How effective are Large Language Models (LLMs) at inferring people's personality based on texts they authored?

Victoria Popa[1,2], Guglielmo Cola[1], Caterina Senette[1,*] and Maurizio Tesconi[1]

[1]Institute of Informatics and Telematics (IIT) CNR Via G.Moruzzi 1 56124 Pisa
[2]University of Pisa, Computer Science Department, Largo Bruno Pontecorvo 3 56127 Pisa

Abstract

Personality refers to *"individual differences in characteristic patterns of thinking, feeling, and behaving"*. It is considered a stable spectrum of enduring traits which remain consistent across different situations and over time. According to the notion that individuals' writing reflects their personality, in the past decade, user-generated text has become a focal point for researchers in computer science and psychology aiming to automate the extraction of personality traits. Concurrently, the advent of Large Language Models (LLMs) has enabled exponential growth in the fields of text mining and natural language processing, making large-scale personality inference a potential area of exploration especially in cybersecurity, where understanding user behavior and enhancing threat detection are critical. In this paper, we investigate the capabilities of three state-of-the-art LLMs in assessing personality traits using text samples from two publicly available datasets. We tested a variety of prompts to better guide the LLMs in extracting personality traits based on the Big Five model, one of the most widely accepted psychological frameworks. The results from a comprehensive set of experiments confirm that inferring psychological dispositions from user-generated text, without explicit training, remains a challenging task for LLMs. Additionally, the work herein described provides a benchmark for comparing proprietary and open LLMs models using accuracy as the performance metric. In this regard, our findings do not show a substantial difference in performance between open-source and proprietary models.

Keywords

Large Language Models, Personality, Big Five, User-generated Text

1. Introduction

According to the American Psychological Association [1], personality refers to *"individual differences in characteristic patterns of thinking, feeling and behaving"*. It is considered a stable spectrum of enduring traits that remain consistent across different situations and over time. This stability is a key aspect of what makes personality distinct from temporary states or moods, and it is the basis of its predictive power regarding life outcomes such as well-being, lifespan, job performance, and societal perspectives [2, 3]. For these reasons, current research focuses on finding strategies to infer individuals' personality traits in profitable domains like marketing [4], criminal justice [5], affective computing [6] etc.

Inferring personality traits through text is essential in cybersecurity for understanding user behavior and improving threat detection. Such insights help identify individuals prone to risky behaviors, such as susceptibility to phishing or potential insider threats [7, 8]. Typical applications include: (i) Behavioral Authentication - Enhancing access control systems by verifying users through behavioral consistency; (ii) Training Personalization - Adapting cybersecurity awareness programs to personality traits for higher effectiveness; (iii) Threat Detection - Identifying insider risks or malicious intent through communication analysis. These techniques are increasingly vital as human factors remain the weakest link in cybersecurity. Psychologists have endeavored to develop theories describing human personality

Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), February 03-8, 2025, Bologna, IT
*Corresponding author.
victoria.popa@phd.unipi.it (V. Popa); guglielmo.cola@iit.cnr.it (G. Cola); caterina.senette@iit.cnr.it (C. Senette); maurizio.tesconi@iit.cnr.it (M. Tesconi)
0009-0007-0862-1820 (V. Popa); 0000-0003-2890-723X (G. Cola); 0000-0002-4411-7134 (C. Senette); 0000-0001-8228-7807 (M. Tesconi)

and assessing it consistently [9, 10]. The field has successfully introduced several strong frameworks with corresponding assessment tools [11, 12]. Some of these have also been provided for self-assessment, eliminating the need for a specialized external observer and thus leading the way for their mass use with associated consequences [13, 14]. At the same time, the feasibility of automatically deriving such psychological assessments is increasingly demonstrated [15]. Additionally, if personality drives our thoughts and behaviors, it is also true that these thoughts and behaviors, along with associated traces we leave behind (digital foot-prints), can provide clues about our personality [16, 17]. In this context, the notion that what we write reveals our psychological disposition is growing more compelling [18].

In the last decade, the use of text as a source of individual personality information has been widely studied thanks to the concurrent development of new text mining strategies and the exponential growth of knowledge in the fields of machine learning and natural language processing [19, 20]. Focusing on user-generated text, current literature shows that it is possible to automatically derive personality traits with sufficient accuracy by focusing on different machine learning (ML) strategies applied to manually labeled datasets including high-quality data meticulously collected for this purpose [21]. Conversely, the same ML algorithms show reduced accuracy with "in the wild" data from diverse sources, especially conversations and posts published on social media [21].

