Systematic Evaluation of GPT-3 for Zero-Shot Personality Estimation

Adithya V Ganesan* **Yash Kumar Lal*** **August Håkan Nilsson**
Stony Brook University Oslo Metropolitan University

H. Andrew Schwartz
Stony Brook University

{avirinchipur, ylal}@cs.stonybrook.edu

Abstract

Very large language models (LLMs) perform extremely well on a spectrum of NLP tasks in a zero-shot setting. However, little is known about their performance on human-level NLP problems which rely on understanding psychological concepts, such as assessing personality traits. In this work, we investigate the zero-shot ability of GPT-3 to estimate the Big 5 personality traits from users' social media posts. Through a set of systematic experiments, we find that zero-shot GPT-3 performance is somewhat close to an existing pre-trained SotA for broad classification upon injecting knowledge about the trait in the prompts. However, when prompted to provide fine-grained classification, its performance drops to close to a simple most frequent class (MFC) baseline. We further analyze where GPT-3 performs better, as well as worse, than a pretrained lexical model, illustrating systematic errors that suggest ways to improve LLMs on human-level NLP tasks. The code for this project is available on Github[1].

1 Introduction

Human-level NLP tasks, rooted in computational social science, focus on the link between social or psychological characteristics and language. Example tasks include personality assessment (Mairesse and Walker, 2006; Kulkarni et al., 2017; Lynn et al., 2020), demographic estimation (Sap et al., 2014; Preotiuc-Pietro and Ungar, 2018), and mental health-related tasks (Coppersmith et al., 2014; Guntuku et al., 2017; Matero et al., 2019). Although using LMs as embeddings or fine-tuning them for human-level NLP tasks is becoming popular (V Ganesan et al., 2021; Butala et al., 2021; Yang et al., 2021b), very little is known about zero-shot performance of LLMs on such tasks.

In this paper, we test the zero-shot performance of a popular LLM, GPT-3, to perform personality trait estimation. We focus on personality traits because they are considered the fundamental characteristics that distinguish people, persisting across cultures, demographics, and time (Costa and McCrae, 1992; Costa Jr and McCrae, 1996). These characteristics are useful for a wide range of social, economic, and clinical applications such as understanding psychological disorders (Khan et al., 2005), choosing content for learning styles (Komarraju et al., 2011) or occupations (Kern et al., 2019), and delivering personalized treatments for mental health issues (Bagby et al., 2016). Focusing on zero-shot evaluation of GPT-3 on these fundamental characteristics forms a strong benchmark for understanding how much and what dimensions of traits GPT-3 encodes out-of-the-box. Further, while fine-tuned LMs have only had mixed success beyond lexical approaches (Lynn et al., 2020; Kerz et al., 2022), using zero-shot capable LLMs could help lead to better estimates.

The NLP community has a growing interest in understanding the capabilities and failure modes of LLMs (Wei et al., 2022a; Yang et al., 2021c), and we explore questions that surround LLMs in the context of fundamental human traits of personality. Zero-shot performance can depend heavily on the explicit information infused in the prompt (Lal et al., 2022). Personality, defined by information in its well-established questionnaire tests, presents new opportunities for information infusion.

Our **contributions** address: **(1)** what information about personality is useful for GPT-3, **(2)** how its performance compares to current SotA, **(3)** the relation between ordinality of outcome labels with performance and **(4)** whether GPT-3 predictions stay consistent given similar external knowledge.

2 Background

Psychological traits are stable individual characteristics associated with behaviors, attitudes, feelings, and habits (APA, 2023). The "Big 5" is a popu-

*These authors contributed equally
[1] github.com/humanlab/gpt3-personality-estimation

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 390–400
July 14, 2023

lar personality model that breaks characteristics into five fundamental dimensions, validated across hundreds of studies across cultures, demographics, and time (Costa and McCrae, 1992; McCrae and John, 1992). The approach is rooted in the *lexical hypothesis* that the most important traits must be encoded in language (Goldberg, 1990). We investigate all five factors from this model: openness to experience (OPE– intellectual, imaginative and open-minded), conscientiousness (CON– careful, thorough and organized), extraversion (EXT– energized by social and interpersonal interactions), agreeableness (AGR– friendly, good natured, conflict avoidant) and neuroticism (NEU– less secure, anxious, and depressive).

