 0.47*** |
| Subjectivity | 0.48 | 0.40 | 19.20 | 0.45 | 4.67*** | 0.32*** |
| Formality | 0.80 | 0.52 | 51.90 | 0.58 | 6.06*** | 0.28*** |
| Politeness | 1.0 | 0.83 | 16.81 | 0.54 | 5.57*** | 0.85*** |
| Empathy | 0.94 | 0.93 | 0.75 | 0.024 | 0.25 | 0.85*** |
| Diversity | 0.36 | 0.38 | -4.87 | -0.98 | -10.10*** | 0.50*** |

with statistical significance (Cohen's $d$=1.17, $t$=12.20, $p$<0.01). We also note that an LLM is more likely to repeat some key entities in the email, and in the context of this dataset, they tend to repeat technical terms such as "URL" or "link". For example, while a human-written email, *"I have experience to get WEB or Binary files by using your service, but I am fail when I try to get file with above URL via email. The program is free, and I can use FTP explorer to get it. What is the problem?"*, has a repeatability score of 0.263, a GPT-generated email response for the same thread, "Thank you for reaching out with your query regarding the difficulty in accessing the file through the provided FTP link via email. It seems that the issue might be related to how FTP links are handled within email clients [...]" has a repeatability score of 0.612 since 'FTP' was mentioned multiple times. This suggests that GPT-4-generated emails, in general, have a less diverse vocabulary in generated texts [5, 48], whereas human-written emails cover a broader vocabulary and a wider range of topics via more varied language.




**(a) Words per Sentence** **(b) Words per Email** **(c) Sent. per Email** **(d) Readability**

**(e) Complexity** **(f) Repeatability** **(g) Polarity** **(h) Politeness**

**(i) Subjectivity** **(j) Formality** **(k) Empathy** **(l) Diversity**

**Figure 2: Comparison of the distribution of linguistic features measures between GPT-4 (LLM) and human-written email responses. Dash lines represent the means of respective distribution.**

## 4.2 Semantic Analyses

*4.2.1 Polarity.* Politeness is a core component of Polarity, a feature in sentiment analysis provided by TextBlob [52]. It quantifies the sentiment of a text on a continuous scale, where values near -1 indicate a highly negative sentiment, 0 is neutral, and values approaching 1 represent a highly positive sentiment. Results shown as a density plot in Figure 2(g) highlight that LLM-generated emails generally carry more positive sentiment and mostly have a positive polarity score with a mean score of 0.231, whereas human-written emails' scores are more spread out with a mean of 0.114. Typically, AI-generated email responses retained a neutral to positive polarity. This can be an artifact of the fact that LLMs may be trained along with red-teaming and moderation efforts [41] to avoid using negative language in their responses. For example, a human-written email, *"Sorry, I only just saw your previous mail.. Clement REALLY ANNOYING ME!!!"* with a polarity score of -0.42, whereas for the same thread, GPT-4 responded with *"I just saw your email, and I understand how frustrating this situation must be for you. I'm here to help resolve this issue as quickly as possible. Could you please provide more details about what's happening, or if there's a specific problem you need assistance with?"*, with a polarity score of 0.16. Furthermore, AI-generated responses with lower scores tend to mention words related to issues and challenges in work, as in *"Hi Geoff, Thank you for reaching out and I'm sorry to hear about the difficulties you're experiencing with downloading the PS file. Here are a few troubleshooting steps that might resolve the issue [...]"*.

*4.2.2 Subjectivity.* Subjectivity estimates the degree to which a piece of text contains personal opinions, emotions, or judgments,

as opposed to factual content. To operationalize subjectivity, we used Textblob [52] that measures the amount of personal opinion and factual information contained in the text, with values in the range of 0.0 to 1.0, where a higher value represents more personal opinions and more subjective. We find that GPT-generated emails (mean=0.48) show 19.2% higher subjectivity than human-written emails (mean=0.40). Interestingly, this indicates that GPT-generated emails are likely to include more "personal opinions" than human-written emails. However, Figure 2i shows that human-written emails tend to have a higher variance than GPT-written emails. For example, one human-written email included, *"Please kindly tell us what the problem is? They're REALLY ANNOYING ME!!!"* which scored 0.90 on subjectivity.

*4.2.3 Formality.* Formality is a semantic measure that estimates the level of sophistication, politeness, and adherence to "established conventions" in communication [49]. For our study, we operationalized formality using a fine-tuned RoBERTa-based formality classifier from prior work [7]. This classifier is built on Grammarly's Yahoo Answers Formality Corpus (GYAFC) [67] and the Online Formality Corpus [63], and it achieves an approximate accuracy of 91% on its benchmark dataset. This model labels text to be formal or informal, with a confidence score between 0 and 1. We find GPT-generated emails (mean=0.80) to be significantly more formal by 51.90% than human-written ones (mean=0.52), with statistically significant differences (Cohen's $d$=0.58, $t$=6.06, $p$<0.01). For example, a human-written email is *"Send only the word help Examples help send help index help advanced help get help xget No need for send or get or something similar."* and it has a formality score of -0.51, compared to the GPT-generated email is *"Thank you for reaching out with your request for assistance regarding the ERROR REPORT. I understand you need command instructions related to the URLs you mentioned, as well as support for handling offline HTML browsing. To provide you with the most effective support, I'll need a bit more detail about the specific issues or errors you're encountering [...]"*, which has a formality score of 0.96. This example illustrates how AI-generated emails maintain a consistently formal tone, while human responses may include grammatical errors or personal emotions, making them appear more informal to the classifier.

