OPEN ACCESS

ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA
PUCRS

PSICO
Psico, Porto Alegre, v. 56, n. 1, p. 1-13, jan.-dez. 2025
e-ISSN: 1980-8623 | ISSN-L: 0103-5371

http://dx.doi.org/10.15448/1980-8623.2025.1.47225

ARTIGOS

Validade de conteúdo de um inventário de personalidade: psicometria assistida por LLMs

Content validity evidence for a personality inventory: LLMs-assisted psychometrics
Validez basada en el contenido de un inventario de personalidad: psicometría asistida por LLMs

José Maurício Haas Bueno[1]
orcid.org/0000-0002-9179-7216
mauricio.bueno@ufpe.br

Ricardo Primi[2]
orcid.org/0000-0003-4227-6745
ricardo.primi@usf.edu.br

Emanuel Duarte de Almeida Cordeiro[3]
orcid.org/0000-0001-6437-3197
emanuel.cordeiro@uesb.edu.br

Ana Deyvis Santos Araújo Jesuíno[4]
orcid.org/0000-0002-7031-7682
ana.deyvis@ufma.br

Monalisa Muniz[5]
orcid.org/0000-0003-1628-6296
monamuniz@ufscar.br

Ana Paula Porto Noronha[2]
orcid.org/0000-0001-6821-0299
ana.noronha@usf.edu.br

Recebido em: 10 dez. 2024.
Aprovado em: 23 out. 2025.
Publicado em: 19 dez. 2025.

Resumo: Os Large Language Models (LLMs) representam um avanço significativo no Processamento de Linguagem Natural (PLN). Este estudo investiga a utilização desses modelos na obtenção de evidências de validade baseadas no conteúdo de um novo instrumento de avaliação dos cinco grandes fatores de personalidade. Os itens do novo instrumento foram criados pelo ChatGPT e analisados semanticamente pelo Gemini, ao lado dos itens do BFI2 (criados por humanos). A análise empregou classificação dos itens via prompt (simulando um juiz especialista) e análise fatorial exploratória dos embeddings dos itens (obtidos via API), propondo uma nova abordagem à psicometria. Os resultados mostraram convergência semântica para neuroticismo, amabilidade, abertura e conscienciosidade, mas maior dispersão nos itens de extroversão. Observou-se também convergência semântica entre itens criados pelo LLMs e por humanos (validade convergente de conteúdo). Conclui-se que os LLMs apresentam bom potencial para contribuir no processo de obtenção de evidências de validade de conteúdo.

Palavras-chave: inteligência artificial; avaliação psicológica; psicometria.

Abstract: Large Language Models (LLMs) represent a significant advancement in Natural Language Processing (NLP). This study investigates the use of these models in gathering content-based validity evidence for a new instrument assessing the Big Five personality factors. Items for the new instrument were created by ChatGPT and semantically analyzed by Gemini, alongside items from the BFI-2 (human-created). The analysis employed item classification via prompt (simulating an expert judge) and Exploratory Factor Analysis of item embeddings (obtained via API), proposing a novel approach to psychometrics. Results showed semantic convergence for neuroticism, agreeableness, openness, and conscientiousness, but greater dispersion for extraversion items. Semantic convergence was also observed between LLM-generated and human-created items (content-convergent validity). It is concluded that LLMs show significant potential to contribute to the process of gathering content-based validity evidence.

Keywords: artificial intelligence; psychological assessment; psychometrics.

Resumen: Los Large Language Models (LLMs) representan un avance en el Procesamiento del Lenguaje Natural (PLN). Este estudio investiga la utilización de MLGEs en la obtención de evidencias de validez basadas en el contenido en la evaluación de los cinco grandes factores. Los ítems del nuevo instrumento fueron creados por ChatGPT y analizados semánticamente por Gemini, junto a los ítems del BFI2 (creados por humanos). El análisis empleó la clasificación de los ítems mediante prompt (juez experto) y el análisis factorial exploratorio de los *embeddings* (API), un nuevo enfoque psicométrico. Los resultados mostraron

[1] Universidade Federal de Pernambuco (UFPE), Recife, Pernambuco, Brasil.
[2] Universidade São Francisco (USF), Campinas, São Paulo, Brasil.
[3] Universidade Estadual do Sudoeste da Bahia (UESB), Vitória da Conquista, Bahia, Brasil.
[4] Universidade Federal do Maranhão (UFMA), São Luís, Maranhão, Brasil.
[5] Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo, Brasil.

convergencia semántica para neuroticismo, amabilidad, apertura y consciencia, pero una mayor dispersión en los ítems de extraversión. Se observó también convergencia semántica entre los ítems creados por el LLMs y por humanos (validez convergente de contenido). Se concluye que los LLMs presentan un buen potencial para contribuir en el proceso de obtención de evidencias de validez de contenido.

Palabras clave: inteligencia artificial; evaluación psicológica; psicometría.

Introduction

Content validity is defined as the degree to which the items, tasks, or questions of a test represent all central elements of the construct intended to be measured (American Educational Research Association et al., 2014). Some authors restrict this process to item analysis conducted by an expert committee (Dempsey & Dempsey, 2000; Fitzner, 2007), while others broaden the concept to include the creation and development of items (Pasquali, 2010).

Both the creation and evaluation processes of items face limitations due to their subjective nature (Alexandre & Coluci, 2011). Additionally, there is often a challenge in finding experts to assess items for a new construct, whether due to the lack of specialists in the field, limited availability, or language barriers, especially when experts are speakers of a language different from the one in which the instrument is being developed.