LLMs like PaLM (Chowdhery et al., 2022) have shown significant improvement in performance on various NLP tasks (Wei et al., 2022b; Suzgun et al., 2022), even without finetuning. There is a growing body of work investigating one of the ubiquitous LLMs, GPT-3, under different settings (Wei et al., 2022a; Shi et al., 2022; Bommarito et al., 2023). Inspired by this, we systematically study the ability of GPT-3 to perform personality assessment under zero-shot setting. Following evidence that incorporating knowledge about the task can improve performance (Vu et al., 2020; Yang et al., 2021b; Lal et al., 2022), we evaluate the impact of three different types of knowledge to determine which type improves personality estimation.

Modeling personality traits through natural language has been extensively studied using a wide range of approaches, from simple count-based models (Pennebaker and Stone, 2003; Golbeck et al., 2011) to complex hierarchical neural networks (Read et al., 2010; Yang et al., 2021a). Finetuning LMs has become the mainstream approach for this task only recently (V Ganesan et al., 2021). With the advent of GPT-3, zero- or few-shot settings have become the primary approach to leverage LLMs in other NLP applications, but are yet untested for personality estimation.

3 Dataset

To get a sample of language associated with personality, we followed the paradigm set forth in Jose et al. (2022) whereby consenting participants shared their own Facebook posts along with taking a battery of psychological assessments, including the big five personality test (Donnellan et al., 2006; Kosinski et al., 2013). The dataset comprises of 202 participants with outcomes of interests who had also shared their Facebook posts. First, we filter the data to only include user posts from the last year of data collection (Eichstaedt et al., 2018). Next, we only retain users for whom we have exactly 20 Facebook posts, similar to the approach described in other human-level NLP works (Lynn et al., 2020; Matero et al., 2021). Finally, we anonymize the data by replacing personable identifiable information using SciPy's (Virtanen et al., 2020) NER model. We also remove phone numbers and email IDs using regular expressions. Finally, we are left with anonymized Facebook posts for 142 users and their associated 5 personality traits. This population (all from US) has a gender ratio of 79:18:3 (female:male:others). The age ranged from 21 to 66 (median=37). The big 5 personality trait scores fall in the continuous range of [1, 5]. We discretize the outcome values into the desired number of bins/classes using a quantile discretizer (in Pandas). We explain why we choose to discretize the outcome values in §4.

4 Experimental Design

In this work, GPT-3 is evaluated in a zero-shot setting. We frame the problem of personality prediction as classifying the degree (i.e. high/low or high/medium/low) to which a person exhibits a trait. Ideally, because the big 5 are considered continuously valued variables (McCrae and Costa Jr, 1989), one would model as a regression task, but we found this simplification to classification necessary to get any meaningful insights from GPT-3's zero-shot capability. We also investigate the degradation of performance for tertiary classification instead of binary in §5.

We devise a simple, reasonable prompt (BASIC)[2] to first estimate the ability of GPT-3 to predict the Big 5 personality traits. Building on this, we investigate whether adding external knowledge about these traits helps the model perform better. We use three types of knowledge: **(1) TEXTBOOK**: a concise definition of these traits from Roccas et al. (2002), **(2) WORDLIST**: frequent and infrequent words[3] used by people exhibiting those traits, and **(3) ITEMDESC**: survey items[4] (a positive and a negative) users responded to, based on which their personality scores were estimated.

[2]Examples of all prompts are in Appendix Figure 2.

[3]We use the wordlist from Schwartz et al. (2013).

[4]See Appendix Table 7 for detailed item descriptions.

Model	OPE	CON	EXT	AGR	NEU	Avg
Benchmarks						
MFC	0.352	0.427	0.411	0.372	0.333	0.379
WT-LEX (Park et al.)	**0.492**	0.393	0.516	**0.609**	**0.578**	**0.518**
Zero-Shot GPT-3						
BASIC	0.329†	0.385	0.521	0.435‡	0.333‡	0.400
TEXTBOOK	0.328†	0.401	0.496	0.506*	0.364‡	0.419
WORDLIST	0.366†	0.457	0.445	0.544	0.393‡	0.441
ITEMDESC	0.342†	**0.521**†	**0.569**	0.488†	0.349‡	0.454

Table 1: MACRO F1 scores for different kinds of knowledge added to the prompt. TEXTBOOK refers to adding the definition of the trait as described in Roccas et al. (2002), WORDLIST refers to adding the top 5 positively and negatively correlated unigrams with the trait reported by Schwartz et al. (2013), ITEMDESC refers to adding the items that were a part of the personality questionnaire (Table 7). WT-LEX refers to the SotA model described in §4. The findings indicate a statistically significant distinction when compared to the WT-LEX model, with significance levels of $p < 0.05$ (*), $p < 0.01$ (†), and $p < 0.001$ (‡).

