LLM embeddings on test items predict post hoc loadings in personality tests.

Monica Casella[1], Maria Luongo[2], Davide Marocco[3], Nicola Milano[4*], Michela Ponticorvo[5]

[1]*University of Naples Federico II, Department of humanistic studies, Natural and artificial cognition laboratory "Orazio Miglino", via Porta di Massa 1, Naples, 80125, Italy*

Abstract

In this article we examine the application of Large Language Models (LLMs) in predicting factor loadings in personality tests through the semantic analysis of test items. By leveraging text embeddings generated from LLMs, we assess the semantic similarity of test items and their alignment with hypothesized factorial structures, without relying on human response data. Our methodology uses embeddings from the Big Five personality test to explore correlations between item semantics and their grouping in factorial analyses. Preliminary results suggest that LLM-derived embeddings can effectively capture semantic similarities among test items, potentially serving as a valid measure for initial survey design and refinement. This approach offers insights into the robustness of embedding techniques in psychological evaluations, indicating a significant correlation with traditional test structures and providing a novel perspective on test item analysis.

Keywords

Embeddings, language models, personality traits, semantic similarity

1. Introduction

In psychological testing and, more in general, in the different contexts that include evaluation, it is very important to assess item quality. This process, known as item analysis, foresees the evaluation of different item characteristics including item difficulty, item discrimination, item and test reliability.

The two main approaches to run item analysis - classical test theory (CTT) and item response theory (IRT) – (see [1]) share the starting point of a person per item matrix: a matrix where examinees are rows and items are columns. This raises some problems, for example the matrix can be sparse if a lot of missing data are present.

Whereas the item formulation relies on procedures that come before the test (and items) administration, for example focus group with experts [2], item analysis is typically based on post hoc analysis that used data coming from the administration to a sample of respondents. In this phase, together with item characteristics, evaluation is run on the test, especially in the framework of TCT, including the study reliability and of test structure in terms of latent variables by the means of factor analysis [3]. This is a consolidated approach.

Factor analysis is used to reduce a large number of variables into fewer numbers of factors, which have a psychological meaning. This technique extracts maximum common variance from all variables and expresses them into a common score, that is used for further analysis. This analysis allows also to calculate factor loadings that are basically the correlation coefficient between the single variable and the factor. Factor analysis has been widely used in psychological research both in cognitive domain (consider the factorial theories of intelligence, [4-5]) and in personality domain. In this latter case, between the

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
* Corresponding author.
nicola.milano@unina.it

theoretical proposals up the 1990s ([6-7]), the Big Five Theory of personality is a notable example of how personality can be conceived and described as a constellation of different dimensions, factors in the terms we have used before. Evidence of this theory has grown over the years and five broad personality traits described by the theory are identified in extraversion, agreeableness, openness, conscientiousness, and neuroticism. This approach was developed in 1949 by [8] and later expanded by other researchers up to the work by [9].

For the development of this approach, a key role has been played by traits measurement. Traditionally, a Big Five personality test is taken with a questionnaire and response on a Likert scale [10]. The Big Five Questionnaire [11], and its newer version BFQ-2 [12], is used in different contexts and represents a golden standard for measuring personality, according to the Big Five theory. It is formed of 132 items. Some questions ask how much a person agrees or disagrees that he or she is someone who exemplifies various specific statements, such as: "Is open to trying new experiences" (for openness, or open-mindedness) or "Is anxious about the future all the time" (for neuroticism, or negative emotionality). The responses, "Strongly agree" to "Strongly Disagree" (with alternatives in between) determines to what degrees the respondent show that specific traits.

In this contribution, we propose a method to understand the strength of the connection between items and factors based only on the item considered as a linguistic material before the test administration. In order to do this, we used LLM, large language models [13], artificial intelligence models that are able to process and generate natural language. In these models, embeddings take a piece of text - a word, sentence, paragraph or even a whole article, and convert that into an array of floating-point numbers, the embedding vector. This way, the artificial neural network can code the linguistic information. Embeddings are a numerical representation of the semantic meaning of the content in a many-multi-dimensional space.

We used this representation to analyze item and check if the proximity of items in this multi-dimensional space corresponds to post-hoc loadings in personality test factor analysis.

2. Materials and Methods: Utilizing Item Embeddings for Semantic Analysis

Advanced large language models (LLMs), have set new benchmarks in processing and generating text that is understandable by humans, seeing widespread application across countless tasks by millions globally. Within the realm of psychology, the ability of LLMs to interpret, contextualize, and extrapolate from human language without prior exposure has sparked considerable interest due to their potential in exploring unseen texts through a zero-shot approach [14-16]. This section outlines the methodology for leveraging LLMs to assess semantic similarities among test items, determining if these similarities align with a hypothesized factorial structure subsequently identified in participant responses. LLMs internally convert input text into vector representations, known as embeddings, through the training phase. Each word or sentence is transformed into a fixed-size vector of floating-point numbers. Research indicates that the space of these embeddings possesses metric characteristics [18], allowing for the mapping of similar concepts, such as colors or synonyms, closer together than disparate ones [18-19]. By utilizing this feature, embeddings serve as a tool for gauging semantic similarity across various domains.

