AI-Driven Classification of Personality Items: A Study on Large Language Models

Carlos Henrique Sancineto da Silva Nunes, Alexandre José de Souza Peres, Leonardo de Barros Mose, Petar Čolović, Ricardo Primi, Vithor Rosa Franco

https://doi.org/10.1590/1982-4327e3609

Submitted on: 2026-05-07
Posted on: 2026-05-08 (version 1)
(YYYY-MM-DD)

Paidéia
2026, Vol. 36, e3609.https://doi.org/10.1590/1982-4327e3609
ISSN 1982-4327 (online version)

Psychological Evaluation

AI-Driven Classification of Personality Items: A Study on Large Language Models

Carlos Henrique Sancineto da Silva Nunes[1] https://orcid.org/0000-0002-7769-6937

Alexandre José de Souza Peres[2] https://orcid.org/0000-0002-3472-6120

Leonardo de Barros Mose[3] https://orcid.org/0000-0002-5328-7442

Petar Čolović[4] https://orcid.org/0000-0003-1212-3131

Ricardo Primi[3 e 5] https://orcid.org/0000-0003-4227-6745

Vithor Rosa Franco[3] https://orcid.org/0000-0002-8929-3238

Abstract: This study investigates the effectiveness of large language models (LLMs) in classifying items that evaluate the Big Five personality dimensions, focusing on performance variations across traits and comparing local and cloud-based models. Five Natural Language Processing (NLP) models were used to classify 385 personality items, employing three levels of detail in prompt design. The results indicate that larger generative models, such as ChatGPT-4o and Gemini 1.5 Pro, outperformed smaller models in terms of accuracy, both for the five personality factors and overall. However, the Llama 3.1 model, run locally, showed adequate results for judge-based analyses, offering a viable alternative for those prioritizing

[1]*Universidade Federal de Santa Catarina, Florianópolis-SP, Brazil.*
[2]*Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande-MS, Brazil.*
[3] *Universidade São Francisco (USF), Campinas-SP, Brazil.*
[4]*University of Novi Sad, Novi Sad, Serbia.*
[5] *EduLab21, Instituto Ayrton Senna.*
Fonte de financiamento: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and National Council for Scientific and Technological Development (CNPq).
Correspondence address: Carlos Henrique Sancineto da Silva Nunes. Universidade Federal de Santa Catarina, Centro de Filosofia e Ciências Humanas - CFH, Departamento de Psicologia. Campus Universitário, Trindade, 88040900 - Florianópolis-SC, Brazil. carlos.nunes@ufsc.br

data privacy. The study highlights the potential of LLMs as complementary tools in the development of psychological assessment instruments.

Keywords: personality, natural language processing, artificial intelligence, psychometrics, validity

IA para Classificação de Itens de Personalidade com Grandes Modelos de Linguagem

Resumo: Este estudo investiga a eficácia dos grandes modelos de linguagem (LLMs) na classificação de itens que avaliam as dimensões de personalidade do Big Five, com foco nas variações de desempenho entre traços e na comparação entre modelos locais e baseados em nuvem. Foram utilizados cinco modelos de Processamento de Linguagem Natural (PLN) para classificar 385 itens de personalidade, empregando três níveis de detalhamento no design dos prompts. Os resultados indicam que modelos generativos maiores, como ChatGPT-4o e Gemini 1.5 Pro, superaram os modelos menores em termos de acurácia, tanto para os cinco fatores de personalidade quanto de forma geral. No entanto, o modelo Llama 3.1, executado localmente, apresentou resultados adequados para análises baseadas em juízes, oferecendo uma alternativa viável para quem prioriza a privacidade de dados. O estudo destaca o potencial dos LLMs como ferramentas complementares no desenvolvimento de instrumentos de avaliação psicológica.

Palavras-chave: personalidade, processamento de linguagem natural, inteligência artificial, psicometria, validade

IA para Clasificación de Ítems de Personalidad con Grandes Modelos de Lenguaje

Resumen: Este estudio investiga la eficacia de los grandes modelos de lenguaje (LLMs) para clasificar ítems que evalúan las dimensiones de personalidad del Big Five, analizando las variaciones en el rendimiento entre rasgos y comparando modelos locales y en la nube. Se utilizaron cinco modelos de Procesamiento del Lenguaje Natural (PLN) para clasificar 385 ítems, empleando tres niveles de detalle en el diseño de las instrucciones. Los resultados indican que modelos generativos mayores, como ChatGPT-4o y Gemini 1.5 Pro, superaron a los modelos más pequeños en términos de precisión, tanto en los cinco factores de personalidad como en general. Sin embargo, el modelo Llama 3.1, ejecutado localmente, mostró resultados adecuados para análisis basados en jueces, ofreciendo una alternativa viable para quienes priorizan la privacidad de los datos. El estudio destaca el potencial de los LLMs como herramientas complementarias en el desarrollo de instrumentos de evaluación psicológica.

Palabras clave: personalidad, procesamiento del lenguaje natural, inteligencia artificial, psicometría, validez

According to the lexical hypothesis (Cutler & Condon, 2023), the most important personality traits are encoded in language, which has been utilized in the development of personality models such as the Big Five. These models are derived from the analysis of trait-descriptive adjectives or constructions (Hilliard et al., 2024), which are textual artifacts that, at least in principle, reflect latent behavioral patterns and attitudinal tendencies. Research and applications in modern Natural Language Processing (NLP), mainly represented by large language models (LLMs) and foundation models (Schneider et al., 2024), have systematically shown (Myers et al., 2024) that these methods are quite efficient at detecting subtle patterns and relationships between words and knowledge of language, including those related to

psychological behavior (Savcisens et al., 2024). The integration of NLP and LLMs offers promising avenues for refining traditional personality models and developing new measurement tools for psychological assessment with reduced costs and increased efficiency. This study aimed to investigate the efficacy of LLMs in classifying item text according to Big Five personality dimensions, examining performance variations across traits, comparing local and cloud-based models, and assessing the impact of prompt design on classification quality.

