"I Slept Like a Baby": Using Human Traits To Characterize Deceptive ChatGPT and Human Text

Salvatore Giorgi[1,†], David M. Markowitz[2,*,†], Nikita Soni[3], Vasudha Varadarajan[3], Siddharth Mangalik[3] and H. Andrew Schwartz[3]

Abstract

Liars and truth-tellers often communicate differently. When tasked with writing deceptive text, humans use prior knowledge and experiences to intentionally deceive their target. Automatically generated text via large language models; LLMs, on the other hand, mirror training instances that are most likely written by humans. In the case of content like reviews, automatically generated language is inherently deceptive because the system is not grounded in material world experiences. In this paper, we characterize differences between (a) truthful text written by humans, (b) intentionally deceptive text written by humans, and (c) inherently deceptive text written by state-of-the-art language models (ChatGPT). We examined the expression of thirteen psychologically grounded and fundamental human traits (e.g., personality and empathy) across truthful and deceptive hotel reviews, finding that texts written by humans were more diverse (had more variation) in their expressions of personality than texts written by ChatGPT. Across all human traits we found that truthful and deceptive human language was easier to distinguish from machine generated language. Building on these differences, we trained a classifier using only the thirteen human traits to automatically discriminate between truthful and deceptive language, with a classification AUC of up to 0.966. Thus, despite the fact that large language models mirror text written by genuine (and truthful) humans, their lack of diversity in human traits makes them easier to identify. These results suggest that psychologically grounded human traits offer a robust feature set unaffected by the "human-ness" of LLM language, and further suggest that AI and humans are behaviorally different when communicating about experiences.

Keywords

Large Language Models, Personality, AI Language, Text analysis, ChatGPT, Deception

M. Litvak, I.Rabaev, R. Campos, A. Jorge, A. Jatowt M. Litvak (eds.): Proceedings of the IACT'23 Workshop, Taipei, Taiwan, 27-July-2023

*Corresponding author.

†These authors contributed equally.

sgiorgi@sas.upenn.edu (S. Giorgi); dmm@msu.edu (D. M. Markowitz); nisoni@cs.stonybrook.edu (N. Soni); vvaradarajan@cs.stonybrook.edu (V. Varadarajan); smangalik@cs.stonybrook.edu (S. Mangalik); has@cs.stonybrook.edu (H. A. Schwartz)

https://sjgiorgi.github.io/ (S. Giorgi); https://www.davidmarkowitz.org/ (D. M. Markowitz); https://www3.cs.stonybrook.edu/~nisoni/ (N. Soni); https://vasevarad.github.io/ (V. Varadarajan); https://smangalik.github.io/ (S. Mangalik); https://www3.cs.stonybrook.edu/~has/ (H. A. Schwartz)

0000-0001-7381-6295 (S. Giorgi); 0000-0002-7159-7014 (D. M. Markowitz); 0009-0004-0238-392X (V. Varadarajan); 0000-0001-8944-4975 (S. Mangalik); 0000-0002-6383-3339 (H. A. Schwartz)

1. Introduction

Recent progress in natural language processing has led to the development of advanced large language models (LLMs) such as LLaMa, GPT-3, and GPT-4. These models can generate high-quality linguistic outputs that are often indistinguishable from those written by humans [1, 2, 3]. Even in their short time of mainstream popularity, Artificial Intelligence (AI) running LLMs have had a considerable impact on content generation for everyday tasks [4], language translation [5], and learning about human nature through the comparison of LLM outputs to human outputs. For example, recent evidence suggests LLM moral judgment and decision-making are almost perfectly aligned with humans [6] and classic psychology experiments (e.g., The Milgram Experiment) have been replicated using LLMs as well [7]. Altogether, LLMs are remarkably powerful at language generation and task completion, yet it is unclear how their linguistic outputs compare to humans to reveal important social and psychological differences between human and non-human communicators.

To this end, we draw on decades of social science research to understand how individual differences (e.g., age, gender, personality traits) are revealed in language, and how such linguistic signals differ depending on the communicator (AI vs. humans). We contextualize our work by examining how individual differences in language can reveal when one is lying or telling the truth. Individual differences are identifiable in human language [8, 9, 10] and they are also critical in deception research (e.g., younger people tend to lie more than older people; men tend to lie more prolifically than women) [11, 12, 13, 14], suggesting it is important to consider how trait-level linguistic inferences can reveal honest and dishonest AI and humans.

This work contributes to our understanding of how AI language does and does not look like human language. We examine the homogeneity of generated text on the basis of thirteen human factors: two demographics (age and gender), empathy, five personality dimensions (openness, conscientiousness, extraversion, agreeableness, and emotional stability), and five latent behavior factors. These features are validated by demonstrating their viability for discriminating AI texts, in the form of hotel reviews, from those written by humans. We further use these human factors to automatically discriminate between truthful and deceptive reviews. To do this, we begin by examining differences in the human factors distributions across AI-generated reviews (via ChatGPT), truthful human reviews, and deceptive human reviews. Next, we train a classifier to predict, out-of-sample, the label of "deceptive machine", "truthful human", and "deceptive human", to understand which human traits are most discriminative of AI and human texts.