The advent of pre-trained Large Language Models (LLMs) has fundamentally transformed text analysis. They have improved natural language understanding, enabling more accurate sentiment analysis and semantic search [22]. Overall, LLMs' versatility and customizability make them indispensable in various applications, revolutionizing how text is processed and understood. At present, researchers are investigating whether LLMs can capture elements of psychology by analyzing linguistic patterns and behaviors expressed in text, especially whether they can infer psychological traits, emotions, and even cognitive states from written content.

In this paper, we investigate the capabilities of three state-of-the-art LLMs (Mixtral, GPT-4, LLAMA3) in assessing personality traits from text samples collected from two publicly available datasets. We tested a variety of prompts to guide the LLMs in extracting personality traits based on the Big Five model, one of the most widely accepted and used psychological frameworks [11]. The main *contribution* of this work lies in having demonstrated that identifying personality traits from text is a challenging task for LLMs. Additionally, we provided a benchmark for comparison between proprietary and open LLMs models in terms of accuracy. In this regard, our findings do not show a substantial difference in performance between open-source and proprietary models for the problem under study. Finally, we address the task of personality trait recognition in an uncontrolled scenario, using texts not specifically written for this purpose. In fact, all the LLMs models were tested under consistent settings and with two types of datasets: one consisting of Facebook user statuses, representing the "in the wild" personality inference scenario, and the other consisting of essays, which are streams of consciousness from psychology students at the request of a professor, substantially outside the social media paradigm (visibility, engagement, network, etc.).

The remainder of this paper is structured as follows: Section 2 reviews related work. In Section 3, we describe the experimental setup. Section 4 presents a summary and discussion of the results, including an examination of limitations. Finally, Section 5 provides conclusions and suggests directions for future research.

2. Related Work

This section reviews the body of literature related to the recognition of personality traits through computational methods, with a particular focus on deep learning and large language models (LLMs).

2.1. Deep Learning for personality recognition

In recent years, the swift advancements in deep learning have led to an increasing number of methods employing deep neural networks for text-based personality recognition, as these networks can automatically extract complex features from user-generated text. For instance, Majumder et al. [23] developed a

deep convolutional neural network using Word2Vec embeddings for personality identification. Xue et al. [24] introduced a two-tier hierarchical neural network to capture the deep semantic features of users' posts for Big Five personality recognition. Lynn et al. [25] applied message-level attention to determine the significance of users' posts in evaluating Big Five personality traits. Other authors utilized a contrastive graph transformer network to learn post embeddings for personality detection achieving excellent results in personality detection [26]. As mentioned when discussing our findings 4, these results represent the best performance achieved in the literature with a supervised approach. In contrast, our goal is to test a completely zero-shot approach.

2.2. LLMs for personality recognition

Large Language Models (LLMs) have currently demonstrated exceptional performance across a variety of Natural Language Processing (NLP) tasks such as machine translation, text generation, and sentiment analysis (to name a few). These outstanding results have prompted researchers to test the ability of LLMs to recognize personality traits in written text. BERT has been the most studied model [22]. Mehta et al [27] conducted extensive experiments with BERT to determine the optimal configuration for personality detection. In Ren et al.'s study [28], BERT was used to generate sentence-level embeddings for personality recognition, also considering sentiment information through a sentiment dictionary. In 2022, Jain et al. [29] proposed a "*personality BERT*", a textual modality-specific deep neural model that fine-tunes a pre-trained bidirectional representation for transformers (BERT) for the personality classification task applied to Kaggle's MBTI dataset and reporting a F1 score of 0.6945.

Today, ChatGPT has garnered significant interest due to its remarkable proficiency in general language processing and an increasing number of researchers are exploring its capabilities in several tasks including text-based personality recognition. GPT-3.5 and GPT-4, have been tested by Zhang et al. [30] in assessing two of the BigFive personality traits (*Extraversion* and *Conscientiousness*) from asynchronous video interviews indicating that LLMs can match or surpass task-specific AI models in zero-shot personality trait prediction. However, they exhibit uneven performance across different traits, lack consistent test-retest reliability, and may introduce biases. Other than using LLMs directly as personality predictors, GPT-3.5 turbo 03-01 has also been fused in different pipelines including NLP models [31] to infer all the Big Five traits of personality. In these cases, researchers successfully leverage ChatGPT verbose responses to bear novel knowledge in affective computing. Even the authors in [32] achieved a GPT-3's zero-shot performance comparable to a state-of-the-art pre-trained model for generic classification, however, for detailed classification, they reported a performance dropping to the level of a simple most frequent class (MFC) model. The authors in [33] tested GPT-3 and GPT-4 on different tasks including personality recognition mainly focusing on prompt design. Results show that ChatGPT models achieved strong performance in sentiment-related problems but not in personality recognition task.

