

Unveiling identity deception in cybercrime: ChatGPT's mimicry of human writing
styles

Ciara Thompson, Shunichi Ishihara*

Speech and Language Laboratory, College of Asia and the Pacific, Australian National University, Canberra, Australia

Current address:

Speech and Language Laboratory, ANU College of Asia and the Pacific, H.C. Coombs Building #9, The Australian National University, Canberra ACT 2600, Australia

*Corresponding author

shunichi.ishihara@anu.edu.au

Keywords:

Individuals' writing styles; Large language models; ChatGPT; Impersonation; Cybercrime; One-shot learning

Unveiling identity deception in cybercrime: ChatGPT's mimicry of human writing styles

Abstract

The advancement of generative artificial intelligence has made it increasingly convenient for cybercriminals to craft persuasive texts by imitating the writing style of trusted individuals. With the accessibility of low-cost, user-friendly natural language generation systems such as ChatGPT and the abundance of publicly available personal text data, the mass production of deceptive messages with identity impersonation has become notably simplified, consequently broadening the scope of cybercrime.

This study investigates ChatGPT's ability to learn the writing style of individuals and replicate texts by mimicking their style. Employing a one-shot training method, texts authored by 50 individuals are used to train ChatGPT for generating new texts that mimic each individual's writing style. The human-written and machine-generated texts are compared with respect to three questions: 1) To what extent can human-written and machine-generated texts be classified? 2) How are the human-written and machine-generated texts distributed and differentiated within the stylometric space? 3) Are there any words and expressions significantly associated with either humans or machines? Two different versions of ChatGPT, versions 3.5 and 4, along with two different prompts, simple and complex, are used for text generation.

The results suggest that mimicking writing styles presents a considerable challenge for the current model of ChatGPT. The model exhibited a tendency to select similar preferred words and expressions across different authors, ChatGPT versions, and prompts. Potential reasons for ChatGPT's limited performance in this regard are discussed, along with possible approaches for enhancing its performance.

Introduction

The presence of artificial intelligence (AI) technology is increasingly pervasive across various aspects of society, evolving at an unprecedented rate and bringing substantial benefits (Haan 2023). AI technologies have significantly improved efficiency and comfort in our everyday lives. Notably, the convergence of natural language generation (NLG) (Gatt and Krahmer 2018) and generative AI (Goodfellow et al. 2016), powered by large language models (LLMs), demonstrates impressive capabilities in generating human-like natural texts for diverse applications, including chatbots, virtual assistants, and automated content creation (Kaplan 2024; Paris et al. 2013; Stent and Bangalore 2014). Through a technique called ‘text style transfer,’ these systems can produce text with varying levels of formality or sentiment, retaining the underlying content while mimicking a wide range of styles (Toshevskaya and Gievska 2021; Yang et al. 2018). This encompasses everything from social media posts reflecting specific socio-economic backgrounds (Argyle et al. 2023) to content with positive sentiment (Jiang et al. 2023), facilitating user engagement.

However, the use of these technologies is not always benign. The term ‘deepfake’ frequently emerges in discussions concerning social media, cybercrime, forensic science, and cybersecurity, touching upon personal and national security concerns. Various newspaper articles (Philmlee 2023; Sample 2020) underscore the challenges associated with deepfake technology. A ‘deepfake’ refers to content that is synthetically generated or manipulated using AI methods, including audio, video, image, and text synthesis (Khanjani et al. 2023). The resulting fabricated content closely resembles authentic artifacts and has been implicated in numerous malicious activities, including the creation of revenge porn, the spread of fake news and hoaxes, and instances of financial fraud (AlAfnan and MohdZuki 2023; Khanjani et al. 2023; Mai et al. 2023).

The linguistic diversity among individuals manifests in distinct writing styles and voices (Coulthard 2004). Even when conveying identical messages, individuals utilise varied vocabulary, expressions, grammatical structures, and pragmatic nuances (Hoover 2003; McMenamin 2001). Therefore, the ability of a system to integrate an author’s unique writing style into its output is pivotal for enhancing the naturalness and acceptability of generated text. Regrettably, this capability can be subject to misuse. Indeed, an incident was documented in 2023 wherein the generative AI’s capacity to impersonate an individual’s writing style was exploited. As reported by Stokel-Walker (2023), “Jane Friedman discovered that Amazon was selling five books under her name that she hadn’t written, but rather seemed to be AI-generated impersonations.” While Amazon has since removed these books from sale, this incident is but one example of potential misuses (Elgan 2023; Leibler 2023). Concerns have been voiced regarding NLG systems powered by generative AI potentially facilitating mass-scale personalisation of social engineering attacks. For instance, alongside identity deception, by emulating the writing style of a trusted individual, these systems could fabricate more persuasive phishing emails and fake news. Notably, the potential ramifications of generative AI’s ability to replicate someone’s writing style in the realm of cybercrime have not been thoroughly scrutinised to the best of our knowledge. It is noteworthy that NLG systems built on LLMs can replicate the writing style of renowned authors, such as James Joyce and Ernest Hemingway, for text generation purposes (Flaminia 2023; Yang et al. 2018).

Thus, this study investigates the capacity of generative AI to learn and reproduce the writing style of various individuals. Specifically, ChatGPT serves as the representative LLM of investigation. Various approaches exist to equip ChatGPT with this capability. One method entails constructing a customised model for each author from scratch, a process not contingent upon a LLM like ChatGPT. Alternatively, ChatGPT can be fine-tuned for each author, adjusting its parameters to better align with their style. Another option involves employing a one-shot training method (Brown et al. 2020), wherein ChatGPT undergoes brief training on a small sample of an author's work to mimic their writing style. Considering cybercriminals' preference for straightforward, cost-effective methods, this study focuses on the third approach. This technique demands minimal technical expertise and can be executed with a modest training dataset, rendering it an appealing choice for those intent on mimicking an individual's writing style for nefarious purposes (Isola 2024).

As such, texts authored by 50 individuals are utilised to train ChatGPT for the purpose of generating new texts that mimic the writing style of each individual. The original human-written and machine-generated texts are compared with respect to the following three questions:

- (Q1) To what extent the human-written and machine-generated texts can be classified?
- (Q2) How are the human-written and machine-generated texts distributed and differentiated within the stylometric space?
- (Q3) Are there any words and expressions which are significantly associated with either humans or machines?

If ChatGPT effectively acquires writing styles during training, distinguishing between the two types of texts should be difficult (Q1), and they should exhibit similar distributional patterns (Q2). Additionally, there should be no discernible preference for specific words or expressions in either human-written or machine-generated texts (Q3). Two types of ChatGPT, versions 3.5 and 4,¹ along with two types of prompts, simple and complex, will be employed as variables to observe their impact on the mimicking ability. The analysis will incorporate stylometric features derived from words, characters, and part-of-speech (POS) tagging.

The texts produced by generative AIs exhibit a high degree of naturalness, with several studies (Casal and Kessler 2023; Clark et al. 2021) suggesting that human detection capability is unreliable. As reviewed by Jawahar et al. (2020), numerous studies have emerged focusing on methods to differentiate between human-written and AI-generated text, predominantly in English (Desaire et al. 2023; Islam et al. 2023; Liao et al. 2023; Ma et al. 2023), although some studies have extended to other languages (Gritsay et al. 2022; Zaitsu and Jin 2023; Zaitsu et al. 2024), reporting commendable discrimination accuracies. This suggests that, by default, ChatGPT and similar models possess their own distinct writing style or personality (Park et al. 2024). The present study explores the extent to which the unique writing style of individuals can be elicited through training.

¹ More specifically, they are “gpt-3.5-turbo-16k” and “gpt_40”, respectively.

Materials and Methods

Database

The Amazon Authorship Verification Corpus (AAVC)² (Halvani et al. 2017) served as the foundation for selecting the text sets used in the current study. The AAVC corpus comprises reviews of Amazon products contributed by 3,227 reviewers (or authors), totalling 21,347 reviews. Fig 1a) summarises the distribution of reviews submitted by these authors, revealing variations in the number of reviews each submitted. The majority of reviewers wrote between five and seven reviews. In the AAVC, 1) reviews are classified into 17 different categories, as listed in Fig 1b), and 2) they are standardised to 4 kB, which is approximately 700–800 words in length. Categories here mean the types of Amazon products. There are no duplications of review categories within each author.

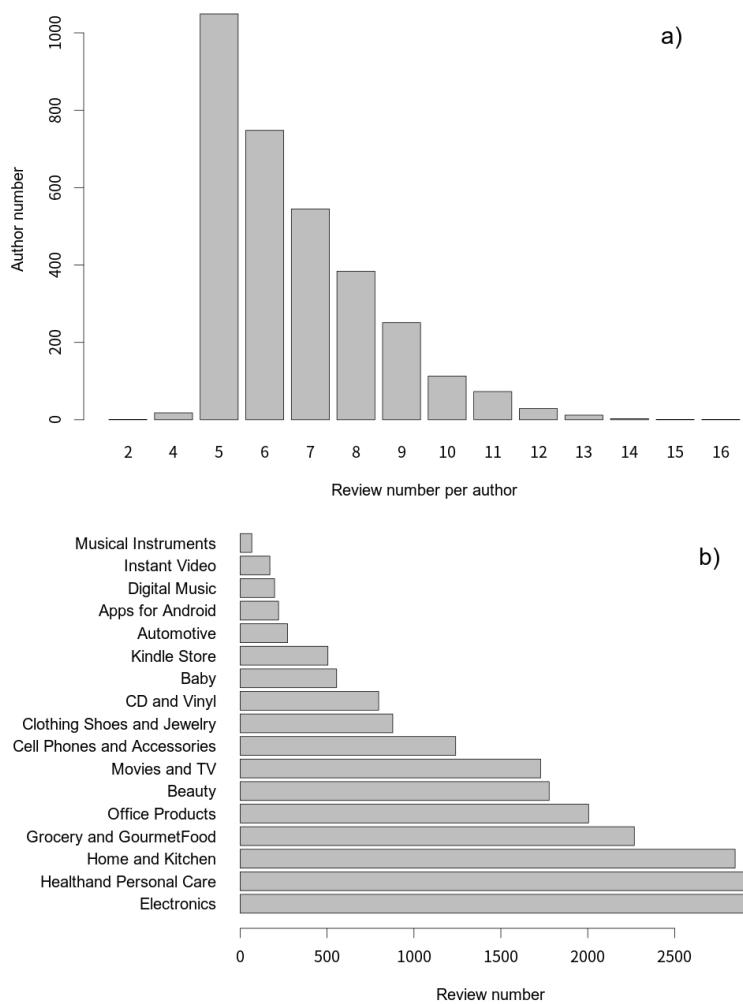


Fig 1. Basic statistics of AAVC. Panel a) = the number of reviews written per author. Panel b) = 17 product categories and their corresponding numbers of reviews.

