The Consistent Lack of Variance of Psychological Factors Expressed by LLMs and Spambots

Vasudha Varadarajan*[,1], Salvatore Giorgi*[,2], Siddharth Mangalik[1]
Nikita Soni[1], David M. Markowitz[3], H. Andrew Schwartz[1]
[1]Department of Computer Science, Stony Brook University
[2]Department of Computer and Information Science, University of Pennsylvania
[3]Department of Communication, Michigan State University
{vvaradarajan,has}@cs.stonybrook.edu, sgiorgi@sas.upenn.edu

Abstract

In recent years, the proliferation of chatbots like ChatGPT and Claude has led to an increasing volume of AI-generated text. While the text itself is convincingly coherent and human-like, the variety of expressed of human attributes may still be limited. Using theoretical individual differences, the fundamental psychological traits which distinguish people, this study reveals a distinctive characteristic of such content: AI-generations exhibit *remarkably* limited variation in inferrable psychological traits compared to human-authored texts. We present a review and study across multiple datasets spanning various domains. We find that AI-generated text consistently models the authorship of an "average" human with such little variation that, on aggregate, it is clearly distinguishable from human-written texts using unsupervised methods (i.e., without using ground truth labels). Our results show that (1) fundamental human traits are able to accurately distinguish human- and machine-generated text and (2) current generation capabilities fail to capture a diverse range of human traits.

1 Introduction

Modern large language models (LLMs; e.g., LLaMA and GPT4) can produce coherent, grammatically sound, and human-like text. These models can also take on human personas (Jiang et al., 2024), reproduce human-like biases (Aher et al., 2023), and may be able to pass a Turing test (Jones and Bergen, 2024). As such, these models are being deployed in real-world situations, such as tutoring (García-Méndez et al., 2024), serving as synthetic patients for training therapists (Wang et al., 2024), and replacing humans in crowdsourcing tasks (Dillion et al., 2023).

These advances have also driven an increase in machine-generated text. While LLMs can be used for innocuous tasks (generating a cover letter for an employment application) they can also be used with malicious intent, such as for phishing attacks, spamming, and disinformation (Crothers et al., 2023). Thus, machine-generated text presents a significant problem for cybersecurity and other social and political contexts.

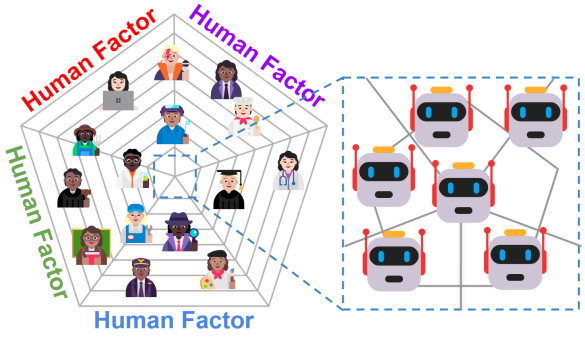

Figure 1: Humans express a range of psychological traits (or human factors) through language. While LLMs and spambots produce fluent text, the psychological traits they express tend to average out across all dimensions, which is uncharacteristic of humans.

Despite their human-like generations, there is mounting evidence that LLMs express a limited range of humanness. LLMs have been shown to reflect Western norms (Havaldar et al., 2023), lean politically left (Feng et al., 2023), and fail to reflect opinions of many sociodemographic groups (Santurkar et al., 2023; Giorgi et al., 2024). In particular, these models are known to generate text according to the average of their training data (i.e., predict the most probable next token), and thus reflect average values and beliefs (Johnson et al., 2022).