Unlike the works mentioned above (currently available only on arXiv), our objective is to compare the performance of multiple LLMs: GPT-4, the best proprietary model, and Mixtral and LLAMA3, two of the best open models (according to ChatBot Arena's leaderboard[1]) in the task of recognizing the BigFive personality traits focusing on text coming from social media. To this end, we defined and tested several prompts, some of which, to the best of our knowledge, are not present in the literature. Moreover, in most cases (3 out of 4), our prompts asked the models to return personality scores, rather than just high/low classifications as commonly proposed in the literature.

2.3. Psychological Framework

Automatically extracting personality traits requires a personality framework of reference. Various frameworks exist in the literature, but the widely recognized and empirically supported model in personality psychology is the Big Five also known as the Five Factor Model (FFM) [11]. It encompasses five broad dimensions of personality: Openness to Experience, Conscientiousness, Extraversion, Agreeableness,

[1] https://chat.lmsys.org/, accessed on 2024-06-15

Table 1
Meaning of y/n labels (classes) for each trait

Trait	Label	Disposition
Openess	y	insightful
	n	unimaginative
Conscientiousness	y	precise
	n	careless
Extraversion	y	extravert
	n	shy
Agreableness	y	friendly
	n	uncooperative
Neurocitism	y	neurotic
	n	secure

Table 2
y/n distribution in Essays and MyPersonality for each trait

	Essays		MyPersonality	
Trait	**y**	**n**	**y**	**n**
O	0.52	0.48	0.71	0.29
C	0.51	0.49	0.52	0.48
E	0.52	0.48	0.39	0.61
A	0.53	0.47	0.54	0.46
N	0.50	0.50	0.39	0.61

and Neuroticism. Each dimension represents a range of behaviors and traits, with Openness involving creativity and curiosity, Conscientiousness referring to organization and dependability, Extraversion relating to sociability and assertiveness, Agreeableness encompassing compassion and cooperativeness, and Neuroticism indicating emotional instability and anxiety. Numerous standardized tools following the Big Five paradigm are available to internet users, among which the IPIP (International Personality Item Pool) is the most used [34]. The tool is freely available for research and has been employed in numerous studies to assess personality traits, providing a robust and reliable measure of the Big Five dimensions. The IPIP's comprehensive item pool allows for flexible and customizable assessments, making it a valuable resource for both academic and applied psychology settings.

In our experiments, the ground truth data is represented by the scores (or both labels and scores) obtained by users, who authored the analyzed texts, from completing the IPIP-100 items test.

3. Experimental Setup

This section details the experimental setup used to assess the capabilities of large language models in recognizing personality traits, outlining the datasets, model configurations, and evaluation metrics employed.

3.1. Datasets

For our experiments, we used two gold standard datasets: *Essays* [18] and *MyPersonality*. More precisely, we used the versions published by Celli et al. [35] in 2013. These datasets were shared in an effort to address some persistent issues in automated personality recognition: (i) researchers often use different datasets for their experiments, making it hard to compare results; (ii) different evaluation methods further complicate performance comparisons; (iii) creating high-quality, gold standard data for personality recognition from text and social network data is both difficult and expensive.

Table 3
Description of Prompts

Prompt	Description
P1	Act as a Big 5 evaluator. For each personality trait, assign a numerical score ranging from 1.0 (very low) to 5.0 (very high) based on user-provided texts.
P2	Act as the user who wrote the texts. For each of the following 50 statements, choose a value ranging from 1 (very inaccurate) to 5 (very accurate) indicating how accurate they are to describe yourself.
P3	Act as the user who wrote the texts. Indicate how accurate the 5 statements that summarize each personality trait are by giving a numerical score ranging from 1.0 (very inaccurate) to 5.0 (very accurate) to describe yourself.
P4	Act as a Big 5 evaluator. Directly label each user as y or n on each personality trait.

Essays It is a substantial collection of stream-of-consciousness texts (approximately 2400, one per author/user), compiled between 1997 and 2004 and annotated with personality classes [18]. The texts were created by students who completed the Big Five personality test. In the version used in this paper, the dataset's publisher [35] assigned labels based on z-scores, converting them into nominal classes using a median split. Scores are not provided.