We propose employing embeddings from established psychological tests to examine item similarities and verify whether the test exhibits the anticipated factorial structure by focusing solely on the items, excluding participant responses. To this end, we employ the Bidirectional Encoder Representations from Transformers (BERT) model [20], a pre-trained, open-source language model developed by Google. BERT, which has been trained on over five billion sentences from Wikipedia and the Google Book Corpus, aims to predict missing words in sentences. Since its introduction in 2018, numerous enhancements to the BERT model have been suggested. Our methodology utilizes roBERTa, a variant that has achieved top results on standard LLM benchmarks [21]. As an initial step, we apply the Big5 personality test [22], mapping each of its 50 items into the embedding space with BERT to achieve a 1024-dimensional vector representation for each item.

To determine the proximity of embeddings for different items within this space, we use cosine similarity, which calculates the cosine of the angle

between two vectors. This measure, dependent solely on the angle and not the magnitude of vectors, is obtained by dividing the dot product of two vectors by the product of their magnitudes.

$$\text{cosinesimilarity}(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \qquad (1)$$

The resulting cosine similarity equation, where A and B represent the embeddings of two different items, produces a similarity matrix. This matrix has various applications, such as clustering to observe if the embeddings align with the hypothesized factorial structure or directly applying principal component analysis. The cosine similarity matrix, particularly when vectors are centered to have a zero mean, equates to the Pearson correlation coefficient. Thus, by analyzing items through the lens of LLMs, we can deduce whether our test's structure conforms to the expected factorial arrangement. Moreover, this approach allows for the modification or elimination of items that either poorly correlate with others or duplicate the same construct, streamlining the test process.

3. Results : Alignment with Human Responses

Our analysis aims to determine whether embeddings derived from large language models can successfully identify semantic similarities among items and predict human responses to them. To this end, we first examine whether there is similarity between construct-related items in the embedding space. Figure 1 shows a t-SNE [23] projection of the 1024-dimensional embeddings of the Big Five items into a two-dimensional space. The colors and letters indicate the factors underlying the items; for example, yellow and 'O' represent openness, with the numbers specifying the particular item. Different categories of items are mapped closer together and occupy substantially different zones of the space. There is some overlap due to single items that are more difficult to classify and to specific constructs that share overlapping meanings, even for humans, such as Agreeableness and Extraversion.

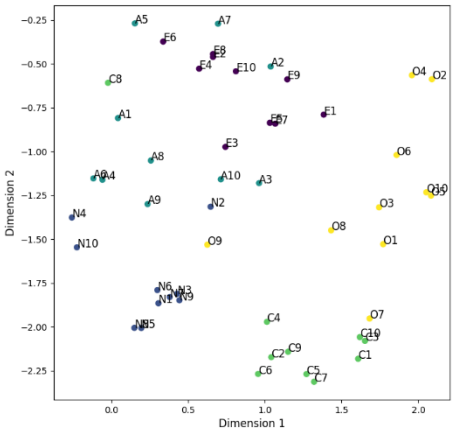

Figure 1. T-SNE projection of the embeddings in a 2-dimensional space. Yellow points are openness (O) related items, Green points are Conscientiousness (C) items, Blue points are Neuroticism (N) items, Black points are extra-version (E) items and violet points are Agreeableness (A) items.

We then applied Principal Component Analysis (PCA) to the cosine correlation matrix of these embeddings, which revealed a latent space that accurately groups items belonging to the same construct under a single principal component. This process not only facilitates the interpretation of outcomes but also confirms the theoretically anticipated structure. The focus now shifts to comparing the latent structure unearthed from the semantic similarity analysis with that derived from actual human responses. Specifically, we aim to investigate whether the item loadings generated from the embedded item representations mirror those obtained from analyzing human responses. By examining the correlation between the two sets of loadings, we gain insight into the extent to which item semantics predict human response patterns. Ideally, a high correlation between construct loadings would indicate not only that related items are grouped accurately but also that both cross-loadings and other factor loadings exhibit similar trends.

For this purpose, we sourced human response data from the Open Psychometrics website for the Big 5 personality test. We then replicated the previously outlined embedding analysis on this response data, starting with the calculation of Pearson correlations for each pair of items, followed by PCA to determine item loadings based on the theoretical number of latent factors. For items phrased in reverse, we

adjusted their scales to align with the correct direction, a necessary step because the embedding model does not differentiate between reversed and non-reversed items. This adjustment ensures that the comparison of loadings between the models remains valid.