Personality traits are reflected in natural language, which refers to the way people communicate in their daily interactions (Goldberg, 1981). Therefore, the development of personality items is grounded in the use of adjectives and expressions commonly used to describe oneself and others, which have been identified as personality-relevant and recorded in dictionaries and lexicons. The Big Five model is based on the premise that five broad traits – Openness to Experience (a similar trait named Intellect or Culture appeared in several bottom-up lexical studies), Conscientiousness, Extraversion, Agreeableness, and Neuroticism – synthesize the most relevant terms for describing human personality (John, 2021). Since personality scales are based on semantic terms extracted from natural language, it is essential that, in test development, each item is carefully evaluated for its ability to adequately represent and discriminate the target personality trait being measured.

One of the most fundamental sources of validity evidence in instrument development is content-based validity evidence (American Educational Research Association [AERA] et al., 2014). Also referred to as content validity, this type of evidence assesses how comprehensively and representatively the items in an instrument cover the domains intended for measurement (Haynes et al., 1995). Researchers can identify and address potential validity issues by evaluating content validity before collecting data to gather other forms of validity evidence. Thus, establishing content-based validity is typically the first step in test development.

Content validity has traditionally been assessed qualitatively or, at best, semi-quantitatively. The most common method of obtaining this type of evidence is expert rating – in other words, asking experts in the domain being measured to evaluate the relevance and congruence of the items with the construct (AERA et al., 2014). Expert ratings have been routinely used in the initial phases of traditional psycholexical studies, although the procedure has not been standardized so far. To assess the degree of agreement and reliability of experts' ratings regarding test content, some indices are frequently used, such as Kappa, Cronbach's alpha, and the intraclass correlation coefficient (ICC) (Polit & Beck, 2006). One way to assess the agreement of each individual item is through coefficients such as the content validity index (CVI) (Lynn, 1986), which quantifies the proportion of experts who judged an item as relevant, and the content validity ratio (CVR) (Lawshe, 1975), whose coefficient is a linear transformation of the percentage of experts who indicated an item as essential.

Although expert agreement is the most common method for collecting evidence based on test content, there are some disadvantages. First, it is necessary to rely on experts qualified to assess the construct. Additionally, the process of item content analysis requires the availability and time of the experts. Lengthier instruments, for example, may require more time for evaluation, making it more challenging to find available experts. It is also worth noting that the literature has criticized the use of some widely utilized indices for assessing expert agreement, such as the CVI and the CVR. Both indices are typically used with small samples of experts and do not adjust estimates for chance agreement among raters (Polit & Beck, 2006). The Kappa statistic, in turn, is sensitive to the number of categories used and is also limited in situations where the sample of raters is small (Wynd et al., 2003). Finally, coefficients such as the ICC and Cronbach’s alpha only assess inter-rater consistency, providing little information about agreement on each individual item (Almanasreh et al., 2019).

Despite the importance of the steps for assessing the validity of the item generation process, it can be quite costly in many aspects (Russell-Lasalandra et al., 2024). An alternative for speeding up the content analysis process can be found in NLP models, through techniques known as text classification. NLP may be used to classify items prior to their presentation to specialists or as an additional expert, helping to optimize the content analysis process, which often requires significant cognitive effort and considerable time. In the field of psychometrics, NLP approaches have been applied for automatic item generation (Circi et al., 2023), predicting job knowledge, skills, and abilities ratings (Putka et al., 2023), automated scoring (Kjell et al., 2024), and using different sources of information in psychological evaluation (Brickman et al., 2025), among other tasks.

NLP is an umbrella term covering approaches to computational linguistics, which has been one of the pillars of development for machine learning and artificial intelligence approaches (Hussain et al., 2024). In a few words, NLP is a set of quantitative and computational methods that can be used to make machines understand, interpret, and generate human language (Chowdhary, 2020). Traditional NLP tasks include text classification, sentiment analysis, and machine translation, among others (Peres, 2021). As NLP evolved and access to large datasets was facilitated, tasks that required some form of computation of the more complex aspects of language started to become more feasible. This led to the development of LLMs (Schneider et al., 2024), conceived with transformer architecture as central to most effective and efficient applications (Vaswani et al., 2017).

The transformer architecture revolutionized NLP by enabling models to process entire sequences of text simultaneously, rather than sequentially, as was the case with earlier architectures such as Recurrent Neural Networks (RNNs) (Chowdhary, 2020). This is achieved through two mechanisms. The first is the use of embeddings to represent meaning. Embeddings are numerical vectors that encode the intricacy of complex concepts that are

difficult, or impossible, to represent using categorical representations of words or sentences (Primi, 2021). The use of embeddings builds on and expands the tradition of using semantic vectors as comprehensive representations of word semantic content (see, e.g., Jurafsky & Martin, 2000). The second is the mechanism called self-attention, which allows the model to weigh the importance of different words in a sentence relative to each other, regardless of their position, effectively creating contextual embeddings (i.e., contextual representations of words). These mechanisms allow for the scalability and efficiency of transformers, resulting in an enhanced ability to capture long-range dependencies in text. Because of this, transformers are the backbone of most state-of-the-art LLMs today (Minaee et al., 2024), including variants such as GPT, Bidirectional Encoder Representations from Transformers (BERT), and Large Language Model Meta AI (Llama), driving advancements across a wide range of NLP tasks.