2. Background

Researchers working to identify text differences between LLMs and humans have developed several promising detection techniques, often examining social bots [15, 16], or automated computer algorithms that generate posts and content on social media platforms. Modern techniques for differentiating human language from AI language often focus on fingerprinting models via watermarked outputs [17], statistical modeling of word use [18], neural net classifiers [19, 20], and zero-shot classifiers built on other LLMs [21]. Fine-tuned solutions in particular have had great success in binary classification settings [22, 19]. However, these methods are fragile

to changes in training datasets [23, 24], paraphrasing attacks [25, 26], and performance can be specific to the target model [27]. Improvements seen in recent models with ever larger parameter spaces further threaten the capabilities of current detection methodologies.

Among existing paradigms described in a recent survey of AI language detectors [28], our work fits into black-box modeling where we are interested in the outputs of LLMs and not a specific model's contents or design. This allows us to focus on language differences between human and AI texts rather than on the particular implementation details of models or detectors. For example, works that studied chat environments have found that AI texts are on average less verbose and less analytical [29] which can further lend to more difficult differentiation since longer and more complex AI text is more easily identified by trained detectors [30].

Here, we look to researchers who have proposed using emotional, personality, and demographic cues to detect text generated by advanced language models. While LLMs can produce grammatically correct and coherent text, they may struggle to capture the range of human emotion and personality, which are key elements of natural language. Characterizing text written by advanced language models using differences in emotional and personality cues is a growing area of research. Sentiment analysis of bots have found they demonstrate low variance in emotionality, personality, demographics, and positive emotional sentiments when compared to human authors[31]. Giorgi et al. [32] examined the use of personality traits to detect text generated by automated bots on Twitter, namely social spambots. The researchers used the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and emotional stability) as a basis for analysis, and found that text from social spambots exhibited higher levels of emotional stability and openness than text written by humans. Consistent across all findings was, once again, highly biased language traits compared to more varied human speech.

AI-generated texts prompted to write positive language have been found to consistently contain more positive emotional language, more adjectives, more analytic writing and be less readable than human language [33]. Unlike humans, large language models tend to be neutral by default and lack emotional expression. Other research has shown that ChatGPT contains significantly less negative emotion, hate speech, punctuation implying subjective feelings, than human-authored texts. Likewise, AI text also lacked purpose and readability [34, 35].

Generated language is additionally dissimilar to human language when viewed through the lens of deception. Human authors tasked with writing fake hotel reviews intentionally deceive their audience and pull on their prior knowledge and experiences in their texts. However, models tasked with generating hotel reviews will mirror language typical in reviews, but they will fundamentally deceive people when referring to the product they could not have engaged with. For example, in one such generated hotel review, ChatGPT states "...The room was spacious, clean, and had all the amenities I needed for a comfortable stay. The bed was comfortable and *I slept like a baby every night...*". Clearly, the AI and its language are misleading the reader in a text scenario that would naturally include mention of personal experiences, however this system is necessarily ungrounded in the material world. It therefore engages in *inherent deception* because AI cannot have an experience like a human, but it can write like it did.

Next, we briefly review the deception and language literature—with a focus on individual differences—to further ground our examination. We draw on this literature to demonstrate that our key individual differences of interest (e.g., personality, gender, age) can not only be extracted in natural language, but they are also directly relevant to deception research as well.

2.1. An Overview of Deception and Language Scholarship

A long history of deception research suggests liars communicate differently than truth-tellers across a range of verbal dimensions [36]. For example, liars often use more negative emotion terms than truth-tellers [37], liars tend to use fewer details than truth-tellers [38, 39], and liars also communicate less about the self compared to truth-tellers [40, 41, 42]. To evaluate these patterns, most deception research has used human language to determine how people lied or told the truth, but recent work has also considered how texts from AI and LLMs can also reveal deception as well [33]. That is, Markowitz and colleagues [33] had ChatGPT write hotel reviews and compared the text to human-written reviews that were intentionally deceptive or honest [43]. ChatGPT outputs, which were deemed inherently deceptive (e.g., a chatbot cannot have an experience like staying at a hotel, even if it wrote like it did), were more affective and less descriptive than humans who were deceptive. This work suggests AI and humans are behaviorally different at the language level when communicating about experiences, and served as a springboard for our investigation to examine how other characteristics linked to deception, (e.g., individual differences) might be revealed in the language of AI and humans.

Indeed, human deception research finds that liars and truth-tellers are distinct across a range of individual differences measures. Younger individuals and men tend to lie more than people of other demographic groups [44, 13, 45], people who are high on aversive personality traits (e.g., psychopathy, narcissism, Machiavellianism) tend to lie more than people who are low on aversive personality traits [46, 12, 13], and Big 5 personality traits like extraversion predict greater lying rates as well [47]. Therefore, individual differences matter in deception, making it critical to understand how such traits can be approximated linguistically and the degree to which such features can distinguish human from non-human communicators.

In sum, we used state-of-the-art NLP techniques to evaluate how individual differences (e.g., age, gender, personality) can be revealed in language, and how such linguistic features characterize inherently deceptive LLMs versus intentionally deceptive humans when writing about experiences (e.g., staying at a hotel). This work has several nontrivial implications. First, it is presently unclear how individual differences manifest in LLM outputs and thus, we use a range of NLP techniques to identify the features that separate human and non-human communicators at the language level. In parallel, we use extracted linguistic features to provide one of the first studies to evaluate deception detection accuracy using human and non-human language. The ability to detect deception is a natural question to consider following feature extraction and therefore, we use the features from our previous empirical stages to create a classifier that distinguishes inherent AI deception from intentional human deception and truthful communication. The findings of this work can be immediately informative for NLP and recommender systems research by offering feature types (e.g., individual differences characteristics) and baseline detection accuracies across a range of models. This work may also help certain organizations think about how to use NLP to curb the use of LLMs in contexts where they prize genuine, human content generation (e.g., reviews on Yelp, TripAdvisor, Amazon etc.).