The results, depicted in Figure 2, show the Spearman correlation coefficient for the Big 5 test examined. We observed a high correlation between the embedding loadings and the human response loadings within the same constructs ($R > 0.4$, p-value $<< 0.001$). Additionally, a significant correlation ($R > 0.4$, p-value $<< 0.001$) was noted between the constructs of Agreeableness and Extraversion. These findings suggest that the semantic similarities among items effectively reflect the relationships among factors as found in human subjects.

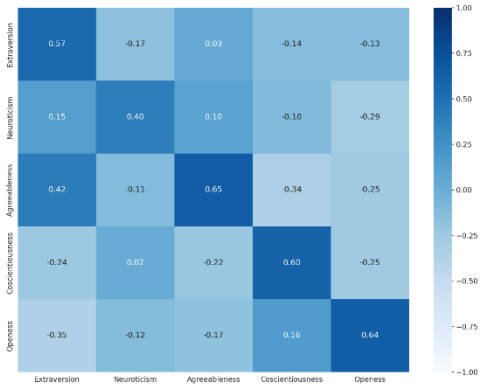

Figure 2. Spearman correlation between Item embedding loadings (x-axis) and human response loadings (y-axis). Sperman r greater than 0.40 shows significative correlation between the loadings (p-value $<< 0.001$)

4. Conclusions

From our analysis we found that the t-SNE projection of the embeddings maps items related to similar constructs close together in the embedding space. Despite some overlap due to ambiguous items or constructs with overlapping meanings, such as Agreeableness and Extraversion, the overall pattern suggests that embeddings derived from large language models capture semantic similarities among items.

Moreover, the application of Principal Component Analysis (PCA) on the cosine correlation matrix of the embeddings reveals a latent space where items belonging to the same construct are grouped under a single principal component. This alignment with the theoretical structure provides further validation of the embedding analysis. Considering the comparison with human response data, the correlation between the item loadings derived from the embedded item representations and those obtained from analyzing human responses indicates a significant relationship. The high correlation observed, particularly within the same constructs, suggests that the semantic similarities captured by the embeddings effectively mirror the relationships observed in human subjects. Notably, a significant correlation is observed between the constructs of Agreeableness and Extraversion, indicating a meaningful association between these factors in both the embedding analysis and human responses. Overall, these findings support the notion that embeddings derived from large language models can successfully identify semantic similarities among items and, so, serve as a valid preliminary measure in the context of survey design.

References

[1] Kline, T. J. (2005). Psychological testing: A practical approach to design and evaluation. Sage publications.

[2] Mallinckrodt, B., Miles, J. R., & Recabarren, D. A. (2016). Using focus groups and Rasch item response theory to improve instrument development. The Counseling Psychologist, 44(2), 146-194.

[3] Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. Journal of consulting and clinical psychology, 55(4), 584.

[4] Sternberg, R. J. (1980). Factor theories of intelligence are all right almost. Educational Researcher, 9(8), 6-18.

[5] Carroll, J. B. (2013). A three-stratum theory of intelligence: Spearman's contribution. In Human abilities (pp. 1-17). Psychology Press.

[6] Zuckerman, M., Kuhlman, D. M., Thornquist, M., & Kiers, H. (1993). Five (or three) robust questionnaire scale factors of personality without culture. Personality and individual Differences, 14(4), 569-578.

[7] Eysenck, H. J. (1953). The structure of human personality (Psychology Revivals). Routledge.

[8] Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. The Journal of Abnormal and Social Psychology, 44(3), 329.

[9] Costa Jr, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. Personality and individual differences, 13(6), 653-665.

[10] Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. Frontiers in psychology, 12, 637547.

[11] Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The "Big Five Questionnaire": A new questionnaire to assess the five factor model. Personality and individual Differences, 15(3), 281-288.

[12] Caprara, G. V., Barbaranelli, C., Borgogni, L., & Vecchione, M. (2008). BFQ-2. Big five questionnaire, 2.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[14] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D : Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. (2020).

[15] Binz, M., & Schulz, E.: Turning large language models into cognitive models. arXiv preprint arXiv:2306.03917. (2023)

[16] Buschoff, L. M. S., Akata, E., Bethge, M., & Schulz, E. : Have we built machines that think like people?. arXiv preprint arXiv:2311.16093. (2023).

[17] Chuang, Y. S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., ... & Rogers, T. T.: Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv preprint arXiv:2311.09618. (2023).

[18] Yan, F., Fan, Q., & Lu, M.: Improving semantic similarity retrieval with word embeddings. Concurrency and Computation: Practice and Experience, 30(23), e4489. (2018).

[19] Colla, D., Mensa, E., & Radicioni, D. P.: Novel metrics for computing semantic similarity with sense embeddings. Knowledge-Based Systems, 206, 106346. (2020).

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).

[21] https://www.sbert.net/docs/pretrained_models.html

[22] McCrae, R. R., & Costa, P. T.: Updating Norman's" adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. Journal of personality and social psychology, 49(3), 710. (1985).

[23] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).