Against this backdrop, the current work leverages the limited diversity in human-like expressions to identify machine-generated text. This is done through the lens of *individual differences* (which we call Human Factors), or fundamental psychological traits (such as personality) known to distinguish people and their outcomes (Caspi et al.,

* equal contribution

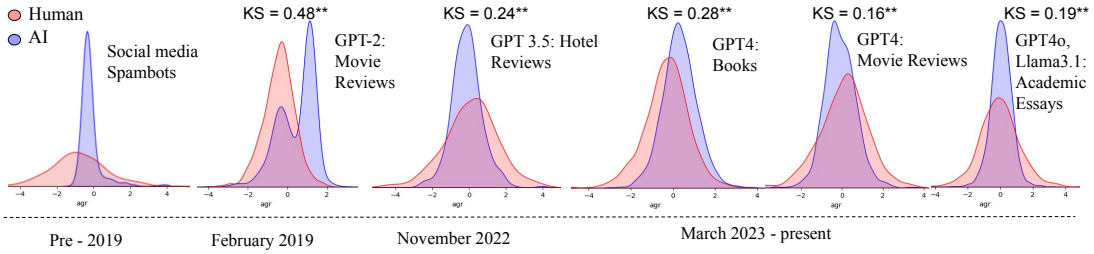

Figure 2: Kernel Density Estimate plot for Agreeableness, a Big-5 Personality trait across various NLP models over the years over multiple domains. Although recent developments seem to have improved AI-generated texts' capacity to display more variability in Agreeableness trait, AI-generated texts are still distinguishable from human-generated texts when analyzing *multiple* such human traits together. KS: two-sample Kolmogorov-Smirnov statistic for a two-sided test; ** : p-value < 0.001.

1997; Perlman et al., 2009; Attig et al., 2017; Anglim et al., 2020). Using preexisting, "off-the-shelf" machine learning models, we estimate individual differences across human- and machine-generated text, representing each text as a small number of interpretable features (e.g., age, personality, and empathy). Across several datasets, ranging from social media bots to academic essays, we see that machine generated text shows a lack of variance in expressed individual differences. Leveraging this lack of variation, these features are then clustered using *unsupervised* methods (i.e., with no human/machine label). Cluster labels are then used to classify the text as machine or human generated. These results show that interpretable, psychologically informed features can be used to identify machine generated text, but also shed light on current text generation capabilities and their lack of diversity in psychological traits.

2 Related Work

Recent LLM research has extensively focused on distinguishing machine-generated text from human writing. Some studies have considered linguistic patterns such as sentence lengths, lexical variations, and richness of vocabulary (Muñoz-Ortiz et al., 2023). Conversely, some prior works focused on emotions (Huang et al., 2023), cultural variations (Havaldar et al., 2023; Das et al., 2024), and psychological factors such as personality (Jiang et al., 2024), and psychometric inventories (Pellert et al., 2024). LLMs have been shown to exhibit an ecological fallacy by treating individual text sequences as independent samples rather than considering the broader context of authorship (Soni et al., 2024), resulting in an averaged representation of writing styles (Johnson et al., 2022) and personalities (Huang et al., 2024) from their training data.

Prior work has leveraged this lack of variance in LLMs-generated text in tasks like authorship attribution in the realm of human versus machine generated texts (Mitchell et al., 2023; Sadasivan et al., 2023; Hu et al., 2023), differentiating human versus bot language (Giorgi et al., 2021). In this study, we further build on past works to show that psychological features can help identify machine generated text.

3 Data

We estimate human factors across four datasets of human/machine text, which span a range of domains and LLMs. Two of the datasets have been used in past work incorporating human factors (summarized here) and the remaining two applications are novel. All datasets used were collected from previous works and our contribution is the application of our methods to these domains. Table 1 summarizes all datasets.

Twitter Spambots This dataset consists of 2,913 genuine (human) Twitter accounts and 2,913 spambots originally collected by Cresci et al. (2017) and analyzed for human traits by Giorgi et al. (2021).[1] These spambots are known as social spambots and differ from traditional bots in that they intentionally try to emulate real humans (Ferrara et al., 2016).[2]

Hotel Reviews This dataset consists of 400 human and 400 machine generated hotel reviews from

[1]Unsupervised classification results using human factors can be found in Giorgi et al. (2021) This dataset is included here to summarize previous work and show how the human factors of machine generated text has evolved over time.

[2]Social media bot accounts are understood to be a mixture of humans (as malicious or unfaithful actors), machines, and human-machine hybrids, and therefore their outputs are not considered purely "machine generations". For this study, we consider social media bots to be non-genuine humans and distinct from real humans, thus closer to machine generations.