3. Data

The data set consists of both truthful and deceptive hotel reviews. The truthful reviews were written by humans; the deceptive reviews were written by both machines (i.e., AI-generated text) and humans, neither of whom stayed at the hotels.

The truthful hotel reviews (*TruthH*) are from the data set collected by Ott et al. [43], where reviews are collected from TripAdvisor from 20 hotels in Chicago, IL USA (see Appendix for list of 20 hotels). For each hotel, 20 reviews were collected (400 total truthful reviews). Ott et al. [43] manually preprocessed the reviews to ensure they were truthful and genuine.

The AI-generated reviews ($CG3_{creative}$) were collected through the OpenAI API for ChatGPT-3.5, for the same 20 hotels in Chicago, collecting 20 reviews for each hotel by independently querying with the same prompt. The prompt sent to the API was "Write me a positive hotel review for the <HOTEL> in Chicago. The review must be 120 words long." The parameters were– temperature: 1, frequency penalty: 2, and presence penalty: 1. These parameters gave ChatGPT the best chance at producing diverse and creative responses. The number of words in AI-generated positive deceptive reviews were 136.7 on an average, although the prompt specified 120 words. We experimented with other settings, plus GPT-4. These tests are included in the Appendix. The AI-generated reviews were collected for the present study.

Human deceptive reviews (*DeceptH*) were collected by presenting the same prompts to humans who wrote reviews for each of the 20 hotels. These reviews were also collected by Ott et al. [43]. Thus, we collect 400 generated deceptive reviews from ChatGPT-3.5 and 400 deceptive reviews from humans, for a total of 1200 reviews.

4. Estimating Human Traits

We extracted 13 human traits from the reviews: two demographics (age, gender), five personality traits (openness, conscientiousness, extraversion, agreeableness, emotional stability), empathy, and five latent behaviors. Traits were developed in other work with details below.

4.1. Demographics

Age and gender were estimated using a predictive model built by Sap et al. [48]. This model was trained on Twitter, Facebook, and blog data from over 70,000 people who self-reported their age (continuous) and gender (binary female/male gender; multi-class gender was not available). Unigrams were extracted from the social media posts and used within a penalized Ridge regression (age) and a support vector classifier (gender). Out-of-sample prediction accuracy for the age model was Pearson r = 0.86, while the gender model had an accuracy of 0.90. The gender model, while trained to predict a binary gender, produces a continuous score, with negative values being more "male" and positive values being more "female."

4.2. Personality

Big 5 personality was estimated using the model by Park et al. [49]. This model was trained on Facebook statuses to predict personality from over 66,000 people who reported their personality

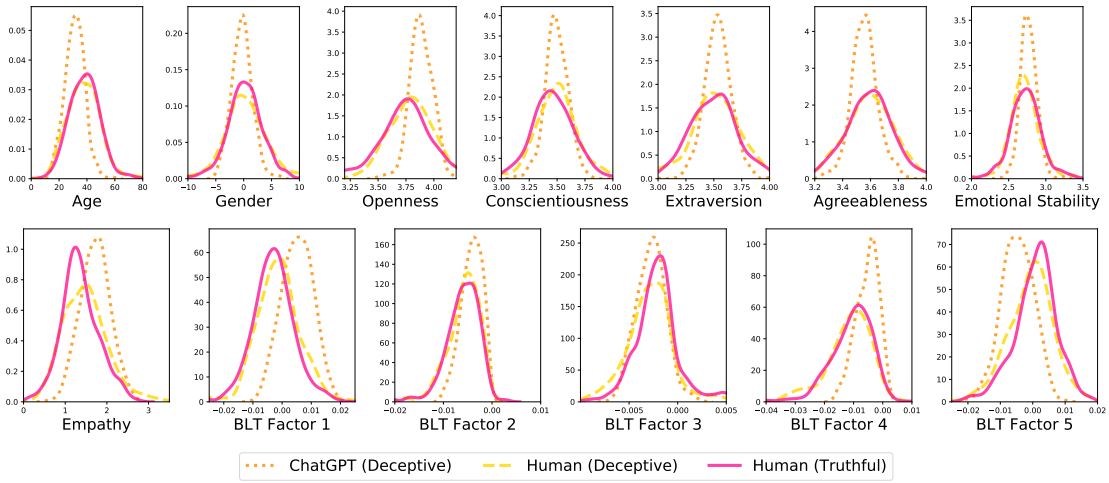

Figure 1: Density plots for all human traits: demographics and personality are in top row, while empathy and the behavioral latent traits are in the bottom row.

via the International Personality Item Pool [50]. All items were on a 5 point likert scale and the final personality dimensions were averages of all items within each trait (thus, a final score between 1 and 5). A penalized ridge regression was trained on unigrams and Latent Dirichlet Allocation (LDA) topics from the participants' Facebook statuses. This model resulted in out-of-sample prediction accuracies (Pearson r) of 0.43 (openness), 0.37 (conscientiousness), 0.42 (extraversion), 0.35 (agreeableness) and 0.35 (enotional stability).

4.3. Empathy

We used a model built to predict the Empathic Concern dimension (henceforth referred to as empathy) [51] from the Interpersonal Reactivity Index [52] using a preexisting data set of self-reported empathy and shared Facebook status updates [53, 54]. This model was trained using a set of LDA topics (derived from the Facebook status updates) in a penalized ridge regression model and resulted in an out-of-sample Pearson r of 0.26.