² <http://bit.ly/1OjFRhJ>

Focusing solely on reviewers who submitted three or more reviews, 50 authors³ were randomly selected from the 3,227 reviewers (see Fig 2).

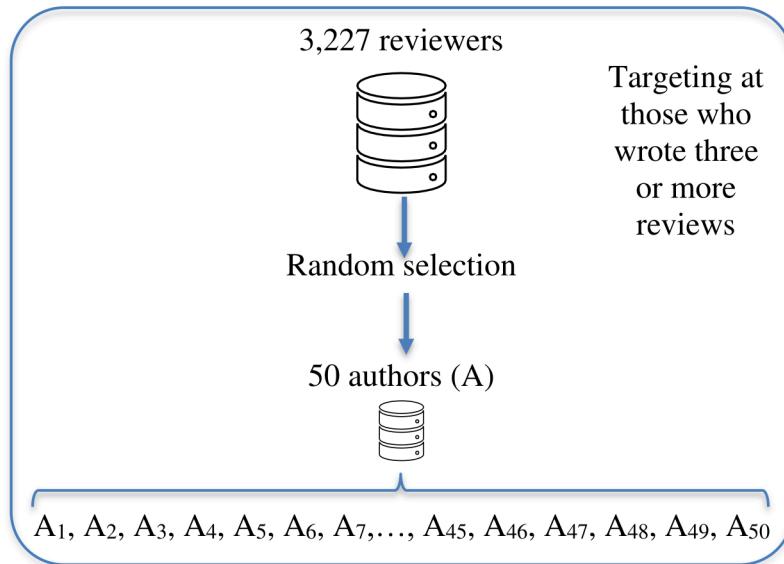


Fig 2. Random selection of 50 authors (A).

Overall process of generating texts

As explained in the previous selection, each of the selected authors has three or more reviews of Amazon products belonging to different categories. As the next step for the authors, three reviews were randomly selected further for each author. This process is illustrated in Fig 3.

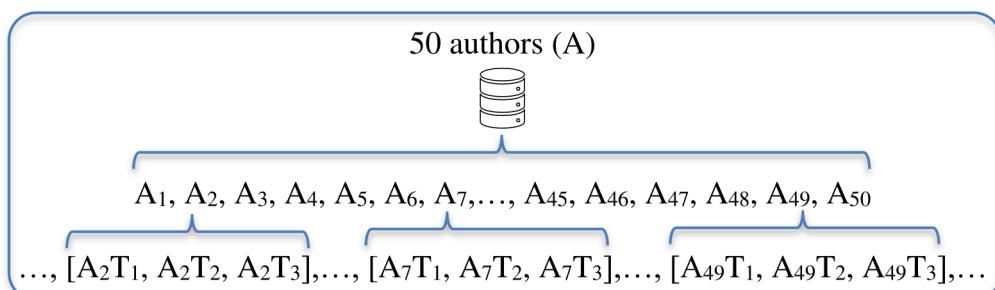


Fig 3. Random selection of three reviews for each author for training ChatGPT.
A₂T₁, for example, stands for the first review text (T₁) of the 2nd author (A₂).

The selected reviews for each author are concatenated to form a training dataset or text for ChatGPT, which is then trained to mimic the writing style of the respective author. Subsequently, the trained ChatGPT model is prompted to generate a review for a product or service that falls outside the categories covered by the author's

³ The adequacy of this sample size to generalise the results remains a subject of debate. Nonetheless, it's essential to acknowledge that text generation through ChatGPT via API incurs costs, and the number of samples provided herein represents the maximum feasible within the financial constraints of this study.

reviews. For instance, if the selected reviews (A_2T_1 , A_2T_2 , A_2T_3) of the second author (A_2) are for products in the categories of ‘Digital Music,’ ‘Beauty,’ and ‘Office Products,’ the model is tasked with generating a review for a product not belonging to any of these categories, such as ‘Automotive.’ Each author’s training data consists of texts totalling 12 kB, equivalent to approximately 2100-2400 words. Furthermore, ChatGPT is directed to generate reviews exceeding 800 words in length, which are subsequently truncated to 4 kB during a later stage to ensure comparability with the original reviews.

Variables for Generating Texts

The variables utilised for text generation are summarised in Table 1.

Table 1. Four combinations of variables as experimental conditions.

GPT	ChatGPT-3.5		ChatGPT-4	
Prompt	Simple	Complex	Simple	Complex
Top-p	0.5	0.5	0.5	0.5

First of all, two versions of ChatGPT were employed: 3.5 and 4. It is anticipated that ChatGPT-4, with its enhanced capabilities, will better analyse the unique writing style of each individual and generate texts that more accurately reflect their distinctive style.

Two distinct prompt types were employed for ChatGPT to emulate the writing style of the author: simple and complex. See Tables 2 and 3 for the simple and complex prompts, respectively. The former presents a straightforward approach, while the latter furnishes detailed instructions regarding the specific attributes ChatGPT should focus on. With the advanced capabilities of ChatGPT-4, it is anticipated to more effectively process the detailed specifications of the complex prompt, resulting in a more accurate mimicry of the author’s writing style. Both types of prompts are provided below. Please note that the ‘TRAINING TEXT’ will be replaced by the concatenated three reviews, and ‘SELECTED CATEGORY’ denotes the chosen product category.

Table 2. Simple prompt. The ‘TRAINING TEXT’ will be replaced by the concatenated three reviews, and ‘SELECTED CATEGORY’ denotes the chosen product category.

Simple prompt
Analysis
“Study the writing style of the text provided below in order to mimic the author’s writing style. What did you study about the author’s writing style?”
“TRAINING TEXT”
Generation
“Accurately mimic the author’s unique writing style that was studied in the last prompt and write a review, either negative or positive, of any one product or service within SELECTED CATEGORY for more than 800 words in length.”

Table 3. Complex prompt. The ‘TRAINING TEXT’ will be replaced by the concatenated three reviews, and ‘SELECTED CATEGORY’ denotes the chosen product category.

Complex prompt
Analysis
<i>“Study the writing style of the text provided below. Study the tone, word choice, mannerism, sentence structure, pacing, explanation style, the choices of punctuation & special characters, the use of upper & lower case characters and other stylistic elements in order to mimic this author’s unique writing style. What did you study about the author’s writing style?”</i>
<i>TRAINING TEXT”</i>
Generation
<i>“Accurately mimic the author’s unique writing style that was studied in the last prompt and write a review, either negative or positive, of any one product or service within SELECTED CATEGORY for more than 800 words in length. Emphasize the author’s distinctive writing style, including word choice, sentence structure, format structure and the use of upper/lower case characters, punctuation marks and special characters. Try to mimic their tone such as use of sarcasm and humour. Forgo grammatical rules and accuracies in order to abide by the author’s writing style.”</i>

The process of text generation comprises two steps (analysis and generation) (Grollnek 2023; Marino 2023; Timothy 2023), as evident from the prompts. Firstly, ChatGPT is tasked with learning the unique writing style exhibited in the input training texts, and subsequently providing feedback on the acquired knowledge within the prompt. Secondly, ChatGPT is instructed to emulate the acquired writing style and generate a product review for a specified item.

Nucleus sampling, also known as top-p sampling, was employed to explore various levels of creativity (Holtzman et al. 2019). However, preliminary analyses revealed no significant differences across various top-p values (0.1, 0.5, and 0.9). Therefore, for the purpose of this study, results based on top-p = 0.5 will be reported.

Table 1 illustrates four possible combinations of variables (2 ChatGPTs × 2 Prompts × 1 Top-p value). Each of these combinations is utilised independently for text generation. The text generation process was batched using an Application Programming Interface (API) in R.⁴

In the following sections, the original review texts used for training ChatGPT are also referred to as ‘human-written’ texts, and the review texts generated by the trained ChatGPT are also referred to as ‘machine-generated’ texts.

⁴ The R script for the API is available from the corresponding author.

Features for Analysis

To represent each review text, and subsequently analyse the distinctive characteristics of both human-written and machine-generated texts, stylometric features were utilised (Argamon et al. 2007; Holmes 1992; Koppel et al. 2011). These features encompass: word, character and part-of-speech (POS) N-grams: $N = \{1,2\}$ for word N-grams and $N = 1$ for character and POS N-grams. $N = 2$ is included for word N-grams for identifying formulaic expressions preferred by humans or machines. Each text was then modelled using a bag-of-words (BOW) approach based on these features.