Recent research has provided a more comprehensive comparison of the performance of state-of-the-art LLMs on text classification tasks, highlighting the strengths and limitations of each model in different contexts. One study compared the agreement of content analyses performed by eight students with classifications made by 11 transformer models on personality items. The overall accuracy of the raters ranged from 68% to 78%, with an average of 71%. However, some transformer models, such as DeBERTa, RoBERTa, and XLNet, surpassed human raters' accuracy, reaching precision levels above 80% (Fyffe et al., 2024). The results are promising, but further studies are needed to ensure the accuracy of language models in classifying items in psychological domains. In particular, it is known that fine-tuning models for specific tasks can improve their performance on these tasks (Russell-Lasalandra et al., 2024).

Current study

Based on the results presented previously, one may safely conclude that LLMs have emerged as potential tools for evaluating psychological phenomena, particularly those dependent on language structures, as is the case with most psychological assessment instruments. This raises the question: Can LLMs serve as reliable “experts” in personality assessment? This study investigates the effectiveness of LLMs in classifying items to evaluate the Big Five personality dimensions, focusing on performance variations across traits and comparing local and cloud-based models. This study also explores whether LLMs can effectively classify personality dimensions, particularly those of the Big Five. The Big Five was chosen as arguably the most popular, widely applied, and replicated personality trait taxonomy (De Raad & Mlacic, 2015).

We further examine potential differences in classification quality across personality dimensions, as some traits may be easier to detect and interpret. These assumptions stem from the varying cross-cultural replicability of the Big Five traits; for example, Openness may be more challenging to replicate in non-WEIRD cultures. Additionally, the study highlights the importance of comparing local, smaller LLMs with cloud-based, larger models, as their performance may differ significantly. Finally, we compare two distinct methodological approaches: fine-tuning the established BERT model and zero-shot and few-shot prompting techniques applied to more contemporary, larger-scale GPT models. Within the latter approach, we investigate the impact of prompt design (also known as prompt engineering) on the quality of LLM-generated classifications, specifically examining how varying levels of detail in prompts related to the Big Five personality dimensions influence the outcomes.

Method

Five NLP models were used to classify a set of 385 items designed to assess personality based on the Big Five model. The classifications from two of these models were

performed using the web pages provided by the respective companies and executed on their servers (OpenAI, 2023; Pichai & Hassabis, 2024), while the classifications from the other three models were carried out locally on a personal computer. Quantized versions of the generative models were used for local execution. Quantization is a technique for compressing model weights, thereby reducing their size and enabling their use on less advanced hardware (Tang et al., 2022). The analyses with the generative models were conducted using three different prompts, which, in addition to the items to be classified, provided varying levels of information about the Big Five.

Personality Item Pool for Classification

The item pool for classification consisted of three personality assessment measures, with all items originally written in Brazilian Portuguese or adapted from English-language versions. All measures have studies providing evidence of content validity – including expert judgment analysis – and internal structure evidence based on factorial methods. The allocation of items to the five factors was carried out by the test authors using both theoretical and empirical approaches. A total of 63 items from the Big Five Inventory-2 (BFI-2) (Pires et al., 2023), 160 items from the Survey on Social and Emotional Skills (SSES) developed by the Organisation for Economic Co-operation and Development (OECD, 2024), and 162 items from a Socioeconomic Status (SES) measure developed by the Ayrton Senna Institute (Primi, Santos, et al., 2021) were used. The item pool classified by the NLP models consisted of 385 items: 27 for Openness, 36 for Agreeableness, 45 for Conscientiousness, 27 for Extraversion, and 27 for Neuroticism. The analyzed items include their content in Portuguese, as well as the assessed dimension, facet, and polarity. The items were randomized to prevent sets from the same dimension from being presented sequentially, and this order was maintained across all models after the randomization process.

Personality Item Pool for Fine-Tuning the BERT Model

The items were drawn from the Adaptive Personality Battery (BAP), a test developed in Brazil (Nunes et al., 2015), which employs computerized adaptive testing (CAT) for personality assessment based on the Big Five model. The 801 items from the BAP have been subjected to research to establish content validity through expert evaluations, as well as to analyses verifying their internal structure using factor analysis. The items were allocated to the Big Five dimensions: 150 for Agreeableness, 166 for Conscientiousness, 153 for Extraversion, 183 for Neuroticism, and 149 for Openness.

NLP Models Used for Personality Item Classification

BERT

The fine-tuning of the multilingual BERT (Bidirectional Encoder Representations from Transformers) model was conducted for the task of personality item classification. Developed by Devlin et al. (2019), BERT is a language model that considers the contextual understanding of words within a sentence. In this study, BERT was implemented using Apple's CoreML framework. The fine-tuning process used the personality items from the BAP and their classifications within the Big Five. In this process, 761 items were used for training and 40 items for validation, allowing for model parameter adjustments to enhance generalization in classifying new items.

Llama 3.1 Instruct

The Llama (Large Language Model Meta AI) version 3.1 with 8 billion parameters was employed to classify the set of 385 personality items. Developed by Meta AI, Llama represents a series of language models designed for NLP tasks. The model was run locally on a microcomputer, specifically the 8-bit quantized version (Q8). The temperature parameter,

which controls the randomness of the model's output, was set to 0.4. The context window was configured to 4096 tokens, and a feature was enabled to discard information from the middle of the context window once the limit was reached, prioritizing both the initial instructions and the most recent information.