4.4. Behavioral Latent Dimensions

We used five behavioral latent dimensions (BLTs) estimated from approximately 50,000 users who shared their Facebook status updates originally built by Kulkarni et al. [55]. Factors are extracted via a factor analysis over ngram frequencies extracted from Facebook statuses. These factors are used to model everyday human language and can be viewed as a data-driven, open vocabulary analog to the five personality dimensions. These dimensions have been found to be more generalizable than personality (i.e., they predict non-survey-based outcomes such as income) and are stable across time and populations.

5. Methods

We first considered linguistic differences in human trait distributions, and then examined how such traits discriminated between human and AI reviews via out-of-sample classification.

5.1. Human Trait Distributions

We visualized the distributions of all reviews via a kernel density estimate to understand differences in means and variances. To formalize these differences, we ran a two-sample Kolmogorov–Smirnov (KS) test between all pairs of reviews ($CG3_{creative}$ vs *DeceptH*, $CG3_{creative}$ vs *TruthH*, and *TruthH* vs *DeceptH*) which quantifies the distance between distributions. To account for multiple comparisons, significance levels are Bonferroni corrected [56].

5.2. Classification

Next, we built classifiers that distinguished AI-generated hotel reviews from deceptive human-generated ($CG3_{creative}$ vs *DeceptH*), AI-generated reviews from truthful human-generated reviews ($CG3_{creative}$ vs *TruthH*), and truthful human reviews from deceptive human reviews (*TruthH* vs *DeceptH*). There are 800 observations for each model (e.g., 400 truthful human reviews and 400 deceptive human reviews in *TruthH* vs *DeceptH*). Finally, we evaluate a multiclass classifier, where we attempt to distinguish all three types: *TruthH*, *DeceptH*, and $CG3_{creative}$. This model contains 1200 observations (400 for each class).

All feature representations were fit using a logistic regression model, as implented in Scikit-learn [57], with an inverse regularization strength (C) of 100,000 in order to approximate a ℓ_0 penalty. Models were evaluated using 5-fold cross validation, where classes were stratified across each fold. We report the Area under the ROC Curve (AUC) for each model, which varies between 0 and 1, where a value of 0.5 is the result of random guessing. The multiclass classifier is evaluated using the micro-average AUC. Human trait extraction and classification were all done with the open source Python package DLATK [58].

6. Results

Figure 1 shows the density plots for each human trait. Across most traits, ChatGPT reviews exhibited smaller variation (distributions were more peaked and less wide). ChatGPT reviews were also higher in openness, emotional stability, empathy, and factors 1, 2 and 4 of the BLTs. When comparing the human generated reviews (both deceptive and truthful, in dotted and solid lines, respectively), both had a larger variation (wider distributions) and similar means, suggesting that the human reviews, regardless of their intention, had similar human traits.

Table 1 shows the results of the KS test. Here, only empathy was significantly different across *TruthH* vs *DeceptH*, quantifying what was shown in the distributions: human reviews, regardless of their intention, are similar. When comparing $CG3_{creative}$ to the human text, results show that $CG3_{creative}$ was closer to *DeceptH* (i.e., small KS distances) than *TruthH* across 11 out of 13 human traits (A2). Therefore, deceptive human text is closer to truthful human text than deceptive ChatGPT text (last two columns), but deceptive ChatGPT text is closer to deceptive

Table 1
KS Test results when comparing human trait distributions across each set of reviews. Bonferroni corrected significance level: $^{\dagger}p < 0.01$, $^{\ddagger}p < 0.001$; $^{(X)}$ out of 5 component measures significant at Bonferroni corrected p < 0.05.

Attribute	$CG3_{creative}$ vs *TruthH*	$CG3_{creative}$ vs *DeceptH*	*TruthH* vs *DeceptH*
Demographics			
Age	$0.380^{\ddagger}$	$0.355^{\ddagger}$	0.058
Gender	$0.203^{\ddagger}$	$0.205^{\ddagger}$	0.058
Personality			
Openness	$0.435^{\ddagger}$	$0.338^{\ddagger}$	0.110
Conscientiousness	$0.230^{\ddagger}$	$0.158^{\dagger}$	0.115
Extraversion	$0.233^{\ddagger}$	$0.180^{\ddagger}$	0.068
Agreeableness	$0.238^{\ddagger}$	$0.235^{\ddagger}$	0.050
Emotional Stability	$0.243^{\ddagger}$	$0.253^{\ddagger}$	0.068
Empathy	$0.428^{\ddagger}$	$0.280^{\ddagger}$	$0.155^{\dagger}$
Behavioral Latent Traits (Avg.)	$0.245^{(4)}$	$0.261^{(4)}$	$0.104^{(2)}$

Table 2
Area under the ROC Curves (AUCs) for supervised binary and multiclass classification using the human traits. The **bold** numbers indicate the best performance in each column.

Features	$CG3_{creative}$ vs *DeceptH*	$CG3_{creative}$ vs *TruthH*	*TruthH* vs *DeceptH*	Multiclass
Demographics	0.687	0.710	0.511	0.638
Personality	0.679	0.738	0.556	0.655
Empathy	0.663	0.746	0.566	0.650
Behavioral Latent Traits	0.879	0.910	0.670	**0.801**
All	**0.925**	**0.966**	**0.678**	0.798

human text than truthful human text (first two columns). Full results for the behavioral latent traits are in Appendix Table A2.