Emerging information technologies offer promising avenues for addressing long-standing challenges in content validation. This study investigated the potential of Large Language Models (LLMs), specifically OpenAI's ChatGPT and Google's Gemini, for supporting content-validity procedures in personality assessment. ChatGPT was used to generate test items, emulating the role of the instrument developer, whereas Gemini evaluated item representativeness for the Big Five constructs, functioning analogously to an expert judge. This division of labor reflected the models' relative strengths at the time of data collection, with ChatGPT excelling in generative tasks and Gemini demonstrating strong semantic-analysis capabilities.

Large Language Models (LLMs) constitute a major advancement in Natural Language Processing (NLP), a field dedicated to enabling computers to understand and generate human language (Demszky et al., 2023). Trained on extensive text corpora to predict subsequent words, LLMs learn contextual and nuanced meanings arising from word interactions (Debelak et al., 2024; Demszky et al., 2023; Pellert et al., 2024)such as social media posts, to infer psychological characteristics, as well as survey and interview analysis. In this tutorial paper, we demonstrate the use of the Python-based natural language processing software package transformers (and related modules from the Hugging Face Ecosystem. This capability stems from the Transformer architecture, which uses a self-attention mechanism to assess the relevance of each word relative to all others in a sequence, allowing the model to capture long-range dependencies and complex sentence structures while processing information in parallel (Debelak et al., 2024; Pellert et al., 2024; Vaswani et al., 2017)such as social media posts, to infer psychological characteristics, as well as survey and interview analysis. In this tutorial paper, we demonstrate the use of the Python-based natural language processing software package transformers (and related modules from the Hugging Face Ecosystem.

LLMs encode semantic information in the form of embeddings: numerical vector representations in multidimensional space, where the proximity of vectors reflects semantic similarity (e.g., "dog" near "cat" or "teacher" near "educator")(Debelak et al., 2024; Pellert et al., 2024)such as social media posts, to infer psychological characteristics, as well as survey and interview analysis. In this tutorial paper, we demonstrate the use of the Python-based natural language processing software package transformers (and related modules from the Hugging Face Ecosystem. These embeddings can be obtained via APIs offered by systems such as ChatGPT and Gemini, enabling users to input text and receive the corresponding vector representation. Converting language into numerical vectors allows the application of statistical methods to evaluate semantic similarity between

words, sentences, or test items (OpenAI, 2023).

This approach has been applied across diverse areas of psychological research, including sentiment analysis (W. Zhang et al., 2023), social psychology (J. Zhang et al., 2023), and mental-health support tools in clinical settings (J. Hu et al., 2024). In psychometrics, LLMs have been explored for assessing psychological attributes of the models themselves (Pellert et al., 2024), developing new assessment paradigms (Kjell et al., 2024), automatically generating test items (Attali et al., 2022), and detecting personality traits in social-media data (L. Hu et al., 2024). Given these technological advances, incorporating LLMs into content validity procedures represents a promising opportunity. Their capacity for text generation and semantic analysis supports both item development and evaluation of theoretical relevance, whether through prompting strategies or embedding-based approaches via APIs.The choice to develop an instrument based on the Big Five model derives from its linguistic foundation in the lexical hypothesis, which posits that socially relevant individual differences become encoded in language over time. Human characteristics are thus expressed through words and phrases, facilitating communication and shared understanding of behavior (Goldberg, 1990).

The choice to develop an instrument based on the Big Five model derives from its linguistic foundation in the lexical hypothesis, which posits that socially relevant individual differences become encoded in language over time. Human characteristics are thus expressed through words and phrases, facilitating communication and shared understanding of behavior (Goldberg, 1990).

Extensive research using dictionary analyses and adjective inventories has consistently identified five broad and universal personality dimensions (extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience) replicated across languages and cultures (McCrae & Costa, 1997; Pires et al., 2023). The lexical hypothesis therefore not only underpins the Big Five framework but also highlights the intrinsic link between natural language and Personality Psychology, making it particularly relevant for studies employing LLMs, which operate through text analysis and processing.

Among self-report instruments assessing the Big Five, the NEO-PI-R (McCrae & Costa, 1997) and the BFI-2 stand out, both operationalizing personality through facets. In BFI-2, each factor is represented by three facets (Soto & John, 2017). Extraversion comprises sociability (desire for social interaction), assertiveness (expressing opinions and pursuing goals in social contexts), and energy level (positive affect and enthusiasm). Agreeableness includes compassion (concern for others), respectfulness (consideration for others' rights and preferences), and trust (positive expectations about others). Conscientiousness encompasses organization (structure and order), productiveness (work ethic and persistence), and responsibility (dependability and obligation fulfillment). Neuroticism is defined by anxiety (tendency toward fear), depression (sadness and low energy), and emotional volatility (mood instability). Openness involves intellectual curiosity (interest in ideas), aesthetic sensitivity (appreciation of art and beauty), and creative imagination (originality and creativity) (Soto & John, 2017).

Research demonstrates broad convergence among facets within factors but also meaningful cross-loadings, reflecting trait complexity. For instance, neuroticism facets have loaded on agreeableness and conscientiousness, and extraversion facets on neuroticism, openness, and agreeableness; assertiveness has shown cross-loadings on several dimensions (McCrae & Costa, 1997; Soto & John, 2017). These findings suggest that, despite the robustness of the Big Five, facets may share semantic and functional overlap across traits.

