Running Head:

PSEUDO FACTOR ANALYSIS

Enhancing Scale Development: Pseudo Factor Analysis of Language Embedding Similarity Matrices

Nigel Guenole[1*], E. Damiano D'Urso[2*], Andrew Samo[3], Tianjun Sun[4**], & Jonas M. B. Haslbeck[5]

[1] *Goldsmiths, University of London*
[2] *Independent Researcher*
[3] *Bowling Green State University*
[4] *Rice University*
[5] *University of Amsterdam*

Author Notes:

[*]These authors contributed equally to this work.

[**]Questions and comments regarding this work can be sent to any of the authors. Correspondence regarding this article can be directed to Tianjun Sun at tianjunsun@rice.edu.

All project materials are available on the Open Science Framework (OSF) at https://osf.io/3mpzb/.

There are no conflicts of interest regarding funding or publication.

Abstract

This article builds and extends on recent work using Large Language Models (LLMs) in psychometrics to generate pseudo-discrimination parameters. While earlier work looked at pseudo-discrimination on an item-by-construct basis, we introduce pseudo-factor analysis to enhance scale design. Pseudo-factor analysis is a data-less, model-based approach to evaluating aspects of a latent construct's measurement model, such as dimensionality and the relations between factors and their indicators. Across two studies using Five and Six-factor personality frameworks, a variety of sentence transformer models, and three encoding approaches (i.e., atomic, atomic reversed, and macro), pseudo-factor analyses recovered theoretically expected structures. These pseudo-factor structures were strongly related to their established empirical factor structures based on factor analyses of human ratings in prior published research. We suggest that Pseudo-Factor Analysis is a viable method for checking and potentially modifying scale items after item generation and before item trialing. We provide a Shiny application for calculating pseudo-factor analysis parameters and related psychometric estimates.

Keywords: *Latent Variables, Large Language Models (LLMs), Factor Analysis, Transformers, Artificial Intelligence, Psychometrics, Personality, IPIP, NEO, HEXACO*

Enhancing Scale Development:

Pseudo Factor Analysis of Language Embedding Similarity Matrices

Measurement is an essential part of psychological science. The dominant method for measuring psychological attributes is multi-item scales. Yet a persistent challenge faced by practitioners of psychological measurement is that the scale development process is expensive and time-consuming. This process, for example, typically involves specifying the construct domain, drafting an initial pool of items that capture domain content, and then iteratively pretesting, piloting, analyzing, and refining the item pool before implementing the survey (e.g., Lambert & Newman, 2023; Schultz et al., 2014). The most resource-intensive phases involve item generation, pre-trial evaluation by subject matter experts, and empirical verification of item properties from trial responses using methods such as item response theory and factor analysis. Methods that reduce the time and resource costs required for scale development without compromising the quality of the resulting scales would, therefore, have considerable utility (Brogden, 1949).

Recent developments in natural language processing (NLP) methods involving transformer-based large language models (LLMs; Vaswani et al., 2017) are now being used in areas of scale development, specifically by taking advantage of natural language generation capabilities for item writing (Fyffe et al., 2023; Hernandez & Nie, 2022; Hommel et al., 2022; Russell-Lasalandra et al., 2024) or natural language understanding for item validation (Arnulf et al., 2021; Cutler & Condon, 2023; Wulff & Mata, 2023). In this article, we continue this tradition of fusing LLMs and psychometrics by building on the recently introduced idea of *pseudo-discrimination* parameters (Guenole et al, 2024). Pseudo-discrimination parameters approximate empirical estimates of item-to-construct associations by assessing how similar an item's

semantic content is relative to its parent construct definition. Importantly, pseudo-parameters do not require empirical data because they are based on relative semantic associations between scale definitions and the items themselves. Here, we generalize the idea of pseudo-discrimination beyond the initially proposed scale-by-scale level to a multidimensional structure with *pseudo-factor analysis*, extending the principles of pseudo-parameter estimation to factor analytic methods.

The present work rests on a simple proposition at the psychometric level: the embedding vector for each personality item can serve as a substitute for an empirical vector of observed responses to personality items. We expect that embedding vectors might serve as substitutes for item response vectors because of conceptual parallels that exist in the creation process of both types of vectors. For instance, both vector formats represent aggregated encodings of personality statements. The item response vector is an aggregate of individuals' personalized encodings of the self-descriptiveness of each statement. The LLM representations of each statement are encodings based on the model training data that, to the extent that the training data represents the test population, can also be considered an aggregated item encoding. We hereafter refer to this proposition as the "*substitutability assumption*".

While this process is clearly extremely complex, and we make no claims to have a full understanding of it, we also do not rely on such an understanding here. We only take the fact that both vectors are encodings of similar constructs as a heuristic motivation to investigate whether there is an alignment between the results of the analysis of embedding vectors and the results of empirical response data. We would like to stress that, at the end of the day, empirical research on real data is always required. We, therefore, do not propose to fully automatize scale construction

with AI. However, if similarities exist between the results of the two methods, we may be able to considerably improve the process of scale construction.

From Pseudo-Discrimination to Pseudo-Factor Analysis (PFA)

The pseudo-discrimination approach analyzes direct relations between item and construct definition embeddings. However, it does not allow examination of how items from different scales "behave" together because the relationship between items and construct definitions is considered one item at a time. An item-by-item or scale-by-scale approach is problematic because it ignores associations across items, which can result in biased psychometric estimates due to statistical issues such as violations of local independence, cross-loadings, and merging or collapsing subdimensions (Aschenbach, 2021; Christensen et al., 2023; Marsh et al., 2010). Considering that many of the constructs in psychological science are multidimensional, it is important to use measurement methods that appropriately address this complexity. Conventional multidimensional measurement models provide information concerning the latent structure of the psychological construct space. This commonly includes features such as determining the number of latent constructs measured by the items, which constructs are measured by which items, the strength of item-to-construct associations, and how constructs are related (Rhemtulla, et al., 2020).