Name	Domain	LLMs	Humans:LLMs	Citation
Academic Essays	English Essays	GPT-3.5-Turbo, GPT-4o, GPT-4o-mini, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini and Claude-3.5	1145:1224	Chowdhury et al. (2025)
	Arabic Essays	GPT-3.5-Turbo, GPT-4o, GPT-4o-mini, Gemini-1.5, Llama-3.1 (8B), Phi-3.5-mini and Claude-3.5	1864:1858	Chowdhury et al. (2025)
Hotel Reviews	Hotel Reviews	GPT4	400:400	Markowitz et al. (2024)
RAID	Abstracts	GPT4	1966:1966	Dugan et al. (2024)
	Books	GPT4	1981:1981	Dugan et al. (2024)
	News	GPT4	1980:1980	Dugan et al. (2024)
	Social Media	GPT4	1979:1979	Dugan et al. (2024)
	Movie reviews	GPT4	1143:1143	Dugan et al. (2024)
	Wiki	GPT4	1979:1979	Dugan et al. (2024)

Table 1: Dataset description. The sample size of each dataset is denoted as the ratio of the number of documents written by humans to those written by LLMs (Humans:LLMs).

20 hotels in Chicago, US (Markowitz et al., 2024). The human reviews were collected from TripAdvisor and the machine reviews were generated by GPT4. The human dataset was collected by Ott et al. (2011) and all texts were analyzed for human traits by Giorgi et al. (2023).

Academic Essays This dataset consists of 3,722 English academic essays and 2,369 Arabic academic essays written by humans and machines (Chowdhury et al., 2025). For machine language, seven different open and closed LLMs were used. For this dataset, both human and LLM English essays were provided alongside Arabic essays. Before running human trait inference all Arabic essays were translated into English using the Google Translate API.

RAID This is a benchmark dataset for machine-generated text detection, which includes 6 million generations across 11 models and 11 domains (Dugan et al., 2024). Because our human factor models were trained on social media data, we dropped domains that we believed were least similar to social media language: recipes, poetry, and code. We also dropped non-English texts. Due to space limitations, we only consider GPT4, with a greedy decoding strategy and no repetition penalty.

4 Methods

We proceed in three steps: (1) estimate human factors from text, (2) visualize the human factor distributions, and (3) cluster the human factors using unsupervised methods (i.e., clustering with no ground truth) to assign human/machine labels. The DLATK package (Schwartz et al., 2017) is used for both human factor estimation and clustering.

4.1 Estimating Human Factors

All human factors are estimated from English text using pre-existing models. High-level details are below, with further details in Appendix A.

Demographics. Age and gender were predicted using a social media-based model trained on unigrams (Sap et al., 2014), achieving high accuracy (product moment correlation = 0.86 for age, 90% accuracy for gender), with gender predictions being output as a continuous score.

Personality. Big Five personality traits (openness, conscientiousness, extraversion, agreeableness, and emotional stability) were predicted by a Ridge regression model trained on annotated Facebook statuses (Park et al., 2015), with prediction accuracies (product moment correlation) ranging from 0.35 to 0.43 across the five traits.

Empathy. Empathy was predicted using a Ridge regression model trained on Facebook data and LDA topics, achieving an out-of-sample product moment correlation of r = 0.26 (Yaden et al., 2023).

Behavioral Linguistic Traits (BLTs). Behavior-based Linguistic Traits were introduced by Kulkarni et al. (2018) as a new set of five human traits derived from unprompted language use on social media through factor analysis of Facebook n-grams. It offers a language-based and open-vocabulary alternative to personality.