*4.2.4 Politeness.* Politeness has been studied as an key aspect of professional communication, especially emails [2, 13]. We employed a pre-trained politeness classification model [80]. This classifier labeled all the AI-generated emails as "polite"—highlighting that GPT-4 model is significantly trained and red-teamed to only generate polite interactions. On the other hand, the mean politeness score of human-written emails is 0.83—with instances of emails with impolite tone, e.g., *"Standardization comes by wittling down the list of contenders, not continuously extending it [..] I have nothing to add to it, maybe Chris or Michael want to revive it. I still believe it addresses the problems that are still evident in more recent proposals."*

*4.2.5 Empathy.* An empathetic tone in emails can often lead to positive communication and building stronger rapport both professionally and personally. Prior work describes empathy as a cognitively complex process in which one can stand in the shoes of another person to understand their perspectives, emotions, and situations they are in [36, 77]. We employed a RoBERTa-based empathy detection model, fine-tuned on a dataset of empathetic interactions [12, 81] to identify empathetic tone in emails. Interestingly, we find no statistically significant difference in the usage of empathetic tone by both human-written (mean=0.94) and AI-generated (mean=0.93) emails. This aligns with recent work about LLMs' ability to *simulating* empathy in interactions [37, 47].

*4.2.6 Diversity.* Linguistic diversity quantifies variation in email responses within a thread using word embeddings and cosine distance. We obtained the vector representation of each email using a pre-trained Word2Vec model [17]. Next, we computed the centroid vector of each thread by averaging all email embeddings within the thread. We then measured the cosine distances between each email and its respective centroid within the thread. Therefore, for a thread, the average of these cosine distances refer to the linguistic diversity measure—ranging between 0 and 1—where higher values indicate more diverse responses. As shown in Table 2, AI-generated emails (mean=0.36) exhibit lower content diversity than human-written emails (mean=0.38) with statistical significance ($t$=-10.10, $p$<0.05). Furthermore, Figure 2l illustrates that AI-generated responses not only have lower diversity but also show much lower variance compared to human-written emails.

## 4.3 Psycholinguistic Analysis

As email communication is between at least two people, we also examined the psycholinguistic features in human-written emails and GPT-4-generated emails by using Linguistic Inquiry and Word Count (LIWC) [82]. LIWC consists of several psycholinguistic categories ranging across—1) Affect, 2) Cognition and Perception, 3) Social and Personal Concerns, 4) Biological Processes, 5) Functional Words, 6) Interpersonal Focus, 7) Temporal References, and 8) Informal Languages. These eight main categories include more than 60 psycholinguistic attributes, and we measured the normalized frequencies for each attribute of AI-generated and human-written emails. Table 3 reports the LIWC comparisons and we summarize our observations below.

*4.3.1 Affect.* We find that AI-generated emails contain significantly fewer affective attributes compared to human-written emails. In particular, AI-generated emails contain a significantly higher quantity of positive emotion words than human-written emails—aligning with our prior analysis on language polarity. Although AI-generated emails do not contain any *anger* and *sadness* keywords, they contain many more *anxiety* keywords with a substantial difference of 2,298%. However, higher usage of anxiety words does not necessarily mean emotion, but a polite response to the other person in the conversation: examples such as "Don't hesitate to let me know if you need my input or there are specific areas where I could provide support or further insights."

*4.3.2 Cognition and Perception.* AI-generated emails have greater occurrences of *perception* (by 1,046%) and *hear* (by 106.4%), such as, *"I'm thrilled to hear about your focus on improving interoperability and adaptability in agent-based systems."* AI-generated emails were less likely to contain strong personal emotion or as a command to another person in the text for professionalism [43]. In contrast,

**Table 3: Comparing LIWC usage between GPT-4-generated Human-written emails, with mean occurrences, percentage difference (Δ%), Cohen's $d$, $t$-test, and Kolmogorov–Smirnov test (KS) test. Statistical significance reported after Bonferroni correction ($^{*}$ $p$<0.05, $^{**}$ $p$<0.01, $^{***}$ $p$<0.001). Only categories with significant differences are reported.**

| LIWC | GPT-4 | Human | Δ% | d | t | KS |
|---|---|---|---|---|---|---|
| **Affect** | | | | | | |
| Anxiety | 0.017 | 0.001 | 2,298.31 | 0.13 | 0.84** | 0.20 |
| Pos. Emo. | 0.005 | 0.032 | 84.42 | 0.73 | 4.63*** | 0.33*** |
| Neg. Emo. | 0.002 | 0.002 | -14.81 | 0.46 | 2.91** | 0.24* |
| **Cognition & Perception** | | | | | | |
| Percept | 0.114 | 0.010 | 1,046.12 | 1.18 | 7.46*** | 0.54*** |
| Hear | 0.004 | 0.002 | 106.35 | 0.25 | 1.56* | 0.34*** |
| **Social & Personal Concerns** | | | | | | |
| Friend | 0.053 | 0.003 | 1,666.08 | 0.60 | 3.82*** | 0.40*** |
| Female | 0.050 | 0.001 | 4,534.26 | 1.07 | 6.79*** | 0.51*** |
| Leisure | 0.002 | 0.003 | -16.57 | 0.41 | 2.57* | 0.44*** |
| Achiev. | 0.034 | 0.016 | 107.00 | 0.25 | 1.59* | 0.36*** |
| **Function Words** | | | | | | |
| Article | 0.053 | 0.059 | -10.63 | -0.30 | -1.91* | 0.26** |
| Preposition | 0.114 | 0.096 | 18.96 | 0.66 | 4.20*** | 0.38*** |
| Conjunction | 0.051 | 0.036 | 43.34 | 1.06 | 6.68*** | 0.59*** |
| Adverb | 0.026 | 0.034 | -23.15 | -0.52 | -3.27** | 0.30** |
| Negate | 0.003 | 0.013 | -81.18 | -1.32 | -8.36*** | 0.69*** |
| Aux. Verb | 0.050 | 0.070 | -28.42 | -0.95 | -6.03*** | 0.53*** |
| Number | 0.098 | 0.004 | 2280.27 | 1.20 | 7.56*** | 0.57*** |
| Quant | 0.041 | 0.016 | 161.99 | 0.93 | 5.89*** | 0.47*** |
| **Interpersonal Focus (Pronouns)** | | | | | | |
| 1st P. Singular | 0.027 | 0.030 | -10.48 | -0.19 | -1.18* | 0.20 |
| 1st P. Plural | 0.013 | 0.005 | 141.59 | 0.70 | 4.40*** | 0.41*** |
| 2nd P. | 0.043 | 0.017 | 150.40 | 1.66 | 10.50*** | 0.56*** |
| **Temporal References** | | | | | | |
| Focus Past | 0.016 | 0.018 | -8.88 | -0.90 | -5.66*** | 0.54*** |
| Focus Present | 0.022 | 0.098 | -77.43 | -0.49 | -3.10** | 0.24* |