Phi-3 Medium Instruct

The Phi-3 Medium model, a 14-billion-parameter language model developed by Microsoft, was employed to classify the personality item pool. The model was executed locally in its 4-bit quantized version (Q4_K_L) with the temperature set to 0.4. The context window was configured to 4096 tokens, with a feature enabled that discards information from the middle of the context window when its limit is reached. The flash attention feature was also activated.

Gemini 1.5 Pro

The Gemini 1.5 Pro model, a LLM developed by Google AI (Pichai & Hassabis, 2024), was utilized through the Google AI Studio website. Given the available context window at the time, which could handle up to 1 million tokens, all items were submitted in a single request. The temperature of the model was set to 0.4. The classification of the items was performed on June 8, 2024.

ChatGPT-4o

The ChatGPT-4 model, a LLM developed by OpenAI based on the GPT-4 architecture (OpenAI, 2023), was employed for analysis. Analyses were conducted using the web interface provided by OpenAI for accessing the model. Due to token limitations imposed by free-tier access, it was necessary to divide the items into two blocks for each prompt used. The classification of the items was performed on June 8 and 9, 2024.

Prompts Used

Three prompts were employed to classify personality items using the four generative models, varying in the level of detail regarding information about the Big Five personality model. All prompts, along with detailed descriptions of each, are available in the supplementary material of the article.

Prompt 1: Zero-Shot

The prompt instructs language models to apply the Big Five personality traits model (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) to evaluate a list of provided sentences (items). It then requests the model to present its findings in a table, excluding the content of the evaluated items. The table must include the following information: Classification – assigning each item to one of the five factors in the Big Five model; Facet identification – mapping each classified item to its corresponding facet according to the NEO-PI-R, a widely used measure based on the Big Five model; and Polarity indication – determining and noting the polarity of each item, that is, whether it indicates a high or low level in the evaluated personality dimension.

Prompt 2: Contextual Few-Shot

The second prompt expands upon the first by providing concise definitions for each Big Five factor, illustrating typical characteristics of individuals with high and low levels in each factor. Furthermore, it lists specific traits associated with each factor. The instructions regarding the presentation of results remained identical to those in the first prompt.

Prompt 3: Few-Shot

The third prompt introduces the Big Five and its dimensions, similar to the first prompt. However, it additionally presents a table containing 60 classified items. The table comprises 12 items for each dimension, with some exhibiting reverse polarity. Each personality facet, as defined by the NEO-PI-R model (Costa & McCrae, 1992), is represented

by two items. The instructions for presenting the results remain consistent with those in the first prompt.

Data Analysis

We computed the F1-score to assess the performance of each model. This performance metric is calculated from the precision and recall of the models. Precision represents the percentage of items that were correctly classified within a specific category, while recall measures the percentage of items from a particular category that were correctly identified by the model. The F1-score is the harmonic mean of precision and recall, providing an overall measure of performance. These metrics are calculated from the confusion matrix, a contingency table representing the relationship between predictions and actual observations.

Additionally, we evaluated the overall accuracy of each model, defined as the percentage of items correctly classified across all dimensions, which provides a comprehensive assessment of the model's ability to classify items across multiple categories. The overall accuracy of the models was compared using Cochran's Q test, followed by pairwise comparisons using the Wilcoxon signed-rank test. P values were adjusted using the FDR method.

Results

Figure 1 presents the F1-scores and overall accuracy for the models and prompts considered in the analyses for each of the Big Five dimensions. The findings show that Prompt 2, which is based on contextual few-shot learning, outperformed the other prompts in most generative models, with higher F1-scores. Specifically, when comparing Prompt 2 to Prompt 1 (zero-shot), Prompt 2 achieved significantly higher overall accuracy in the local models Llama 3.1 ($p = 0.002$) and Phi-3 ($p = 0.029$). This improvement likely stems from the provision of additional information that expands the search space of internal representations,

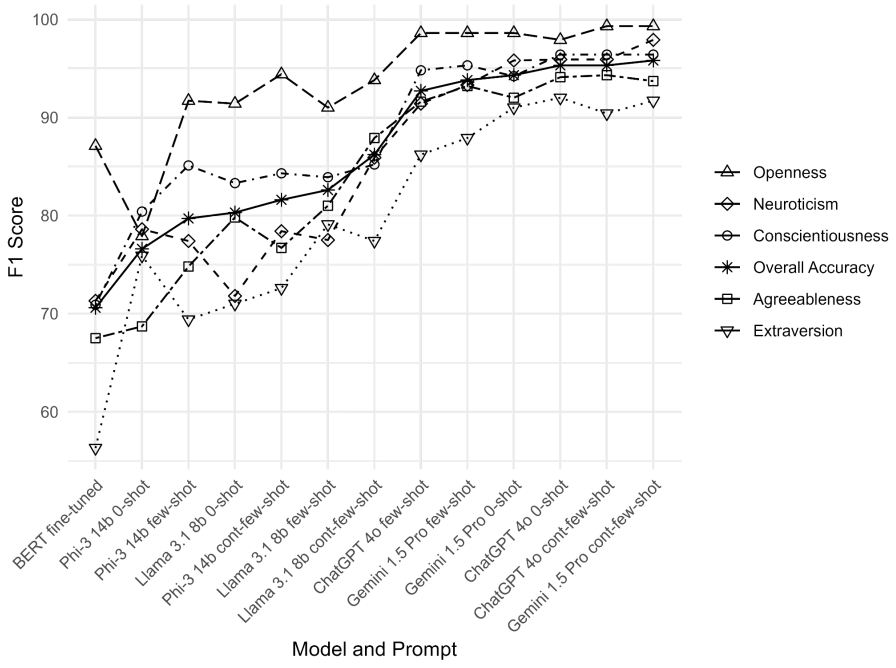
thus overcoming the limitations imposed by the relatively small number of parameters in these models (8 billion and 14 billion, respectively).