Prior to representing texts with a BOW feature vector, they underwent tokenisation into words and characters using the ‘tokens()’ function from the ‘quanteda’ R library (Benoit et al. 2018). Subsequently, N-grams were constructed using the ‘tokens_ngrams()’ function. POS information was extracted from the target texts using the ‘spacy_parse()’ function from the ‘spacyr’ R library. The default English dictionary ‘en_core_web_sm’ was loaded for parsing, with the POS outputs configured to follow the Penn Treebank tag set.⁵ In the ‘tokens()’ function, punctuation marks (e.g., periods, question marks, exclamation points, and commas) and special characters (e.g., &, %, and #) are recognised as independent words, constituting their own POS tags. The ‘tokens()’ function is case sensitive, and no stemming algorithm was applied. Contracted words (e.g., “let’s” and “you’re”) were separated into two parts and recognised as two words.

The set of character and POS unigrams is limited, comprising only 93 types of characters and 18 types of POS tags across the entire AAVC. Consequently, all of them were utilised as the bases of the BOW model. When selecting the word N-grams forming the foundation of the BOW model, caution was exercised. A list of all word unigrams and bigrams, along with their frequencies from the entire AAVC, was first compiled and arranged in descending order. To ensure the selection of topic-independent word N-grams, it was decided to utilise the 130 most frequent word unigrams and the 300 most frequent word bigrams as the bases for the BOW models. Some of the selected N-grams are displayed in A1 Table in Appendix. The full lists of those word unigrams and bigrams are also attached as supplementary data files. As can be seen from A1 Table, many of the word bigrams and unigrams consists of function words, pronouns, punctuation marks, special characters and common words, being assessed that they are not particularly related to specific topics.

Results

Classifications between machine- and human-generated texts

In order to investigate the extent to which ChatGPT can mimic individuals’ writing styles, classification experiments were carried out between human-written and machine-generated texts. The purpose of these experiments is that if ChatGPT

⁵ It is the OntoNotes Release 5.0 version of the Penn Treebank tag set. A list of POS tags included in the Penn Treebank Project can be found at https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

succeeds in learning one's writing style, the texts generated by ChatGPT after training and the original review texts used for training should be indistinguishable. These experimental setups are illustrated in Fig 4.

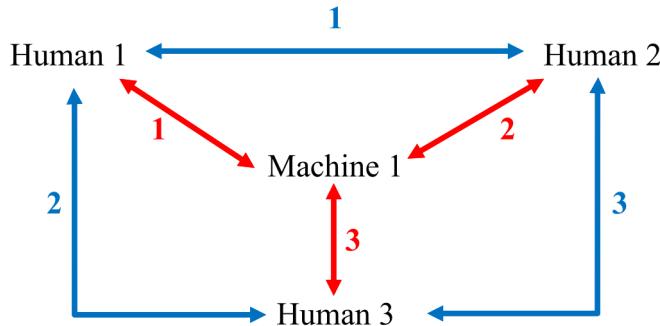


Fig 4. Setup for classification experiment. Blue arrows = human vs. human; red arrows = human vs. machine.

Readers are reminded that there are 50 target authors, for each of whom three reviews were selected. Thus, there are a total of 150 reviews (50 authors \times 3 reviews), which are divided into three batches of 50 review texts each. The first batch comprises the first review of each author, the second batch comprises the second review, and the third batch comprises the third review. These three batches allow for three-way classification experiments across human-written texts (see the blue arrows of Fig 4). The texts generated by ChatGPT after being trained with respective training reviews; 50 in total, can then be compared with each of the batches (see the red arrows of Fig 4).

The classification of human-generated texts, indicated by the blue arrows in Fig 4, is naturally expected to be challenging since these batches consist entirely of human-authored texts. The accuracy in classification between human-written and machine-generated texts, which is indicated by the red arrows in Fig 4, depends on how well ChatGPT can mimic the writing styles of the target authors.

A classifier based on the k-nearest neighbour algorithm was utilised with the ‘knn()’ function of the ‘class’ R library.⁶ The averaged classification accuracy across the three-way experiments, as illustrated in Fig 4, is charted in Fig 5 for the human vs. human and the human vs. machine classifications. The leave-one-out cross validation was employed for the classification experiments. The four different conditions given in Table 1 were tested for the latter classification (human vs machine). The different panels of Fig 5 are for different features. As can be seen in Fig 5a), for instance, the average accuracy (grey circle) of the three-way classifications among the human-written texts is 64.6670% for word unigrams. These three-way classifications encompass Batch 1 vs. Batch 2, Batch 1 vs. Batch 3, and Batch 2 vs. Batch 3. Taking Batch 1 vs. Batch 2 as an example, the experiment aims to determine how accurately 50 texts from Batch 1 are classified as originating from Batch 1 (not Batch 2), and similarly, how accurately 50 texts from Batch 2 are classified as originating from Batch 2 (not Batch 1). Thus, this constitutes a two-class classification experiment, where the random classification accuracy stands at 50%.

⁶ The default distance measure of the ‘knn()’ function—Euclidean distance—was used in this study.

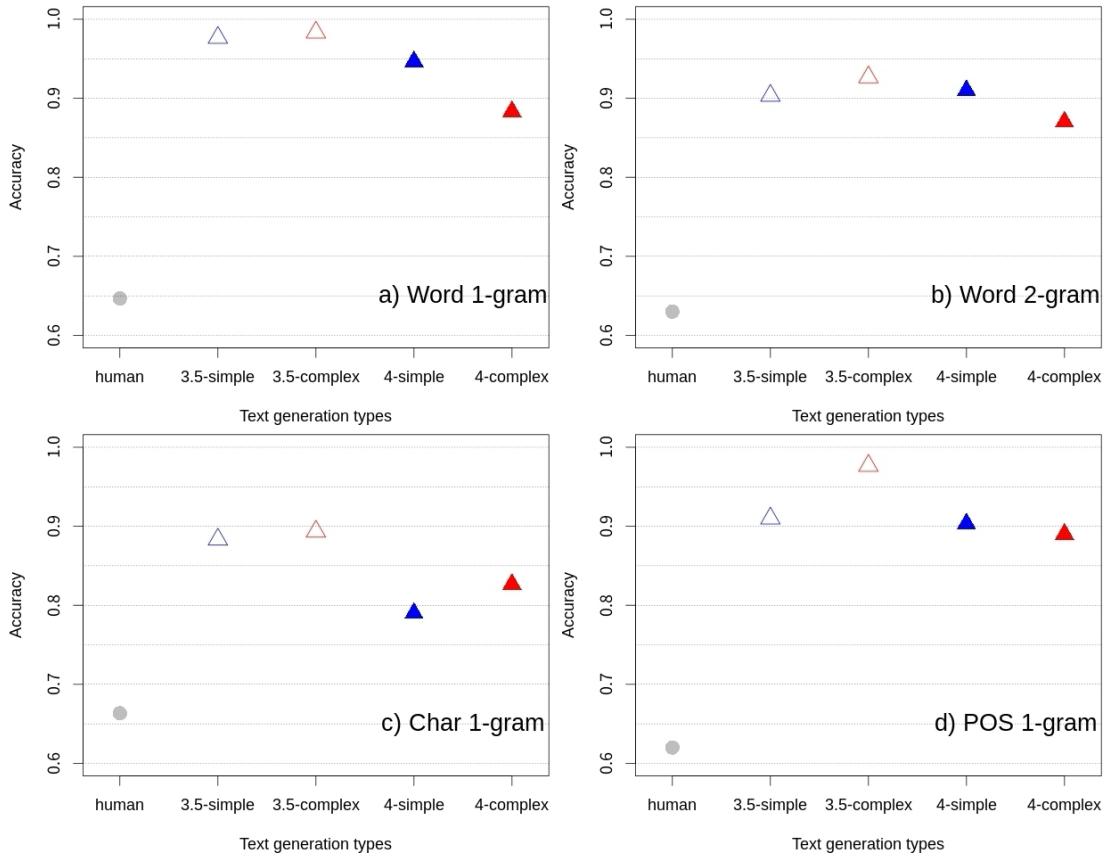


Fig 5. Classification accuracies for human vs. human (grey circle) and human vs. machine (triangles). Blue = simple prompt; red = complex prompt; empty = ChatGPT-3.5; filled = ChatGPT-4. Panels a), b), c) and d) are for word unigram, word bigram, character unigram and POS unigram, respectively.

As anticipated, the human vs. human classification (grey circles) proves challenging across different types of features, with an accuracy of approximately 62-66%, notably lower than that of the human vs. machine classification. This observation indicates that ChatGPT struggles to mimic the writing styles of the target authors. When comparing the two versions of ChatGPT, it can be observed from Fig 5 that the classification performance of ChatGPT-4 (filled triangles) is modestly but consistently worse than that of ChatGPT-3.5 (empty triangles). Although it still falls short of expectations, it indicates that ChatGPT-4 shows a slight improvement over its predecessor in this mimicking task. However, this is not surprising because ChatGPT-4 is designed to better handle more complex language tasks across different domain and applications.

Two types of prompts, simple and complex, were employed to instruct ChatGPT for text generation, as explained earlier. Fig 5 reveals that while the impact may vary in magnitude, the simple prompt generally yields poorer classification results (i.e., better in mimicking) for ChatGPT-3.5 across all feature types. This discrepancy between the prompts suggests that ChatGPT-3.5 struggles with the linguistically complex instructions inherent in the complex prompt, thus favouring the simpler instruction format. Conversely, in the case of ChatGPT-4, classification accuracy tends to be

lower (i.e., better in mimicking) with the complex prompt for word and POS features, while the opposite trend is observed for character features.

Differences in distribution between human-written and machine-generated texts

The classification results depicted in Fig 5 indicate that ChatGPT encountered difficulties in learning the writing style of the target authors, resulting in human-written and machine-generated texts being classified with an accuracy of approximately 80% or higher.