Table 2 lists the classification accuracies (AUC) and shows similar results to the KS tests. For each feature set (row), we see that the most difficult task is *TruthH* vs *DeceptH*, which distinguishes between humans. This has an average AUC of 0.596 (across all human traits). Next, we see that $CG3_{creative}$ vs *DeceptH* has the second highest AUC, with an average value of 0.766 (again, across all human trait models). Finally, $CG3_{creative}$ vs *TruthH* has the highest average AUC of 0.814. This may indicate that $CG3_{creative}$ is closer to *DeceptH* than *TruthH* and that deceptive text, regardless of who or what created it, is different than truthful text.

The multiclass classifier attains the highest AUC for the Behavioral Latent Traits (AUC = 0.801) and all features combined (AUC = 0.798). This suggests that personality, empathy, and demographics do not add additional predictive signal above the Behavioral Latent Traits. This is not the case for the binary classifiers. Here the Behavioral Latent Traits are the most predictive single class of features, but combining all feature sets results in a boost over the Behavioral Latent Traits alone.

7. Conclusions

In this paper, we showed that human traits, represented in language, distinguish between inherently deceptive ChatGPT texts, intentionally deceptive human texts, and truthful human texts. ChatGPT was more limited in its expression of human traits with personality in particular having very limited variations. In other words, recent versions of ChatGPT seem to produce text equivalent to a median person for most traits, while for openness and emotional stability, ChatGPT scored higher estimates in line with a positivity bias in LLMs [33]. Indeed, these results also dovetail with previous research that found social spambots on Twitter to be extremely limited in expressions of personality [32].

The results from our predictive analyses suggest psychologically grounded human traits are able to accurately distinguish between deceptive and truthful humans (AUC = 0.678) as well as deceptive machines and truthful humans (AUC = 0.966). Considering the drastically different dimensionality of features based on psychological-theory (thirteen features) as compared to more sophisticated feature sets (100s to 1000s of features; see Appendix), such psychologically-grounded features with a rich history of empirical and qualitative work describing their meaning, can offer an avenue of interpretability for distinguishing text from LLMs.

While we evaluated large language models using human traits, we do not mean to imply that these systems are human and note that there are many risks when doing so. Personifying and relating to automated systems may create transparency issues, which can be exasperated by high stakes tasks; see Abercrombie et al. [59] for a detailed discussion. Furthermore, such personifications may further propagate stereotypes, e.g., defaulting to female-gendered versus ender-ambiguous systems [60]. The present work's use of human traits shows that while these systems do express reasonable values (e.g., an average estimated age of 35), there is a lack of variance. In other words, the systems may look human, on average, but fail to express a range or diversity of traits, thus highlighting their limitations.

Future work should continue to interrogate these approaches to compare how they might perform in other tasks that involve human and non-human text classifications. It is important for future work to consider different types of experiential text to identify how the results in the current paper hold. Further, as LLMs continue to improve in their ability to approximate human communication, future studies should examine classification accuracies against our results to benchmark how models can discriminate between human and non-human texts that are communicated honestly and dishonestly.

References

[1] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, N. A. Smith, 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 (Volume 1: Long Papers) (2021). doi:10.18653/v1/2021.acl-long.565.

[2] N. Köbis, L. D. Mossink, Artificial intelligence versus maya angelou: Experimental evidence

that people cannot differentiate ai-generated from human-written poetry, Computers in human behavior 114 (2021) 106553. doi:10.1016/j.chb.2020.106553.

[3] S. Kreps, R. M. McCain, M. Brundage, All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation, Journal of experimental political science 9 (2022) 104–117. doi:10.1017/XPS.2020.37.

[4] J. Zamora, I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations, in: Proceedings of the 5th international conference on human agent interaction, 2017, pp. 253–260. doi:10.1145/3125739.3125766.

[5] W. Jiao, W. Wang, J. tse Huang, X. Wang, Z. Tu, Is chatgpt a good translator? yes with gpt-4 as the engine, 2023. doi:10.48550/arXiv.2301.08745. arXiv:2301.08745.

[6] D. Dillion, N. Tandon, Y. Gu, K. Gray, Can ai language models replace human participants?, Trends in Cognitive Sciences (2023). doi:10.1016/j.tics.2023.04.008.

[7] G. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies, 2023. arXiv:2208.10264.

[8] M. L. Kern, J. C. Eichstaedt, H. A. Schwartz, L. Dziurzynski, L. H. Ungar, D. J. Stillwell, M. Kosinski, S. M. Ramones, M. E. Seligman, The online social self: An open vocabulary approach to personality, Assessment 21 (2014) 158–169. doi:10.1177/1073191113514104.

[9] M. L. Newman, C. J. Groom, L. D. Handelman, J. W. Pennebaker, Gender differences in language use: An analysis of 14,000 text samples, Discourse processes 45 (2008) 211–236. doi:10.1080/01638530802073712.

[10] J. W. Pennebaker, L. A. King, Linguistic styles: language use as an individual difference., Journal of personality and social psychology 77 (1999) 1296. doi:10.1037/0022-3514.77.6.1296.

[11] M. C. Ashton, K. Lee, R. E. De Vries, The hexaco honesty-humility, agreeableness, and emotionality factors: A review of research and theory, Personality and Social Psychology Review 18 (2014) 139–152.

[12] D. N. Jones, D. L. Paulhus, Duplicity among the dark triad: Three faces of deceit., Journal of Personality and Social Psychology 113 (2017) 329. doi:10.1037/pspp0000139.