human-written emails used such words more frequently, for example, *"If we can reach agreement in this two area, I hopeful that we will leave Dallas with broad agreement on all the main topics."*

*4.3.3 Social and Personal Concerns.* Social and personal concerns are experessed in email communication to refer to another person [97]. To highlight, LLM-generated emails have noticeably more uses of attributes like *friend* (1,666%), *female* (4,534%), *achievement* (107%), and *achievement* (107%). In fact, the attributes of *friend* and *female* are commonly seen in LLM-generated emails due to the template of generated emails, such as "Hi [Colleague's Name]" and "Dear Sir or Madam". LLMs also use *achievement* more frequently in showing politeness and encouragement, such as in, *"I'm looking forward to discussing this in greater depth and exploring your insights [..]"* Nevertheless, we find no mention of *family, home, religion,* and *motion* in AI-generated emails, given that the LLMs were prompted to write emails in a professional setting [62]. In contrast, human-written emails would often contain "folks", to refer to a group of people. This observation again reveals that LLMs are more consistent with language use and tone, and aligns with our *formal* feature observation in semantic analyses.

*4.3.4 Function Words.* For this category, AI-generated emails have substantially more uses of *number* and considerably more uses

**Table 4: Comparing the occurrences of LIWC categories across several LLMs, along with Kruskal-Wallis $H$ tests. $p$-values reported after Bonferroni corrections ($^{*}$ $p$<0.05, $^{**}$ $p$<0.01, $^{***}$ $p$<0.001).**

| LIWC Category | GPT-4 | GPT-3.5 | Llama2 | Mistral | H-stat. |
|---|---|---|---|---|---|
| **Affect** | | | | | |
| Affect | 0.026 | 0.026 | 0.030 | 0.024 | 16.74*** |
| Anxiety | 0.017 | 0.015 | 0.016 | 0.015 | 4.33 |
| Pos. Emo. | 0.005 | 0.004 | 0.004 | 0.005 | 15.13** |
| Neg. Emo | 0.002 | 0.002 | 0.001 | 0.002 | 18.01*** |
| **Cognition & Perception** | | | | | |
| Insight | 0.001 | 0.000 | 0.000 | 0.001 | 9.58* |
| Cog. Process | 0.001 | 0.002 | 0.001 | 0.001 | 6.18 |
| Percept | 0.114 | 0.131 | 0.157 | 0.133 | 38.01**** |
| Hear | 0.004 | 0.011 | 0.018 | 0.008 | 94.51**** |
| **Social & Personal Concerns** | | | | | |
| Friend | 0.053 | 0.079 | 0.076 | 0.064 | 74.80**** |
| Female | 0.050 | 0.074 | 0.071 | 0.059 | 71.35**** |
| Male | 0.000 | 0.000 | 0.000 | 0.001 | 1.76 |
| Leisure | 0.002 | 0.002 | 0.001 | 0.001 | 23.99**** |
| Space | 0.012 | 0.014 | 0.008 | 0.010 | 29.70**** |
| Time | 0.007 | 0.009 | 0.003 | 0.006 | 37.90**** |
| Achiev. | 0.034 | 0.043 | 0.035 | 0.036 | 9.81* |
| Reward | 0.019 | 0.014 | 0.016 | 0.019 | 16.08** |
| **Function Words** | | | | | |
| Article | 0.053 | 0.052 | 0.056 | 0.051 | 3.08 |
| Preposition | 0.114 | 0.126 | 0.110 | 0.106 | 28.19**** |
| Conjunction | 0.051 | 0.048 | 0.040 | 0.045 | 31.55**** |
| Adverb | 0.026 | 0.020 | 0.018 | 0.023 | 21.86**** |
| Negate | 0.003 | 0.003 | 0.004 | 0.005 | 17.47*** |
| Aux. Verb. | 0.050 | 0.054 | 0.056 | 0.059 | 13.06** |
| Number | 0.098 | 0.112 | 0.109 | 0.110 | 20.74*** |
| Quant | 0.041 | 0.041 | 0.041 | 0.034 | 17.93*** |
| **Interpersonal Focus (Pronouns)** | | | | | |
| 1st P. Singular | 0.027 | 0.040 | 0.037 | 0.035 | 43.86**** |
| 1st P. Plural | 0.013 | 0.007 | 0.009 | 0.011 | 20.67*** |
| 2nd P. | 0.043 | 0.057 | 0.062 | 0.055 | 36.85**** |
| 3rd P. Plural | 0.001 | 0.001 | 0.002 | 0.002 | 12.23** |
| Impersonal Pronoun | 0.031 | 0.033 | 0.027 | 0.035 | 12.43** |
| **Temporal References** | | | | | |
| Focus Past | 0.016 | 0.013 | 0.009 | 0.008 | 52.42**** |
| **Informal** | | | | | |
| Netspeak | 0.002 | 0.002 | 0.002 | 0.002 | 9.03* |
| Assent | 0.001 | 0.000 | 0.001 | 0.000 | 11.13* |

of *quantity* as functional words. These could be seen in examples like *"Thank you once again for your thorough guidance and support."* where the keyword "once" is used and is recognized as the number attribute. AI-generated emails also show greater use of prepositions (by 19%) and conjunctions (by 43%), but lower use of articles (by -11%) and adverbs (-23%) than human-written emails.