Given the empirical strength of the Big Five and its grounding in the lexical hypothesis, applying Large Language Models (LLMs) to examine content-based validity in this framework is compelling. Accordingly, this study developed a Big Five inventory (BF_BR) and sought content-validity evidence using LLMs. ChatGPT was employed for item generation and Gemini for content evalua-

tion, representing, to our knowledge, the first application of this methodology in psychometrics.

Method

The search for content-based validity evidence was conducted using two approaches: one involving the use of Gemini's prompt to request item classification into facets and factors, and the other employing the Gemini API to obtain and analyze the embeddings of the items.

Instruments

Personality Inventory Based on the Big Five Factors (BF_BR)

All items in this instrument were generated by ChatGPT, version 3.5 (OpenAI, 2023), under the following system-prompt instructions: "Consider that the trait extraversion is composed of the facets sociability, assertiveness, and energy level. Write 8 statements for each facet (24 in total) that can be used as test items. The statements should be simple, concise, and understandable for individuals with low education levels. Do not use the names of the facets in the statements (sociability, assertiveness, and energy level)". Items for the other traits and their facets were created in the same way, replacing trait and facet names accordingly. A zero-shot prompting strategy was adopted because it aligns with a purely generative task, allowing the model to describe behaviors typical of each facet with minimal instructions and without excessive technical elaboration. As a result, the BF_BR comprises 120 self-report items in Brazilian Portuguese assessing the Big Five personality traits (agreeableness (A), conscientiousness (C), extraversion (E), neuroticism (N), and openness (O)) with three facets per trait and eight items per facet, mirroring the BFI-2 facet structure (Soto & John, 2017).

Unlike the BFI-2, which includes items representing both poles of each trait, the BF_BR contains only positively keyed statements. Conceptually, agreeableness reflects compassion, trust, and respectfulness toward others, expressed through items such as "I often care about other people's feelings," "I generally believe in people's good intentions," and "I believe everyone deserves to be treated with dignity." Conscientiousness encompasses organization, productiveness, and responsibility, represented by items like "I like to keep things tidy and orderly," "I enjoy staying busy and getting things done," and "I fulfill my commitments and promises." Extraversion represents assertiveness, sociability, and energy level, reflected in items such as "I am not afraid to express my opinion," "I enjoy being in the company of others," and "I have plenty of energy throughout the day." Neuroticism captures tendencies toward anxiety, depression, and emotional volatility, exemplified by the statements "Sometimes I worry a lot about things that might go wrong," "I often feel sad and down," and "My emotions can change rapidly." Finally, openness reflects intellectual curiosity, creative imagination, and aesthetic sensitivity, illustrated by "I like learning new things whenever I can," "I enjoy inventing new ways of doing things," and "I often appreciate beauty in art and nature." This strategy ensured conceptual fidelity to Big Five theory while maintaining linguistically simple and behavior-focused phrasing suitable for a broad range of educational backgrounds.

Brazilian Portuguese Version of the Big-Five Inventory Version 2 (BFI-2)

This version comprises 60 items, evenly distributed across the same 15 facets. However, for each facet, there were two items in the direction of the trait and two items in the opposite direction (e.g., I am compassionate, have a soft heart, and care about others, enjoying helping them for the compassion facet of agreeableness trait, and I don't care much about others, and I can be indifferent, cold, and distant for the opposite direction) (Pires et al., 2023; Soto & John, 2017).

Procedures

This study did not require ethics approval because no human data were collected. The procedures comprised item generation by ChatGPT, item classification by Gemini, extraction of

embeddings via Gemini's API, and subsequent processing and analysis of those embeddings.

Item Classification via Gemini Prompt

Items generated by ChatGPT were evaluated by Gemini 1.5 Pro (Google, 2024a), which classified each item according to the Big Five factors and facets. The following instruction was provided verbatim to the model: "You will act as an expert judge in personality theories, specifically regarding the Big Five model, one of the trait theories. Consider that the Big Five factors are: extraversion, agreeableness, conscientiousness, neuroticism, and openness. These traits are subdivided into three facets each: extraversion: sociability, assertiveness, and energy level; agreeableness: compassion, respectfulness, and trust; conscientiousness: organization, productiveness, and responsibility; neuroticism: anxiety, depression, and emotional volatility; and openness: intellectual curiosity, aesthetic sensitivity, and creative imagination. Make two classifications for each item: one of the five factors and one of the three facets of the classified factor. Present the results in a table with three columns: items, factor classification, facet classification." The resulting classifications were exported to R for subsequent analysis.

Embeddings were then obtained using Gemini's API with the *httr* (Wickham, 2023) and *jsonlite* (Ooms, 2014) R packages. Using the *models/text-embedding-004* model, 768-dimensional vectors were generated for each item (120 BF_BR items and 60 BFI-2 items), yielding a 180 × 768 embedding matrix for analysis.

Unlike the item-generation phase performed by ChatGPT, item classification required specialized analytical judgment. Accordingly, a role-prompting strategy was used, specifying that Gemini should act as an expert judge to emulate the function of a human specialist while leveraging the model's semantic capabilities.

Embeddings were then computed for the full text of each item using R (R Core Team, 2023) to interface with the Google Gemini API via the *httr* (Wickham, 2023) and *jsonlite* (Ooms, 2014) packages. The *models/text-embedding-004* model (Google, 2024) was used, and because the API accepts one input per request, a loop iterated over all 180 items. Each POST call to the endpoint *...models/text-embedding-004:embedContent* returned a 768-dimension embedding vector, which was extracted and concatenated into a final 180 × 768 matrix. The full R script is available from the authors upon request.