In contrast, this significant difference was not observed in larger models such as Gemini 1.5 Pro and ChatGPT-4o. These models, with their substantially higher parameter counts, showed no statistically significant improvement when using Prompt 2 over Prompt 1. This suggests that the benefits of contextual few-shot prompting may be more pronounced for smaller, locally run models compared to their larger counterparts.

In Tables S1 and S2 of the supplementary material for this article, we provide detailed results of each model's performance. Table S1 includes the results for all items, while Table S2 distinguishes the results based on positive and negative items. Figure S1 of the supplementary material presents the confusion matrix for the models and prompts considered in the analyses for each of the Big Five dimensions.

Figure 1

F1-score performance metric for each model and prompt



Note. Higher F1-scores values indicate better performance.

Prompt 3, which included examples of classified items, showed significantly lower overall accuracy compared to the other prompts in the ChatGPT-4o model. This result suggests that the model already had extensive knowledge of the Big Five dimensions, and the examples provided may have introduced bias, causing items with different formats or content to be misclassified. In the Gemini 1.5 Pro model, the differences in overall accuracy between Prompt 3 and the other prompts were not statistically significant, although there was a borderline *p* value when compared to Prompt 2 ($p = 0.053$).

Among the personality factors, Openness items were most accurately classified by all LLMs across the three prompts, with one exception: the Phi-3 model using the zero-shot prompt, in which Conscientiousness achieved a higher F1-score. Extraversion consistently showed the poorest performance for the ChatGPT-4o and Gemini 1.5 Pro models across all prompts. This trend was also observed in BERT, Llama 3.1, and Phi-3 models when using Prompt 2. Notably, the differences in F1-scores across dimensions were more pronounced in models with fewer parameters: BERT, Llama, and Phi-3. This suggests that model size may influence the consistency of performance across different personality factors.

The fine-tuned BERT model produced the poorest classification results across all personality dimensions, despite having undergone fine-tuning with exposure to 801 personality items and their classifications. This outcome may be attributed to BERT being an older and smaller model compared to the others. The overall accuracy of BERT was significantly lower than that of the other models across all three prompts, except for Phi-3 with the zero-shot prompt. These findings suggest that while fine-tuning can enhance a model's capabilities, the baseline architecture and size of the model play crucial roles in determining its overall performance. The results highlight the rapid advancements in language model capabilities, with newer, larger models outperforming older, smaller ones even after task-specific fine-tuning. These findings also underscore the paradigm shift away from the

pre-GPT era, in which fine-tuning language models for specific tasks generally yielded superior results. In contrast, the advent of GPT-3 and subsequent models has demonstrated the efficacy of few-shot techniques, often referred to as in-context learning, which are currently achieving optimal performance.

The ChatGPT-4o and Gemini 1.5 Pro models demonstrated the highest overall accuracy levels, with statistically significant differences when compared to the local models ($p < 0.001$ in all pairwise comparisons). These models also achieved the highest F1-scores. These results can be attributed to their execution on a much more advanced computational infrastructure, their significantly larger number of parameters compared to the locally run models, and the absence of quantization, which maximally preserves model precision.

Table 1 reports the items that had a high incidence of misclassification by the models. It is important to note that, considering the four LLMs used and three different prompts, in addition to the fine-tuned version of the BERT model, each item was classified 13 times. Table 1 presents the items that had three or fewer correct classifications across all models and prompts. In total, 11 out of 385 items had three or fewer correct classifications. The Extraversion domain had the highest number of problematic items (i.e., items with three or fewer correct classifications), with five of these items belonging to the positive pole of the trait. Only two negatively keyed items were frequently misclassified.

Table 1

Big Five items most frequently misclassified by the LLMs

Item	Classification in the instrument	Polarity	Correct Class.	Most Class. Domain by LLMs (Count)
I enjoy life.	Neuroticism	Positive	1	Extraversion (12)

Tell the teacher you didn't understand an explanation so he can explain it again.	Extraversion	Positive	2	Agreeableness (6)
Get along with colleagues.	Extraversion	Positive	1	Agreeableness (12)
Overcome difficulties.	Neuroticism	Positive	3	Conscientiousness (10)
I avoid calling attention.	Agreeableness	Positive	2	Extraversion (9)
I am slow to start in the morning.	Extraversion	Negative	2	Neuroticism (7)
I like to show off.	Agreeableness	Negative	0	Extraversion (13)
Ask questions to the teacher during class.	Extraversion	Positive	2	Openness (11)
Avoid drawing attention.	Agreeableness	Positive	0	Extraversion (13)
Ask for help from teachers when you experience difficulties.	Extraversion	Positive	0	Agreeableness (11)
I insist on doing things my own way.	Extraversion	Positive	3	Conscientiousness & Agreeableness (4)

Note. Correct Class. = Number of model-prompt combinations that classified the item within its theoretical-empirical domain; Most Class. Dom. = Domains most frequently assigned by the models (number of classifications in parentheses). A total of 13 conditions were tested.