In this article, we use the similarity between empirical correlations and semantic similarity estimates to propose "pseudo-factor analysis" (PFA). This allows us to evaluate all these features based on the similarities between the embedding vectors, which serve as substitutes for corresponding empirical data. Under PFA, the substitutability assumption of item response vectors and statement embedding vectors is generalized such that an item embedding cosine similarity matrix may stand as a substitute for the empirical item response covariance

matrix used in conventional factor analysis during the item validation process. We aim to show that viewing embeddings in terms of their similarities to one another creates a bridge between AI and conventional modeling approaches. If this approach were to work, an enhanced scale development workflow would involve checking the expected factor structure using PFA *after* item pool creation and *before* empirical trialing to see if the item improves the results obtained in item trialing.

The Current Project

We present two new studies of PFA that compare the embedding-based pseudo-factor structures for two widely used personality inventories with previously published empirical factor structures. The two models capture alternate five- and six-factor structures with the Big Five IPIP NEO-300 and the Big Six IPIP HEXACO-24. Personality structures appear to be well suited to novel explorations of LLM-based measurement approaches because, theoretically, there is a rich tradition of natural language-based measurement with the lexical approach (Allport & Odbert, 1936; Cutler & Condon, 2020) and, empirically, there is a *relatively* standard structure of personality within the established taxonomic frameworks (Mõttus et al., 2020; Thielmann et al., 2022).

In the current project, we compare the pseudo-factor structures that emerge from each transformer model we study and the empirical factor structures from development samples for the IPIP-300 (Johnson, 2014) and for the HEXACO-240 (Ashton & Lee, 2007). The primary contribution of this article is introducing PFA, a new way to improve scale development processes. Although we introduce PFA in the context of personality, we highlight that this method is extremely general and, therefore, may have a large, beneficial impact on scale development more broadly. To move towards the goal of helping improve scale development

more generally, we also provide an easy-to-use Shiny app that also allows researchers without any programming experience to use PFA. We conclude by outlining directions for future research to develop the potential of PFA and extensions such as pseudo-structural equation modeling.

Method

Openness and Transparency

All relevant data, code, and the associated Shiny App is publicly available on the OSF and GitHub, respectively. The study was not preregistered.

Data

No new data were collected or analyzed for this research. For the LLM sections of our manuscript, we used the natural language content of the Big Five and Big Six IPIP-NEO and IPIP-HEXACO personality inventories. The IPIP (International Personality Item Pool) is a large, open-source collection of personality content (i.e., items, scales) that includes large datasets that are freely available for research (http://ipip.ori.org/; Goldberg et al., 2006). We used the factor structure results already published for the IPIP-NEO-300 (Johnson, 2014) and HEXACO-240 (Goldberg, 1999).

Measures

NEO personality model (Study 1). The measure for Study 1 was the NEO-IPIP (Goldberg, 1999). This assessment measures the factors of the five-factor model along with the 30 facets that fall beneath the top five domains. The NEO model captures the personality traits (facets in parentheses) of Neuroticism (i.e., anxiety, anger, depression, self-consciousness, immoderation, and vulnerability), Extraversion (i.e., friendliness, gregariousness, assertiveness, activity level, excitement-seeking, and cheerfulness), Openness (i.e., imagination, artistic

interests, emotionality, adventurousness, intellect, and liberalism), Agreeableness (i.e., trust, morality, altruism, cooperation, modesty, and sympathy), and Conscientiousness (i.e., self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness). Each facet is measured by 10 items there are 300 items in total.

HEXACO personality model (Study 2). For Study 2, we used the HEXACO model, which proposes six factors of personality with related reinterpretations to those of the Big Five, along with an additional factor, Honesty-Humility (Ashton & Lee 2007). HEXACO broadly captures the domains (facets in parentheses) of Honesty-Humility (i.e., sincerity, fairness, greed avoidance, and modesty), Emotionality (i.e., fearfulness, anxiety, dependence, sentimentality), Extraversion (i.e., social self-esteem, social boldness, sociability, and liveliness), Agreeableness (i.e., forgivingness, gentleness, flexibility, and patience), Conscientiousness (i.e., organization, diligence, perfectionism, and prudence), and Openness to Experience (i.e., aesthetic appreciation, inquisitiveness, creativity, and unconventionality). We used the IPIP open-source version of the HEXACO assessment (Goldberg, 1999).

Language Modeling

With two-tier measurement structures where items form factors without an intermediary facet structure, we would generate embeddings for each item for factor analysis. Yet here we are factor analyzing facets because each of the current personality frameworks incorporates a three-tier structure: items form facets, which in turn form factors. We need a way to obtain a single embedding vector for each facet because facets are represented by a single “response” vector in three-tier empirical data when facets are factor-analyzed. We explored embedding items separately and pooling the item embeddings by averaging, as well as pooling the items for each facet by concatenating items and embedding the concatenated item string for each facet. We call

the former approach “atomic” and the latter approach “macro”. For a second atomic approach, we tried reversing item embeddings prior to aggregation. We do this by multiplying the embeddings by their theoretically expected item signs. Figure 1 shows a visual depiction of the natural language processing workflow.

Language models can be based on the encoder (e.g., BERT; RoBERTa), the decoder (e.g., GPT, LlaMa), or the encoder-decoder (e.g., BART, T5) style transformer architectures. Very broadly, encoders excel at “understanding” text input by encoding them as embedding vector representations. In contrast, decoders excel at “generating” text by predicting sequential words from embedding representations. Although decoder-style models have grown in popularity for reasons out of the scope of this paper, encoder-only models still excel at tasks involving embeddings-based semantic similarity.