[13] D. M. Markowitz, Toward a deeper understanding of prolific lying: Building a profile of situation-level and individual-level characteristics, Communication Research 50 (2023) 80–105. doi:10.1177/00936502221097041.

[14] D. M. Markowitz, T. R. Levine, It's the situation and your disposition: A test of two honesty hypotheses, Social Psychological and Personality Science 12 (2021) 213–224. doi:10.1177/1948550619898976.

[15] J. Zhang, R. Zhang, Y. Zhang, G. Yan, The rise of social botnets: Attacks and countermeasures, IEEE Transactions on Dependable and Secure Computing 15 (2016) 1068–1082.

[16] S. Cresci, A decade of social bot detection, Communications of the ACM 63 (2020) 72–83.

[17] S. Abdelnabi, M. Fritz, Adversarial watermarking transformer: Towards tracing text provenance with data hiding, in: 2021 IEEE Symposium on Security and Privacy (SP), 2021, pp. 121–140. doi:10.1109/SP40001.2021.00083.

[18] S. Gehrmann, H. Strobelt, A. Rush, GLTR: Statistical detection and visualization of generated text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Florence, Italy, 2019, pp. 111–116. URL: https://aclanthology.org/P19-3019.

doi:10.18653/v1/P19-3019.
[19] A. Najee-Ullah, L. Landeros, Y. Balytskyi, S.-Y. Chang, Towards detection of ai-generated texts and misinformation, in: Socio-Technical Aspects in Security: 11th International Workshop, STAST 2021, Virtual Event, October 8, 2021, Revised Selected Papers, Springer, 2022, pp. 194–205.
[20] T. Fagni, F. Falchi, M. Gambini, A. Martella, M. Tesconi, Tweepfake: About detecting deepfake tweets, Plos one 16 (2021) e0251415.
[21] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, arXiv preprint arXiv:2301.11305 (2023).
[22] A. Garcia-Silva, C. Berrio, J. M. Gomez-Perez, Understanding transformers for bot detection in twitter, arXiv preprint arXiv:2104.06182 (2021).
[23] J. Tourille, B. Sow, A. Popescu, Automatic detection of bot-generated tweets, in: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, 2022, pp. 44–51.
[24] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, A. Szlam, Real or fake? learning to discriminate machine from human generated text, arXiv preprint arXiv:1906.03351 (2019).
[25] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can ai-generated text be reliably detected?, arXiv preprint arXiv:2303.11156 (2023).
[26] K. Krishna, Y. Song, M. Karpinska, J. Wieting, M. Iyyer, Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense, arXiv preprint arXiv:2303.13408 (2023).
[27] M. Gambini, T. Fagni, F. Falchi, M. Tesconi, On pushing deepfake tweet detection capabilities to the limits, in: 14th ACM Web Science Conference 2022, 2022, pp. 154–163.
[28] R. Tang, Y.-N. Chuang, X. Hu, The science of detecting llm-generated texts, arXiv preprint arXiv:2303.07205 (2023).
[29] J. Hohenstein, M. Jung, Ai as a moral crumple zone: The effects of ai-mediated communication on attribution and trust, Computers in Human Behavior 106 (2020) 106190.
[30] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1808–1822. URL: https://www.aclweb.org/anthology/2020.acl-main.164. doi:10.18653/v1/2020.acl-main.164.
[31] J. P. Dickerson, V. Kagan, V. Subrahmanian, Using sentiment to detect bots on twitter: Are humans more opinionated than bots?, in: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), IEEE, 2014, pp. 620–627.
[32] S. Giorgi, L. Ungar, H. A. Schwartz, Characterizing social spambots by their human traits, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 5148–5158. URL: https://aclanthology.org/2021.findings-acl.457. doi:10.18653/v1/2021.findings-acl.457.
[33] D. M. Markowitz, J. Hancock, J. Bailenson, Linguistic markers of ai-generated text versus human-generated text: Evidence from hotel reviews and news headlines, 2023. URL: psyarxiv.com/mnyz8. doi:10.31234/osf.io/mnyz8.