MyPersonality It is a dataset of personality scores and Facebook profile data collected by David Stillwell and Michal Kosinski through a Facebook application[2] that administers the Big Five test, among other psychological assessments. Since the original dataset is no longer shared, we used the sample published by Celli et al. [35], consisting of 250 users and approximately 9900 status updates. This subset includes raw text from Facebook statuses and gold standard personality labels (y/n label assigned for each personality trait) derived from self-assessments using a 100-item version of the IPIP[3] personality questionnaire. Each label (y/n) has a different meaning depending on the considered personality trait, as shown in Table 1. Table 2 presents the distribution of classes across the two gold standard datasets.

3.2. LLM setting and prompting

We tested three Large Language Models (LLMs): two open-source models, Mixtral-8x7b and LLAMA3, and one proprietary, OpenAI GPT-4. To ensure more deterministic and focused responses, we set the *temperature* parameter to 0.2 and *top-p* value to 1.0. To assess the models' understanding and portrayal of personality traits, we employed four distinct prompting strategies. These strategies were designed to explore how well each model could recognize and articulate personality traits based on the Big Five personality framework. We employed two distinct approaches to define the prompts: asking the LLMs to act as expert evaluators of the Big Five personality traits, and having them impersonate the user who wrote the texts. Table 3 provides a brief description of our prompts. Table 4 also shows the content of Prompt 2 (P2) as it was provided to LLMs, with only a few of the 50 statements shown for brevity.

As already mentioned, MyPersonality includes both scores and personality y/n labels. Therefore, we were able to derive the thresholds applied on each trait's score for classification. For the Essays dataset, since scores are not available, the thresholds were found using a median split on our results.

For prompts P1 and P3 the inferred scores were transformed into classes based on the threshold values. For P2, we simulated a personality test for each user using 50 items from the IPIP inventory. This included 10 items for each of the Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), with 5 items positively correlated and 5 negatively correlated with each trait. Scores for the items negatively correlated with each personality trait were inverted (scores 1, 2, 3, 4, 5 were transformed to 5, 4, 3, 2, 1 respectively). After inverting the scales, scores for each personality trait were summed, and the average score was calculated and then transformed into classes

[2] http://mypersonality.org/
[3] https://ipip.ori.org/newMultipleconstructs.html

Table 4
Prompt P2 detailed description

```
You must act as a facebook user and you wrote the following list of different statuses "text_to_analyze" . Each
      status is separated by the '~|~' separator.
 Based on what you wrote and on the way you wrote your facebook status (words, intonation, sentiment etc.),
      your JOB is to choose a value ranging from 1 (very inaccurate) to 5 (very accurate) to accurately
      describe yourself. Indicate for each statement whether it is 1 (Very Inaccurate), 2 (Moderately
      Inaccurate), 3 (Neither  Accurate Nor Inaccurate), 4 (Moderately Accurate), or 5 (Very Accurate as a
      description of you).
Return values for each statement in a JSON like:
        {
    'EXT1': I am the life of the party: <chosen value>
    'AGR1': I feel little concern for others: <chosen value>
    'CON1': I am always prepared: <chosen value>
    'NEU1': I get stressed out easily: <chosen value>
    'OPN1': I have a rich vocabulary: <chosen value>
    'EXT2': I do not talk a lot: <chosen value>
    ...
    ...
    ...
    'OPN10': I am full of ideas: <chosen value>

 }
You must meet ALL these requirements ONLY:
 - Return only the values and don't describe or motivate the process of choosing. Avoid motivation and
      description about your choice.
- No additional new columns should be created beyond those listed above.

You are capable. You can reason step by step, check that every requirement is met and therefore answer
     correctly.
```

Table 5
Classification accuracy on **MyPersonality**. The best model results are in bold.

Model	Prompt	O	C	E	A	N	Average
Mixtral	P1	0.52	0.48	0.58	0.54	0.57	0.538
	P2	0.44	0.48	0.60	0.57	0.45	0.508
	P3	0.49	0.56	**0.65**	0.53	0.57	0.560
	P4	0.63	0.54	0.54	0.51	0.52	0.548
LLAMA3	P1	0.57	**0.60**	0.42	0.53	0.49	0.522
	P3	0.48	0.54	0.43	0.53	0.44	0.484
	P4	0.62	0.50	0.50	**0.58**	**0.60**	0.560
GPT-4	P1	0.56	0.52	0.55	0.52	0.51	0.532
	P2	0.61	0.58	0.55	0.57	0.50	**0.562**
	P3	0.60	0.58	0.59	0.52	0.52	**0.562**
	P4	**0.67**	0.57	0.43	0.55	0.55	0.554
SOTA supervised [26]	-	0.82	0.83	0.83	0.73	0.83	0.805

based on the thresholds. For P4, we employed a binary classification approach, asking the model to directly label each user's personality trait with 'y' or 'n'. Consequently, the models for P4 provided the classes directly, eliminating the need to transform scores into classes.