Discussion

Can AI-based approaches assist or augment the laborious and costly process of expert judgment in classifying Big Five personality items? Our objective was to test a range of models and methods (small, large, commercial, and freely available for local use, using fine-tuning and few-shot learning approaches) to address this question. The results demonstrate that LLMs can indeed be assigned this task and produce satisfactory outcomes.

Among the locally executed models, Llama 3.1 with the contextual few-shot prompt achieved a statistically higher overall accuracy than Phi-3 (across all three prompts) and BERT. It is important to note that, although Llama 3.1 has fewer parameters (8B) compared to Phi-3 (14B), the latter required a more compact version (4-bit quantization), which likely limited its precision. Differences in the training methods and architectures of the models may also explain the discrepancies in their results.

It is worth noting that the performance of Llama 3.1 with Prompt 2 reached levels that can be considered adequate for judge-based analyses, with F1-scores ranging from 77.4% for Extraversion to 93.8% for Openness, and an overall accuracy of 86.2%. Given the model's smaller size, which enables local use on equipment with relatively common configurations in research labs, this characteristic makes it an attractive option for researchers who prefer not to send their test items to cloud-based model servers (such as those of OpenAI and Google, in the case of this study). However, if data privacy concerns regarding the transmission of items to company servers are not an issue, the larger models demonstrated statistically superior overall accuracy compared to Llama 3.1 (8B) across all prompts ($p < .001$ for all pairwise comparisons).

Additionally, we can observe that generative models (such as ChatGPT-4o and Gemini 1.5 Pro) performed better in the text classification task compared to BERT in a fine-tuning approach. Furthermore, the use of specific prompts (contextual few-shot) improved model accuracy in most cases. The results are similar to those reported by Fyffe et al. (2024), who compared the accuracy of DeBERTa and GPT-3. In that study, GPT-3 achieved an accuracy greater than 70% with only two few-shot examples, whereas DeBERTa required 40 few-shot examples to reach the same level of accuracy. Moreover, our results align with those of Cao and Kosinski (2024), reinforcing the notion that LLMs possess inherent knowledge of the Big Five personality framework. Since our findings also indicate that generative models perform

better in text classification, we recommend the use of generative models alongside few-shot prompts for users aiming to classify items using an NLP approach.

Finally, an examination of the most frequently misclassified items suggests that the use of specific adjectives and verbs in the item wording may have contributed to their misclassification. For instance, the item "I like to show off" was developed to measure a specific facet of Agreeableness, modesty, as a reverse-keyed item; however, it was classified as Extraversion 13 times. In a study involving 140 men and 140 women evaluated by 10 raters, the adjective "show-off" had a strong association with the Extraversion domain, achieving a high factor loading of .68 (John, 1990). Conversely, the item "Get along with colleagues," intended to assess Extraversion, was classified as Agreeableness 12 times. The ability to "get along" with others is often linked to Agreeableness, as higher scores in this trait are associated with prosocial and altruistic behaviors, which are essential for fostering positive interpersonal relationships (de Vries et al., 2020). Another item, "I enjoy life," which theoretically assesses low levels of Neuroticism, was misclassified as Extraversion 12 times. Given that greater life satisfaction and well-being are associated with Extraversion (Kuijpers et al., 2022), it is possible that LLMs recognized expressions of positive emotions toward life as indicative of this trait. These examples highlight the inherent ambiguity of these items, even for human classification, suggesting that the LLMs' classifications may, in fact, be justifiable.

The findings support the use of NLP models as a complementary source of information in the classification of personality items by experts (content validity). We propose that the classifications generated by such models can serve as preliminary data for test developers, aiding expert analyses. This approach would allow for the revision of items that exhibit significant discrepancies between their content and the intended dimension of the test prior to submission to human experts. Alternatively, these model-generated classifications

could be utilized alongside expert judgments, functioning as an additional evaluator in the assessment process.

It is important to emphasize that the classification of items into the five factors by the test developers served as the gold standard against which LLM classifications were evaluated. This classification was based on both theoretical criteria – which point to the association between item content and the Big Five factors – and empirical criteria, mainly derived from factor analyses. However, although these items belong to well-established international tests with robust validity evidence, this specific subset of items consistently misclassified by LLMs may have limitations and could potentially be rewritten to more effectively assess their intended traits in Portuguese. Whether LLM classification performance relates to item-level psychometric properties remains an open question. It is plausible that items with higher misclassification rates display poorer factor-analytic indicators, such as low loadings or high complexity. As the final factor solutions for the reference scales were unavailable (OECD, 2024; Primi, Santos, et al., 2021), we suggest that future research formally examine these potential associations.

This study aimed to evaluate the feasibility of employing NLP models for the classification of personality items based on their content. The findings suggest that certain models achieved high accuracy in classifying personality items according to the Big Five personality framework. Notably, larger generative models, characterized by a higher number of parameters and deployed on commercial servers (e.g., Chat GPT-4o and Gemini 1.5 Pro), outperformed other approaches in terms of classification accuracy. However, smaller models executed locally, such as Llama 3.1, also produced results that support their potential usefulness for this task. The choice between prioritizing classification accuracy, as offered by more sophisticated generative models, and ensuring data privacy through locally operated

models should be carefully considered by researchers developing personality assessment tools.

It is important to acknowledge some limitations that affect the generalizability of the findings of this study. First, the personality items were written in Brazilian Portuguese. While the models utilized were multilingual, their training predominantly focused on English-language materials, resulting in variable performance across different languages. Additionally, for models executed locally on a personal computer, hardware limitations – particularly memory constraints – impacted the selection of model size, both in terms of parameter count and quantization. On more advanced systems, less constrained models could be employed, which might lead to improved outcomes.