Baseline and Evaluation. The baseline, WT-LEX, is a ridge regression model from Park et al. 2015 trained on dimensionally reduced feature set of n-grams and LDA-based topics extracted from Kosinski et al. (2013) Facebook data. The number of parameters in this model is orders of magnitude less than GPT-3. Even complex neural models (Lynn et al., 2020) have been unsuccessful to surpass its performance. WT-LEX also produces predictions in the continuous scale within the range of [1, 5]. In order to make a fair comparison with GPT-3, we perform the quantile discretization described in §3 and calculate MACRO F1. We evaluate the predictions using macro F1 scores.

5 Results

Table 1 shows GPT-3's performance on different personality traits, with and without knowledge. We find that ITEMDESC prompts the best performance with GPT-3 on average. Surprisingly, the model is able to directly use survey items (ITEMDESC) to predict EXT and CON the best. Utilizing these is hard since it requires relating abstract concepts described in these survey items to the ecological language in the posts. The top frequent and infrequent words (WORDLIST) help model perform the most on AGR, OPE and NEU. We hypothesize that simple, lexical cues are more helpful here since it is easier to draw relations from the surface form in posts. We also note that estimating NEU is difficult for the model, which also is difficult for humans to estimate in zero-acquaintance contexts, (Kenny, 1994), including estimating neuroticism from Facebook profiles. Overall, GPT-3's predictions are heavily biased towards predicting individuals to be high openness and low in neuroticism.

We also tried incorporating all types of knowledge into a prompt and found that performance dropped below BASIC. However, combining knowledge types involves non-trivial decisions such as the order of knowledge types and its composition. We leave this to future work.

Using ITEMDESC, we establish the best possible GPT-3 performance for personality estimation. Although GPT-3's average performance over all traits is still lower than WT-LEX, it outperforms the MFC baseline. Prior work (V Ganesan et al., 2022; Matero et al., 2022) has shown dimensions of mental health constructs and personality traits being captured through language use patterns in LMs. GPT-3's performance in zero-shot setting provides reasonable evidence to believe that language patterns associated with these traits are encoded in its embedding space as well.

6 Analysis

To better understand the utility of GPT-3 for personality estimation, we analyze the effect of **(1)** problem framing, and **(2)** effect of survey items. Furthermore, we perform error analysis of GPT-3 to suggest avenues for improvement.

Problem Framing. When personality estimation is framed as a binary classification, GPT-3 is worse than SoTA on average in a zero-shot setting. Upon looking closer, we note that it is the best model for 2 out of the 5 traits. However, these observations are made in a simplified two-class setting, whereas the big 5 personality model produces a real valued outcome. In order to assess GPT-3's practical viability, we prompt it (ITEMDESC) to provide more fine-grained predictions by presenting trait estimation as a three-class classification problem.

# class	OPE	CON	EXT	AGR	NEU	Avg
2	0.342	0.521	0.569	0.488	0.349	0.454
3	0.141	0.288	0.240	0.160	0.320	0.230

Table 2: MACRO F1 scores of classifying the outcomes into varying number of classes using GPT-3. We find a sharp drop in performance on increasing the number of classes from 2 to 3. Hence, framing personality estimation as a binary classification is the simplest for GPT-3

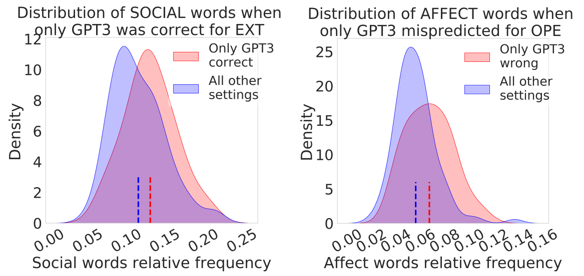
Table 2 shows that problem framing has a major impact on GPT-3 performance for all traits. Three class framing of the problem is harder than the binary framing which is evident from GPT-3's drop in performance (0.229) to close to MFC (0.212). This trend indicates that GPT-3 is ineffective in performing more fine-grained prediction tasks and consequently regression, which is the natural way to estimate the Big 5 traits. Clearly, GPT-3 is yet unsuited for fine-grained personality estimation.