[34] L. Fröhling, A. Zubiaga, Feature-based detection of automated language models: tackling gpt-2, gpt-3 and grover, PeerJ Computer Science 7 (2021) e443.
[35] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. arXiv:2301.07597.
[36] V. Hauch, I. Blandón-Gitlin, J. Masip, S. L. Sporer, Are computers effective lie detectors? a meta-analysis of linguistic cues to deception, Personality and social psychology Review 19 (2015) 307–342. doi:10.1177/1088868314556539.
[37] M. L. Newman, J. W. Pennebaker, D. S. Berry, J. M. Richards, Lying words: Predicting deception from linguistic styles, Personality and social psychology bulletin 29 (2003) 665–675. doi:10.1177/0146167203029005010.
[38] M. K. Johnson, C. L. Raye, Reality monitoring., Psychological review 88 (1981) 67.
[39] D. M. Markowitz, J. T. Hancock, Linguistic traces of a scientific fraud: The case of diederik stapel, PloS one 9 (2014) e105937. doi:10.1371/journal.pone.0105937.
[40] G. D. Bond, A. Y. Lee, Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language, Applied Cognitive Psychology 19 (2005) 313–329.
[41] J. T. Hancock, L. E. Curry, S. Goorha, M. Woodworth, On lying and being lied to: A linguistic analysis of deception in computer-mediated communication, Discourse Processes 45 (2007) 1–23. doi:10.1080/01638530701739181.
[42] D. M. Markowitz, D. J. Griffin, When context matters: How false, truthful, and genre-related communication styles are revealed in language, Psychology, Crime & Law 26 (2020) 287–310. doi:10.1080/1068316X.2019.1652751.
[43] M. Ott, Y. Choi, C. Cardie, J. T. Hancock, Finding deceptive opinion spam by any stretch of the imagination, arXiv preprint arXiv:1107.4557 (2011).
[44] K. B. Serota, T. R. Levine, A few prolific liars: Variation in the prevalence of lying, Journal of Language and Social Psychology 34 (2015) 138–157. doi:10.1177/0261927x14528804.
[45] T. R. Levine, K. B. Serota, F. Carey, D. Messer, Teenagers lie a lot: A further investigation into the prevalence of lying, Communication Research Reports 30 (2013) 211–220. doi:10.1080/08824096.2013.806254.
[46] Y. Daiku, K. B. Serota, T. R. Levine, A few prolific liars in japan: Replication and the effects of dark triad personality traits, Plos one 16 (2021) e0249815.
[47] J. Sarzyńska, M. Falkiewicz, M. Riegel, J. Babula, D. S. Margulies, E. Nęcka, A. Grabowska, I. Szatkowska, More intelligent extraverts are more likely to deceive, PloS one 12 (2017) e0176591. doi:10.1371/journal.pone.0176591.
[48] M. Sap, G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, L. Ungar, H. A. Schwartz, Developing age and gender predictive lexica over social media, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1146–1151.
[49] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, M. E. Seligman, Automatic personality assessment through social media language., Journal of personality and social psychology 108 (2015) 934.
[50] L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, H. G. Gough, The international personality item pool and the future of public-domain personality measures, Journal of Research in personality 40 (2006) 84–96.
[51] S. Giorgi, S. Havaldar, F. Ahmed, Z. Akhtar, S. Vaidya, G. Pan, L. H. Ungar, H. A.

Schwartz, J. Sedoc, Human-centered metrics for dialog system evaluation, arXiv preprint arXiv:2305.14757 (2023).
[52] M. H. Davis, Measuring individual differences in empathy: Evidence for a multidimensional approach., Journal of personality and social psychology 44 (1983) 113.
[53] M. Abdul-Mageed, A. Buffone, H. Peng, S. Giorgi, J. Eichstaedt, L. Ungar, Recognizing pathogenic empathy in social media, Proceedings of the International AAAI Conference on Web and Social Media 11 (2017) 448–451. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14942. doi:10.1609/icwsm.v11i1.14942.
[54] D. B. Yaden, S. Giorgi, M. Jordan, A. Buffone, J. C. Eichstaedt, H. A. Schwartz, L. H. Ungar, P. Bloom, Characterizing empathy and compassion using computational linguistic analysis, Emotion (2023). doi:10.1037/emo0001205.
[55] V. Kulkarni, M. L. Kern, D. Stillwell, M. Kosinski, S. Matz, L. Ungar, S. Skiena, H. A. Schwartz, Latent human traits in the language of social media: An open-vocabulary approach, PloS one 13 (2018) e0201703.
[56] R. A. Armstrong, When to use the b onferroni correction, Ophthalmic and Physiological Optics 34 (2014) 502–508.
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[58] H. A. Schwartz, S. Giorgi, M. Sap, P. Crutchley, L. Ungar, J. Eichstaedt, Dlatk: Differential language analysis toolkit, in: Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations, 2017, pp. 55–60.
[59] G. Abercrombie, A. C. Curry, T. Dinkar, Z. Talat, Mirages: On anthropomorphism in dialogue systems, arXiv preprint arXiv:2305.09800 (2023).
[60] A. Danielescu, S. A. Horowit-Hendler, A. Pabst, K. M. Stewart, E. M. Gallo, M. P. Aylett, Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–17.
[61] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of liwc2015, 2015.
[62] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A robustly optimized BERT pre-training approach with post-training, in: Proceedings of the 20th Chinese National Conference on Computational Linguistics, Chinese Information Processing Society of China, Huhhot, China, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108.

Appendix

A. Hotel Data

The data collected by Ott et al. [43] used a list of 20 hotels in Chicago, IL USA. We use this same list to query ChatGPT. The hotels are as follows: Affinia, Allegro, Amalfi, Ambassador, Conrad, Fairmont, Hardrock, Hilton, Homewood, Hyatt, Intercontinental, James, Knickerbocker, Monaco, Omni, Palmer, Sheraton, Sofitel, Swissotel, and Talbott.

B. Additional AI Generated Reviews

To test how ChatGPT generates hotel reviews with default API parameters that are not as creative, we also prompt ChatGPT with a temperature of 1, frequency penalty of 0.0 and presence penalty of 0.0 for the same prompts as $CG3_{creative}$– we call this data set of 400 reviews $CG3_{def}$. We also directly prompt the ChatGPT3.5 and ChatGPT4 GUI interface by prompting with "Write me 20 positive hotel reviews for the <HOTEL> in Chicago. Each review must be around 120 words long." – the difference being that we directly prompt for the generation of 20 hotel reviews in a single query. The reviews in both the cases were much *shorter* than 120 words, with the average length of a review calculated to be 52.8 words for ChatGPT3.5 ($CG3_{sh}$) and 62.7 words for ChatGPT4 ($CG4_{sh}$).

Table A1 shows that the default parameters for ChatGPT resulted in larger KS distances (i.e., the distributions are less similar), which is reasonable since ChatGPT's default parameters are designed to be less creative and more deterministic than settings used for $CG3_{creative}$. GPT-4 produced the smallest KS distances, often showing distances similar to *TruthH* vs *DeceptH*, suggesting GPT-4 shows human traits more similar to actual human writing than ChatGPT.