Another limitation of the present study involves the replicability of the obtained results. During the period when we performed item classifications using the described methods, the literature on strategies for obtaining deterministic responses from LLMs was incipient (Atil et al., 2024). Our decision to reduce the models' temperature to 0.4 decreased response randomness compared to the models' default values, typically 1.0, without, however, eliminating it completely. We acknowledge that, while the temperature parameter was set to 0.4 rather than 0.0, the high F1, recall, and precision scores indicate consistent classification performance, suggesting that this level of randomness did not substantially affect the models' outputs. Given that temperature modulates the probability distribution over candidate tokens during generation (Atil et al., 2024), a setting of 0.4 may have primarily influenced the classification of instances in which the probability distributions across two or more personality factor classes were relatively similar, allowing for greater variability in token selection in such ambiguous cases.

It is important to consider that more recent studies indicate that even adjusting the temperature to 0.0 is not sufficient to achieve complete consistency in LLM responses, even

when using identical prompts (Klishevich et al., 2025). Current studies also indicate that, in addition to reducing temperature to 0.0, it is necessary to adjust other parameters (e.g., setting the seed value, top-P, and top-L) to obtain deterministic responses when using LLMs (Larsen, 2025). Thus, future studies should follow current recommendations to ensure the replicability of LLM responses in item classification tasks.

Finally, the rapid evolution of NLP models means that the versions used in this study may quickly become outdated. The field is characterized by continuous development and refinement of models. As such, the performance metrics obtained in this study using current model versions may be surpassed by subsequent iterations or novel architectures in the near future.

These limitations underscore the need for cautious interpretation of the results and suggest avenues for future research, including cross-linguistic validation and the application of more advanced computational resources. On the other hand, our results are also promising, indicating that the future of psychological assessment, in combination with the evolution of NLP and AI, holds innovations with strong potential to improve the field and enhance how we measure our phenomena of interest.

References

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214-221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., & Baldwin, B. (2024). *LLM stability: A detailed analysis with some surprises*. arXiv. https://arxiv.org/html/2408.04667v2

Brickman, J., Gupta, M., & Oltmanns, J. R. (2025). Large language models for psychological assessment: A comprehensive overview. *Advances in Methods and Practices in Psychological Science, 8*(3), 1–26. https://doi.org/10.1177/25152459251343582

Cao, X., & Kosinski, M. (2024). *ChatGPT can accurately predict public figures' perceived personalities without any training*. ResearchGate. https://www.researchgate.net/publication/379641669_ChatGPT_Can_Accurately_Predict_Public_Figures'_Perceived_Personalities_Without_Any_Training

Chowdhary, K. R. (2020). *Fundamentals of artificial intelligence*. Springer.

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education, 8*, 858273. https://doi.org/10.3389/feduc.2023.858273

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEP Five Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology, 125*(1), 173–197. https://doi.org/10.1037/pspp0000443

De Raad, B., & Mlacic, B. (2015). Big five factor model, theory and structure. In J. D. Wright. (Ed.), *International encyclopedia of the social & behavioral sciences* (2th ed., pp. 559–566). Elsevier.

de Vries, R. E., Pronk, J., Olthof, T., & Goossens, F. A. (2020). Getting along and/or getting ahead: Differential HEXACO personality correlates of likeability and popularity

among adolescents. *European Journal of Personality*, *34*(2), 245-261. https://doi.org/10.1002/per.224

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, *1*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

Fyffe, S., Lee, P., & Kaplan, S. (2024). "Transforming" personality scale development: Illustrating the potential of state-of-the-art natural language processing. *Organizational Research Methods*, *27*(2), 265-300. https://doi.org/10.1177/10944281231155771

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Sage.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024). *Eliciting personality traits in large language models*. arXiv. https://doi.org/10.48550/arXiv.2402.08341

Hussain, Z., Binz, M., Mata, R., & Wulff, D. U. (2024). A tutorial on open-source large language models for behavioral science. *Behavior Research Methods, 56*, 8214-8237. https://doi.org/10.3758/s13428-024-02455-8

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). Guilford Press.

John, O. P. (2021). History, measurement, and conceptual elaboration of the Big-Five trait taxonomy: The paradigm matures. In O. P. John & R. W. Robins (Eds.), *Handbook of personality: Theory and research* (4th ed., pp. 35–82). Guilford Press.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall.

Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, language models are poised for psychological assessment. *Psychiatry Research, 333*, 115667. https://doi.org/10.1016/j.psychres.2023.115667

Klishevich, E., Denisov-Blanch, Y., Obstbaum, S., Ciobanu, I., & Kosinski, M. (2025). *Measuring determinism in large language models for software code review*. arXiv. https://arxiv.org/abs/2502.20747

Kuijpers, E., Pickett, J., Wille, B., & Hofmans, J. (2022). Do you feel better when you behave more extraverted than you are? The relationship between cumulative counterdispositional extraversion and positive feelings. *Personality and Social Psychology Bulletin*, *48*(4), 606-623. https://doi.org/10.1177/01461672211015062

Larsen, E. (2025). *The instability of safety: How random seeds and temperature expose inconsistent LLM Refusal Behavior*. arXiv. arXiv:2512.12066. a

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382-386. https://journals.lww.com/nursingresearchonline/citation/1986/11000/DeterminationandQuantificatonOfContent.17.aspx

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2402.06196

Myers, D., Mohawesh, R., Chellaboina, V. I., Sathvik, A. L., Venkatesh, P., Ho, Y. H., Henshaw, H., Alhawawreh, M., Berdik, D., & Jararweh, Y. (2024). Foundation and large language models: Fundamentals, challenges, opportunities, and social impacts. *Cluster Computing, 27*, 1-26. https://doi.org/10.1007/s10586-023-04203-7

Nunes, C. H. S. S., Spenassato, D., Bornia, A. C., & Primi, R. (2015). Testes adaptativos computadorizados - CAT [Computerized adaptive tests - CAT]. In M. C. R. Silva, D. Bartholomeu, C. M. M. Vendramini, & J. M. Montiel (Eds.), *Aplicações de métodos estatísticos avançados à avaliação psicológica e educacional* [Applications of advanced statistical methods to psychological and educational assessment] (pp. 37-76). Vetor Editora.