4. Results and Discussion

Table 5 and Table 6 present the experimental results in terms of classification accuracy on MyPersonality and Essays. For completeness, we included the performance of state-of-the-art (SOTA) supervised strategies for both datasets, further discussed in Section 4.1.1.

Results of these experiments show that LLMs have limited efficacy in deriving the Big Five traits from

Table 6
Classification accuracy on **Essays**. The best model results are in bold.

Model	Prompt	O	C	E	A	N	Average
Mixtral	P1	0.58	**0.59**	0.55	0.56	0.58	0.572
	P2	0.58	0.53	0.52	0.49	0.44	0.512
	P3	0.54	0.55	0.55	0.58	0.57	0.558
	P4	0.57	0.54	0.53	**0.59**	0.52	0.550
LLAMA3	P1	0.58	0.57	0.56	0.58	0.58	0.574
	P3	0.54	0.55	0.57	0.56	**0.61**	0.566
	P4	**0.6**	0.55	0.52	0.57	0.57	0.562
GPT-4	P1	0.55	**0.59**	**0.59**	0.58	0.6	**0.582**
	P2	0.56	0.57	0.56	0.57	0.41	0.534
	P3	0.52	**0.59**	0.58	0.57	0.58	0.568
	P4	0.53	0.57	0.58	0.58	0.56	0.564
SOTA supervised [26]	-	0.82	0.80	0.81	0.81	0.82	0.809

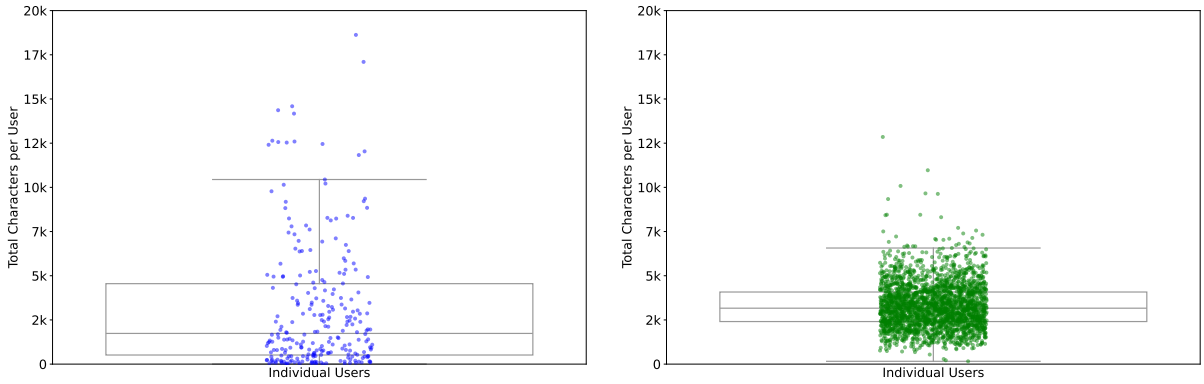
the considered gold standard datasets. Across the diverse prompts, all models, both closed and open-source, achieved accuracy results close to random guessing. This suggests that, despite their potential, general-purpose LLMs may lack the contextual knowledge required to reliably infer personality traits. To further investigate these results, we also found Pearson correlation between the scores found in the MyPersonality dataset and the actual scores. Results indicate lack of correlation, with r values mostly below 0.1. This confirms that the derived scores fail to capture the underlying behavioral patterns in the gold standard dataset. As mentioned earlier, MyPersonality y/n labels were derived from scores using the same thresholds used in the gold standard dataset. Instead, for the Essays dataset we used a median split on our results, as the original dataset does not include scores but only y/n labels. Nevertheless, both approaches led to similar results, close to random behavior.