Consistency with Survey Items. The standard questionnaire used to create the dataset had a total of 4 survey items per trait (2 positive and 2 negative). For ITEMDESC, we use one positive and one negative item to describe each trait (see Figure 2). To investigate whether GPT-3 performance can be attributed to specific items in the prompt, we perform ITEMDESC with all possible combinations of a positive and a negative survey item for all traits.

	Avg
ITEMDESC	0.454
BOTHALTITEMS	0.448
ALTPOS	0.430
ALTNEG	0.448

Table 3: MACRO F1 scores for different pairs of positive and negative survey items combinations. Table 7 in Appendix contains the survey items that correspond to these four combination labels.

Table 3 shows that there is no meaningful difference in performance when provided different item combinations. This shows that GPT-3 is not sensitive to the items of the personality questionnaire. This is in line with data in Table 8, which shows that factor loading values (Fabrigar and Wegener, 2011) of these item combinations have similar powers to distinguish the corresponding traits.



(a) (b)

Figure 1: (a) SOCIAL words distributions compared for GPT-3 and WT-LEX under two prediction settings: (1) only GPT-3 correct, and WT-LEX incorrect, and (2) both models correct or GPT-3 incorrect. (right) AFFECT words distributions compared for GPT-3 and WT-LEX under two prediction settings: (1) only GPT-3 incorrect, and WT-LEX correct, and (2) both models incorrect or GPT-3 correct.

Error Analysis. Finally, we examine the linguistic variables that account for the errors in GPT-3 and the areas where it excels as compared to a traditional, lexical-based technique WT-LEX. Figure 1a shows the distributions of SOCIAL words (Tausczik and Pennebaker, 2010) between users that were correctly predicted by only GPT-3 and the users that were either misclassfied by GPT-3 or correctly predicted by WT-LEX for EXT task. SOCIAL words are better captured by LLMs probably owing to its ability to produce contextualized embeddings. Figure 1b depicts the distributions of AFFECT words between the users that were misclassified only by GPT-3 and the users that were either correctly classified by GPT-3 or WT-LEX misclassifies for OPE task[5].

7 Conclusion

We performed a systematic investigation of GPT-3's zero-shot performance on personality estimation. While using a simple prompt did not yield strong performance, injecting knowledge about the traits themselves led to significant improvement. Even so, it falls short of using a strong, extensively-trained, supervised model (WT-LEX). Further, we find that it is much harder for GPT-3 to provide more fine-grained predictions (when asked to select between 3 labels instead of 2), suggesting that LLMs may not be as capable at making dimen-

[5]We also looked at the differences in other LIWC categories for EXT and OPE tasks measured using Cohen's d (Diener, 2010) and logs odds ratio with informative dirichlet prior (Monroe et al., 2008) that offers more explanations for the errors and correctness of GPT-3 in Appendix C.

sional estimates about personality. Our systematic investigation helps understand GPT-3's zero-shot capabilities for a human-level NLP task, contextualizing its failure modes and showing avenues for LLM improvements.

Ethics Statement

Our work seeks to advance interdisciplinary NLP-psychology research for understanding human attributes associated with language. This research is intended to inform Computational Social Science researchers about the ability of LLMs to estimate psychological rating scales as well as for LLM researchers to understand types of psychological information that LMs capture. We intend for our work on personality trait assessments to have an impact on social, NLP, and clinical use cases to improve the well-being of people. We strongly condemn malevolent adoption of these technologies for targeted advertising, directed misinformation campaigns, and other malicious acts that could have potential harms on mental health.

If used for clinical practice, we strongly recommend that any use of LLM-based personality estimates be overseen by clinical psychology experts. During trials, models should be extensively tested for their failure mode rates (e.g. False-positive vs False-negative rates), and error disparities (Shah et al., 2020).

This interdisciplinary computer science, psychology, and health study had extensive privacy & ethical human subjects research protocols. All procedures were approved by an academic institutional review board. All contributors are certified to perform human subject research, and took steps and precautions while collecting and analyzing data to keep participants protected. The Facebook posts shared by consenting users were anonymized as described in §3 to prevent the participants from being identified.

Limitations

The Big 5 personality trait model measures the fundamental dimensions of human on a continuous scale. This real valued representation preserves more information and is more descriptive of inter-individual differences. While we acknowledge that the binary classification of Big 5 traits fails the purpose of the model, it is a necessary simplification to understand the ability of LLMs to perform personality assessment. Our investigation shows potential to improve the practical utility of LLMs in personality estimation.