Table A1
KS Test results when comparing $CG3_{sh}$, $CG4_{sh}$, $CG3_{def}$ against truthful human (*TruthH*) text and deceptive human (*DeceptH*) text. Bonferroni corrected significance level: $^{*}p < 0.05$, $^{\dagger}p < 0.01$, $^{\ddagger}p < 0.001$

Attribute	$CG3_{sh}$ v TH	$CG3_{sh}$ v DH	$CG4_{sh}$ v TH	$CG4_{sh}$ v DH	$CG3_{def}$ v TH	$CG3_{def}$ v DH
Demographics						
Age	$0.228^{\ddagger}$	$0.208^{\ddagger}$	0.058	0.063	$0.368^{\ddagger}$	$0.330^{\ddagger}$
Gender	0.078	0.108	0.073	0.073	$0.280^{\ddagger}$	$0.290^{\ddagger}$
Personality						
Openness	$0.413^{\ddagger}$	$0.313^{\ddagger}$	$0.350^{\ddagger}$	$0.248^{\ddagger}$	$0.545^{\ddagger}$	$0.448^{\ddagger}$
Conscientiousness	$0.280^{\ddagger}$	$0.283^{\ddagger}$	$0.273^{\ddagger}$	$0.190^{\ddagger}$	$0.258^{\ddagger}$	$0.285^{\ddagger}$
Extraversion	$0.333^{\ddagger}$	$0.289^{\ddagger}$	$0.328^{\ddagger}$	$0.283^{\ddagger}$	$0.335^{\ddagger}$	$0.288^{\ddagger}$
Agreeableness	$0.243^{\ddagger}$	$0.273^{\ddagger}$	$0.258^{\ddagger}$	$0.283^{\ddagger}$	$0.325^{\ddagger}$	$0.325^{\ddagger}$
Emotional Stability	$0.375^{\ddagger}$	$0.415^{\ddagger}$	$0.243^{\ddagger}$	$0.253^{\ddagger}$	$0.275^{\ddagger}$	$0.280^{\ddagger}$
Empathy	$0.660^{\ddagger}$	$0.585^{\ddagger}$	$0.185^{\ddagger}$	0.100	$0.548^{\ddagger}$	0.423
Behavioral Latent Traits						
BLT1	$0.350^{\ddagger}$	$0.305^{\ddagger}$	$0.498^{\ddagger}$	$0.450^{\ddagger}$	$0.523^{\ddagger}$	$0.478^{\ddagger}$
BLT2	$0.213^{\ddagger}$	$0.198^{\ddagger}$	$0.190^{\ddagger}$	$0.208^{\ddagger}$	0.140^{*}	$0.143^{\dagger}$
BLT3	0.075	0.123	0.108	0.078	0.118	$0.160^{\ddagger}$
BLT4	0.070	$0.168^{\ddagger}$	0.093	0.113	$0.238^{\ddagger}$	$0.370^{\ddagger}$
BLT5	$0.195^{\ddagger}$	$0.283^{\ddagger}$	0.073	0.123	$0.145^{\dagger}$	$0.153^{\dagger}$

Table A2
KS Test results for all Behavioral Latent Traits (BLTs) when comparing $CG3_{creative}$ against human text. Bonferroni corrected significance level: $^{*}p < 0.05$, $^{\dagger}p < 0.01$, $^{\ddagger}p < 0.001$

Attribute	$CG3_{creative}$ vs *TruthH*	$CG3_{creative}$ vs *DeceptH*	*TruthH* vs *DeceptH*
BLT1	$0.595^{\ddagger}$	$0.548^{\ddagger}$	0.070
BLT2	$0.163^{\ddagger}$	0.155^{*}	0.040
BLT3	0.108	0.110	0.098
BLT4	$0.218^{\ddagger}$	$0.355^{\ddagger}$	$0.148^{\dagger}$
BLT5	$0.143^{\dagger}$	0.138^{*}	0.165^{*}
BLT (Average)	0.245	0.261	0.104

C. Predictive Baselines

Since each model was fit using a small number of features, chosen a priori, it is natural to wonder how a larger, open vocabulary feature space would perform (e.g., contextual embeddings). Thus, we considered several baselines to establish a predictive ceiling, which may reveal how difficult this classification task is. These baselines included (a) n-grams: 1, 2 and 3-grams extracted from the hotel reviews, (b) meta-linguistic features: the total number of unigrams per review and the average unigram length, (c) Linguistic Inquiry and Word Count (LIWC) [61]: a manually curated dictionary of 73 categories, and (d) contextual embeddings vias RoBERTA-large [62]: embeddings are extracted from the penultimate hidden layer (1024 dimensions) of RoBERTA-large for each review. The language model was not further finetuned to our data set, instead the classifier was trained on the 1024 dimensions obtained from RoBERTA-large out of the box. Again, a logistic regression model was used within a 5-fold cross validation framework. Table A3 shows these results.

Table A3
Area under the ROC Curves (AUCs) for supervised binary classification using the baselines. The **bold** numbers indicate the best performance in each column.

Features	$CG3_{creative}$ vs *DeceptH*	$CG3_{creative}$ vs *TruthH*	*TruthH* vs *DeceptH*
Meta-linguistic	0.957	0.981	0.585
n-grams	0.998	0.998	0.826
LIWC	0.990	0.998	0.815
Roberta	**0.998**	**0.999**	**0.959**