Data Processing and Analysis: The data were processed and analyzed using R (R Core Team, 2023) packages ltm (Rizopoulos, 2006) and psych (Revelle, 2007). The analyses conducted will be described below.

Data Analysis

Concordance between Gemini's classifications and the theoretical classifications was evaluated using standard performance metrics for classifier assessment (Sokolova & Lapalme, 2009). Confusion matrices were generated, and accuracy, precision (the proportion of correctly identified positive cases, sensitive to false positives), recall (the proportion of true positive cases correctly identified, sensitive to false negatives), and F1-score (the harmonic mean of precision and recall) were computed. These analyses were performed at both the factor and facet levels using the *caret* package in R (Kuhn, 2008).

The analysis based on embeddings was *procedurally* more complex. After obtaining the embeddings from both instruments, the first step was to identify those most likely to contain relevant information about any of the Big Five factors. This procedure aimed to pinpoint "expert embeddings", akin to identifying expert judges in traditional content validity studies.

Dummy variable columns were created, assigning a value of 1 to indicate the factor of interest (e.g., extraversion) and 0 for the other factors, applying the same method to all factors. Point biserial correlations were then computed between the embeddings and the dummy variables for each factor. Finally, embeddings with correlations that deviated by at least two standard deviations from the mean, either above or below, in at least one factor, were selected. Using this method,

140 columns were identified: “V1”, “V6”, “V7”, “V8”, “V19”, “V22”, “V27”, “V34”, “V35”, “V47”, “V49”, “V58”, “V67”, “V70”, “V72”, “V73”, “V76”, “V82”, “V97”, “V99”, “V101”, “V104”, “V107”, “V112”, “V119”, “V121”, “V122”, “V125”, “V128”, “V130”, “V131”, “V134”, “V135”, “V137”, “V142”, “V146”, “V147”, “V157”, “V158”, “V162”, “V174”, “V179”, “V187”, “V191”, “V193”, “V198”, “V204”, “V212”, “V227”, “V228”, “V230”, “V247”, “V253”, “V254”, “V256”, “V261”, “V262”, “V263”, “V267”, “V269”, “V286”, “V289”, “V300”, “V306”, “V307”, “V310”, “V318”, “V321”, “V325”, “V326”, “V331”, “V335”, “V343”, “V344”, “V354”, “V356”, “V357”, “V371”, “V397”, “V399”, “V402”, “V405”, “V408”, “V414”, “V423”, “V424”, “V426”, “V437”, “V440”, “V444”, “V446”, “V449”, “V453”, “V461”, “V475”, “V476”, “V483”, “V495”, “V501”, “V504”, “V509”, “V512”, “V518”, “V519”, “V525”, “V526”, “V534”, “V535”, “V561”, “V569”, “V575”, “V580”, “V582”, “V590”, “V591”, “V617”, “V618”, “V633”, “V637”, “V661”, “V670”, “V675”, “V676”, “V677”, “V681”, “V683”, “V690”, “V691”, “V693”, “V702”, “V705”, “V711”, “V714”, “V720”, “V732”, “V737”, “V739”, “V748”, “V751”, “V768”.

The decision to analyze a subset of embeddings rather than the full embedding matrix was driven by methodological and statistical considerations. Methodologically, the goal was to isolate semantic signal relevant to the Big Five constructs from linguistic noise, using exploratory factor analysis (EFA) to support this selection. Statistically, it is important to distinguish the present EFA from traditional applications aimed at establishing internal-structure validity. Here, the analysis targeted content validity within a semantic (not behavioral) space. Thus, selecting “extreme” embeddings based on point-biserial correlations with dummy variables (M ± 2SD) and discarding dimensions with null associations parallels identifying expert versus non-expert judges in conventional content-validity procedures. Although such filtering would be inappropriate in respondent-based EFAs, it is a purposeful and necessary step when evaluating semantic representations.

While this dummy-variable–based filtering introduces potential circularity, falsifiability is preserved, as no factor structure was imposed a priori and embedding dimensions could meaningfully associate with multiple theoretically distinct traits (e.g., agreeableness and neuroticism). Accordingly, the subsequent EFA should be interpreted not as uncovering latent structure but as clustering items within a theoretically informed semantic space. The partial circularity and risk of overfitting are deliberate tradeoffs to isolate psychometric signal from linguistic noise.

The 140 selected embeddings were arranged in rows and the 180 items in columns and analyzed via EFA using the *psych* package, extracting five factors (as indicated by parallel analysis) with minimum residuals (MinRes) estimation and Promax rotation, allowing factor correlations. Loadings ≥ .40 were considered salient. Additionally, factor congruence coefficients were computed to compare the five empirical factors with 15 dummycoded theoretical facets. Because this comparison contrasted semantic embeddings with theoretical targets, rather than two empirical structures, a .30 threshold (analogous to minimum acceptable loading criteria) was adopted as an indicator of meaningful abovechance semantic correspondence (Lorenzo-Seva & Ten Berge, 2006).Results

Results

The first step of this study consisted of conducting the content validity analysis using the Gemini prompt. The classification diagnostic metrics (Precision, Recall, and F1-Score) for the five factors and fifteen facets are presented in Table 1.