Despite the strong results from existing works in support of in-context learning and larger message history for better performance, we were limited by the significant multiplicative cost these experiments entailed, as the GPT-3 API is billed based on token usage. Further, since each user's post history is typically long, it is infeasible to experiment with all in-context learning options due to GPT-3's context window size limitation. This is worthy of exploration, to understand the sample efficiency of GPT-3 and the impact of post history on its performance.

Acknowledgement

This work wouldn't have been possible without the support of AVG's and YKL's dear friends, Swanie Juhng, Aakanksha Rajiv Kapoor, Aravind Parthasarathy, Aditya Krishna, Akshay Bharadhwaj, and Somadutta Bhatta, who provided OpenAI API keys for running the experiments. We would also like to extend our gratitude to Matthew Matero, Niranjan Balasubramanian, Harsh Trivedi and Sid Mangalik for providing valuable feedback. AVG, AHN, and HAS were supported in part by NIH grant R01-AA028032. YKL was supported in part by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003 and in part by the NSF under the award IIS #2007290.

References

American Psychological Association APA. 2023. Psychology.

R Michael Bagby, Tara M Gralnick, Nadia Al-Dajani, and Amanda A Uliaszek. 2016. The role of the five-factor model in personality assessment and treatment planning. *Clinical Psychology: Science and Practice*, 23(4):365.

Jillian Bommarito, Michael Bommarito, Daniel Martin Katz, and Jessica Katz. 2023. Gpt as knowledge worker: A zero-shot evaluation of (ai) cpa capabilities. *arXiv preprint arXiv:2301.04408*.

Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea. 2022. Using paraphrases to study properties of contextual embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4558–4568, Seattle, United States. Association for Computational Linguistics.

Yash Butala, Kanishk Singh, Adarsh Kumar, and Shrey Shrivastava. 2021. Team phoenix at WASSA 2021: Emotion analysis on news stories with pre-trained language models. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 274–280, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Paul T Costa and Robert R McCrae. 1992. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5.

Paul T Costa Jr and Robert R McCrae. 1996. Mood and personality in adulthood. In *Handbook of emotion, adult development, and aging*, pages 369–383. Elsevier.

Marc J. Diener. 2010. *Cohen's d*, pages 1–1. John Wiley and Sons, Ltd.

M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. 2006. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Leandre R Fabrigar and Duane T Wegener. 2011. *Exploratory factor analysis*. Oxford University Press.

Jennifer Golbeck, Cristina Robles, and Karen Turner. 2011. Predicting personality with social media. *CHI '11 Extended Abstracts on Human Factors in Computing Systems*.

Lewis R Goldberg. 1990. An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59(6):1216.

Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

R Jose, M Matero, G Sherman, B Curtis, S Giorgi, HA Schwartz, and LH Ungar. 2022. Using facebook language to predict and describe excessive alcohol use. *Alcoholism, Clinical and Experimental Research*, 46(5):836–847.

David A Kenny. 1994. *Interpersonal perception: A social relations analysis.* Guilford Press.

Margaret L Kern, Paul X McCarthy, Deepanjan Chakrabarty, and Marian-Andrei Rizoiu. 2019. Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, 116(52):26459–26464.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 182–194, Dublin, Ireland. Association for Computational Linguistics.

Amir A Khan, Kristen C Jacobson, Charles O Gardner, Carol A Prescott, and Kenneth S Kendler. 2005. Personality and comorbidity of common psychiatric disorders. *The British Journal of Psychiatry*, 186(3):190–196.

Meera Komarraju, Steven J Karau, Ronald R Schmeck, and Alen Avdic. 2011. The big five personality traits, learning styles, and academic achievement. *Personality and individual differences*, 51(4):472–477.

M. Kosinski, D. Stillwell, and T. Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110:5802 – 5805.

Vivek Kulkarni, Margaret L. Kern, David Stillwell, Michal Kosinski, Sandra C. Matz, Lyle H. Ungar, Steven Skiena, and H. A. Schwartz. 2017. Latent human traits in the language of social media: An open-vocabulary approach. *PLoS ONE*, 13.

Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why-questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Veronica Lynn, Niranjan Balasubramanian, and H. Andrew Schwartz. 2020. Hierarchical modeling for user personality prediction: The role of message-level attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5306–5316, Online. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88, New York City, USA. Association for Computational Linguistics.