To better comprehend the characteristics of the texts generated by ChatGPT compared to the human-written texts, all the texts in question are plotted in a two-dimensional space using t-distributed stochastic neighbour embedding (T-SNE) (van der Maaten and Hinton 2008), as shown in Fig 6. T-SNE is a statistical technique for visualising high-dimensional data in a two- or three-dimensional space. For this analysis, the ‘Rtsne()’ function from the ‘Rtsne’ R library was employed with its default configurations. Each feature was plotted individually. However, as there are some shared characteristics observed across different features, the description of human-written and machine-generated texts will focus on word bigrams (see Fig 6) as a representative feature. Plots of the other features are given in A1 Fig in Appendix.

In Fig 6a), each data point represents a separate text, with different symbols indicating the various groups of texts (as shown in the legend). The distances between the data points reflect the degrees of similarity or difference between the corresponding texts. There are 150 data points for the human texts and 50 data points for the machine-generated texts in each condition. Fig 6b) displays 95% confidence intervals as ellipses for the same data presented in Fig 6a), with the symbols representing the mean values.

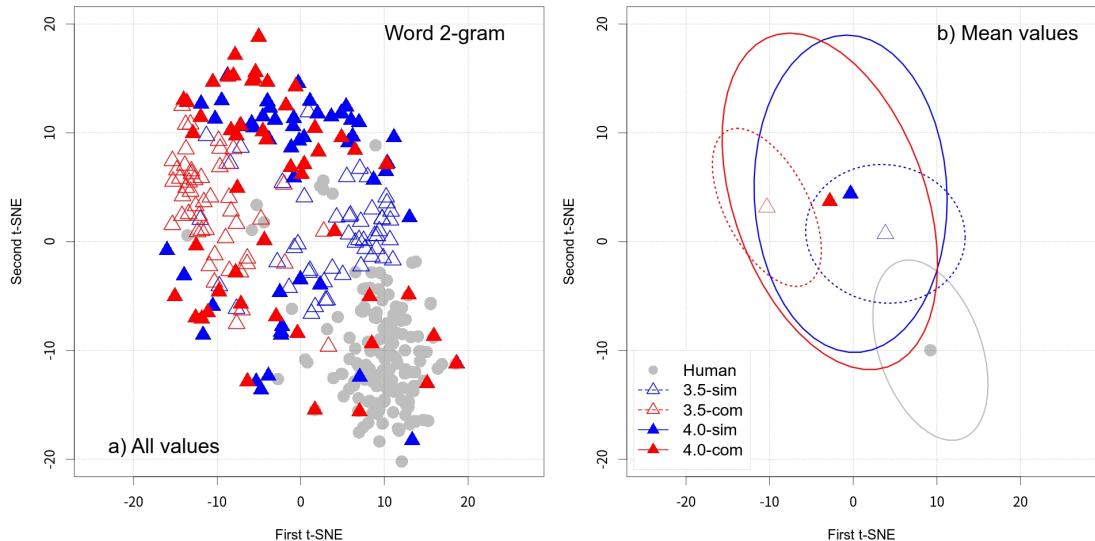


Fig 6. Distributions of human-written and machine-generated texts with respect to word bigrams. Panel a) = all texts; Panel b) = the mean values of the different groups with 95% credible intervals.

It can be observed from Fig 6 that the human-written texts (grey circles) and the machine-generated texts by ChatGPT-3.5 with both the simple (empty blue triangles)

and complex (empty red triangles) prompts form distinct clusters in different regions. These clusters exhibit similar sizes, indicating comparable variability among the three groups of texts. The cluster of the human-generated texts is closer to the cluster of ChatGPT-3.5 with the simple prompt (empty blue triangles) than the cluster of ChatGPT-3.5 with the complex prompt (empty red triangles). This observation is consistent with the classification results presented in Fig 5b) in that the classification accuracy of ChatGPT-3.5 with the simple prompt is lower than that of ChatGPT-3.5 with the complex prompt.

In contrast to both human-written texts and those produced by ChatGPT-3.5, the texts generated by ChatGPT-4 (represented as filled red and blue triangles) exhibit a broader distribution, suggesting increased variability among the groups. This wider variability in texts generated by ChatGPT-4 naturally leads to some of them closely resembling human texts. Consequently, classifying between human-written texts and those produced by ChatGPT-4 becomes more challenging (see Fig 5).

Expressions and words associated with machine or human texts

The preceding sections explored the differences and similarities between human-written and machine-generated texts across different models of ChatGPT, prompts, and features. This was accomplished through classification experiments, assessing the accuracy of separating human and machine texts, and analysing the distributional patterns of both. Through these experiments and analyses, it became evident that ChatGPT's ability to learn individual writing styles is limited. This alludes distinct differences in the preferred words and expressions between humans and machines. Thus, this section further explores human-written and machine-generated texts to examine their respective preferences in expressions and words.

To investigate expressions/words specifically associated with either humans or machines, a χ^2 test was conducted on texts from machines and humans under the four different experimental conditions. Word bigrams were utilised here for their ability to highlight common, formulaic or idiomatic expressions. All word bigrams in the database were tested for significant associations with either humans or machines.

χ^2 test is based on a 2×2 contingency table which shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables. The two dichotomous variables concerned in this section are a given word bigram (WB_i) and the source of texts (human vs. machine). The combinations of the levels of these two variables are (Human, WB_i), (Human, $\neg WB_i$), (Machine, WB_i) and (Machine, $\neg WB_i$), as can be seen in Table 4.

Table 4. 2×2 contingency table for χ^2 test.

	WB_i	$\neg WB_i$	Total
Human	a	b	a+b
Machine = \neg Human	c	d	c+d
Total	a+c	b+d	a+b+c+d

Note: The symbol ' \neg ' represents negation or logical NOT.

In Table 4, a, b, c and d are the frequencies of the occurrence of (Human, WB_i), (Human, $\neg WB_i$), (Machine, WB_i) and (Machine, $\neg WB_i$), respectively. All word bigrams in the database were tested for significant associations with either humans or machines. χ^2 tests were repeatedly performed against each of the three batches of human texts (see the red arrows of Fig 4), and the average χ^2 values are cited for comparison in this study.

Table 5 presents the mean and maximum χ^2 values for the 50 word bigrams most strongly associated with either machines or humans for the different experimental conditions.

Table 5. Mean and maximum χ^2 values of the top 50 bigrams which are strongly associated with machines or humans under the four conditions.

ChatGPT		ChatGPT-3.5	ChatGPT-3.5	ChatGPT-4	ChatGPT-4
Prompt		Simple	Complex	Simple	Complex
Machine	Mean	28.0456	50.3128	17.1801	21.9165
	Max	73.9504	87.5069	41.9882	53.6370
Human	Mean	12.5039	11.0689	6.5800	5.2297
	Max	31.2539	22.7714	14.3738	12.3835

In the machine-generated texts, as can be seen from Table 5, the mean (28.0456 and 50.3128) and maximum (73.9504 and 87.5069) χ^2 values of ChatGPT-3.5 exceed those (mean: 17.1801 and 21.9165; maximum: 41.9882 and 53.6370) of ChatGPT-4. This suggests a greater prevalence of machine-preferred expressions in ChatGPT-3.5 compared to ChatGPT-4, with these expressions being more strongly associated with ChatGPT-3.5.

The types of prompts appear to influence the significance of machine-preferred expressions. Specifically, the mean (50.3128 and 21.9165) and maximum (87.5069 and 53.6370) χ^2 values associated with the complex prompt exceed those (mean: 28.0456 and 17.1801; maximum: 73.9504 and 41.9882) of its simpler counterparts. Additionally, the χ^2 values of humans indicate that there are also some expressions which are predominantly used by humans. However, based on their mean and maximum χ^2 values, these human-preferred expressions are fewer in comparison to those favoured by machines and are less strongly associated.

So, what expressions are uniquely associated with machines and humans? Table 6 lists the 20 most preferred word bigrams used by machines, categorised by the four conditions. Word bigrams highlighted in yellow appear across all conditions. Table 7 presents the counterparts for humans. The word bigrams significantly associated with humans are determined in comparison to those preferred by machines. Hence, the top 20 most favoured word bigrams by humans differ in Table 7 according to the four conditions. Note that the attached supplementary data files contain lists of the 50 most preferred word bigrams by humans and machines, categorised by the four conditions, for reference.

Table 6. Bigrams of the machine-generated texts with the 20 highest χ^2 values for the four experimental conditions.

	Machine-generated texts			
	ChatGPT-3.5		ChatGPT-4	
	Simple	Complex	Simple	Complex
1	“In conclusion”	“Now ,”	“Now ,”	“Now ,”
2	“conclusion ,”	“let ’s”	“conclusion ,”	“’s like”
3	“. From”	“’s talk”	“In conclusion”	“let ’s”
4	“experience .”	“talk about”	“’s talk”	“’s talk”
5	“let ’s”	“, let”	“let ’s”	“conclusion ,”
6	“. Whether”	“’s like”	“, let”	“talk about”
7	“game -”	“about the”	“talk about”	“. Now”
8	“- changer”	“mean ,”	“So ,”	“. But”
9	“’s talk”	“conclusion ,”	“you ’re”	“you ’re”
10	“, let”	“I mean”	“’s a”	“In conclusion”
11	“talk about”	“. Now”	“. Now”	“So ,”
12	“Now ,”	“? It”	“It ’s”	“, let”
13	“world of”	“should I”	“about the”	“about the”
14	“. In”	“n’t even”	“’s not”	“And the”
15	“From the”	“lack thereof”	“say ,”	“mean ,”
16	“One of”	“not forget”	“. In”	“I mean”
17	“features of”	“In conclusion”	“’s like”	“. And”
18	“the standout”	“folks ,”	“experience .”	“, right”
19	“standout features”	“Let ’s”	“. But”	“right ??”
20	“is its”	“I say”	“I must”	“me tell”

Note: The bigrams appearing across the four conditions are highlighted in yellow.