*4.3.5 Interpersonal Focus (Pronouns).* Interpersonal focus is a key aspect in email writing centered around the use of pronouns [16]. We find contrasting trends in the use of first-person pronouns—AI-generated emails contain lower occurrences of first-person singular pronouns (by -10.5%) but higher occurrences of first-person plural pronouns (by 142%) than human-written emails. Per prior work [65], first-person singular pronouns are indicative of personal narratives, whereas first-person plural pronouns are indicative of collective

identities and a polite tone, e.g., as a GPT-generated email contained, *"Looking forward to your response so we can get this sorted out promptly."* In contrast, human-written email responses tend to frequently use phrases such as, "I look forward to it." This highlights a key difference in writing style—LLMs tend to adopt a collective voice using "we/us" whereas human writers are more likely to express themselves from an individual perspective using "I".

*4.3.6 Temporal References.* Temporal references are associated with different behaviors in an organization [8]. We find that AI-generated emails show fewer mentions for all temporal occurrences—past (-8.88%), present (-77.43%), and future (-100%) than human-written emails. Temporal focus is an indicator of recollection of events and personal narratives [18, 65]. However, the lack of these references indicate that LLMs cannot bring in temporal references and past contexts in the emails, for which the LLMs would need to incorporate a significant amount of information. The AI-generated responses did not contain any word in the *focus future* subcategory, which is fairly commonly used human-written responses, *"Well, I am planning something based around the get or send command, which basically does a text version of a web page."*

## 5 Qualitative Examination

To explore the social impact of AI-generated emails, we conducted a manual evaluation by collecting feedback on AI-generated and human-written emails through a survey completed by 41 participants who are university students. We also conducted detailed one-on-one interviews with four of the 41 participants. Our survey and interviews found that a majority of our participants correctly recognized emails as AI-generated. They described AI-generated emails as "wordy" and "formal."

Within the survey, the participants were presented with two different sets of email conversations, each with three responses—one human-written, one GPT-3.5-generated, and another GPT-4-generated. This survey was structured in three distinct phases—1) the first phase kept participants unaware of the AI-generated nature of some emails; 2) in the second phase, participants were asked if they felt one or more than one of the responses were AI-generated, informed about the presence of AI-generated emails, and asked to guess which response emails as AI-generated; and 3) in the third phase, selective participants were invited for an interview to discuss their perceptions of the differences between emails composed by GPT and those written by humans. For participants who participated in the interview, we organized interviews following each phase to gather their in-depth perspectives.

### 5.1 Phase 1: Evaluating Email Perceptions

In the first phase, participants were presented with three types of responses for each of the two email threads, and were asked to vote for their top choices on which email was the most—1) professional, 2) clear, 3) unclear, 4) confident, 5) accommodating, 6) helpful, 7) unnecessarily wordy, and 8) best writing. Participants could skip a question if they did not have a preference. Figure 3 provides an overview of our survey results, which we describe further.

Figure 3 indicates that AI-generated emails consistently outperform human-authored ones in terms of professionalism, with GPT-4 receiving the highest ratings (56.1% on average) followed by GPT-3.5 (37.8%), while human-written emails were rated as most professional by only 6.1% of participants. However, this professional tone comes at a cost: GPT-4-generated emails were overwhelmingly considered as "most unnecessarily wordy" (82.9% on average), significantly higher than both human-written emails (both 8.6%).

GPT-3.5 emerged as particularly effective in balancing different aspects of communication. It received the highest ratings for clarity (59.8%) and confidence (46.4%) and scored well in writing quality (54.9%). This balanced performance suggests that GPT-3.5 might be more effective in matching human expectations for email communication. Human-written emails showed a distinct pattern—although these emails scored lowest in professionalism and were often rated as most unclear (61.0%), they closely matched AI-generated emails in aspects such as wordiness, helpfulness, and confidence—suggesting a more direct and concise communication style. The helpfulness ratings reveals that although GPT-4's high scores in professionalism and willingness to help, it only slightly outperformed GPT-3.5 in actual helpfulness (39.0% vs 37.8%). This suggests that while AI-generated emails may appear more professional and accommodating, this does not necessarily translate into greater practical value for recipients.

### 5.2 Phase 2 & 3: Deeper Understanding

In the second phase, we investigated the participants' ability to identify AI-generated emails and their perceptions of AI versus human authorship. After the initial evaluation, participants were informed that some responses were generated by an LLM (GPT), without specifying which ones. Of the 41 participants surveyed, the majority (85.4%) suspected AI involvement in email responses before being told—63.4% were highly suspicious, 22.0% had mild doubts, and only 14.6% did not suspect AI involvement at all. This high rate of suspicion suggests that current AI-generated text may have detectable patterns that alert readers to its artificial nature. When asked to identify which emails were AI-generated, participants showed varying levels of accuracy. For the first email thread, GPT-4's response was most accurately identified as AI-generated (78% accuracy), followed by GPT-3.5's response (68% accuracy). However, 24% participants incorrectly identified the human-written response as AI-generated. For the second email thread, participants showed high accuracy in identifying GPT-4's (82%) and GPT-3.5's (71%) response. The human-written response was correctly identified as human-written by 85% participants.

Finally, in the third phase, we informed the participants about which emails were AI-generated and asked them to compare them with human-written emails. Here again, participants characterized AI-generated emails as more wordy and using less common words. They noted that AI-generated emails tend to be more formal and polite, whereas human-written emails are more direct. These observations align with our quantitative examinations.