4.2 Human Factor Distributions

Here we plot the density distribution of the human factors, for both human and machine generations, to visually inspect distributional differences, as past work has shown that humans and machines differ on these human factors (Giorgi

Domain	Personality					Empathy	Behavioral Linguistic Traits					Demographics	
	Ope	Con	Ext	Agr	Emo		F1	F2	F3	F4	F5	Age	Gender
RAID													
Abstracts	.18***	.13***	.05*	.06**	.06**	.22***	.05**	.07***	.18***	.31***	.25***	.05**	.29***
Books	.31***	.10***	.09***	.26***	.18***	.11***	.07***	.05*	.55***	.31***	.20***	.07***	.16***
News	.34***	.05*	.04*	.13***	.07***	.11***	.05*	.08***	.48***	.22***	.09***	.16***	.18***
Reddit	.36***	.13***	.14***	.13***	.09***	.13***	.16***	.07***	.48***	.31***	.06***	.10***	.25***
Reviews	.42***	.22***	.20***	.13***	.15***	.17***	.08***	.04	.50***	.55***	.28***	.13***	.41***
Wiki	.30***	.08***	.05*	.12***	.03	.07***	.09***	.09***	.41***	.11***	.13***	.15***	.12***

Table 2: Kolmogorov-Smirnov test comparing the human and GPT4 distributions across all RAID domains. Benjamini-Hochberg corrected significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

et al., 2023). For both the Twitter Spambot and Hotel Reviews datasets, past work has shown that humans exhibit a larger variation in human traits (wider distributions), while machines tend to have less variance but still exhibit a human-like range in values (e.g., the "age" of social spambots are still within an acceptance human-like range, with no negative values or extreme outliers). We also perform a Kolmogorov-Smirnov test, a non-parametric statistical test, across the human and machine distributions to assess whether they differ.

4.3 Unsupervised Classification

The 13 estimated human factors from the texts are then clustered into two clusters, since we are concerned with human/machine binary classification and each dataset has roughly a 50/50 split of human/machine text. We use spectral clustering with radial basis function (RBF) kernel for capturing the concentric geometry akin to Figure 2 but across 13 human factors. Spectral clustering was used with a gamma parameter of 0.5 with 2 dimensions used to calculate the spectral embedding. The affinity matrix was constructed considering 10 nearest neighbors. Clusters were assigned with column-pivoted QR factorization.

For labeling as human/machine, the intra-cluster spread is calculated for cluster by averaging the distance of all the points from the cluster's center. The cluster with the higher intra-cluster spread has higher variability in the human traits – and hence is more likely to contain human-written text. All texts in this cluster are labeled as human (0) and all texts from the other cluster are assigned a machine label (1). All labels are thus assigned in a completely unsupervised fashion (without the use of ground truth human/machine labels).

Baseline For a baseline comparison, we extract unigrams from each dataset, encode them via their relative frequency within each document, and consider the 10,000 most frequent unigrams. We then project the 10,000 unigrams down to 13 dimen-

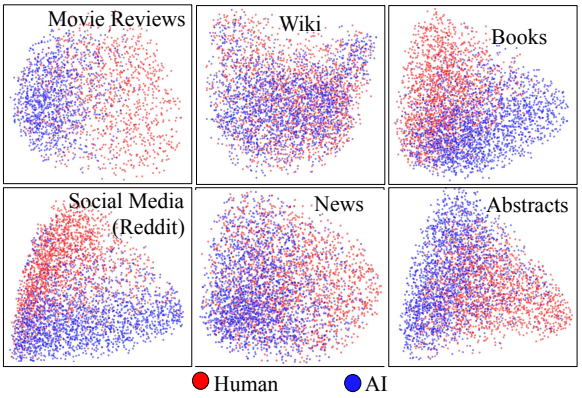

Figure 3: Plot of each human (red) and AI (blue) document in 2-D using spectral embeddings. In reduced dimensions, we see a separation between the human and GPT4 generated text in the RAID dataset.

	13-D Proj. Unigrams			13 Human Factors			All Unigrams (Upper Bound)
	F1	Prec	Rec	F1	Prec	Rec	F1
Hotel Reviews	.55	.64	.49	**.59**	.60	.58	.56
Acad. Essays							
English	.52	.52	.52	**.78**	.71	.87	.52
Arabic	.55	.55	.54	**.63**	.58	.70	.52
RAID							
Abstracts	.62	.61	.64	**.65**	.48	.98	**.87**
Books	.49	.46	.53	**.66**	.63	.69	**.75**
News	.51	.50	.52	**.68**	.58	.80	**.68**
Reddit	.27	.50	.18	**.65**	.50	1.00	.35
Reviews	.54	.52	.56	**.81**	.75	.89	**.84**
Wiki	.50	.53	.46	**.54**	.53	.56	**.86**

Table 3: Classification metrics for Unsupervised classification of machine-text for all the tasks. To make a fair comparison the 10,000 unigrams were projected to 13 dimensions using a random linear projection. The F1 score with all the unigrams as input is given in the rightmost column, as an upper-bound. **Bold** represents the higher F1 among 13-D unigrams and 13 human factors, and underline represents 13 features performing better than the full set of unigrams.