Table 1 - *Classifications*

Level	Class (Factor/Facet)	Precision	Recall	F1-Score
Factor	Agreeableness (A)	1.000	0.958	0.979
Facet	Compassion	0.889	1.000	0.941

Level	Class (Factor/Facet)	Precision	Recall	F1-Score
Facet	Trust	1.000	1.000	1.000
Facet	Respectfulness	1.000	0.750	0.857
Factor	Conscientiousness (C)	0.960	1.000	0.980
Facet	Organization	0.800	1.000	0.889
Facet	Productiveness	0.833	0.625	0.714
Facet	Responsibility	0.667	0.750	0.706
Factor	Extraversion (E)	1.000	0.958	0.979
Facet	Assertiveness	1.000	0.875	0.933
Facet	Energy Level	0.875	0.875	0.875
Facet	Sociability[1]	0.875	0.875	0.875
Factor	Neuroticism (N)	0.960	1.000	0.980
Facet	Anxiety	0.889	1.000	0.941
Facet	Depression	1.000	1.000	1.000
Facet	Emotional Volatility	1.000	1.000	1.000
Factor	Openness (O)	1.000	1.000	1.000
Facet	Intellectual Curiosity[2]	0.889	1.000	0.941
Facet	Creative Imagination[3]	1.000	0.875	0.933
Facet	Aesthetic Sensitivity[4]	1.000	1.000	1.000

At the factor level, Precision coefficients ranged from 0.960 to 1.000, whereas Recall coefficients ranged from 0.958 to 1.000. F1-Scores were uniformly high (≥ 0.979). The analysis demonstrated near-perfect agreement (Overall Accuracy = 98.3%). Two items migrated across factors: "I always try to be fair and impartial in my judgments," theoretically expected to load on Agreeableness but classified under Conscientiousness; and "Sometimes I feel anxious if I do not have something to do," expected to load on Extraversion but classified under Neuroticism.

At the facet level, Precision coefficients ranged from 0.667 to 1.000, whereas Recall coefficients ranged from 0.625 to 1.000. F1-Scores showed greater variability (≥ 0.706), with the lowest values observed for Responsibility (0.706) and Productiveness (0.714). The analysis revealed high agreement (Overall Accuracy = 90.8%). Eleven cross-facet migrations occurred: "I feel energized when I am in a social group" migrated from Sociability to Energy Level (both Extraversion); "I rarely hesitate to start a conversation or interaction with strangers" migrated from Assertiveness to Sociability (both Extraversion); "Sometimes I feel anxious if I do not have something to do" migrated from Energy Level (Extraversion) to Anxiety (Neuroticism); "I believe we should treat people with kindness" migrated from Respectfulness to Compassion (both Agreeableness); "I always try to be fair and impartial in my judgments" migrated from Respectfulness (Agreeableness) to Responsibility (Conscientiousness); "I make a consistent effort to complete my tasks on time" migrated from Productiveness to Responsibility (both Conscientiousness); "I usually finish what I start" migrated from Productiveness (Conscientiousness) to Responsibility (both Conscientiousness); "I plan my schedule to make the most of my day, avoiding idle time" migrated from Productiveness to Organization (both Conscientiousness); "I fulfill my commitments and promises" migrated from Responsibility to Productiveness (both Conscientiousness); "I keep my personal belongings organized" migrated from Responsibility to Organization (both Conscientiousness); and "I enjoy experimenting with new ideas and approaches" migrated from Creative Imagination

to Intellectual Curiosity (both Openness).

The second step of this study was to analyze item content through embeddings obtained via the Gemini API. This analysis was conducted through an exploratory factor analysis applied to a matrix consisting of 140 rows (embeddings) and 180 columns (BF_BR and BFI2 items). The factor loadings from this analysis are summarized (average item loadings for each facet, by instrument) in Table 2.

Table 2 - *Factor loadings averages*

Test	Factor	Facet	F1	F2	F3	F4	F5
BF_BR	A	Compassion	0,10	**0,67**	0,01	0,01	0,08
BF_BR	A	Trust	0,01	**0,80**	0,09	0,00	-0,06
BF_BR	A	Respectfulness	-0,01	**0,73**	0,06	0,03	0,03
BFI2	A	Compassion	0,33	**0,55**	-0,02	-0,08	0,05
BFI2	A	Trust	0,18	**0,73**	-0,07	0,01	-0,02
BFI2	A	Respectfulness	0,05	**0,65**	0,00	0,05	0,11
BF_BR	C	Organization	0,07	-0,03	0,14	**0,66**	0,01
BF_BR	C	Productiveness	0,04	0,04	0,07	**0,64**	0,11
BF_BR	C	Responsibility	0,11	0,34	-0,01	**0,48**	-0,02
BFI2	C	Organization	0,04	0,07	0,03	**0,68**	0,03
BFI2	C	Productiveness	0,15	0,19	-0,03	**0,58**	-0,04
BFI2	C	Responsibility	0,31	0,22	-0,10	**0,43**	0,03
BF_BR	E	Assertiveness	0,18	0,27	0,03	0,07	**0,35**
BF_BR	E	Energy level	**0,50**	-0,09	0,05	0,21	0,23
BF_BR	E	Sociability	0,12	0,20	0,13	-0,06	**0,52**
BFI2	E	Assertiveness	0,16	**0,44**	0,03	0,05	0,22
BFI2	E	Energy level	**0,46**	-0,09	0,03	0,17	0,27
BFI2	E	Sociability	**0,33**	0,16	-0,02	0,10	0,26
BF_BR	N	Anxiety	**0,78**	0,13	0,05	0,05	-0,12
BF_BR	N	Depression	**0,84**	0,02	0,06	-0,02	-0,08
BF_BR	N	Emotional volatility	**0,81**	-0,03	0,02	0,00	0,07
BFI2	N	Anxiety	**0,64**	0,14	-0,08	0,20	-0,01
BFI2	N	Depression	**0,57**	0,18	-0,06	0,08	0,08
BFI2	N	Emotional volatility	**0,74**	-0,02	-0,04	0,09	0,10
BF_BR	O	Intellectual curiosity	0,05	0,13	**0,67**	-0,03	0,03
BF_BR	O	Creative imagination	0,10	-0,02	**0,80**	0,01	-0,05
BF_BR	O	Aesthetic sensitivity	-0,11	0,15	**0,53**	0,21	0,10
BFI2	O	Intellectual curiosity	0,14	0,10	**0,57**	-0,06	0,10
BFI2	O	Creative imagination	0,22	-0,16	**0,72**	0,09	-0,04
BFI2	O	Aesthetic sensitivity	0,06	0,15	**0,47**	0,00	0,12