We also asked participants' stance on receiving emails that appear to be predominantly authored by AI, we found varying responses. Only a minority of the participants (12%) voiced a negative perception, sharing that it reflected a lack of effort or authenticity from the sender. For example, some participants mentioned that AI-generated texts are "robotically warm hearted" or "excessively polite and positive." A significant majority of the participants (76%)

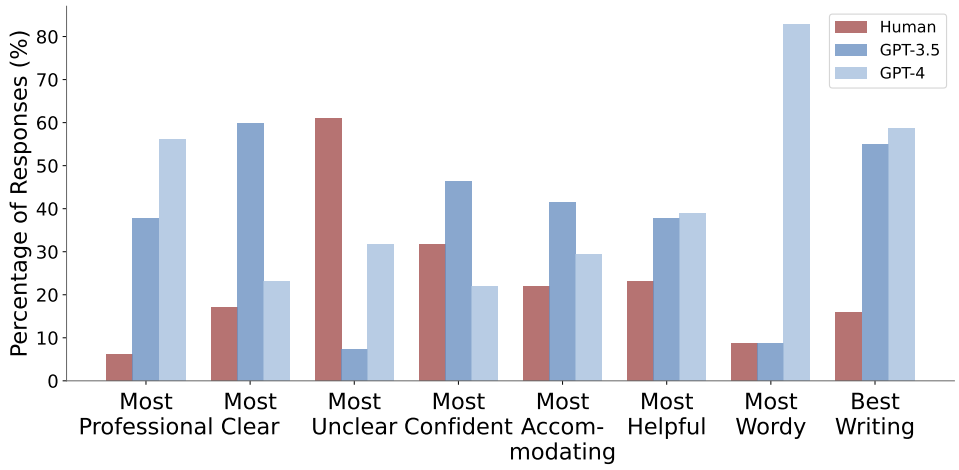



**Figure 3: Comparing user responses about human-written, GPT-3.5-generated, and GPT-4-generated email responses. The percentages are the proportion of participants (N=41).**

were neural or mixed about receiving AI-generated emails, while they have some negative perception related to the artificial nature or perceived lack of genuine effort in AI-generated emails but also recognizing the consistency and politeness in AI-generated texts, such as "Content generated by GPT is the most complete version in terms of formality, but are not necessarily the most pleasant for reading." They also viewed that while GPT may enhance professionalism and politeness, it could also make emails to be less personal, leading them to have mixed feelings about AI-generated emails. These observations support and add to the insights found by Liu et al. on how people's perception of trust decreased as the perceived sender of the email shifted from human to AI [51].

## 6 Robustness of Analyses

We conducted further robustness tests of our analyses to ensure that our results were not sensitive to one type of LLM. We applied the same pipeline of analyses on three other LLMs, GPT-3.5, Llama-2 [88], and Mistral-7B [42]. We conducted pairwise *t*-tests for LLM-generated emails against human-written emails, and a Kruskal-Wallis test across all the generated emails. Table 5 provides an overview of our comparative examination.

First, examining the LLMs vs. human pairwise *t*-test columns in Table 5, we observe that the majority of t-tests for each metric exhibit consistent trends. This suggests that regardless of the type of LLM, they exhibit similar patterns when compared to human-written emails. One exception is in the case of verbosity—while GPT-4 tended to use more words per sentence, all other LLMs tended to have fewer words per sentence than human-written emails—Llama-2 being the least wordy. In terms of readability, we find that all LLMs show significantly higher readability than human-written (6.82) emails, with GPT-4 (11.30) showing the highest, and Mistral-7B (7.97) showing the lowest readability. All the LLMs also show higher complexity and repeatability than human-written emails.

All the LLMs also show higher polarity, subjectivity, formality, and politeness than human-written emails. In terms of empathy, while most LLM-generated emails show no statistically significant differences compared to human-written emails, Llama-2 demonstrates a weak but statistically significant 5.37% higher empathy than human-written emails. Finally, we also see that all LLMs show lower diversity than human-written emails, highlighting LLMs' tendency to repurpose content in email writing.

Together, our robustness analysis revealed that LLMs are not only similar to one another in generating emails, but also exhibit comparable differences when compared to human-written emails.

## 7 Discussion and Conclusion

Emails are a unique and highly contextual form of communication, whose quality depends on striking an optimal balance of being formal and professional, as well as being brief and personal. This work takes a step towards understanding how AI-generated emails compare against human-written ones, providing meaningful insights into the linguistic characteristics and perceptions surrounding both types of emails, shedding light on how each performs across various parameters important for communication in both personal and professional settings. We first computationally compared email threads based on linguistic features of syntax (verbosity, readability, repeatability, and complexity), semantics (polarity, subjectivity, formality, politeness, empathy, and diversity), and psycholinguistics. We found that AI-generated emails tend to be more readable, but also more verbose, complex, and use repeatable language. Further, AI-generated emails tend to be more positive, subjective, and formal, but less diverse than human-written emails. Interestingly, AI uses comparable empathetic tone like humans in emails. We followed up our quantitative examination through user studies via surveys and interviews, which supported our quantitative examinations, as well as unveiled deeper insights. For instance, our participants looked for professionalism and clarity in emails and were typically able to detect AI-generated content. Although some participants perceived AI-generated writing as lacking authenticity, the majority expressed neutral or positive perceptions about AI-generated or AI-assisted email compositions.