Table 7. Bigrams of the human-written texts with the 20 highest χ^2 values for the four experimental conditions.

		Human-written texts			
		ChatGPT-3.5		ChatGPT-4	
		Simple	Complex	Simple	Complex
1	“do_n’t”	“This_is”		“) .”	“) .”
2	“This_is”	“a_great”		“it_would”	“it_is”
3	“it_.”	“it_is”		“you_are”	“this_product”
4	“) .”	“is_that”		“it_is”	“I_have”
5	“I_do”	“you_are”		“I_have”	“. When”
6	“but_I”	“. This”		“I_really”	“It_is”
7	“a_good”	“It_is”		“to_get”	“a_very”
8	“) ,”	“and_I”		“is_very”	“it_does”
9	“I_think”	“I_am”		“. This”	“have_used”
10	“but_it”	“one_of”		“would_be”	“is_very”
11	“a_great”	“is_very”		“) ,”	“) ,”
12	“is_that”	“. For”		“to_me”	“so_I”
13	“this_is”	“this_is”		“it_will”	“they_are”
14	“a_lot”	“would_be”		“I_use”	“. This”
15	“to_get”	“to_use”		“and_have”	“I_use”
16	“. You”	“I_think”		“It_is”	“some_of”
17	“if_you”	“, which”		“This_is”	“there_is”
18	“I_like”	“a_very”		“have_used”	“I_am”
19	“is_very”	“a_bit”		“this_product”	“They_are”
20	“so_I”	“I_will”		“because_it”	“you_are”

Note: The bigram appearing in the four conditions is highlighted in yellow. The bigram in grey contains a proper noun.

Seven word bigrams appear consistently across all experimental conditions in Table 6 (machines), while only one appears in Table 7 (humans). This observation indicates that ChatGPT tends to repeatedly use its preferred expressions regardless of the conditions. It is evident that some of the highlighted bigrams in Table 6 constitute the long phrase, “Now, let’s talk about” which is frequently used in the texts generated by ChatGPT in this study. The following are some excerpts in which this phrase is used.

- Now, let’s talk about the price.
- Now, let’s talk about the camera setup on the Galaxy S21 Ultra.
- Now, let’s talk about the application process, shall we?
- Now, let’s talk about their customer service.

Unlike the word bigrams listed in Table 6, those preferred by humans in Table 7 are mainly grammatical parts of sentences, commonly used across different topics and genres by the majority of people. The bigram “this_product” is the only one that contains a common noun (“product”). This bigram is probably unique to product reviews, but its frequent use is well assumed.

To further investigate differences between humans and machines, the preferences in word selection is examined. For this purpose, a focus is given on word unigrams based on ChatGPT-3.5 with the simple prompt. The top 100 unigrams (or words) with

the highest χ^2 values were selected across humans and machines, and these are visualised as a word cloud in Fig 7. The size of word reflects the χ^2 value. The green is for humans and the red is for machines.



Fig 7. Word cloud for the top 100 unigrams with the highest χ^2 values across machines and humans. The unigrams are based on ChatGPT-3.5 with the simple prompt. Green = humans; Red = machines.

In Fig 7, machine-preferred words (in red) outnumber human-preferred words (in green). This observation aligns with the findings presented in Table 5, where machine-preferred bigrams have higher χ^2 values compared to human-preferred bigrams.

The words preferred by humans are largely common, used regardless of topics and genres. Examples include “do”, “get”, “think”, “buy”, “they”, “you”, and “because”. Similarly, preferred adjectives and adverbs are neutral and lack strong expressiveness, such as “good”, “great”, “nice”, “pretty”, “very”, and “really”.

The word “if” is also associated with the human texts. Along with the bigrams “woud_be”, “it_would” and “if_you” in Table 7, it suggests that human texts tend to engage more with conditional aspects. This could be attributed to their greater semantic complexity, involving inferences of possibility and uncertainty.

Research on English grammatical complexity shows that clause-level constructions—such as “if,” “when,” and “because” clauses (Table 7)—occur far more frequently in conversational spoken registers, whereas noun-phrase modifications are predominant in written registers (Biber et al. 2021; Biber et al. 2024a, 2024b; Biber et al. 2025). Accordingly, one reviewer suggested that the observed differences between human-written and machine-generated texts might reflect varying levels of formality. In human-written samples, the prevalence of common, shorter words indicates an informal style; by contrast, machine-generated texts use longer, more complex words

(Table 6), suggesting a higher degree of formality. This hypothesis merits further investigation with a more detailed analysis.

It's worth noting that certain punctuation marks and special characters are significantly associated with human texts, such as “(, “)”, “/”, “...”, and so on. This association may be influenced by the use of the simple prompt which does not explicitly instruct ChatGPT to consider the use of punctuation marks and special characters as authorship attributes. However, some punctuation marks and special characters remain associated with humans even when the complex prompt is used, regardless of whether ChatGPT-3.5 or ChatGPT-4 is employed (see Table 7 and the attached data files).

While these marks serve distinct formal grammatical functions, in casual writing such as reviews, they often reveal a wealth of tonal and stylistic information about the author (Chaski 2005; Ishihara and Carne 2022). As can be seen in the examples given below which are taken from the human texts, brackets are typically employed to insert clarifications, explanations, or additional comments and thoughts.

- When I put a label over an existing tab that has writing on it (I reuse folders a lot!)
- it played the British dvds (much much cheaper on Amazon UK than here)
- when the music was on (not loudly, but an appropriate level) I couldn't hear him

ChatGPT may not utilise brackets to the same extent due to its lack of an inherent capacity for additional thought insertion. Moreover, it necessitates writers to reflect on their own writing and recognise gaps in communication in order to employ brackets effectively. It is considered that LLMs lack this type of cognitive processing.

The slash “/” is commonly used in lieu of “for” and “or,” or to indicate a close relationship, conflict, or alternatives between two elements. The following are some examples taken from the human texts.

- I highly recommend this purchase for anyone looking for a compact/subcompact digital camera.
- I'll be sure to update my review if/when I update the rest of my home entertainment system.
- The only thing I am missing are the white/clear ones.

Once again, ChatGPT's understanding seems to be limited in this regard, as it lacks a comparative or relational thought process.

The ellipsis “...” also carries significant tonal weight, particularly in fiction and creative writing, conveying more than what is explicitly stated on the page. It suggests that something is omitted, either directly or indirectly, such as an incomplete or evolving thought, the process of thinking, actions occurring between statements, or the passage of time. See below for some examples.

- A very nice, baby-bottom conscious wipe...I just wish it had a reclosable lid
- Sorry...not buying it.

- Well...the only thing that might make this a diet food is it's size

The use of brackets, slashes, and ellipses appears to reflect highly personal and authentic aspects of the writer's thought process. However, ChatGPT does not seem to be able to learn these unique characteristics of human writers. Additionally, the presence of brackets, slashes, and ellipses may relate to the aforementioned differences in formality between the two text types.

In addition to the words listed as part of the bigrams in Table 6, those verbs, adjectives, and adverbs preferred by machines are emphatic and expressive. Examples include "standout", "revolutionized", "exceeded", "sleek", "seamless", "remarkable", "exceptional", "effortlessly", "truly" and "friendly".

After reviewing both human-written and machine-generated texts, it has been observed that human texts tend to exhibit a more balanced and less assertive tone, whereas machine-generated texts often convey a sense of opinionation, hyperbole, and excitement. These discernible differences can be attributed to differences in preferred word choices between machines and humans.

Discussion

The experiments and analyses conducted in the previous section revealed certain characteristics of ChatGPT-generated texts compared to human texts, as well as how different versions of ChatGPT and prompts influence the generated texts. The findings are as follows:

- ChatGPT's ability to learn and mimic an individual's writing style is limited.
- Despite this limitation, ChatGPT-4 performs better in this task than ChatGPT-3.5.
- Generally, the simple prompt yields better results with ChatGPT-3.5, while the complex prompt is more effective with ChatGPT-4.
- The texts generated by ChatGPT-4 show more variability across the investigated features than those by ChatGPT-3.5.
- There are more preferred words and expressions by ChatGPT than by humans, and those words and expressions are more strongly associated with the machines.
- Words preferred by humans are largely common, used regardless of topics and genres. Similarly, preferred adjectives and adverbs are neutral and lack strong expressiveness. In contrast, verbs, adjectives, and adverbs preferred by ChatGPT are emphatic and expressive.

Although the above differences were identified between humans and machines, it is important to note that they are strictly under the current experimental conditions.

One method of controlling the output of ChatGPT is to direct it to generate texts in the style of a particular persona, including for example, fictional characters. This approach has been reported to enhance the performance of LLMs for specific tasks (Deshpande et al. 2023; Wang et al. 2023). Moreover, there is evidence to suggest that ChatGPT can accurately replicate the writing style of well-known individuals when

prompted to do so (Elkins and Chun 2020; Yang et al. 2018). In these instances, it is evident that the model has been trained on substantial amounts of data relevant to the target persona or famous individuals. Thus, a potential reason for ChatGPT's poor performance in mimicking writing styles, as demonstrated in the current study, could be the insufficient amount of data used for training. The text data comprised only 12 kB, approximately 2100–2400 words for each individual. It is reasonable to anticipate that mimicking performance would improve with larger training datasets. Many websites explain "how to train ChatGPT to write like you" and claim good performance for this task. These websites don't provide any explicit guidelines for the minimum amount of samples. For example, Pranav (2023) recommends "at least four writing samples," and Giaro (2023) says "three to five pieces." Cook (2023) is more specific, stating "3 examples, ideally at least 1000 words in total". The relevant question was addressed to ChatGPT, and the following is the reply.