Choice of Embedding Encoders. To encode the items, we opted for a range of sentence encoders, including a non-transformer, base sentence transformers, distilled models, and a fine-tuned model specific to psychology content. For each item or string of items, the encoding process obtains a vector of *N* entries where *N* is the number of embedding dimensions. As a baseline non-transformer architecture, we used the Universal Sentence Encoder with Deep Average Networks (USE-DAN) developed at Google (Cer et al., 2018). While such a model may account for context to some degree, it is not as powerful as transformer models with their self-attention mechanisms. For base transformer models, we included models based on the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018). We chose the Masked and Permuted Pre-training for Language Understanding (MPNet base v2: Song et al., 2020) and Sentence-T5 (Ni et al., 2021), an encoder-decoder approach that can take text as input and produce text as output. Sentence-T5-base has sentence encoding capability and

has been reported as performing well in psychometric research (Hernandez & Nie, 2022). Distilled transformers are transformers derived from more complex models. One that we consider here is "MiniLM-L12-H384-uncased" from researchers at Microsoft (Wang et al., 2020). We also considered a distilled version of BERT transformers including the Robustly optimized BERT approach, including "all-distilroberta-base-v2" (Liu et al., 2019). Finally, we included one transformer specifically adapted for psychology applications by Wulff & Mata (2023). All models were obtained from the Hugging Face website[1].

Estimating Similarity Matrices. We created cosine similarity matrices from the cosine similarity of every possible pairing of *N*-dimensional embeddings per transformer model. While different transformers produce encodings of different dimensions, the dimensionality of the similarity matrix across transformers is always the same; it is a symmetrical matrix with the same number of rows and columns as there are facets in the model. In other words, the aggregation of raw embeddings to cosine similarities allows combining embedding models of different dimensionality (e.g., 376-d vs. 768-d) into a standardized dimensionality. This allowed us to create an aggregated similarity matrix that was the simple average of every cosine similarity across every sentence encoder that was used.

Factor Analyses

Exploratory Factor Analysis. We used exploratory factor analysis to analyze the cosine similarity matrices (Lawley & Maxwell, 1971). We report maximum likelihood factor extraction solutions for the theoretical models (NEO; HEXACO). We used a fully specified oblique target

[1] DistilRoberta: https://huggingface.co/sentence-transformers/all-distilroberta-v1
miniLM: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
mpnet: https://huggingface.co/sentence-transformers/all-mpnet-base-v2
t5: https://huggingface.co/google-t5/t5-base
Wulf & Matta: https://huggingface.co/dwulff/mpnet-personality

rotation (D'Urso et al., 2023). We extracted the theoretically expected number of factors for each model (i.e., 5 and 6) and adopted a new Dominant (i.e., largest) Average Absolute Loading (DAAL) approach to systematically interpret factor solutions. We calculated the average absolute loading for all facets theoretically expected to be within a factor across all empirical factors. We assigned a label to a factor if the DAAL across all theoretical facets was highest on that factor. In some cases, factors merged (i.e., the DAAL for more than one theoretical factor loaded on the same empirical factor). In such cases, we considered the factors "unassigned". In other cases, the DAAL for a set of theoretical facets was highest on a factor but lower than the same set of facets on another factor. Hence, it is possible that factors are not uniquely assigned, in which case, factors were also designated "unassigned".

Comparing Factor Structures. To identify how well pseudo-factor analysis was able to approximate the empirical factor structures, we considered four metrics. First, we examined the number/proportion of fully recovered factor solutions according to the DAAL criteria. Second, we examined the number of times each specific factor was recovered across each of the encoding methods. Third, we examined the similarity of the loadings using Tucker's congruence coefficient (Tucker, 1951). Lorenzo-Seva and ten Berge (2006) suggested that values over .85 indicate fair similarity, while values over .95 indicate excellent similarity. We also explored the Pearson correlations between corresponding factors in the embedding and empirical factor analysis approaches.

Results

For each of the two studies, the IPIP-NEO and HEXACO, we discuss global factor structure recovery rates, specific factor recovery rates, empirical and embedding-based loading

matrices, and the congruence and correlation coefficients with the originally reported empirical structures.

Study 1: NEO personality inventory

NEO global factor structure recovery rates. All methods performed well overall in terms of global factor recovery. The mean proportion of factors recovered across transformers was 71% of factors for the atomic method, 94% of factors for the atomic reversed method, and 96% of factors recovered for the macro encoding method. When transformer embedding results were considered in aggregate, the transformer models also performed better than the simpler USE-DAN approach. for each encoding approach there were multiple transformers that performed as well or better than USE-DAN in terms of the proportion of factors recovered. The best-performing transformers led to 100% recovery rates across the three encoding approaches, while USE-DAN recovery rates ranged from 60% to 100%. Specific results for each transformer are shown in Table S1 in the supplementary materials.

NEO factor structure recovery for specific factors. As well as considering factor recovery rates overall across every factor, it is important to see which individual factors tended to be well-recovered across the three embedding methods. The lowest recovery rate was for the openness factor under the atomic embedding method, where 57% of transformer embeddings led to factor recovery. All remaining factor recovery rates were 71% or higher. Detailed tables are available in Table S2 in the supplementary materials.

NEO empirical and embedding-based loading matrices. Table 1 shows the loading matrix from factor analysis of the average of all transformer-based similarity matrices, showing clearly interpretable five-factor solutions, particularly for atomic reversed and macro embedding methods. This was not always the case for USE-DAN, where factors either merge, indicated by

two letter factor codes in the column heading; or are unassigned using the dominant average absolute loading approach, indicated by U in a column heading. Factor structures for all transformer and encoding method combinations are available in the supplementary materials.