**Table 5: Summary of comparing human-written and multiple LLM-generated emails, including paired $t$-tests in comparison with human-written emails, and a Kruskal-Wallis $H$-test across the four LLM-generated emails. ($*p < 0.05, **p < 0.01, ***p < 0.001$).**

| | Human | GPT-4 | | GPT-3.5 | | Llama-2 | | Mistral-7B | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Metric** | **Mean** | **Mean** | **t-test** | **Mean** | **t-test** | **Mean** | **t-test** | **Mean** | **t-test** | **H-stat.** |
| **Syntax** | | | | | | | | | | |
| Verbosity: | | | | | | | | | | |
| Words per Sentence | 18.00 | 19.70 | 0.19* | 16.70 | -7.65*** | 15.80 | -2.59* | 16.55 | -1.57*** | 130.00*** |
| Words per Email | 68.80 | 193.00 | 19.50*** | 112.00 | 10.39*** | 95.60 | 5.54*** | 145.39 | 10.72*** | 217.00*** |
| Sentences per Email | 4.14 | 9.81 | 17.80*** | 6.58 | 11.97*** | 5.90 | 6.95*** | 8.08 | 11.32*** | 187.00*** |
| Readability | 6.82 | 11.30 | 14.00*** | 10.2 | 10.54*** | 9.29 | 7.70*** | 7.97 | 2.38* | 128.00*** |
| Complexity | 4.05 | 4.85 | 10.60*** | 4.74 | 9.13*** | 4.56 | 6.80*** | 4.29 | 2.39* | 108.00*** |
| Repeatability | 0.31 | 0.48 | 12.20*** | 0.43 | 8.17*** | 0.39 | 4.92*** | 0.49 | 9.45*** | 85.80*** |
| **Semantics** | | | | | | | | | | |
| Polarity | 0.11 | 0.23 | 0.47*** | 0.35 | 12.97*** | 0.43 | 15.24*** | 0.25 | 6.95*** | 158.00*** |
| Subjectivity | 0.40 | 0.48 | 4.67*** | 0.51 | 5.67*** | 0.45 | 2.52* | 0.44 | 1.87 | 28.10*** |
| Formality | 0.52 | 0.80 | 6.06*** | 0.90 | 8.44*** | 0.91 | 8.76*** | 0.79 | 4.99*** | 169.00*** |
| Politeness | 0.83 | 1.0 | 5.57*** | 1.0 | 5.57*** | 1.0 | 5.55*** | 0.87 | 0.97 | 82.49*** |
| Empathy | 0.93 | 0.94 | 0.25 | 0.97 | 1.93 | 0.98 | 2.53* | 0.97 | 1.84 | 44.43*** |
| Diversity | 0.38 | 0.36 | -10.13*** | 0.35 | -12.08*** | 0.36 | -8.97*** | 0.34 | -6.34*** | 53.80*** |

This study highlights the need for further research into human-AI collaboration, particularly in more specialized professional settings, and how AI-generated emails are perceived in high-stakes contexts, where nuances in tone, trust, and authenticity may carry greater weight. We recognize that AI-generated is not necessarily the same as AI-assisted content—which is likely to be closer to a real-world scenario, i.e., an individual asking for email reply suggestions from an AI chatbot, and then modifying the content as they see appropriate. However, this study only considered AI-generated content, showcasing how LLMs can incorporate context and cohesiveness of information in email writing. Therefore, our work motivates the design of tools that can assist professional writing. Given the potential strengths of AI-generated email writing as highlighted in this work, tools that assist email writing by incorporating complementary strengths of AI-generation as well as human intelligence can even be prevalently used. Email interfaces, such as those in widely-used platforms like Outlook or Gmail, could benefit from deeper integration with generative AI technologies, offering users intelligent text suggestions that enhance both the quality and efficiency of email writing without sacrificing the personal touch that human writers bring to communication.

As AI-assisted writing tools become more common, there are broader implications for workplace automation and the skillsets required of professionals [46, 72]. Rather than focusing purely on writing, employees may need to learn how to edit and refine AI-generated content effectively. Our work also inspires further examinations into understanding how AI-generated emails align with human-values [78], where AI essentially needs to reflect the writer's (and also the recipient's) intent, tone, and values. In particular, if the AI's content generation does not align with the desired values and needs, it can lead to misunderstandings not only between human-and-AI but also between human-human communication, especially when used in professional settings. Emails also need to adapt to the sensitivity of context and how organizational power manifests in email communication, e.g., how to communicate to a team member, a client, or a superior in an organization—and there are subtle cues (beyond politeness and tone) that we are able to incorporate with experience and contexts, but an AI may not be able to incorporate. Accordingly, AI-driven email assistants should also balance autonomy—individuals should be able to be in control and incorporate their own customizations as needed when prompting the email assistant. However, prompting cloud-based LLM services with sensitive organizational or personal data from email content can lead to privacy and security-related risks. Therefore, the potential ethical concerns around trust and transparency in AI-generated communication will require careful consideration, particularly in high-stakes professional contexts.

**Ethical Implications.** Our work uses a publicly accessible email dataset (W3C Corpus) and did not require any direct interactions with individuals, so it did not qualify for ethics board approval. However, we recognize the ethical implications and want to caution against our work being interpreted as advocating for using LLMs for email generation as that can involve different kinds of risks, including plagiarism and privacy sensitivity. In some sense, our work also calls for the establishment of practices and standards to guide the ethical use of LLMs in writing assistance. As LLMs increasingly become a part of professional and personal communication, it is critical to ensure that their development and deployment adhere to ethical guidelines, including, but not limited to the cases of high-stakes and professional settings.

**Limitations and Future Directions.** Given that we conducted a preliminary exploration, albeit thorough data analysis, our study has limitations, many of which suggest interesting directions for future research. Our work only provides initial insights into how LLM-generated emails compare against human-written ones. Future work can adapt a multitude of prompt engineering, fine-tuning, and Retrieval Augmented Generation (RAG)-based approaches to generate more specific types of emails. Additionally, we conducted a manual evaluation with a small group of participants, primarily university students, which introduces sample bias in their perspectives. Future work can include larger-scale surveys with a more

diverse participant pool to enhance the evaluation. Our study focuses solely on an email dataset from work settings within the technology industry, leaving room for future exploration of diverse email conversation contexts. Future work can further expand the understanding of the social impacts of AI-generated texts by extending beyond the scope of our current research, by targeting more specific fields, such as looking into the effectiveness of AI-generated customer service emails, evaluation of AI-generated content moderation on social media, and so on.