	Demog.	Empathy	Pers.	BLTs	13 Human Factors
Hotel Reviews	.52	.43	.54	**.59**	**.59**
Acad. Essays					
English	.41	.40	.66	.74	**.78**
Arabic	.50	.49	.54	.59	**.63**
RAID					
Abstracts	.57	.41	.53	.34	**.65**
Books	.50	.54	.62	.65	**.66**
News	.49	.45	.56	**.68**	**.68**
Reddit	.55	.64	**.66**	.52	.65
Reviews	.54	.55	.56	**.81**	**.81**
Wiki	**.54**	.50	.50	.45	**.54**

Table 4: F1 scores for classification for the Human Factors separately: demographics (demog.), empathy, personality (pers.), and behavioral linguistic traits (BLTs).

sions, using a random linear transformation. This was done (1) since unigrams were used as input when estimating the human factors (and thus all methods begin with similar raw linguistic information), (2) to keep in number of input features identical to the number of human factors, and (3) since the human factors (e.g., personality) were historically derived via an empirical factor analysis (i.e., a linear transformation; Roccas et al., 2002). These 13 dimensions are also similarly clustered and labeled as described in §4.3. We also consider a non-transformed version of the unigrams and cluster all 10,000 unigram frequencies. We consider this baseline a rough upper bound on classification accuracy (since it uses more features) and is thus able to better learn cluster differences as compared to the 13 human factors.

5 Results

Distributions In Figure 2 we show the distribution of agreeableness across each dataset. We see that machine text (blue) has much smaller variation than human (red) text across multiple domains and models. Table 2 shows the full results of a two-sample Kolmogorov-Smirnov test for the RAID dataset, where we find that the human and machine distributions statistically differ for each domain.

Unsupervised Classification Performance Figure 3 shows that the spectral embedding space of human factors produces a clear separation between human and machine text across several domains in RAID. Of these, Wiki seems to be the most difficult domain for Human Factors to differentiate machine-generated texts, indicated by low separation in the human and machine text. This is likely because the dataset consists of Wikipedia articles, which are crowd-sourced from multiple authors. This could lend Wikipedia articles an *averaged voice* that we usually find in machine-generated texts.

Table 3 shows the unsupervised classification results for all datasets. For both Hotel Reviews and Academic Essays, the Human Factors outperform both the reduced unigram factors and the full set of 10,000 unigrams. In RAID, Human Factors outperform reduced unigram factors across all the domains, and the full unigram feature set outperforms the Human Factors on all but one domain: Reddit. We note that the 13 Human Factors were trained on social media data and, thus, these models may generalize to Reddit more than other domains in RAID. Table 4 shows the results for each dataset broken down by specific human factors. Here we see both personality and BLTs generally outperforming all other Human Factors. However, inclusion of all the Human Factors generally yields the best performance across all the domains.

The results show that the human factors are a meaningful factorization of the language and, in some cases, this factorization contains more information than the 10,000 raw linguistic features.

6 Conclusions

We have shown that individual differences — fundamental psychological traits that distinguish humans — can also distinguish humans from machines. Unlike human traits, the values for these dimensions are so consistently average for machines, that it is unusual for a person to have them. Specifically, across multiple bots and generative LLMs, datasets, and domains, machine-generated text exhibits smaller variations in expressed human factors than human-generated text. This enables *unsupervised* classifiers using a handful of interpretable features (those that can theoretically distinguish people) to distinguish bots from people well beyond baseline models.These results also give insight into how current generation methods, such as LLMs aligned with RLHF, generate human-like text that nonetheless lacks a diverse range of human traits. This dovetails with a growing line of research showing that LLMs fail to generate diverse cultural values, beliefs, and attitudes (Hovy and Yang, 2021; Havaldar et al., 2023). These weaknesses underscore limitations in training data quality and generation methods as well as the opportunities for integrating psychological theories of individual differences to improve LLMs.