NEO congruence between empirical and embedding-based loading matrices. Table 2 presents Tucker's congruence coefficient across encoding approaches for each factor. These reveal that, in general, the loading similarity was not sufficient to be equivalent across the embedding and empirical approaches, despite the configural factor structures themselves being relatively well recovered by earlier criteria. Table 3 shows the correlations between factor loadings under the two approaches. Despite showing congruence values indicating the loadings were not equivalent, the correlations between corresponding factors, particularly for the atomic encoding method, were strong, indicating embedding-based loadings are informative about empirical loadings. Figure 2 presents scatter plots for macro encoding, which performs best overall for the NEO instrument.

Study 2: HEXACO personality inventory

HEXACO *global factor structure recovery rates.* In the case of the HEXACO model, the atomic performance outperformed the macro approach, which was, in turn, considerably better than the atomic reversed approach. The transformer average embedding approach also performed similarly to the simpler USE-DAN approach to embeddings. Nonetheless, the overall average recovery rate of 98% of factors recovered for the atomic method indicates that the model recovery was effective. See the supplementary materials Table S3 for a breakdown of factor recovery for the HEXACO model overall by different transformer methods and encoding variations.

HEXACO *factor structure recovery for specific factors*. The atomic reversed approach consistently failed to recover the emotionality factor of the HEXACO model. Conversely, the atomic and macro embedding approaches tended to recover all factors well in a consistent fashion. See the supplementary material Table S4 for a breakdown of specific factor recovery rates for the HEXACO model across different embedding methods and encoder combinations.

HEXACO *empirical and embedding-based loading matrices*. Table 4 shows that in line with the NEO results, atomic and macro embedding approaches recover the HEXACO structure well, whereas the atomic reversed blends the H and E factors. These methods also performed better than USE-DAN, which can be seen in the tables in the supplementary materials.

HEXACO *congruence between empirical and embedding-based loading matrices*. Table 5 presents Tucker's congruence coefficients for the HEXACO factors across encoding methods. These show that for all factors but conscientiousness, the congruence coefficients are not satisfactory and are often < .85. However, the conscientiousness factor has satisfactory congruence across atomic and macro methods for most transformers.

Table 6 shows the correlations between factor loadings under the two approaches. Despite the factors showing lower congruence values than needed for equivalence, they all showed strong correlations between corresponding factors but not between non-corresponding factors. Figure 3 presents scatter plots for the atomic average encoding, which performs best overall for the HEXACO instrument.

Discussion

Building on the principle that the embedding of an item can serve as a substitute for a vector of observed responses in the scale development process, we generalized earlier work that uses item embeddings to acquire pre-knowledge of empirical item discrimination characteristics

(e.g., Guenole et al., 2024). Whereas earlier work looked at item embeddings and construct definition embeddings, in this article, we represented entire matrices of empirical item correlations with matrices of item embedding similarities. This analysis gives rise to Pseudo-Factor Analysis (PFA), whose primary use case we see in streamlining scale development.

Across two studies that used two well-established theoretical models of trait-based personality today (i.e., the NEO and HEXACO), we showed that factor structures could be recovered accurately only using the natural language features represented in LLM embeddings of personality statements. Specifically, while the lower range for the worst performing transformers were 60% and 83%, respectively, for the NEO and HEXACO models, we found that PFA was able to recover 100% of the NEO and HEXACO empirical factor domains using numerous transformers and several encoding techniques. Transformer models were also somewhat more successful at recovering these structures than less complex methods, in particular, the non-transformer architecture USE-DAN – which is unsurprising given the superior natural language understanding of transformers.

The lowest recover rates for specific factors were for openness in the NEO using the atomic method and extraversion for the HEXACO using atomic reversed. In addition, while macro was most effective at factor recovery for the NEO, atomic was most effective for factor recovery of the HEXACO. It would be early to speculate about the reasons that different encoding approaches are more or less effective for different assessments, but as more applications appear, we expect a clearer picture regarding the effectiveness of atomic and macro encoding and which traits are more or less well represented in LLM encodings will begin to emerge.

Consistent with earlier work on pseudo-item discrimination, our expectations were that the two forms of discrimination are distinct quantities but would be informative about each other. Exploring corresponding factor loading correlations indeed revealed substantial associations, albeit they could not be considered equivalent. These results suggest that the embedding can serve as a substitute for its empirical response counterpart in the early stages of scale development. Our earlier suggestion appears reasonable in the traditional scale development workflow; checking the expected empirical factor structure *after* item pool creation but *before* actual data collection using PFA and embeddings can yield valuable insights that will enhance scale design.

Shiny Pseudo Factor Analysis application

The models and methods presented here rely on multiple programming languages (e.g., R and Python) and methodologies that may not be immediately familiar to applied psychologists and practitioners. To facilitate the application of embedding psychometrics to this audience, including PFA and pseudo-discrimination, we developed a shiny app freely accessible at http://drdami.shinyapps.io/nlp-psych1/.

Limitations and Suggestions for Future Research

The ideas in this article are based on a substitutability assumption, namely, that item embeddings can serve a similar purpose to actual response vectors during scale development. The key limitation of the proposed approach is the extent to which the cosine similarity measures are, in fact, approximating the empirical correlation between the corresponding pair of items. To the extent that LLMs represent a global encoding based on large corpora of human natural language training data and item responses represent the aggregated individual encodings of similar individuals to those that created the LLM training data, we anticipate the substitutability

assumption is reasonable. Our results offer reason for optimism. However, it will be important to check how this generalizes to other measurement tools.

An important point is that we cannot easily adapt the population used to generate the LLM embeddings, and it is unclear how similar these two data-generating populations need to be (content generators and test responders) for the results to produce similar (or dissimilar) factor structures. Assuming the training data that LLM distributions are based on are representative of the broad collection of natural language available online, the embeddings would reflect that particular distribution. So, as a subsample of real human test respondents differs more and more from that "internet corpus population", then the embeddings may be less representative. At all times, therefore, we acknowledge and recommend that the ultimate criterion for structure and discrimination is empirical discrimination in the target population.