## References


[1] Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. ChatGPT: Fundamentals, applications and social impacts. In *SNAMS*.

[2] Mohammad Awad AlAfnan. 2014. Politeness in business writing: The effects of ethnicity and relating factors on email communication. *Open Journal of Modern Linguistics* 2014 (2014).

[3] Mohammad Awad AlAfnan and Siti Fatimah MohdZuki. 2023. Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial intelligence and technology* (2023).

[4] Sakhar Alkhereyf and Owen Rambow. 2020. Email classification incorporating social networks and thread structure. In *LREC*.

[5] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*.

[6] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[7] Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't Lose the Message While Paraphrasing: A Study on Content Preserving Style Transfer. In *Natural Language Processing and Information Systems*, Elisabeth Métais, Farid Meziane, Vijayan Sugumaran, Warren Manning, and Stephan Reiff-Marganiec (Eds.). Springer Nature Switzerland, Cham, 47–61.

[8] Matthew Russell Barnes, Mladen Karan, Stephen McQuistin, Colin Perkins, Gareth Tyson, Matthew Purver, Ignacio Castro, and Richard G Clegg. 2024. Temporal Network Analysis of Email Communication Patterns in a Long Standing Hierarchy. In *ICWSM*.

[9] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[10] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2024. How Knowledge Workers Use and Want to Use LLMs in an Enterprise Context. In *CHI Ext. Abstracts*.

[11] Michael J Brzozowski. 2009. WaterCooler: exploring an organization through enterprise social media. In *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 219–228.

[12] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4758–4765.

[13] Ulla Bunz and Scott W Campbell. 2004. Politeness accommodation in electronic mail. *Communication Research Reports* 21, 1 (2004), 11–25.

[14] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *CHI*.

[15] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@ IUI*.

[16] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.

[17] Kenneth Ward Church. 2017. Word2Vec. *Natural Language Engineering* (2017).

[18] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.

[19] Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* (1975).

[20] Anthony W Coman and Peter Cardon. 2024. Perceptions of Professionalism and Authenticity in AI-Assisted Writing. *Business and Professional Communication Quarterly* (2024), 23294906241233224.

[21] Nick Craswell, Arjen P De Vries, and Ian Soboroff. 2005. Overview of the TREC 2005 Enterprise Track.. In *Trec*, Vol. 5. 1–7.

[22] Vedant Das Swain and Koustuv Saha. 2024. Teacher, Trainer, Counsel, Spy: How Generative AI can Bridge or Widen the Gaps in Worker-Centric Digital Phenotyping of Wellbeing. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. 1–13.

[23] Vedant Das Swain, Qiuyue "Joy" Zhong, Jash Rajesh Parekh, Yechan Jeon, Roy Zimmerman, Mary Czerwinski, Jina Suh, Varun Mishra, Koustuv Saha, and Javier Hernandez. 2025. AI on My Shoulder: Supporting Emotional Labor in Front-Office Roles with an LLM-based Empathetic Coworker. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

[24] Munmun De Choudhury and Scott Counts. 2013. Understanding affect in the workplace via social media. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 303–316.

[25] Joan DiMicco, David R Millen, Werner Geyer, Casey Dugan, Beth Brownholtz, and Michael Muller. 2008. Motivations for social networking at work. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 711–720.

[26] Joan Morris DiMicco, Werner Geyer, David R Millen, Casey Dugan, and Beth Brownholtz. 2009. People sensemaking and relationship building on an enterprise social network site. In *2009 42nd Hawaii International Conference on System Sciences*. IEEE, 1–10.

[27] Kate Ehrlich and N Sadat Shami. 2010. Microblogging inside and outside the workplace. In *ICWSM*.

[28] Rosta Farzan, Joan M DiMicco, David R Millen, Casey Dugan, Werner Geyer, and Elizabeth A Brownholtz. 2008. Results from deploying a participation incentive mechanism within the enterprise. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 563–572.

[29] Tira Nur Fitria. 2021. Grammarly as AI-powered English writing assistant: Students' alternative for writing English. *Metathesis: Journal of English Language, Literature, and Teaching* 5, 1 (2021), 65–78.

[30] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[31] Werner Geyer, Casey Dugan, Joan DiMicco, David R Millen, Beth Brownholtz, and Michael Muller. 2008. Use and reuse of shared lists as a social content type. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1545–1554.

[32] Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1037–1046.

[33] Amelia Glaese, Nat McAleese, Maja Trkebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

[34] Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv preprint arXiv:2310.08433* (2023).

[35] Steven M Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N Horne, Michal Lahav, Robert MacDonald, Rain Breaw Michaels, Ajit Narayanan, et al. 2022. Lampost: Design and evaluation of an ai-assisted email writing prototype for adults with dyslexia. In *ASSETS*.

[36] Ilona Herlin and Laura Visapää. 2016. Dimensions of empathy in relation to language. *Nordic Journal of linguistics* 39, 2 (2016), 135–157.

[37] Michael Inzlicht, C Daryl Cameron, Jason D'Cruz, and Paul Bloom. 2023. In praise of empathic AI. *Trends in Cognitive Sciences* (2023).

[38] Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030* (2022).

[39] Michal Jacovi, Ido Guy, Shiri Kremer-Davidson, Sara Porat, and Netta Aizenbud-Reshef. 2014. The perception of others: inferring reputation from social media in the enterprise. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 756–766.

[40] Lori E James, Deborah M Burke, Ayda Austin, and Erika Hulme. 1998. Production and perception of" verbosity" in younger and older adults. *Psychology and aging* (1998).

[41] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *NeurIPS* (2024).

[42] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint* (2023).

[43] Marina Jovic and Salaheddine Mnasri. 2024. Evaluating AI-Generated Emails: A Comparative Efficiency Analysis. *World Journal of English Language* (2024).

[44] Melissa A Kacena, Lilian I Plotkin, and Jill C Fehrenbacher. 2024. The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports* (2024).