Due to the extensive nature of presenting factor loadings for all items, Table 2 displays only the average factor loadings, by facet and test. For instance, the first row displays the average factor loadings of the compassion/agreeableness facet/factor items of BF_BR, while the fourth row

shows the loadings of the same facet/factor of BFI2.

The five extracted factors explained 68% of the total variance, with F1 accounting for 21%, F2 for 17%, F3 for 14%, F4 for 12%, and F5 for 4%. It was observed that the neuroticism facets showed the highest loadings on F1, agreeableness facets on F2, openness facets on F3, and conscientiousness facets on F4.

The extraversion facets exhibited a more complex and diffuse distribution. The highest loadings on Factor 5 corresponded to the extraversion facets from both instruments. However, despite representing the strongest loadings on F5, only the sociability facet from the BF-BR (Extraversion) exceeded the 0.40 threshold. Additionally, although the assertiveness facet from the BF-BR yielded a loading of 0.35 (below 0.40), it was nonetheless the highest loading for that facet relative to the other factors. The three extraversion facets from the BFI-2 showed their highest loadings on factors associated with neuroticism (energy level and sociability) and agreeableness (assertiveness).

Additionally, factor congruence coefficients were calculated between the obtained loadings and the dummy variables for the Big Five factors. The results are shown in Table 3.

Table 3 - *Factor Congruence Indices*

	F1	F2	F3	F4	F5
Compassion	0,11	**0,43**	0,00	-0,02	0,09
Trust	0,04	**0,53**	0,03	0,00	-0,06
Respectfulness	0,00	**0,48**	0,03	0,03	0,07
Organization	0,04	0,00	0,09	**0,56**	0,02
Productiveness	0,05	0,06	0,03	**0,52**	0,07
Responsibility	0,10	0,20	-0,03	**0,39**	0,00
Assertiveness	0,10	0,22	0,03	0,05	**0,38**
Energy level	**0,30**	-0,06	0,03	0,16	**0,30**
Sociability	0,12	0,13	0,07	-0,01	**0,54**
Anxiety	**0,44**	0,09	0,01	0,08	-0,11
Depression	**0,45**	0,05	0,02	0,01	-0,03
Emotional volatility	**0,48**	-0,02	0,00	0,03	0,10
Intellectual curiosity	0,05	0,08	**0,52**	-0,03	0,06
Creative imagination	0,09	-0,04	**0,63**	0,03	-0,06
Aesthetic sensitivity	-0,03	0,10	**0,42**	0,11	0,13

Considering correlations above 0.30, the neuroticism, agreeableness, openness, conscientiousness, and extraversion facets showed the highest congruence with factors F1, F2, F3, F4, and F5, respectively. Furthermore, the energy level (E) facet also demonstrated congruence with F1.

Discussion

The prompt-based classification analysis provided strong evidence of content validity, with 99.2% of items correctly classified at the factor level. Although several facets achieved perfect F1-scores (1.00), diagnostic challenges emerged mainly within Conscientiousness, particularly for Productivity (Recall = 0.625) and Responsibility (Precision = 0.667), whose overlap yielded the lowest F1 values (0.714 and 0.706). The eleven unexpected classifications reflect not flaws in the models but ambiguities likely introduced by the minimal zero-shot prompting instructions,

which may have been insufficient to elicit fully precise item content. For instance, "I usually finish what I start," expected under Productivity, was classified under Responsibility, and "I feel energized when I am in a social group," expected in Sociability, was assigned to Energy Level, as Gemini's attention mechanism emphasized "energized" over "social group" (Rogers et al., 2020; Vaswani et al., 2017). These findings illustrate the diagnostic value of LLM-based semantic analysis for detecting ambiguous items and suggest that future item-generation studies may benefit from more constrained prompts to improve content validity (Roebianto et al., 2023).

Gemini classified items based on semantic probabilities derived from its internal representations, yet the decision processes behind these classifications remain opaque (Rogers et al., 2020). This opacity arises from the complexity of transformer architectures and the proprietary nature of models such as ChatGPT and Gemini, whose training data and weights are not publicly accessible. As a result, although these systems effectively perform content-validity tasks, their decision mechanisms cannot be externally verified, limiting transparency and reproducibility. This limitation also prevents precise estimation of semantic distances between items and facets, as an item may belong to multiple semantic fields simultaneously (Ethayarajh, 2019). Moreover, because Gemini produced a single classification per item, alternative semantic relationships could not be examined, an inherent constraint of attention-based transformer models that weigh linguistic elements differently during classification (Vaswani et al., 2017).