Matthew Matero, Albert Hung, and H. Andrew Schwartz. 2022. Evaluating contextual embeddings and their extraction layers for depression assessment. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 89–94, Dublin, Ireland. Association for Computational Linguistics.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Matero, Nikita Soni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2021. MeLT: Message-level transformer with masked document representations as pre-training for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2959–2966, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Gregory Park, H. A. Schwartz, J. Eichstaedt, M. Kern, M. Kosinski, D. Stillwell, Lyle H. Ungar, and M. Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108 6:934–52.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291.

Daniel Preotiuc-Pietro and Lyle H. Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *International Conference on Computational Linguistics*.

Stephen J Read, Brian M Monroe, Aaron L Brownstein, Yu Yang, Gurveen Chopra, and Lynn C Miller. 2010. A neural network model of the structure and dynamics of human personality. *Psychological review*, 117(1):61.

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.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen 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)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

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, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *ArXiv*, abs/2210.09261.

Y. Tausczik and J. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24 – 54.

Adithya V Ganesan, Matthew Matero, Aravind Reddy Ravula, Huy Vu, and H. Andrew Schwartz. 2021. Empirical evaluation of pre-trained transformers for human-level NLP: The role of sample size and dimensionality. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4515–4532, Online. Association for Computational Linguistics.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H. Andrew Schwartz. 2022. WWBP-SQT-lite:

Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258, Seattle, USA. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. 2020. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1512–1524, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Feifan Yang, Xiaojun Quan, Yunyi Yang, and Jianxing Yu. 2021a. Multi-document transformer for personality detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14221–14229.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021b. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021c. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI Conference on Artificial Intelligence*.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

A GPT-3

A.1 GPT-3 settings

We used a temperature of 0.0 for all the experiments to select the most likely token at each step, as this setting allows for reproducibility.

```
response = openai.Completion.create(
            model="text-davinci-003",
            prompt=prompt,
            temperature=0,
            max_tokens=1,
            top_p=1.0,
            frequency_penalty=0.1,
            presence_penalty=0.0
)
```

We restricted the model outputs to just one token. Only “Yes" or “No" are considered valid answers for our binary classification task. For the 3-class classification, “High", “Medium" and “Low" are considered valid answers.

For one data point in the WORDLIST EXT experiment, the model output was a newline character instead of Yes/No. By adding another newline to the prompt, we were able to get it to generate an answer (in this case, No). For one data point in the BASIC OPE experiment, the model output contained irrelevant tokens instead of High/Medium/Low. By adding another 2 newlines to the prompt, we were able to get it to generate an answer (in this case, High).

A.2 Prompt Design

For our binary classification task, we used the following prompt template:

```
Read the stream of Facebook posts from a
user below. Each newline represents a new
post. The posts are in order of date, the
last one is the most recent.
{messages}
{knowledge} Given these messages from a
user, is this user {trait} according to
the Big 5 personality traits? Select
between yes or no
```

A user’s posts are concatenated with the most recent post presented at the end to fill the *messages* field. Options for *trait* are agreeable, extraverted, open to experiences, neurotic, and conscientious.

For our 3-class problem framing, we used the following prompt template:

```
Read the stream of Facebook posts from a
user below. Each newline represents a new
post. The posts are in order of date, the
last one is the most recent.
{messages}
{knowledge} Given these messages from a
user, rate their {trait}. The options on
the scale are low, medium, high.
{trait}:
```

Options for *trait* are agreeableness, extraversion, openness to experiences, neuroticism, and conscientiousness. The different types of knowledge injected into the prompt for each personlity trait can be found in Figure 2.

B Glossary

We include the survey items from the questionnaires used in the study to collect data from consenting users along with their associated personality trait in Table 7, as well as the categories of language from the LIWC error analysis model in Table 4.

Category Abbrev	Category	Examples
NUMBER	Numbers	second, thousand
SOCIAL	Social Processes	mate, talk, they
AFFILIATION	Affiliation	ally, friend, social
YOU	2nd Person	you, your, thou
TIME	Time	end, until, season
FAMILY	Family	daughter, dad, aunt
PPRON	Personal Pronoun	I, them, her
POSEMO	Positive Emotion	love, nice, sweet
AFFECT	Affective Processes	happy, cried
FRIEND	Friends	neighbor, buddy
THEY	3rd Person plural	they, their, they’d
FOCUSPAST	Past Focus	ago, did, talked
ACHIEVE	Achievement	success, win, better
SHEHE	3rd person singular	she, him, her
NEGATE	Negation	not, never, no
PRONOUN	Total Pronouns	I, them, itself

Table 4: LIWC glossary to map the category abbreviation with its full form and a few examples for each row.