[45] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169* (2023).

[46] Pranav Khadpe, Lindy Le, Kate Nowak, Shamsi T Iqbal, and Jina Suh. 2024. DISCERN: Designing Decision Support Interfaces to Investigate the Complexities of Workplace Social Decision-Making With Line Managers. In *CHI*.

[47] William Kidder, Jason D'Cruz, and Kush R Varshney. 2024. Empathy and the Right to Be an Exception: What LLMs Can and Cannot Do. *arXiv preprint* (2024).

[48] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *PNAS* 118 (2021).

[49] Tove Larsson and Henrik Kaatari. 2020. Syntactic complexity across registers: Investigating (in) formality in second-language writing. *JEAP* 45 (2020), 100850.

[50] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* (2023).

[51] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI console me when I lose my pet? Understanding perceptions of AI-mediated email writing. In *CHI*.

[52] Steven Loria et al. 2018. textblob Documentation. *Release 0.15* 2, 8 (2018), 269.

[53] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored mondays and focused afternoons: The rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3025–3034.

[54] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *Comput. Surveys* 56, 2 (2023), 1–40.

[55] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

[56] Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *CHI*.

[57] Tanushree Mitra and Eric Gilbert. 2012. Have you heard?: How gossip flows through workplace email. In *ICWSM*.

[58] Tanushree Mitra, Michael Muller, N Sadat Shami, Abbas Golestani, and Mikhil Masli. 2017. Spread of Employee Engagement in a Large Organizational Network: A Longitudinal Analysis. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 81.

[59] Nhan Nguyen and Sarah Nadi. 2022. An empirical evaluation of GitHub copilot's code suggestions. In *MSR*.

[60] OpenAI. 2023. OpenAI. https://www.openai.com/. Accessed: 2024-01-21.

[61] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv preprint arXiv:2309.05196* (2023).

[62] Akshay Patil, Juan Liu, Jianqiang Shen, Oliver Brdiczka, Jie Gao, and John Hanley. 2013. Modeling attrition in organizations from email communication. In *2013 International Conference on Social Computing*.

[63] Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *TACL* (2016).

[64] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine* 6, 1 (2023), 210.

[65] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[66] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186–195.

[67] Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535* (2018).

[68] 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*.

[69] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).

[70] Koustuv Saha, Yoshee Jain, Chunyu Liu, Sidharth Kaliappan, and Ravi Karkar. 2025. AI vs. Humans for Online Support: Comparing the Language of Responses from LLMs and Online Communities of Alzheimer's Disease. *ACM Transactions on Computing for Healthcare* (2025).

[71] Koustuv Saha, Manikanta D Reddy, Vedant das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, Shayan Mirjafari, Raghu Mulukutla, et al. 2019. Imputing missing social media data stream in multisensor studies of human behavior. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 178–184.

[72] Koustuv Saha, Manikanta D Reddy, Stephen Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. 2019. Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance. *PACM HCI* CSCW (2019).

[73] Koustuv Saha, Asra Yousuf, Louis Hickman, Pranshu Gupta, Louis Tay, and Munmun De Choudhury. 2021. A Social Media Study on Demographic Differences in Perceived Job Satisfaction. *PACM HCI (CSCW)* (2021).

[74] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525* (2023).

[75] N Sadat Shami, Michael Muller, Aditya Pal, Mikhil Masli, and Werner Geyer. 2015. Inferring employee engagement from social media. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3999–4008.

[76] N Sadat Shami, Jeffrey Nichols, and Jilin Chen. 2014. Social media participation and performance at work: a longitudinal study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 115–118.

[77] Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP*.

[78] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. 2024. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions. *arXiv preprint* (2024).

[79] Jitesh Shetty and Jafar Adibi. 2004. The Enron email dataset database schema and brief statistical report. (2004).

[80] Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A Dataset for Politeness Classification in Nine Typologically Diverse Languages. In *EMNLP*.

[81] Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, João Sedoc, Sven Buechel, and Alexandra Balahur. 2021. Wassa 2021 shared task: predicting empathy and emotion in reaction to news stories. In *WASSA, EACL*.

[82] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[83] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint* (2023).

[84] Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI alignment in the design of interactive AI: Specification alignment, process alignment, and evaluation support. *arXiv e-prints* (2023).

[85] Lukas Teufelberger, Xintong Liu, Zhipeng Li, Max Moebus, and Christian Holz. 2024. LLM-for-X: Application-agnostic Integration of Large Language Models to Support Personal Writing Workflows. *arXiv preprint arXiv:2407.21593* (2024).

[86] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).

[87] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[88] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[89] University of Maryland Institute for Advanced Computer Studies. 2005. Parsed W3C Corpus. https://tides.umiacs.umd.edu/webtrec/trecent/parsed_w3c_corpus.html. Accessed: 2024-01-21.

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

[91] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS* (2022).

[92] Yiheng Xu, Hongjin Su, Chen Xing, Boyu Mi, Qian Liu, Weijia Shi, Binyuan Hui, Fan Zhou, Yitao Liu, Tianbao Xie, et al. 2023. Lemur: Harmonizing natural language and code for language agents. *arXiv preprint arXiv:2310.06830* (2023).

[93] Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving Long Story Coherence With Detailed Outline Control. In *ACL*.

[94] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating Longer Stories With Recursive Reprompting and Revision. In *EMNLP*.

[95] Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. GhostWriter: Augmenting Collaborative Human-AI Writing Experiences Through Personalization and Agency. *arXiv preprint arXiv:2402.08855* (2024).

[96] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *IUI*.

[97] Justine Zhang, James Pennebaker, Susan Dumais, and Eric Horvitz. 2020. Configuring audiences: A case study of email communication. *PACM HCI* CSCW1 (2020).

[98] Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. Emailsum: Abstractive email thread summarization. *arXiv preprint arXiv:2107.14691* (2021).

[99] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).