A second line of research that would push embedding psychometrics forward would be to generalize the measurement modeling approaches presented here to full structural models and even network models. Some recent work has been done in these directions (see Feraco & Toffalini, 2025, and Russel-Lalasandra et al., 2024). Any model based on observed item covariance can, in principle, be fit to a matrix of embedding similarities. For example, if PFA is used to estimate latent pseudo-scores in the measurement model, then those latent pseudo-constructs can be modeled together to uncover the pseudo-structural model. This pseudo-structural approach is building on early work with the Semantic Theory of Survey Response (STSR: Arnulf et al. 2021).

Finally, we expect that the fundamental debate on the merits of out-of-the-box models versus fine-tuned models will be particularly important in embedding psychometrics. In these studies, we incorporated one fine-tuned psychology model that performed reasonably well,

although not as well as the best base models we studied. Fine-tuning LLM sentence transformers on psychometrically relevant content may well improve the accuracy of the sorts of results that we have presented here and should be investigated in future research.

References

Achenbach, T. M. (2021). Hierarchical dimensional models of psychopathology: Yes, but…. *World Psychiatry*, *20*(1), 64.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. Psychological Monographs, 47(1), i–171. https://doi.org/10.1037/h0093360

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Nimon, K. F. (2021). Editorial: Semantic algorithms in the assessment of attitudes and personality. *Frontiers in Psychology*, 12, Article 720559. https://doi.org/10.3389/fpsyg.2021.720559

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, 11(2), 150-166. DOI: 10.1177/1088868306294907

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*.

Cer, D., Yang, Y., Kong, S.-Y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for English. In E. Blanco & W. Lu (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169-174). Association for Computational Linguistics. doi: 10.18653/v1/D18-2029

Christensen, A. P., Garrido, L. E., & Golino, H. (2023). Unique variable analysis: A network psychometrics method to detect local dependence. Multivariate Behavioral Research, 58(6), 1165–1182. https://doi.org/10.1080/00273171.2023.2194606

Cutler, A., & Condon, D. (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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805

D'Urso, E. D., Tijmstra, J., Vermunt, J. K., & De Roover, K. (2023). Awareness is bliss: How acquiescence affects exploratory factor analysis. *Educational and Psychological Measurement*, 83(3), 433-472. doi: 10.1177/00131644221089857

Fyffe, S., Lee, P., & Kaplan, S. (2023). "Transforming" personality scale development: Illustrating the potential of state-of-the-art natural language processing. *Organizational Research Methods*. Advance online publication. https://doi.org/10.1177/10944281231155772

Feraco, T., & Toffalini, E. (2025). Sembeddings: How to evaluate model misfit before data collection using large-language models. *Frontiers in Psychology*, 15, Article 1433339. https://doi.org/10.3389/fpsyg.2024.1433339

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7(1), 7-28.

Guenole, N., Samo, A., & Sun, T. 2024. Pseudo-Discrimination Parameters from Language Embeddings. Open Science Framework. https://doi.org/10.31234/osf.io/9a4qx

Hernandez, I., & Nie, W. (2022). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12543

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2), 749-772. DOI: 10.1007/s11336-021-09823-9.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78-89. https://doi.org/10.1016/j.jrp.2014.05.003

Lambert, L. S., & Newman, D. A. (2023). Construct development and validation in three practical steps: Recommendations for reviewers, editors, and authors*. Organizational Research Methods, 26(4), 574–607. https://doi.org/10.1177/10944281221115374

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. American Elsevier Publishing Company.

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, 2(2), 57-64. https://doi.org/10.1027/1614-2241.2.2.57

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://arxiv.org/abs/1907.11692

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. Psychological Assessment, 22(3), 471–491. https://doi.org/10.1037/a0019227

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C.,

Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. European Journal of Personality, 34(6), 1175–1201. https://doi.org/10.1002/per.2311

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, 25(1), 30-45. https://doi.org/10.1037/met0000220

Russell-Lasalandra, L. L., Christensen, A. P., & Golino, H. (2024). Generative psychometrics via AI-GENIE: Automatic item generation and validation via network-integrated evaluation. Open Science Framework. https://doi.org/10.31234/osf.io/fgbj4

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2014). Measurement theory in action: Case studies and exercises (2nd ed.). Routledge/Taylor & Francis Group.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*. DOI: 10.5555/3495724.3497138

Thielmann, I., Moshagen, M., Hilbig, B., & Zettler, I. (2022). On the comparability of basic personality models: Meta-analytic correspondence, scope, and orthogonality of the Big Five and HEXACO dimensions. European Journal of Personality, 36(6), 870-900.

Tucker, L. (1951). A Method for Synthesis of Factor Analysis Studies.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1706.03762

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33, 5776-5788. DOI: https://doi.org/10.48550/arXiv.2002.10957

Woo, S. E., Tay, L., & Oswald, F. (2024). Artificial intelligence, machine learning, and big data: Improvements to the science of people at work and applications to practice. *Personnel Psychology*. Advance online publication. https://doi.org/10.1111/peps.12643

Wulff, D. U., & Mata, R. (2023). Automated jingle–jangle detection: Using embeddings to tackle taxonomic incommensurability. PsyArXiv. https://doi.org/10.31234/osf.io/9h7aw

Table 1. NEO embedding factor structures using transformer averages for macro encoding.