C Error Analysis

We examine where GPT-3 differs from WT-LEX: **(1)** performing better on EXT in Table 5, and **(2)** predicting OPE worse in Table 6. Results from Table 5 suggest that GPT-3 encodes language categories[6] (Tausczik and Pennebaker, 2010) highly predictive of EXT such as social processes (SOCIAL), group identification (AFFILIATION), and use of second person pronoun (YOU), all of which have been shown to have strong significant association with this trait (Schwartz et al., 2013). GPT-3 can disambiguate common social lexicons occurring in different contexts (Burdick et al., 2022) (e.g., "party" in the context of gathering vs political ideology), which count-based lexical models can’t do.

[6]See Table 4 for details on LIWC categories

	Textbook	Wordlist	Itemdesc
OPE	Note that individuals who are open to experiences tend to be intellectual, imaginative, sensitive and open-minded while individuals that are not open to experiences tend to be down-to-earth, insensitive and conventional.	Note that individuals who are open to experiences tend to use words like universe, art, writing, soul, music while individuals that are not open to experiences tend to use words like cant, dont, gud, nite, 2day.	Note that individuals who are open to experiences tend to have a vivid imagination while individuals that are not open to experiences tend to avoid philosophical discussions.
CON	Note that individuals who are conscientious tend to be careful, thorough, organized and scrupulous while individuals that are not conscientious tend to be irresponsible, disorganized and unscrupulous.	Note that individuals who are conscientious tend to use words like blessed, ready, thankful, relaxing, vacation while individuals that are not conscientious tend to use words like fucking, pokemon, shit, gay, youtube.	Note that individuals who are conscientious tend to complete tasks successfully while individuals that are not conscientious tend to need a push to get started.
EXT	Note that individuals who are extraverted tend to be sociable, talkative, assertive and active while individuals that are not extraverted tend to be retiring, reserved and cautious.	Note that individuals who are extraverted tend to use words like party, girls, baby, gettin, chillin while individuals that are not extraverted tend to use words like anime, manga, internet, japanese, drawing.	Note that individuals who are extraverted tend to make friends easily while individuals that are not extraverted tend to avoid contact with others.
AGR	Note that individuals who are agreeable tend to be good-natured, compliant, modest, gentle, and cooperative while individuals that are not agreeable tend to be irritable, ruthless, suspicious and inflexible.	Note that individuals who are agreeable tend to use words like excited, blessed, great, wonderful, amazing while individuals that are not agreeable tend to use words like fuck, shit, bitch, damn, hell.	Note that individuals who are agreeable tend to believe that others have good intentions while individuals that are not agreeable tend to hold a grudge.
NEU	Note that individuals who are neurotic tend to be anxious, depressed, angry and insecure while individuals that are not neurotic tend to be calm, poised and emotionally stable.	Note that individuals who are neurotic tend to use words like fucking, depression, pissed, anymore, lonely while individuals that are not neurotic tend to use words like success, lakers, basketball, workout, beach.	Note that individuals who are neurotic tend to get stressed out easily while individuals that are not neurotic tend to feel comfortable with themselves.

Figure 2: Different types of knowledge used for each trait in the prompt.

Category	d	OR_{IDP}
NUMBER	0.699	0.140
SOCIAL	0.595	0.191
AFFILIATION	0.459	0.140
YOU	0.451	0.132
TIME	0.448	0.115
FAMILY	0.395	0.108
PPRON	0.359	0.104
POSEMO	0.341	0.102
AFFECT	0.242	0.061
FRIEND	0.217	0.057

Table 5: Lexical categories that are more prevalent when GPT-3 performs better than WT-LEX that explain their EXT predictions. d: Cohen's d – standardized difference in means (Diener, 2010); OR_{IDP}: log **o**dds **r**atio with informative dirichlet prior (Monroe et al., 2008).

Table 6 indicates that GPT-3 fails for OPE on language reflective of social processes (SOCIAL) and affect (AFFECT). Previous work on lexical correlates of personality showed that these categories are discussed more for users low in openness (Yarkoni, 2010), suggesting (together with our result) that GPT-3 misses the connection between these categories of language and personality. These are areas to improve the human-level capabilities of GPT-3.