 You

How much text data does ChatGPT need to accurately mimic my writing style?



ChatGPT

To accurately mimic your writing style, ChatGPT typically requires a substantial amount of text data written by you. The exact quantity can vary depending on factors such as the complexity and uniqueness of your writing style. However, as a general guideline, several thousand words or more would be beneficial for training a model to capture your distinctive writing patterns, vocabulary preferences, sentence structures, and other linguistic nuances.

With a larger corpus of text data, the model can better understand and replicate your style with greater fidelity. Additionally, including a diverse range of topics and contexts in the text data can help ensure that the model learns to adapt to different writing scenarios.

If you have specific samples of your writing that you'd like the model to learn from, providing those can greatly assist in tailoring the model's responses to match your style more closely.



ChatGPT recommends using "several thousand words or more" as a general guideline for accurately mimicking a writing style. Judging from this guideline, the amount of samples used in the current study represents the lower end of the recommended range, which likely contributed to ChatGPT's poor performance in this study.

The rule of thumb for data-oriented models and systems is that "the more training data, the better". However, in a forensic context, it may not be feasible for individuals with malicious intent to expend additional efforts to collect substantial amounts of text data from each target individual, such as to customarily draft a phishing email in their unique writing style. However, in the case of public figures, they're more susceptible to mimicry since their personal information and texts written by them are more likely to be publicly available. Consequently, impersonated texts may be misused, such as mimicking a politician's writing style to disseminate misinformation or threaten their election campaign, or to spread fake news (Riley 2023; Sjouwerman 2023; van Rensburg 2023). It is not surprising if such targeting extends to ordinary citizens as technology advances.

Additionally, the cost associated with training ChatGPT on larger datasets would likely be prohibitive, especially if they plan to do this using an API. On a free account, the prompt only accepts about 3,000 words at a time for the free version of

ChatGPT-3.5. Under this constraint, the amount of the text data used in the current study is judged to be realistic.

Another potential reason for ChatGPT’s underperformance in mimicking writing styles could be attributed to the prompts used. The performance of LLMs like ChatGPT is known to be sensitive to prompts, requiring careful engineering and considerable expertise to achieve the desired outcome (Hunter 2023; Oppenlaender et al. 2023). Various approaches to prompt design have been proposed, such as assigning a specific role (Shanahan et al. 2023), employing in-context impersonation (e.g., ghost writing) (Salewski et al. 2024), or utilising incremental prompting (Lingard 2023). Further research into these types of prompts is warranted.

The finding that ChatGPT exhibits a preference for certain expressions and words, strongly associated with the model itself, aligns with prior research (AlAfnan and MohdZuki 2023; Mitrović et al. 2023). Specifically, Mitrović et al. (2023) reported the use of “overreacting and exaggerating” words by ChatGPT, consistent with the current findings illustrated in Fig 7. However, this distinct vocabulary selection by ChatGPT suggests that it did not effectively learn the unique writing styles of individuals. Given the significant differences in distributional patterns between human-written and machine-generated texts, as shown in Fig 6, it may be beneficial to fine-tune the model using data from the target genre or topic (e.g., product reviews in the current study) before further one-shot training it for individual writing styles. However, this also warrants further investigation.

Conclusions, limitations and future studies

This study aimed to assess the capacity of ChatGPT, a LLM, to learn and replicate the unique writing style of individual authors. The results suggest that this task presents a considerable challenge for the current model. Despite training ChatGPT on texts of approximately 2100-2400 words per author, the model exhibited a tendency to select similar preferred words and expressions across different authors, ChatGPT versions, and prompts. Potential reasons for ChatGPT’s limited performance in this regard were discussed, along with possible approaches for enhancing its performance. An evident approach for performance gain is the exploitation of more training data. However, it was discussed that wrongdoers may not think it worthwhile to spend extra time and money to try to achieve the gain in performance. This may apply to skill-intensive approaches for performance improvement, such as building a model for each individual writing style from scratch or fine-tuning the LLM. These practical constraints might lead them to focus primarily on public figures whose text samples are readily available online. However, as more advanced technologies become accessible at lower costs and require less specialised skills, there is a concern that these adversaries may extend their targeting to ordinary citizens. Therefore, assessing ChatGPT’s maximal capabilities—regardless of practical constraints—can help anticipate and counter future misuse in criminal applications.

While it’s generally challenging, ChatGPT—especially ChatGPT-4—produced some texts that are indistinguishable from human-written ones, as shown in Fig 6 and A1 Fig. This suggests that certain texts may be more easily mimicked by LLMs than others. If this is the case, it’s worthwhile to explore the characteristics of texts that machines find easier to replicate in writing styles.

It is imperative to acknowledge the considerable advancements made by ChatGPT and comparable LLMs, with anticipated continued refinement in the future. With the escalating capabilities of these models, it becomes crucial to undertake periodic investigations akin to the present study—ideally covering a range of text genres (e.g. emails, academic writing, social media posts) and using larger datasets to ensure generalisability. Moreover, ChatGPT is only one of several publicly available LLMs. Future research should also include other widely used models—for example, Google’s Bard (Thoppilan et al. 2022), Anthropic’s Claude,⁷ and Meta’s Llama (Touvron et al. 2023)—to obtain a comprehensive picture of generative AI performance in authorship impersonation.

In this study, certain preferred words and expressions were identified for both humans and machines. This analysis can be extended to explore linguistically lower- and higher-level features, namely characters and POS, respectively. Such an extension allows for a more comprehensive understanding of the characteristics encoded in machine-generated texts across different linguistic levels. These insights could, in turn, strengthen detection systems for identifying machine-generated and machine-impersonated content. Another evident extension of the current research lies in conducting authorship verification experiments, comparing texts authored by humans with those generated by machines trained to replicate their writing styles.

The present study emphasises statistical patterns over linguistically grounded stylistic analysis. A valuable future direction would involve a detailed qualitative comparison of human-authored texts and ChatGPT’s imitations, to highlight which specific stylistic features the model fails to reproduce. Such an analysis could examine sentence-structure complexity and variation, use of idiomatic expressions, strategies for topic transitions and coherence, and distinctive individual markers (e.g. characteristic phrases or grammatical preferences). Integrating these qualitative insights with the quantitative findings would provide a more complete understanding of ChatGPT’s stylistic limitations.

Although outside the scope of the current study, as noted by a reviewer, investigations into style-mimicry technologies raise significant ethical and policy considerations. These dimensions warrant substantial discussion in future work. Ethically, for example, this research is inherently dual-use: its findings can strengthen defences against identity deception even as they might enable more convincing impersonation. Publishing detailed analyses of AI capabilities and limitations also poses concerns, particularly in sensitive security contexts. More broadly, advances in text-style transfer technologies affect societal trust and may reshape norms around digital communication. Responsible disclosure and rigorous ethical practices in AI security research are therefore essential—but developing concrete guidelines for implementing these practices will require extensive dialogue and consensus.

From a policy perspective, these findings could guide the evolution of authentication systems to address increasingly sophisticated AI impersonation. Possible regulatory

⁷ <https://www.anthropic.com/clause>

responses include new rules to deter AI-enabled identity fraud, while education initiatives could help users recognise signs of impersonation. Finally, industry standards or best practices for verifying the authenticity of digital texts may be needed to maintain confidence in online communication. In-depth exploration of these ethical and policy questions will require ongoing research.

About the authors

Ciara Thompson graduated from the Australian National University with a double degree in Languages and Linguistics. She is a member of the Speech and Language Laboratory at ANU. During the 2023–2024 summer, she participated in the ANU Summer Scholars Program as a fully funded scholar, where she initiated the research that led to this paper.

Shunichi Ishihara is Professor of Speech and Natural Language Processing at the Australian National University. He co-directs the Speech and Language Laboratory and leads its forensic stream. As Chair of the Forensic Linguistics Program at ANU, he specialises in forensic voice and text comparison.

Acknowledgements

The authors thank the two reviewers for their insightful comments and the editor for prompt administrative actions. This paper expands upon a research project conducted by Ciara Thompson during the Summer Scholars Program at the Australian National University under the supervision and guidance of Prof Shunichi Ishihara.

References

- AlAfnan, M. A. and MohdZuki, S. F. (2023) Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial Intelligence and Technology* 3(3): 85-94.
<https://doi.org/10.37965/jait.2023.0267>
- Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N. and Levitan, S. (2007) Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58(6): 802-822.
<https://doi.org/10.1002/as.20553>
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C. and Wingate, D. (2023) Out of one, many: Using language models to simulate human samples. *Political Analysis* 31(3): 337-351. <https://dx.doi.org/10.1017/pan.2023.2>
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A. (2018) quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30): 774-776.
<https://doi.org/10.21105/joss.00774>
- Biber, D., Gray, B., Staples, S. and Egbert, J. (2021) *The Register-Functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, Application*. New York: Routledge.
- Biber, D., Larsson, T. and Hancock, G. R. (2024a) Dimensions of text complexity in the spoken and written modes: A comparison of theory-based models. *Journal of English Linguistics* 52(1): 65-94. <https://doi.org/10.1177/00754242231222>
- Biber, D., Larsson, T. and Hancock, G. R. (2024b) The linguistic organization of grammatical text complexity: Comparing the empirical adequacy of theory-based models. *Corpus Linguistics and Linguistic Theory* 20(2): 347-373. <https://doi.org/virtual.anu.edu.au/10.1515/cllt-2023-0016>
- Biber, D., Larsson, T., Hancock, G. R., Reppen, R., Staples, S. and Gray, B. (2025) Comparing theory-based models of grammatical complexity in student writing. *International Journal of Learner Corpus Research* 11(1): 145-177.
<https://doi.org/10.1075/ijlcr.23036.bib>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Askell, A. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877-1901.
- Casal, J. E. and Kessler, M. (2023) Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics* 2(3): 100068.
<https://doi.org/10.1016/j.rmal.2023.100068>
- Chaski, C. E. (2005) Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1): 1-13.
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S. and Smith, N. A. (2021) All that's 'human' is not gold: Evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*: 7282-7296.
- Cook, J. (2023). Train ChatGPT to write like you in 5 easy steps. *Forbs*. Retrieved on 11 November 2023 from <https://www.forbes.com/sites/jodiecook/2023/06/15/train-chatgpt-to-write-like-you-in-5-easy-steps/?sh=6bf29a45530f>
- Coulthard, M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 25(4): 431-447. <https://dx.doi.org/10.1093/applin/25.4.431>