		O	C	E	A	N
NEO MEAN TRANSFORMER	Ima	.65	.15	.01	-.10	.10
	Art	.86	-.26	.16	.19	-.22
	Emo	.42	-.22	-.11	.27	.53
	Adv	.54	.16	.41	-.24	-.05
	Int	.55	-.01	.16	-.03	.12
	Lib	.53	-.01	-.09	.41	-.10
	Sel	-.19	.65	-.05	.13	.34
	Ord	.24	.45	.30	-.11	-.06
	Dut	-.12	.72	-.12	.57	-.27
	Ach	-.13	.79	.11	.08	.02
	Sel	-.19	.85	.28	-.27	.14
	Cau	.25	.50	.10	-.07	.10
	Fri	-.34	-.18	.94	.26	.18
	Gre	.20	-.03	.84	-.04	-.14
	Ass	-.19	.38	.44	.30	-.05
	Act	.09	.50	.37	-.26	.18
	Exc	.54	.02	.40	-.08	-.08
	Che	.27	-.07	.33	.20	.09
	Tru	.10	-.12	.13	.66	.06
	Mor	-.07	.49	-.05	.78	-.27
	Alt	.01	-.08	.24	.76	-.03
	Co	.13	.07	-.04	.48	.22
	Mod	.10	.17	.16	.43	.01
	Sym	.27	-.21	-.05	.79	.06
	Anx	.11	-.03	-.06	-.01	.89
	Ang	.11	.09	-.15	.08	.74
	Dep	.12	.00	.10	.01	.60
	Sel	-.23	-.08	.56	.05	.55
	Imm	.32	.27	-.04	-.01	.28
	Vul	.03	.05	-.18	-.03	1.02

Notes. O=Openness; C=Conscientiousness; E=Extraversion; Agreeableness; N=Neuroticism; Ima=Imagination; Art=Artistic Interests; Emo=Emotionality; Adv=Adventurousness; Int=Intellect; Lib=Liberalism; Sel=Self.Efficacy; Ord=Orderliness; Dut=Dutifulness; Ach=Achievement Striving; Sel=Self Discipline; Cau= Cautiousness; Fri=Friendliness; Gre=Gregariousness; Ass=Assertiveness; Act=Activity Level; Exc=Excitement Seeking; Che=Cheerfulness; Tru=Trust; Mor=Morality; Alt=Altruism; Co=Cooperation; Mod=Modesty; Sym=Sympathy; Anx=Anxiety; Ang=Anger; Dep=Depression; Sel=Self Consciousness; Imm=Immoderation; Vul=Vulnerability.

Table 2. Congruence between NEO empirical and NEO embedding transformer models.

	Atomic congruence					Atomic reverse congruence					Macro congruence					M
Model	O	C	E	A	N	O	C	E	A	N	O	C	E	A	N	M
D_RoBERTa	.68	.77	.66	.73	.76	.65	.84	.77	.43	.86	.76	.80	.57	.73	.73	.72
MiniLM	.77	.71	NA	NA	.67	.47	NA	.35	.01	NA	.77	.77	.41	.75	.71	.58
MPNet	.57	.77	.62	.74	.72	.78	.46	.74	.62	.83	.52	.80	.50	.73	.68	.67
T5	.76	.70	.63	.74	.74	.66	.55	.70	.47	.89	.69	.65	.48	.65	.44	.65
Wulf Matta	NA	.73	.71	NA	.75	.79	.79	.87	.91	.79	.80	.57	.60	.84	.63	.75
USE-DAN	NA	.75	NA	.72	.73	.64	.56	.65	.57	.52	NA	.75	.45	.72	NA	.64
Transformers	NA	.76	.68	.75	NA	.67	.77	.78	.76	.89	.76	.77	.61	.79	.74	.75
Overall M	.70	.74	.66	.74	.73	.67	.66	.69	.54	.80	.72	.73	.52	.74	.66	.69

Notes. D_Roberta = all-distilroberta-base-v2; MiniLM = MiniLM-L6-v2; MPNet = MPNet base v2; W&M = Wulff & Mata (2023) fine-tuned transformer; USE-DAN = Universal Sentence Encoder with Deep Average Networks; Transformers = Mean of all transformers; Overall M = Mean of all studied encoding approaches; O = Openness; C = Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism. NA = factor not recovered by LLM factor analysis for this method and transformer combination.

Table 3. Empirical Correlations between NEO empirical and NEO embedding transformer models.

		Atomic congruence					Atomic reverse congruence					Macro congruence				
		AO	AC	AE	AA	AN	RO	RC	RE	RA	RN	OO	OC	OE	OA	ON
Empirical	EO	.20	-.37	-.21	.29	.14	.58	-.02	.42	.18	.00	.74	-.40	-.20	.00	-.17
	EC	-.45	.72	.01	-.15	-.56	-.09	.73	-.04	.00	-.55	-.38	.72	.11	-.11	-.33
	EE	-.46	-.28	.64	.07	-.22	.33	.15	.76	-.12	-.33	.04	-.20	.56	.05	-.38
	EA	-.45	-.01	-.20	.67	.21	-.14	.35	-.34	.75	-.38	-.19	-.07	-.17	.74	-.30
	EN	.71	-.35	-.22	-.23	.56	.08	-.56	-.22	-.32	.89	.10	-.40	-.29	-.15	.71

Notes. A =. Atomic; R = Atomic Reversed; M = Macro/One-pop; O = Openness; C = Conscientiousness; E=Extraversion; A=Agreeableness; N=Neuroticism.

Table 4. HEXACO embedding factor structures using transformer averages for atomic encoding.