To better understand these relationships, we conducted an embedding-based analysis, identifying vector dimensions most relevant to the Big Five traits by correlating them with dummy variables. This approach parallels the use of expert judges in psychometrics, as it isolates "specialist embeddings" aligned with specific constructs. While effective, this method could be refined through advanced machine learning techniques and applied to other domains, such as analyzing personality traits in social media language, thereby extending its versatility.

The factor loadings indicated that F1 corresponded primarily to Neuroticism, F2 to Agreeableness, F3 to Openness, F4 to Conscientiousness, and F5 to Extraversion, revealing semantic convergence between facets of the BF-BR and BFI-2 instruments. This convergence provides content-validity evidence (Alexandre & Coluci, 2011, 2011; Fitzner, 2007; Haynes et al., 1995; Roebianto et al., 2023; Slaney, 2017). However, a methodological limitation lies in modeling only the factor level, without testing whether the semantic space of embeddings reproduces the hierarchical structure of facets nested within factors. Future research should apply hierarchical analyses to assess semantic congruence at this second-order level.

Unexpectedly, Neuroticism, Agreeableness, Conscientiousness, and Extraversion all showed positive loadings on F1, unlike traditional analyses with human data where Neuroticism correlates negatively with other traits (McCrae & Costa, 1997; Oliveira, 2019; Soto & John, 2017). This pattern reflects how LLM embeddings capture linguistic co-occurrence rather than emotional valence: phrases like "I am sad" and "I am not sad" occupy adjacent positions in the semantic space (Debelak et al., 2024; Devlin et al., 2019; Vaswani et al., 2017). Thus, LLM-based factor analysis reveals semantic, rather than psychological, relationships. Cross-loadings above 0.30 between facets further illustrate semantic overlap among traits.

The first factor (F1) captured, predominantly, anxiety, depression, and emotional volatility (Neuroticism), but also items from other traits expressing excessive concern, as seen in Compassion (e.g., "I often worry about other people's feelings"). This overlap suggests emotional tone influences meaning beyond the nominal construct. Extraversion facets were dispersed across F1 (Neuroticism) and F2 (Agreeableness), reflecting shared elements of emotionality and sociability (Fors Connolly & Johansson Sevä, 2021). Factor F2 primarily represented Agreeableness (compassion, respectfulness and trust) but also included secondary loadings from responsibility (C) and

assertiveness and sociability (E), highlighting semantic intersections related to cooperation and interpersonal harmony.

The interfaces among Neuroticism (F1), Agreeableness (F2), and Conscientiousness (F4) align with DeYoung's (2010) neurobiological model, in which serotonergic systems regulate both emotional stability and behavioral control, explaining the linguistic similarity among those traits. Likewise, Extraversion (F5) and Neuroticism (F1) share semantic proximity due to their conceptual opposition (approach vs. avoidance) reflected in language that expresses contrasting but adjacent meanings (Devlin et al., 2019; Vaswani et al., 2017). Although all loadings were positive, traditional data typically show negative correlations, suggesting that LLM embeddings cluster opposite constructs within shared linguistic fields.

Factors F3 and F4 corresponded to Openness and Conscientiousness, respectively, confirming convergence between human- and LLM-generated items. In F5, Extraversion facets showed higher loadings, but only BF-BR Assertiveness and Sociability achieved their peaks there; other facets loaded more strongly on F1 or F2, consistent with previous semantic interfaces.

Overall, these results indicate that: (1) content validity evidence can be obtained using both prompting and API-based methods, though embeddings offer richer semantic insights; (2) selecting expert embeddings improves focus on construct-relevant semantics; (3) semantic convergence between human- and LLM-generated items supports convergent content validity; (4) traits such as Neuroticism, Agreeableness, Conscientiousness, and Openness are represented by distinct but overlapping semantic fields; and (5) Extraversion items exhibit greater semantic ambiguity, showing stronger associations with Neuroticism and Agreeableness. Secondary loadings should not be viewed as invalidating but as reflecting the interconnected nature of personality traits, which form a complex, interdependent network.

Ultimately, ensuring that items align with theoretical trait definitions remains central to establishing content validity (Alexandre & Coluci, 2011; Haynes et al., 1995; Roebianto et al., 2023). Semantic proximity among items enhances the likelihood that they represent complementary aspects of the same construct, though empirical studies with participant data are still required to confirm whether these semantic correspondences translate into robust psychometric properties.