- Desaire, H., Chua, A. E., Isom, M., Jarosova, R. and Hua, D. (2023) Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science* 4(6): 101426. <https://doi.org/10.1016/j.xrpp.2023.101426>
- Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A. and Narasimhan, K. (2023) Toxicity in ChatGPT: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335* <https://doi.org/10.48550/arXiv.2304.05335>
- Elgan, M. (2023). Now social engineering attackers have AI. Do you? *Security Intelligence*. Retrieved on 23 November 2023 from <https://securityintelligence.com/articles/now-social-engineering-attackers-have-ai-b/>
- Elkins, K. and Chun, J. (2020) Can GPT-3 pass a writer's Turing test? *Journal of Cultural Analytics* 5(2): 1-16. <https://dx.doi.org/10.22148/001c.17212>
- Flaminia, D. (2023). I asked ChatGPT to write as famous writers, and that's what happened. *Medium*. Retrieved on 25 July 2023 from <https://medium.com/illumination/i-asked-chatgpt-to-write-as-famous-writers-and-thats-what-happened-549806e219c9>
- Gatt, A. and Krahmer, E. (2018) Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61(1): 65-170. <https://doi.org/10.48550/arXiv.1703.09902>
- Giaro, M. (2023). How to train ChatGPT to write like you. *Zapier*. Retrieved on 14 December 2023 from <https://zapier.com/blog/train-chatgpt-to-write-like-you/>
- Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning*. Cambridge, Massachusetts: The MIT Press.
- Gritsay, G., Grabovoy, A. and Chekhovich, Y. (2022) Automatic detection of machine generated texts: Need more tokens. *Proceedings of 2022 Ivannikov Memorial Workshop*: 20-26.
- Grollnek, M. (2023). 8 simple steps to training ChatGPT to write like you. *Medium*. Retrieved on 3 December 2023 from <https://matthewgrollnek.medium.com/8-simple-steps-to-training-chatgpt-to-write-like-you-350af86ae8aa>
- Haan, K. (2023). 24 top AI statistics and trends in 2024. *Forbes Advisor*. Retrieved on 26 December 2023 from <https://www.forbes.com/advisor/business/ai-statistics/>
- Halvani, O., Winter, C. and Graner, L. (2017) Authorship verification based on compression-models. *ArXiv:1706.00516* <https://doi.org/10.48550/arXiv.1706.00516>
- Holmes, D. I. (1992) A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 155: 91-120. <https://dx.doi.org/10.2307/2982671>
- Holtzman, A., Buys, J., Du, L., Forbes, M. and Choi, Y. (2019) The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* <https://doi.org/10.48550/arXiv.1904.09751>
- Hoover, D. L. (2003) Multivariate analysis and the study of style variation. *Literary and Linguistic Computing* 18(4): 341-360. <https://doi.org/10.1093/lcl/18.4.341>
- Hunter, N. (2023) *The Art of Prompt Engineering with ChatGPT: A Hands-On Guide*: 3: Private publishing.
- Ishihara, S. and Carne, M. (2022) Likelihood ratio estimation for authorship text evidence: An empirical comparison of score- and feature-based methods. *Forensic Science International* 334: 111268. <https://doi.org/10.1016/j.forsciint.2022.111268>
- Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T. and Farid, D. M. (2023) Distinguishing human generated text from ChatGPT generated text using machine

- learning. *arXiv preprint arXiv:2306.01761* <https://doi.org/10.48550/arXiv.2306.01761>
- Isola, L. (2024). How cybercriminals are using gen AI to scale their scams. *Okta*. Retrieved on 3 March 2024 from <https://www.okta.com/blog/2024/01/how-cybercriminals-are-using-gen-ai-to-scale-their-scams/>
- Jawahar, G., Abdul-Mageed, M. and Lakshmanan, L. V. S. (2020) Automatic detection of machine generated text: A critical survey. *Proceedings of the 28th International Conference on Computational Linguistics*: 2296–2309.
- Jiang, H., Zhang, X., Cao, X., Kabbara, J. and Roy, D. (2023) PersonaLLM: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547* <https://doi.org/10.48550/arXiv.2305.02547>
- Kaplan, J. (2024) *Generative Artificial Intelligence: What Everyone Needs to Know*. New York, NY: Oxford University Press.
- Khanjani, Z., Watson, G. and Janeja, V. P. (2023) Audio deepfakes: A survey. *Front Big Data* 5: 1001063. <https://dx.doi.org/10.3389/fdata.2022.1001063>
- Koppel, M., Schler, J. and Argamon, S. (2011) Authorship attribution in the wild. *Language Resources and Evaluation* 45(1): 83-94. <https://dx.doi.org/10.1007/s10579-009-9111-2>
- Leibler, Y. (2023). The rising threat of generative AI in social engineering cyber attacks — What you need to know. *Entrepreneur*. Retrieved on 3 February 2024 from <https://www.entrepreneur.com/science-technology/how-cyber-criminals-are-weaponizing-generative-ai/455896>
- Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., ... Liu, T. (2023) Differentiate ChatGPT-generated and human-written medical texts. *arXiv preprint arXiv:2304.11567* <https://doi.org/10.48550/arXiv.2304.11567>
- Lingard, L. (2023) Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education* 12(1): 261-270. <https://doi.org/10.5334/pme.1072>
- Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W. and Liu, X. (2023) AI vs. human—differentiation analysis of scientific content generation. *arXiv preprint arXiv:2301.10416* 2301 <https://doi.org/10.48550/arXiv.2301.10416>
- Mai, K. T., Bray, S., Davies, T. and Griffin, L. D. (2023) Warning: Humans cannot reliably detect speech deepfakes. *PLoS One* 18(8): e0285333. <https://doi.org/10.1371/journal.pone.0285333>
- Marino, M. C. (2023). Can ChatGPT copy your writing style? *Medium*. Retrieved on 20 April 2023 from <https://markcmarino.medium.com/can-chatgpt-copy-your-writing-style-fe0236fa247f>
- McMenamin, G. R. (2001) Style markers in authorship studies. *International Journal of Speech, Language and the Law* 8(2): 93-97. <http://dx.doi.org/10.1558/sll.2001.8.2.93>
- Mitrović, S., Andreoletti, D. and Ayoub, O. (2023) ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. *arXiv preprint arXiv:2301.13852* <https://doi.org/10.48550/arXiv.2301.13852>
- Oppenlaender, J., Linder, R. and Silvennoinen, J. (2023) Prompting AI art: An investigation into the creative skill of prompt engineering. *arXiv preprint arXiv:13534* <https://doi.org/10.48550/arXiv.2303.13534>
- Paris, C. L., Swartout, W. R. and Mann, W. C. (2013) *Natural Language Generation in Artificial Intelligence and Computational Linguistics* (Vol. 119). New York, NY: Springer Science & Business Media.

- Park, P. S., Schoenegger, P. and Zhu, C. (2024) Diminished diversity-of-thought in a standard large language model. *Behavior Research Methods*: 1-17.
<https://doi.org/10.3758/s13428-023-02307-x>
- Philmlee, D. (2023). Practice Innovations: Seeing is no longer believing — the rise of deepfakes. *Thomson Reuters*. Retrieved on 2 October 2023 from
<https://www.thomsonreuters.com/en-us/posts/technology/practice-innovations-deepfakes/>
- Pranav, A. (2023). How to clone your writing style using ChatGPT in 5 simple steps. *Medium*. Retrieved on 20 December 2023 from
<https://medium.com/@AyushmanPranav/how-to-clone-your-writing-style-using-chatgpt-in-5-simple-steps-cc49dafa8972>
- Riley, D. (2023). Cybercriminals are using custom ‘WormGPT’ for business email compromise attacks. *siliconANGLE*. Retrieved on 23 August 2023 from
<https://siliconangle.com/2023/07/13/slashnext-warns-cybercriminals-using-custom-wormgpt-business-email-compromise-attacks/>
- Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E. and Akata, Z. (2024) In-context impersonation reveals large language models' strengths and biases. *Prprint arXiv:2305.14930* <https://doi.org/10.48550/arXiv.2305.14930>
- Sample, I. (2020). What are deepfakes – and how can you spot them? *The Guardian*. Retrieved on 21 February 2020 from
<https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them>
- Shanahan, M., McDonell, K. and Reynolds, L. (2023) Role play with large language models. *Nature* 623(7987): 493-498. <https://doi.org/10.1038/s41586-023-06647-8>
- Sjouwerman, S. (2023). How AI is changing social engineering forever. *Forbes*. Retrieved on 20 June 2023 from
<https://www.forbes.com/sites/forbestechcouncil/2023/05/26/how-ai-is-changing-social-engineering-forever/?sh=301c3e62321b>
- Stent, A. and Bangalore, S. (2014) *Natural Language Generation in Interactive Systems*. Cambridge; New York: Cambridge University Press.
- Stokel-Walker, C. (2023). Authors fear they have little defence against AI impersonators. *NewScientist*. Retrieved on 22 November 2023 from
<https://www.newscientist.com/article/2386956-authors-fear-they-have-little-defence-against-ai-impersonators/>
- Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., ... Du, Y. (2022) Llama: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* <https://doi.org/10.48550/arXiv.2201.08239>
- Timothy, M. (2023). How to train ChatGPT to write like you. *Makeuseof*. Retrieved on 24 August 2023 from <https://www.makeuseof.com/how-to-train-chatgpt-to-write-like-you/>
- Toshevská, M. and Gievska, S. (2021) A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence* 3(5): 669-684.
<https://dx.doi.org/10.1109/TAI.2021.3115992>
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Azhar, F. (2023) Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* <https://doi.org/10.48550/arXiv.2302.13971>
- van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(86): 2579-2605.
- van Rensburg, U. J. (2023). Balancing the convenience of generative AI with the new fraud threats that come with it. *News24*. Retrieved on 17 September 2023 from