	H	E	X	A	C	O
Sinc	.75	-.08	.10	.16	.07	-.14
Fair	.67	.19	-.38	.08	.27	.07
Gree	.58	-.16	.15	-.18	.17	.26
Mode	.68	-.14	.17	.03	.02	.08
Fear	.12	.48	.13	-.14	.05	.30
Anxi	-.36	.54	.13	.31	.25	.15
Depe	.25	.24	.13	.32	-.21	.15
Sent	-.12	.46	.00	.46	-.25	.48
Expr	.03	.03	.95	-.02	-.13	.01
SocB	.17	.03	.86	-.11	-.10	.03
Soci	.06	.04	.60	.18	-.09	.10
Live	-.20	.24	.58	.11	.11	.03
Forg	.30	.23	-.20	.78	-.06	-.09
Gent	.13	.15	.07	.84	.02	-.25
Flex	.06	.23	.14	.72	.02	-.18
Pati	-.35	.34	.20	.71	.20	-.12
Orga	.01	-.16	.02	-.01	.88	.05
Dili	.06	-.12	.12	.03	.82	-.12
Perf	.12	-.31	-.19	.09	.97	.10
Prud	.33	.34	-.20	-.01	.49	.01
AesA	-.01	.28	-.22	-.04	-.07	1.01[2]
Inqu	-.02	.28	.07	-.24	.09	.77
Crea	.01	.31	.22	-.23	.13	.53
Unco	.33	.19	.13	-.15	-.05	.50

Notes. H=Honesty; E=Emotional Stability; X=Extraversion; A=Agreeabelness; C=Conscientiousness; O=Openness; Sinc=Sincerity; Fair=Fairness; Gree=Greeed Avoidance; Mode=Modesty; Fear=Fearfulness; Anxi=Anxiety; Depe=Dependence; Sent=Sentimentality; Expr=Expressiveness; SocB=Social Boldness; Soci=Socability; Live=Liveliness; Forg=Forgiveness; Gent=Gentleness; Flex=Flexibility; Pati=Patience; Orga=Organization; Dili=Diligence; Perf=Perfectionism; Prud=Prudence; AesA=Aesthetic Appreciation; Inqu= Inquisitiveness; Crea= Creativity; Unco= Unconventionality.

[2] We are using an oblique (target) rotation, which is why we can end up with loadings > 1. Also, we factor analyzed a semantic (cosine) similarity matrix, which is not completely standardized like a traditional (empirical) correlation matrix.

Table 5. Congruence between HEXACO empirical and HEXACO embedding transformer models.

	Atomic congruence						Atomic reverse congruence						Macro congruence						M
Model	H	E	X	A	C	O	H	E	X	A	C	O	H	E	X	A	C	O	M
D_RoBERTa	.75	.49	.62	.74	.91	.75	NA	NA	.60	.64	.83	.51	.82	NA	.64	NA	.89	.71	.71
MiniLM	.70	.46	.73	.73	.88	.66	NA	NA	.55	.47	.57	.59	.73	.46	.79	.72	.92	.68	.67
MPNet	.72	.47	.43	.87	.86	.67	NA	.60	.68	.65	.86	NA	.75	.55	.61	.77	.89	.58	.69
T5	.77	NA	NA	.80	.89	.63	NA	NA	.71	.65	.71	.42	.75	NA	NA	.83	.90	.71	.73
Wulf & Matta	.73	.66	.76	.83	.88	.73	NA	.63	.62	NA	.89	.68	.81	.56	.83	.91	.86	.61	.75
USE-DAN	.79	.47	.64	.75	.86	.48	NA	NA	.56	.44	.68	.69	.70	.58	NA	.82	.86	NA	.67
Transformers	.87	.73	.81	.93	.91	.77	NA	.65	NA	.72	.85	.61	.84	.48	.80	.85	.92	.68	.78
Overall M	.76	.55	.67	.81	.88	.67	NA	.63	.62	.60	.77	.58	.77	.53	.73	.82	.89	.66	.70

Notes. D_Roberta = all-distilroberta-base-v2; MiniLM = MiniLM-L6-v2; MPNet = MPNet base v2; W&M = Wulff & Mata (2023) fine-tuned transformer ; USE-DAN = Universal Sentence Encoder with Deep Average Networks; Transformers = Mean of all transformers; M = Mean of all studied encoding approaches; H=Honesty-Humility; E=Emotionality; X=Extraversion; A=Agreeableness; C=Conscientiousness; O=Openness. NA = factor not recovered by LLM factor analysis for this method and transformer combination.

Table 6. Correlations between HEXACO empirical and HEXACO embedding transformer models.

		Atomic congruence						Atomic reverse congruence						Macro congruence					
		AH	AE	AX	AA	AC	AO	RH	RE	RU1	RA	RC	RO	OH	OE	OX	OA	OC	OO
Empirical	EH	.82	-.19	-.41	-.10	.04	-.13	-.45	-.22	.39	-.02	.06	-.16	.78	-.36	-.38	-.11	.03	-.08
	EE	-.07	.67	.00	-.25	.00	-.16	-.09	.62	-.20	-.14	-.10	-.25	-.16	.38	-.10	.11	-.21	.14
	EX	-.19	.12	.76	-.15	-.41	-.12	.47	.11	-.30	.10	-.24	.23	-.22	-.15	.77	-.08	-.35	-.07
	EA	-.06	-.29	-.19	.85	-.27	-.19	.12	-.18	.14	.74	-.29	-.33	-.05	.14	-.17	.74	-.18	-.39
	EC	.07	.04	-.44	-.30	.88	-.15	-.28	-.05	.48	-.39	.82	-.22	.04	-.39	-.43	-.22	.90	-.15
	EO	-.12	-.17	-.16	-.20	-.25	.77	.14	-.22	-.05	-.23	-.25	.70	-.07	.01	-.04	-.40	-.14	.71

Notes. A = Atomic; R = Atomic Reversed; M = Macro/One-pop; H=Honesty-Humility; E=Emotionality; X=Extraversion; A=Agreeableness; C= Conscientiousness; O=Openness; U=Unassigned.