Category	d	OR_{IDP}
THEY	0.701	0.126
FOCUSPAST	0.692	0.166
AFFECT	0.676	0.132
ACHIEVE	0.629	0.104
SOCIAL	0.608	0.168
SHEHE	0.588	0.172
PPRON	0.559	0.139
NEGATE	0.517	0.082
PRONOUN	0.510	0.105
POSEMO	0.482	0.118

Table 6: Lexical categories that are more prevalent when GPT-3 performs worse for the OPE task than WT-LEX. d: Cohen's d – standardized difference in means of errors (Diener, 2010); OR_{IDP}: log **o**dds **r**atio with informative dirichlet prior (Monroe et al., 2008) on errors.

Trait	Survey Item	Polarity	ItemDesc	AltPos	AltNeg	BothAltItems
OPE	Have a vivid imagination	+	✓		✓	
	Avoid philosophical discussions	-		✓	✓	
	Enjoy wild flights of fantasy	+		✓		✓
	Do not like poetry	-	✓			✓
CON	Complete tasks successfully	+	✓	✓		
	Need a push to get started	-	✓		✓	
	Am always prepared	+			✓	✓
	Shirk my duties	-		✓		✓
EXT	Do not mind being the centre of attention	+			✓	✓
	Make friends easily	+	✓	✓		
	Keep in the background	-		✓	✓	
	Avoid contact with others	-	✓			✓
AGR	Hold a grudge	-	✓	✓		
	Believe that others have good intentions	+	✓		✓	
	Cut others to pieces	-			✓	✓
	Am easy to satisfy	+		✓		✓
NEU	Feel comfortable with myself	-	✓	✓		
	Often feel blue	+		✓		✓
	Get stressed out easily	+	✓		✓	
	Am not easily bothered by things	-			✓	✓

Table 7: Survey items from the questionnaires answered by people for Big 5 personality assessment along with the combination labels these items were a part of (referenced in Table 3.

Trait	**Item Combination**	**Positive Item**	**Negative Item**	**Factor Loading**	**Macro F1**
OPE	ItemDesc	Have a vivid imagination	Do not like poetry	0.703	0.335
	AltNeg	Have a vivid imagination	Avoid philosophical discussions	0.714	0.342
	AltPos	Enjoy wild flights of fantasy	Avoid philosphical discussions	0.720	0.342
	BothAltItems	Enjoy wild flights of fantasy	Do not like poetry	0.787	0.374
CON	ItemDesc	Complete tasks successfully	Need a push to get started	0.781	0.521
	AltNeg	Am always prepared	Need a push to get started	0.800	0.457
	AltPos	Complete tasks successfully	Shirk my duties	0.821	0.476
	BothAltItems	Am always prepared	Shirk my duties	0.837	0.481
EXT	ItemDesc	Make friends easily	Avoid contact with others	0.766	0.569
	AltNeg	Do not mind being the centre of attention	Keep in the background	0.843	0.528
	AltPos	Make friends easily	Keep in the background	0.846	0.551
	BothAltItems	Do not mind being the centre of attention	Avoid contact with others	0.860	0.523
AGR	AltPos	Am easy to satisfy	Hold a grudge	0.725	0.501
	ItemDesc	Believe that others have good intentions	Hold a grudge	0.741	0.488
	AltNeg	Believe that others have good intentions	Cut others to pieces	0.809	0.509
	BothAltItems	Am easy to satisfy	Cut others to pieces	0.813	0.523
NEU	AltPos	Often feel blue	Feel comfortable with myself	0.697	0.333
	AltNeg	Get stressed out easily	Am not easily bothered by things	0.804	0.364
	ItemDesc	Get stressed out easily	Feel comfortable with myself	0.829	0.349
	BothAltItems	Often feel blue	Am not easily bothered by things	0.835	0.333

Table 8: Comparison of factor loading values of the aggregation of a positive item and a negative item from the Big 5 personality questionnaire and the performance of GPT-3 (ItemDesc) for the corresponding Itemdesc pairs. The factor loadings were calculated on an exeternal dataset (Kosinski et al., 2013) with larger number of samples (N=741). There's very little difference in the factor loading values (distinguisginh power) over the four combinations for almost all traits, which is in line with the minor performance differences observed in the consistency experiments explained in §section 6