Among the various prompts, the simpler prompt, P1, achieved slightly more consistent results across the different scenarios. The primary difference observed between open-source and proprietary models concerns prompt handling. Prompt P2, which presents a significant challenge as it requires the LLM to respond to an IPIP-50 for each user, was not adequately processed by LLAMA3, which was not able to respond to all items for most users. Mixtral performed substantially better in P2 handling, but was still unable to respond to all 50 items for about 4% of users, both in the MyPersonality and Essays datasets. Instead, GPT-4 skipped some Essays' users only due to its content safety policy (around 3% of users with P1, P2, and P3, around 1.5% with P4). Additionally, LLAMA3 struggled to process the entire text for each user when the word length was high, having difficulty maintaining context and acting unpredictably. Therefore, it was necessary to limit the text size to 600 words, based on the average word count per user in both datasets (MyPersonality: 572 words; Essays: 652 words).

4.1. Data quality issues

As far as we know, the datasets we selected represent the best publicly available choice for testing the ability to infer personality traits based on texts produced in an uncontrolled environment, such as social media posts. However, although these datasets are optimal for traditional NLP and ML applications for which they were specifically created, they may not be as optimal for use with LLMs. Unlike traditional methods, which often show clearer relationships between input features and predictions, LLMs act as black boxes. This complexity makes it challenging to assess the decision-making process and how LLMs rate personality traits. In this regard, we conducted specific experiments asking LLMs to justify their responses, in an attempt to enhance their attention mechanism. This approach yielded mixed results and did not substantially improve the findings presented in this paper.

The distribution of total characters per user in MyPersonality and Essays is shown in Figure 1a and



(a) MyPersonality (b) Essays

Figure 1: Total characters per user in the two datasets.

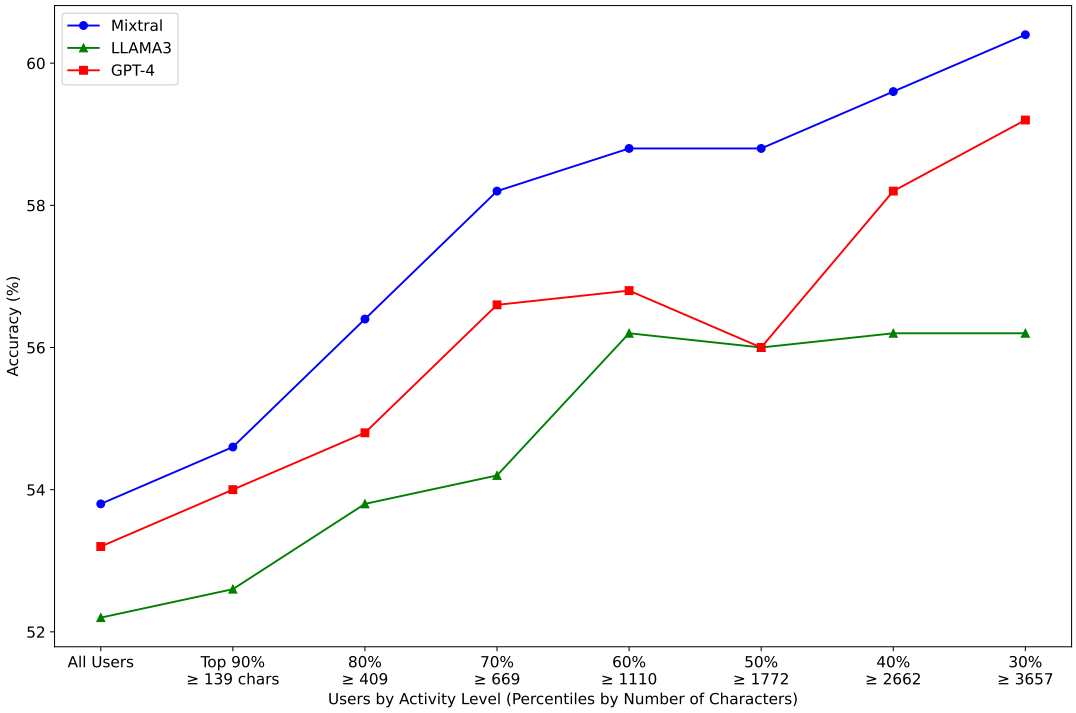


Figure 2: Accuracy for MyPersonality users filtered by activity level (character count) and using prompt P1.

Figure 1b, respectively. In MyPersonality there is a much greater variability, denoted by a greater IQR and the presence of several outliers. In particular, while the median is around 1,700 characters, there are 10% of users with less than 140 characters as well as over 5% of users with more than 10,000 characters. In comparison, Essays (Figure 1b) shows a more consistent length across users, as 99% of users produced more than 1,000 characters. The variability in MyPersonality has important implications for classification performance. This is shown in Figure 2, which depicts how the classification performance of the three LLMs and P1 is influenced by the available characters per user. Results suggest that accuracy could be improved by having enough social media data from each user, as the increase in total character length correlates with higher scores. In contrast, we found no noticeable improvement in selecting the users with longer texts from Essays. Moverover, long texts may even cause LLMs to struggle with maintaining context. As already mentioned, this issue was particularly evident with LLAMA3.