7 Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #202222072200005, and a grant from the NIH-NIAAA (R01 AA028032). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, any other government organization, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

8 Limitations

There are several limitations of this study. First, we only consider English text, as our human factor models are all trained on English data. This limits their application and, in the case of translating other languages to English (as we did with Arabic), this assumes that linguistic expressions of human factors are invariant across cultures, which they are not (Smith et al., 2016). Similarly, the human factor models were all trained on social media data and thus may not generalize to other domains (such as reviews and academic essays). Next, some of the human trait models have lower predictive accuracy (with product moment correlation in the range of 0.26 to 0.43). While these accuracies are near state-of-the-art within their respective domains, low accuracies could produce more noisy estimates, especially when models are applied out of domain. Finally, the demographic model only considers binary expressions of gender as male/female, which may incorrectly characterize non-binary authors.

9 Ethical Considerations

Depending on the setting a classifier is deployed in, misclassifications of human and machine generated text could be high risk. For example, labeling genuine academic essays as machine generated may have serious negative repercussions for students and researchers. It has already been shown that current detection methods are biased against non-native speakers (Liang et al., 2023). Similarly, mislabeling social media bots as human users could enhance the trust and accessibility given to bot accounts used to spread disinformation or hate.

It is crucial to avoid anthropomorphizing LLMs, as doing so can create challenges with transparency and trust, particularly in high-stakes scenarios (see Abercrombie et al. (2023) for a detailed discussion). While we propose evaluation metrics based on human psychology, this does not imply that these systems resemble humans, should be perceived as human, or are human.

References

Muhammad Abdul-Mageed, Anneke Buffone, Hao Peng, Salvatore Giorgi, Johannes C Eichstaedt, and Lyle H Ungar. 2017. Recognizing pathogenic empathy in social media. In *ICWSM*, pages 448–451.

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Jeromy Anglim, Sharon Horwood, Luke D Smillie, Rosario J Marrero, and Joshua K Wood. 2020. Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological bulletin*, 146(4):279.

Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing personality differences in human-technology interaction: an overview of key self-report scales to predict successful interaction. In *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19*, pages 19–29. Springer.

Avshalom Caspi, Dot Begg, Nigel Dickson, HonaLee Harrington, John Langley, Terrie E Moffitt, and Phil A Silva. 1997. Personality differences predict health-risk behaviors in young adulthood: evidence from a longitudinal study. *Journal of personality and social psychology*, 73(5):1052.

Shammur Absar Chowdhury, Hind AL-Merekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, Georgios Mikros, and Firoj Alam. 2025. Genai content detection task 2: Ai vs. human – academic essay authenticity challenge. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Abu Dhabi, UAE. Association for Computational Linguistics.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The

paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972.

Evan N Crothers, Nathalie Japkowicz, and Herna L Viktor. 2023. Machine-generated text: A comprehensive survey of threat models and detection methods. *IEEE Access*, 11:70977–71002.

Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Communications of the ACM*, 59(7):96–104.

Silvia García-Méndez, Francisco de Arriba-Pérez, and María del Carmen Somoza-López. 2024. A review on the use of large language models as virtual tutors. *Science & Education*, pages 1–16.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.

Salvatore Giorgi, David M Markowitz, Nikita Soni, Vasudha Varadarajan, Siddharth Mangalik, and H Andrew Schwartz. 2023. " i slept like a baby": Using human traits to characterize deceptive chatgpt and human text. In *IACT@ SIGIR*, pages 23–37.

Salvatore Giorgi, Lyle Ungar, and H. Andrew Schwartz. 2021. Characterizing social spambots by their human traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5148–5158, Online. Association for Computational Linguistics.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. 2006. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1):84–96.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in Neural Information Processing Systems*, 36:15077–15095.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Cameron R Jones and Benjamin K Bergen. 2024. People cannot distinguish gpt-4 from a human in a turing test. *arXiv preprint arXiv:2405.08007*.