References

Alexandre, N. M. C., & Coluci, M. Z. O. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciência & Saúde Coletiva*, *16*(7), 3061–3068. https://doi.org/10.1590/S1413-81232011000800006

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

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 903077. https://doi.org/10.3389/frai.2022.903077

Debelak, R., Koch, T. K., Aßenmacher, M., & Stachl, C. (2024). *From Embeddings to Explainability: A Tutorial on Transformer-Based Text Analysis for Social and Behavioral Scientists*. https://doi.org/10.31234/osf.io/bc56a

Dempsey, P. A., & Dempsey, A. D. (2000). *Using Nursing Research: Process, Critical Evaluation, and Utilization* (5th ed.). Lippincott Williams & Wilkins.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., Jones Mitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*, 688–701. https://doi.org/10.1038/s44159-023-00241-5

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*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Ethayarajh, K. (2019). *How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings* (arXiv:1909.00512). arXiv. http://arxiv.org/abs/1909.00512

Fitzner, K. (2007). Reliability and Validity A Quick Review. *The Diabetes Educator*, *33*(5), 775–780. https://doi.org/10.1177/0145721707308172

Fors Connolly, F., & Johansson Sevä, I. (2021). Agreeableness, extraversion and life satisfaction: Investigating the mediating roles of social inclusion and status. *Scandinavian Journal of Psychology*, *62*(5), 752–762. https://doi.org/10.1111/sjop.12755

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Google. (2024). *Gemini (Modelo models/text-embedding-004)* [Large language model]. Google. https://ai.google.dev/gemini-api/docs/embeddings

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

Hu, J., Dong, T., Gang, L., Ma, H., Zou, P., Sun, X., Guo, D., & Wang, M. (2024). *PsycoLLM: Enhancing LLM for Psychological Understanding and Evaluation* (Versão 2). arXiv. https://doi.org/10.48550/ARXIV.2407.05721

Hu, L., He, H., Wang, D., Zhao, Z., Shao, Y., & Nie, L. (2024). LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 18234–18242. https://doi.org/10.1609/aaai.v38i16.29782

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

Kuhn, M. (2008). Building Predictive Models in *R* Using the caret Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's Congruence Coefficient as a Meaningful Index of Factor Similarity. *Methodology*, *2*(2), 57–64. https://doi.org/10.1027/1614-2241.2.2.57

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509–516. https://doi.org/10.1037/0003-066X.52.5.509

Oliveira, J. P. (2019). Psychometric Properties of the Portuguese Version of the Mini-IPIP five-Factor Model Personality Scale. *Current Psychology*, *38*(2), 432–439. https://doi.org/10.1007/s12144-017-9625-5

Ooms, J. (2014). *The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects* (Versão 1). arXiv. https://doi.org/10.48550/ARXIV.1403.2805

OpenAI. (2023). *ChatGPT (Versão 3.5, consulta de setembro)* [Large language model]. OpenAI. https://chat.openai.com

Pasquali, L. (2010). *Instrumentação Psicológica: Fundamentos e Práticas*. Artmed.

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*, *19*(5), 808–826. https://doi.org/10.1177/17456916231214460

Pires, J. G., Nunes, C. H. S. D. 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

R Core Team. (2023). *R: A Language and Environment for Statistical Computing* (Vienna, Austria). R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2007). *psych: Procedures for Psychological, Psychometric, and Personality Research* (p. 2.4.6.26) [Dataset]. https://doi.org/10.32614/CRAN.package.psych

Rizopoulos, D. (2006). ltm: An *R* Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5). https://doi.org/10.18637/jss.v017.i05

Roebianto, Roebianto, Savitri, Aulia, Suciyana, & Mubarokah. (2023). Content validity: Definition and procedure of content validation in psychological research. *Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Slaney, K. (2017). *Validating Psychological Constructs*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-38523-9

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *31st Conference on Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wickham, H. (2023). *httr: Tools for Working with URLs and HTTP* (Versão 1.4.6) [Software]. https://CRAN.R-project.org/package=httr

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2023). *Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View* (Versão 3). arXiv. https://doi.org/10.48550/ARXIV.2310.02124

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). *Sentiment Analysis in the Era of Large Language Models: A Reality Check* (Versão 1). arXiv. https://doi.org/10.48550/ARXIV.2305.15005

José Maurício Haas Bueno

Doutor, afiliado à Universidade Federal de Pernambuco.

Ricardo Primi

Doutor, com afiliação institucional na Universidade São Francisco.

Emanuel Duarte de Almeida Cordeiro

Doutor, vinculado à Universidade Estadual do Sudoeste da Bahia.

Ana Deyvis Santos Araújo Jesuíno

Doutora, atua institucionalmente na Universidade Federal do Maranhão.

Monalisa Muniz

Doutora, afiliada à Universidade Federal de São Carlos.

Ana Paula Porto Noronha

Doutora, com afiliação institucional é com a Universidade São Francisco.

Endereço para correspondência

José Maurício Haas Bueno
Universidade Federal de Pernambuco
Av. da Arquitetura, s/n, 7º Andar
Cidade Universitária, 50740-550
Recife, PE, Brasil

Ricardo Primi
Universidade São Francisco
Rua Waldemar César da Silveira, 105
Jardim Cura D'Ars (Swift), 13045-510
Campinas, SP, Brasil

Emanuel Duarte de Almeida Cordeiro
Universidade Estadual do Sudoeste da Bahia
Estrada Bem Querer, Km-04
Bairro Universitário, 45083-900
Vitória da Conquista, BA, Brasil

Ana Deyvis Santos Araújo Jesuíno
Universidade Federal do Maranhão
Avenida dos Portugueses, 1966
Bacanga, 65080-805
São Luís, MA, Brasil

Monalisa Muniz
Universidade Federal de São Carlos
Rodovia Washington Luís, km 235
13565-905
São Carlos, SP, Brasil

Ana Paula Porto Noronha
Universidade São Francisco
Rua Waldemar César da Silveira, 105
Jardim Cura D'Ars (Swift), 13045-510
Campinas, SP, Brasil

Os textos deste artigo foram revisados por Bruno Schroeder dos Santos e submetidos para validação dos autores antes da publicação.