OpenAI. (2023). *ChatGPT*. https://www.openai.com/chatgpt

Organization for Economic Co-operation for Better Lives. (2024). *OECD survey on social and emotional skills*. https://www.oecd.org/en/about/programmes/oecd-survey-on-social-and-emotional-skills.html

Pichai, S., & Hassabis, D. (2024). *Our next-generation model: Gemini 1.5*. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#sundar-note

Peres, A. J. S. (2021). Processamento da linguagem natural: Modelagem de tópicos [Natural language processing: Topic modeling]. In C. Faiad, M. N. Baptista, & R. Primi

(Orgs.), *Tutoriais em análise de dados aplicados a psicometria* [Tutorials in data analysis applied to psychometrics] (pp. 436–459). Vozes.

Pires, J. G., Nunes, C. H. S. S., Nunes, M. F. O., & Primi, R. (2023). Preliminary validity for the Big Five Inventory-2 in Brazilian adults. *Psico-USF*, *28*(1), 91-102. https://doi.org/10.1590/1413-82712023280108

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Primi, R. (2021). Uso do word-to-vec (word embeddings) para análise de textos [Use of word-to-vec (word embeddings) for text analysis]. In C. Faiad, M. N. Baptista, & R. Primi (Orgs.), *Tutoriais em análise de dados aplicados a psicometria* [Tutorials in data analysis applied to psychometrics] (pp. 460–476). Vozes.

Primi, R., Santos, D., John, O. P., & De Fruyt, F. (2021). SENNA inventory for the assessment of social and emotional skills in public school students in Brazil: Measuring both identity and self-efficacy. *Frontiers in Psychology*, *12*, 716639. https://doi.org/10.3389/fpsyg.2021.716639

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, *38*, 385-410. https://doi.org/10.1007/s10869-022-09824-0

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024). *Generative psychometrics via AI-GENIE: Automatic item generation and validation via network-integrated evaluation*. PsyArxiv. https://osf.io/preprints/psyarxiv/fgbj4

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2024). Using sequences of life-events to predict human

lives. *Nature Computational Science, 4*, 43-56. https://doi.org/10.1038/s43588-023-00573-5

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models: A new paradigm for artificial intelligence. *Business & Information Systems Engineering, 66*, 221-231. https://doi.org/10.1007/s12599-024-00851-0

Tang, H., Zhang, X., Liu, K., Zhu, J., & Kang, Z. (2022*). MKQ-BERT: Quantized bert with 4-bits weights and activations*. arXiv. https://arxiv.org/pdf/2203.13483

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *NeurIPS Proceedings*, 1-11. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*(5), 508-518. https://doi.org/10.1177/019394590325299

Research Data Availability

The supplementary material accompanying this article includes the complete set of prompts utilized and a comprehensive analysis of model performance. Due to licensing restrictions, access to the specific item content must be requested directly from the respective intellectual property owners: the Organisation for Economic Co-operation and Development (OECD) and the Instituto Ayrton Senna. All supplementary data are available at https://drive.google.com/file/d/1eDXLam0t9hTdKJq5MAsq_ClelXwmmP5s/view?usp=share_link

Conflict of interest

The authors have no conflicts of interest to declare.

AI Use Disclosure

The authors declare that no artificial intelligence tools were used in the writing or editing of this manuscript.

Carlos Henrique Sancineto da Silva Nunes is a Professor of the Universidade Federal de Santa Catarina, Florianópolis-SP, Brazil.

Alexandre José de Souza Peres is a Professor of the Universidade Federal de Mato Grosso do Sul (UFMS), Campo Grande-MS, Brazil.

Leonardo de Barros Mose is a Postdoctoral researcher of the Universidade São Francisco (USF), Campinas-SP, Brazil.

Petar Čolović is a Professor of the University of Novi Sad, Novi Sad, Serbia.

Ricardo Primi is a Professor of the Universidade São Francisco (USF), Campinas-SP, Brazil.

Vithor Rosa Franco is a Professor of the Universidade São Francisco (USF), Campinas-SP, Brazil.

Authors' Contribution:

All authors made substantial contributions to the conception and design of this study, to data analysis and interpretation, and to the manuscript revision and approval of the final version. All the authors assume public responsibility for the content of the manuscript.

Associate editor:

Alexsandro Luiz de Andrade

Received: Nov. 11, 2024

1st Revision: Oct. 6, 2025

Approved: Mar. 9, 2026

How to cite this article:

Nunes, C. H. S. S., Peres, A. J. S., Mose, L. B., Čolović, P., Primi, R., & Franco, V. R. (2026). AI-Driven Classification of Personality Items: A Study on Large Language Models. *Paidéia (Ribeirão Preto), 36,* e3609. https://doi.org/10.1590/1982-4327e3609

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a Creative Commons CC-BY license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.