Vivek Kulkarni, Margaret L Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H Andrew Schwartz. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PloS one*, 13(11):e0201703.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

David M. Markowitz, Jeffrey T. Hancock, and Jeremy N. Bailenson. 2024. Linguistic markers of inherently false ai communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1):63–82.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Susan B Perlman, James P Morris, Brent C Vander Wyk, Steven R Green, Jaime L Doyle, and Kevin A Pelphrey. 2009. Individual differences in personality predict how people look at faces. *PloS one*, 4(6):e5952.

Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Maarten Sap, Greg Park, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, Lyle H Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, gender, and age in the language of social media: The Open-Vocabulary approach. *PLoS ONE*.

H Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. Dlatk: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60.

Laura Smith, Salvatore Giorgi, Rishi Solanki, Johannes Eichstaedt, H Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle Ungar. 2016. Does 'well-being'translate on twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2042–2047.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

David B. Yaden, Salvatore Giorgi, Matthew Jordan, Anneke Buffone, Johannes C. Eichstaedt, H. Andrew Schwartz, Lyle H. Ungar, and Paul Bloom. 2023. Characterizing empathy and compassion using computational linguistic analysis. *Emotion*.

A Further Details on Human Factor Estimation

To estimate each human factor, we extract the required linguistic features for each document in each dataset. All of the models listed below use some combination of 1-, 2-, and 3-grams (encoded as relative frequencies) and a set of 2,000 LDA topics. The LDA topics were derived in previous work (Schwartz et al., 2013). Topic loads in this work are calculated via a weighted sum of unigram frequencies, where weights were derived via the LDA process (i.e., the conditional probability of the topic given the unigram). We then apply the trained human factor models (e.g., Ridge regression for personality or a factor reduction for Behavioral Linguistic Factors) to the extracted features, producing 13 human factor scores for each document.

Demographics Age and gender were predicted using a model developed by Sap et al. (2014). This model was trained on data from over 70,000 users of Twitter, Facebook, and blogs, who self-reported their age (continuous) and gender (binary male/female; multi-class gender data was unavailable). Unigrams were extracted from social media posts, which were then used in penalized Ridge regression for age prediction and a support vector classifier for gender prediction. The model achieved a product moment correlation of r = 0.86 for age and an accuracy of 90% for gender. Although the gender model was designed to predict binary outcomes, it produces a continuous score, where negative values align with "male" and positive values with "female."

Personality Personality traits were assessed using a model by Park et al. (2015), trained on Facebook status updates from over 66,000 individuals who reported their personality via the International Personality Item Pool (Goldberg et al., 2006). Responses were recorded on a 5-point Likert scale, with trait scores calculated as averages of corresponding items, resulting in final scores ranging from 1 to 5. The model employed penalized Ridge regression using 1-, 2-, and 3-grams and Latent Dirichlet Allocation (LDA) topics derived from the posts. Out-of-sample prediction accuracies (product moment correlations) were 0.43 for openness, 0.37 for conscientiousness, 0.42 for extraversion, 0.35 for agreeableness, and 0.35 for emotional stability.

Empathy Empathic Concern (referred to as empathy) was predicted using a model trained on data from the Interpersonal Reactivity Index (Davis, 1983) combined with Facebook status updates from prior datasets (Yaden et al., 2023; Abdul-Mageed et al., 2017). LDA topics derived from the posts were incorporated into a penalized Ridge regression model, yielding an out-of-sample product moment correlation of r = 0.26.

Behavioral Linguistic Factors Behavioral Linguistic Factors were estimated using a dataset of Facebook status updates from approximately 50,000 users, leveraging a model originally developed by (Kulkarni et al., 2018). N-gram frequencies (1-, 2-, and 3-grams) from these updates underwent factor analysis to derive the dimensions, which serve as a data-driven, open-vocabulary analog to the Big Five personality traits. These dimensions have demonstrated broader applicability, predicting outcomes such as income, and have been shown to be stable across time and diverse populations.