Finally, there is a potential lack of consistency between the text content and the user's self-reported personality scores, particularly given that the ground-truth data was obtained through self-assessment.

While the reliability of the psychometric tool in self-assessment mode is not in question, a *social desirability bias* [36] is likely present in the text that users compose about themselves in a Facebook status. This may also be true for the Essays dataset, which consists of texts written by psychology students at the request of a professor. Similarly, this bias could create some disparities between the psychometric tool's revelations about their personalities and the content of their texts.

4.1.1. Comparison with Supervised State-of-the-Art (SOTA) Methods

As shown in Table 5 and Table 6, the state-of-the-art method presented in [26], which is fully supervised and task-specific, achieved a significantly higher performance compared to our approach based on LLMs. However, it is worth emphasizing again the advantages of an unsupervised approach, which justify ongoing research in this area, despite the current poor performance of LLMs on this specific task. In fact, compared to traditional deep learning techniques, LLMs offer a combination of advanced pre-training, contextual understanding, efficient feature extraction, and scalability, making them more effective and practical for personality recognition tasks. Their potential to achieve superior performance with less domain-specific data, combined with ongoing advancements in ethical AI and interpretability, positions LLMs as a preferred choice for modern NLP applications.

5. Conclusion

This paper evaluates the current capabilities of three cutting-edge LLMs (Mixtral, GPT-4, LLAMA3) in deriving personality traits from text samples sourced from two publicly available datasets. Experiments were conducted with various prompts to possibly enhance the LLMs' ability to identify personality traits according to the Big Five model, a widely recognized and utilized psychological framework. While all three LLMs generally exceeded random baseline performance on both datasets, they did not demonstrate substantial predictive effectiveness. These accuracy scores are comparable to those found in a recent study [33] that considered two models from OpenAI (GPT-4 and older GPT-3.5). A key insight is that, despite improvements in the latest versions of both proprietary and open-source models, recognizing personality traits from text remains a complex task for LLMs. Further research is essential to address the ethical implications associated with the use of LLMs in personality inference. Future efforts should prioritize prompt engineering, testing new models, and enhancing data quality. Given the vast amounts of data accumulated by social media companies, testing larger datasets is crucial to fully understand what these companies may infer about their users, raising significant privacy and ethical concerns.

Acknowledgments

This work is supported by (i) Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021; (ii) Project SERICS (PE00000014) under the NRRP MUR program funded by the EU – NGEU.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. R. VandenBos, APA dictionary of psychology., American Psychological Association, 2007.