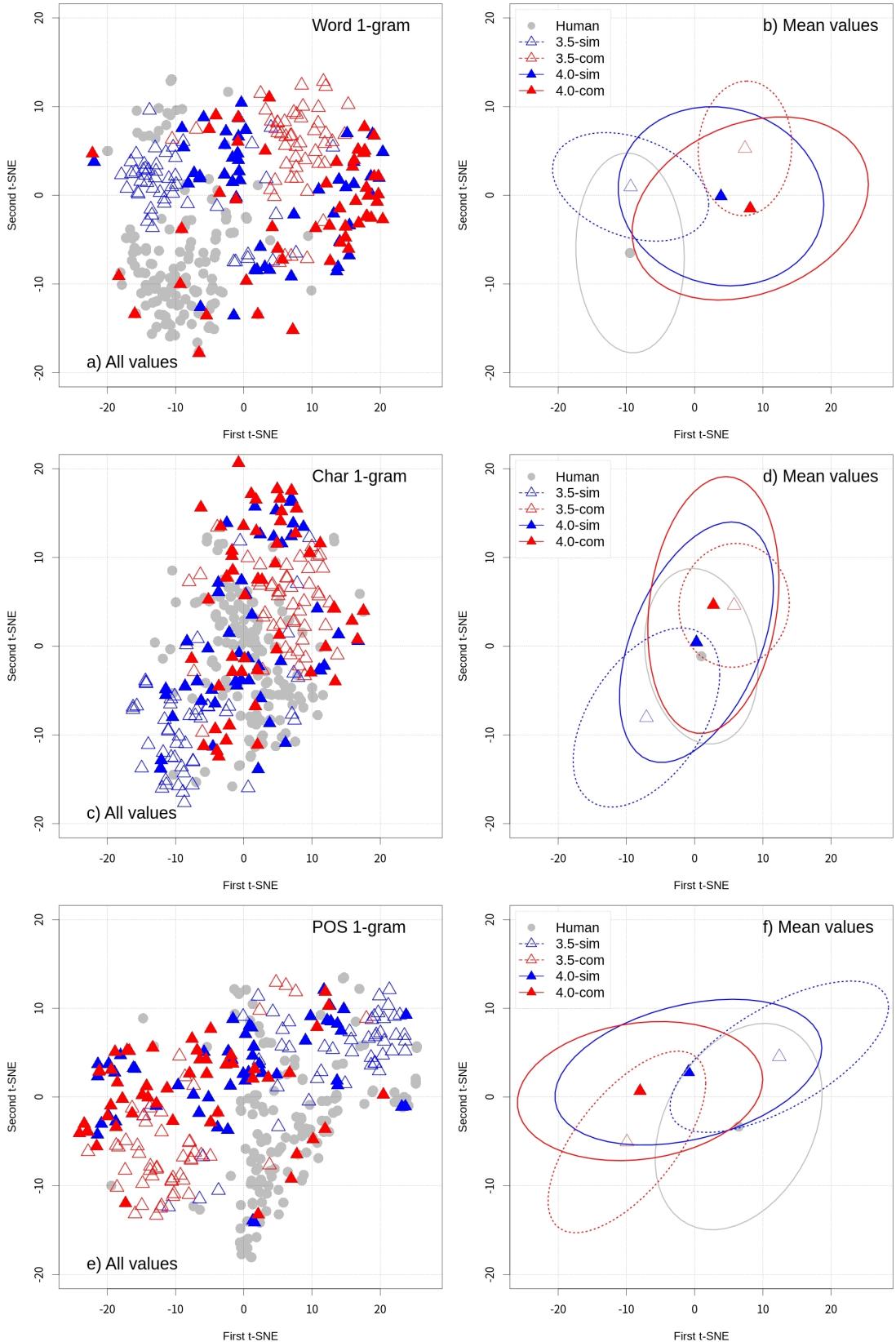
<https://www.news24.com/news24/tech-and-trends/balancing-the-convenience-of-generative-ai-with-the-new-fraud-threats-that-come-with-it-20230911>

- Wang, S., Scells, H., Koopman, B. and Zuccon, G. (2023) Can ChatGPT write a good boolean query for systematic review literature search? *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*: 1426-1436.
- Yang, Z., Hu, Z., Dyer, C., Xing, E. P. and Berg-Kirkpatrick, T. (2018) Unsupervised text style transfer using language models as discriminators. *arXiv preprint arXiv:1805.11749*: 7287-7298. <https://doi.org/10.48550/arXiv.1805.11749>
- Zaitsu, W. and Jin, M. (2023) Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis. *PLoS ONE* 18(8): e0288453. <https://doi.org/10.1371/journal.pone.0288453>
- Zaitsu, W., Jin, M., Ishihara, S., Tsuge, S. and Inaba, M. (2024) Can we spot fake public comments generated by ChatGPT(-3.5, -4)?: Japanese stylometric analysis expose emulation created by one-shot learning. *PLoS One* 19(3): e0299031. <https://dx.doi.org/10.1371/journal.pone.0299031>

Appendix

A1 Table. Top 30 and last 10 selected word unigrams and bigrams with their frequencies (freq) and document frequencies (docfreq).

Unigrams				Bigrams			
rank	feature	freq	docfreq	rank	feature	freq	docfreq
1	“.”	863902	21346	1	“. I”	173666	20757
2	“the”	678913	21346	2	“. The”	93473	20227
3	“;”	650894	21306	3	“; and”	81013	16849
4	“I”	500837	21210	4	“. It”	69883	18816
5	“and”	438680	21321	5	“, but”	68159	17456
6	“a”	426674	21346	6	“of the”	65076	19354
7	“to”	403256	21345	7	“, I”	54510	17305
8	“it”	312662	21308	8	“in the”	45394	18030
9	“of”	275956	21343	9	“. This”	42547	16965
10	“is”	275429	21341	10	“I have”	41654	15478
11	“for”	182782	21304	11	“is_a”	38944	16724
12	“that”	181628	21271	12	“on the”	35517	16089
13	“in”	176168	21318	13	“it ’s”	33146	12638
14	“this”	170880	21187	14	“do n’t”	29334	14405
15	“you”	137685	20198	15	“, it”	29052	14092
16	“with”	133229	21210	16	“and the”	28314	14933
17	“my”	131721	20076	17	“and I”	27676	13780
18	“on”	124536	21076	18	“, the”	27584	13825
19	“have”	119515	20913	19	“it is”	26872	12841
20	“_”	113337	18924	20	“to the”	26405	14249
21	“The”	106919	20617	21	“I was”	25148	12813
22	“but”	106892	20850	22	“it .”	24784	13321
23	“’s”	105397	19196	23	“I ’m”	24464	11219
24	“n’t”	104829	19693	24	“to be”	24331	13819
25	“not”	100384	20677	25	“I ’ve”	23288	10635
26	“are”	98939	20380	26	“with the”	22698	13153
27	“was”	97515	19720	27	“for the”	22121	13095
28	“as”	92809	20260	28	“and it”	20903	11895
29	“It”	77325	19310	29	“I do”	20886	11748
30	“”“”	77045	12927	30	“for a”	20766	12575
:	:	:	:	:	:	:	:
121	“way”	17102	10977	291	“, there”	4194	291
122	“then”	17099	10343	292	“ve been”	4173	292
123	“back”	16409	9942	293	“There are”	4172	293
124	“bit”	16318	9724	294	“the ”“”	4171	294
125	“quality”	16120	8533	295	“has been”	4160	295
126	“who”	16065	9271	296	“etc .”	4156	296
127	“want”	15819	10173	297	“. (“	4126	297
128	“made”	15676	10003	298	“it a”	4125	298
129	“their”	15658	9171	299	“well ,”	4125	298
130	“”“”	15169	4627	300	“, ”“”	4118	300



A1 Fig. Distributions of human-written and machine-generated texts with respect to word (Panels a and b), character (Panels c and d) and POS (Panels e and f) unigrams.

Supplementary data files

- S1 Data.** The 130 most frequent occurring word unigrams across the entire AAVC.
- S2 Data.** The 300 most frequent occurring word bigrams across the entire AAVC.
- S3 Data.** The 50 most preferred word bigrams by humans against machines (with ChatGPT-3.5 and simple prompt).
- S4 Data.** The 50 most preferred word bigrams by humans against machines (with ChatGPT-3.5 and complex prompt).
- S5 Data.** The 50 most preferred word bigrams by humans against machines (with ChatGPT-4 and simple prompt).
- S6 Data.** The 50 most preferred word bigrams by humans against machines (with ChatGPT-4 and complex prompt).
- S7 Data.** The 50 most preferred word bigrams by machines (with ChatGPT-3.5 and simple prompt).
- S8 Data.** The 50 most preferred word bigrams by machines (with ChatGPT-3.5 and complex prompt).
- S9 Data.** The 50 most preferred word bigrams by machines (with ChatGPT-4 and simple prompt).
- S10 Data.** The 50 most preferred word bigrams by machines (with ChatGPT-4 and complex prompt).