[2] B. W. Roberts, N. R. Kuncel, et al., The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes, Perspectives on Psychological Science 2 (2007) 313–345. doi:10.1111/j.1745-6916.2007.00047.x.
[3] S. Jaiswal, S. Song, M. Valstar, Automatic prediction of depression and anxiety from behaviour and personality attributes, in: 2019 8th international conference on affective computing and intelligent interaction (acii), IEEE, 2019, pp. 1–7.
[4] Z. Liu, W. Xu, W. Zhang, Q. Jiang, An emotion-based personalized music recommendation framework for emotion improvement, Information Processing & Management 60 (2023) 103256.
[5] J. Clark, M. T. Boccaccini, B. Caillouet, W. F. Chaplin, Five factor model personality traits, jury selection, and case outcomes in criminal and civil cases, Criminal Justice and Behavior 34 (2007) 641–660.
[6] N. Majumder, S. Poria, H. Peng, N. Chhaya, E. Cambria, A. Gelbukh, Sentiment and sarcasm classification with multitask learning, IEEE Intelligent Systems 34 (2019) 38–43.
[7] J. J. Sirasapalli, R. M. Malla, A deep learning approach to text-based personality prediction using multiple data sources mapping, Neural Computing and Applications 35 (2023) 20619–20630.
[8] S. Tardelli, M. Avvenuti, M. Tesconi, S. Cresci, Characterizing social bots spreading financial disinformation, in: International conference on human-computer interaction, Springer International Publishing Cham, 2020, pp. 376–392.
[9] R. R. McCrae, P. T. Costa, Human nature and individual differences: The evolution of basic traits, Journal of Personality 72 (2004) 173–189. doi:10.1111/j.0022-3506.2004.00262.x.
[10] O. P. John, R. W. Robins, L. A. Pervin, Handbook of personality: Theory and research, 3rd ed. ed., Guilford Press, New York, 2008.
[11] P. T. Costa, R. R. McCrae, A five-factor theory of personality, The five-factor model of personality: Theoretical perspectives 2 (1999) 51–87.
[12] I. B. Myers, A guide to the development and use of the Myers-Briggs type indicator: Manual, Consulting Psychologists Press, 1985.
[13] S. C. Matz, M. Kosinski, G. Nave, D. J. Stillwell, Psychological targeting as an effective approach to digital mass persuasion, Proceedings of the national academy of sciences 114 (2017) 12714–12719.
[14] A. B. Grubb, M. A. McDaniel, Validity of self-assessment in personality, skills, and knowledge, Journal of Applied Psychology 93 (2008) 59–74. doi:10.1037/0021-9010.93.1.59.
[15] W. Youyou, M. Kosinski, D. Stillwell, Computer-based personality judgments are more accurate than those made by humans, Proceedings of the National Academy of Sciences 112 (2015) 1036–1040.
[16] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Proceedings of the national academy of sciences 110 (2013) 5802–5805.
[17] D. Azucar, D. Marengo, M. Settanni, Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis, Personality and individual differences 124 (2018) 150–159.
[18] J. W. Pennebaker, L. A. King, Linguistic styles: language use as an individual difference., Journal of personality and social psychology 77 (1999) 1296.
[19] J. Golbeck, C. Robles, M. Edmondson, K. Turner, Predicting personality from twitter, in: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, IEEE, 2011, pp. 149–156.
[20] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, L. H. Ungar, Personality, gender, and age in the language of social media: The open-vocabulary approach, PloS one 8 (2013) e73791. doi:10.1371/journal.pone.0073791.
[21] N. Akrami, J. Fernquist, T. Isbister, L. Kaati, B. Pelzer, Automatic extraction of personality from text: Challenges and opportunities, in: 2019 IEEE international conference on big data (big data), IEEE, 2019, pp. 3156–3164.
[22] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of naacL-HLT, volume 1, 2019, p. 2.
[23] N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for

personality detection from text, IEEE Intelligent Systems 32 (2017) 74–79.
[24] D. Xue, L. Wu, Z. Hong, S. Guo, L. Gao, Z. Wu, X. Zhong, J. Sun, Deep learning-based personality recognition from text posts of online social networks, Applied Intelligence 48 (2018) 4232–4246.
[25] V. Lynn, N. Balasubramanian, H. A. Schwartz, 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, 2020, pp. 5306–5316.
[26] Y. Zhu, L. Hu, N. Ning, W. Zhang, B. Wu, A lexical psycholinguistic knowledge-guided graph neural network for interpretable personality detection, Knowledge-Based Systems 249 (2022) 108952.
[27] Y. Mehta, N. Majumder, A. Gelbukh, E. Cambria, Recent trends in deep learning based personality detection, Artificial Intelligence Review 53 (2020) 2313–2339.
[28] M. Venugopalan, D. Gupta, An enhanced guided lda model augmented with bert based semantic strength for aspect term extraction in sentiment analysis, Knowledge-based systems 246 (2022) 108668.
[29] D. Jain, A. Kumar, R. Beniwal, Personality bert: a transformer-based model for personality detection from textual data, in: Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021, Springer, 2022, pp. 515–522.
[30] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi, R. E. de Vries, Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns, IEEE Transactions on Affective Computing (2024).
[31] M. M. Amin, E. Cambria, B. W. Schuller, Can chatgpt's responses boost traditional natural language processing?, IEEE Intelligent Systems 38 (2023) 5–11.
[32] A. V. Ganesan, Y. K. Lal, A. H. Nilsson, H. A. Schwartz, Systematic evaluation of gpt-3 for zero-shot personality estimation, arXiv preprint arXiv:2306.01183 (2023).
[33] M. M. Amin, R. Mao, E. Cambria, B. W. Schuller, A wide evaluation of chatgpt on affective computing tasks, arXiv preprint arXiv:2308.13911 (2023).
[34] 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.
[35] F. Celli, F. Pianesi, D. Stillwell, M. Kosinski, Workshop on computational personality recognition: Shared task, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 7, 2013, pp. 2–5.
[36] G. Seidman, Self-presentation and belonging on facebook: How personality influences social media use and motivations, Personality and individual differences 54 (2013) 402–407.