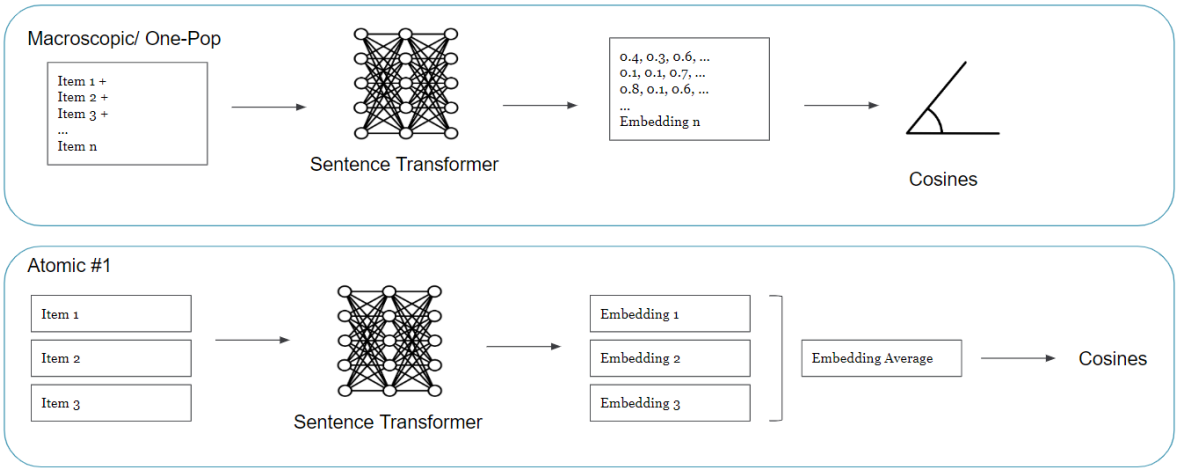

Figure 1. Visual depiction of Natural Language Processing workflow

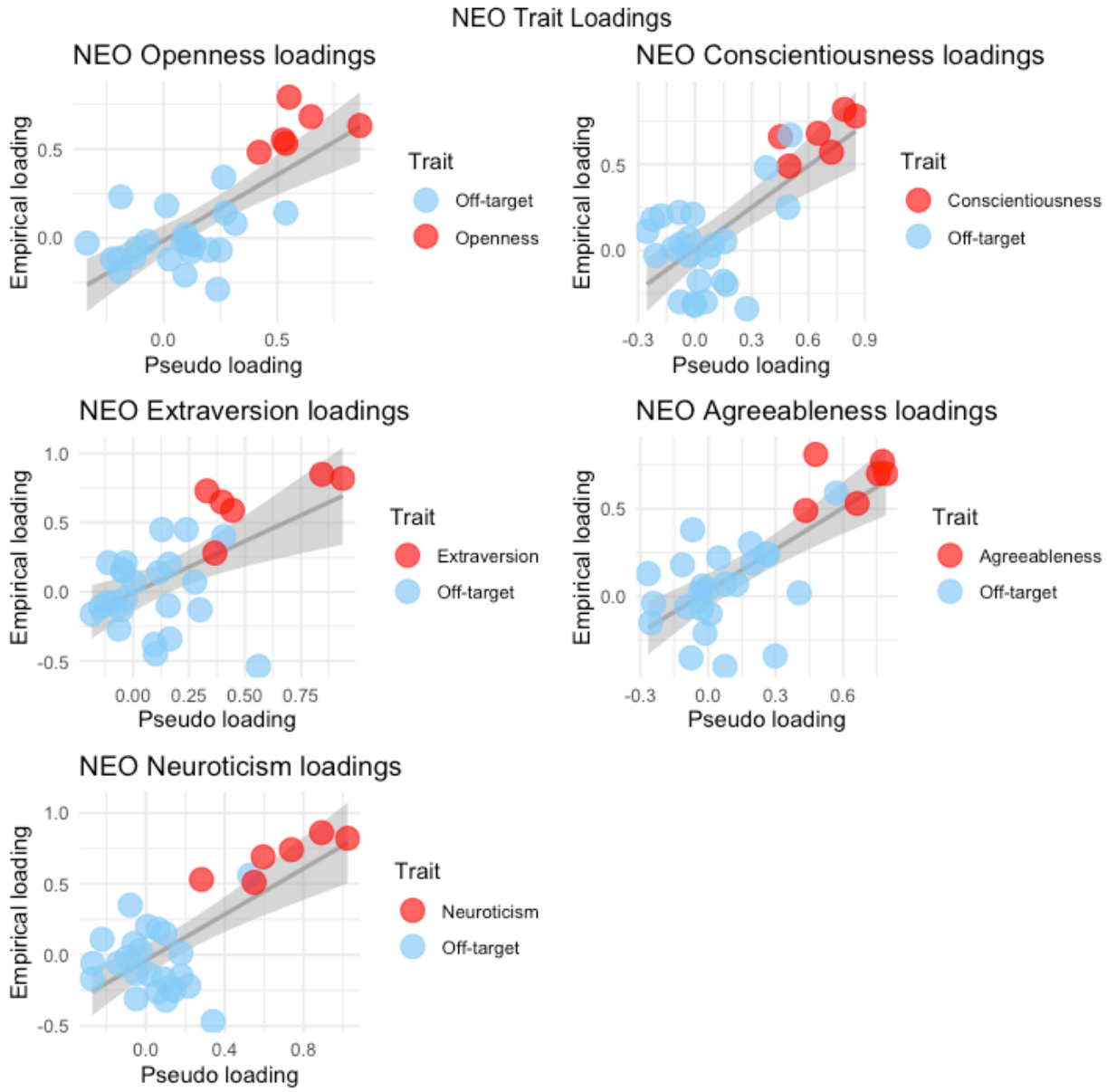

Figure 2. NEO Macro embedding loadings versus empirical loading plots

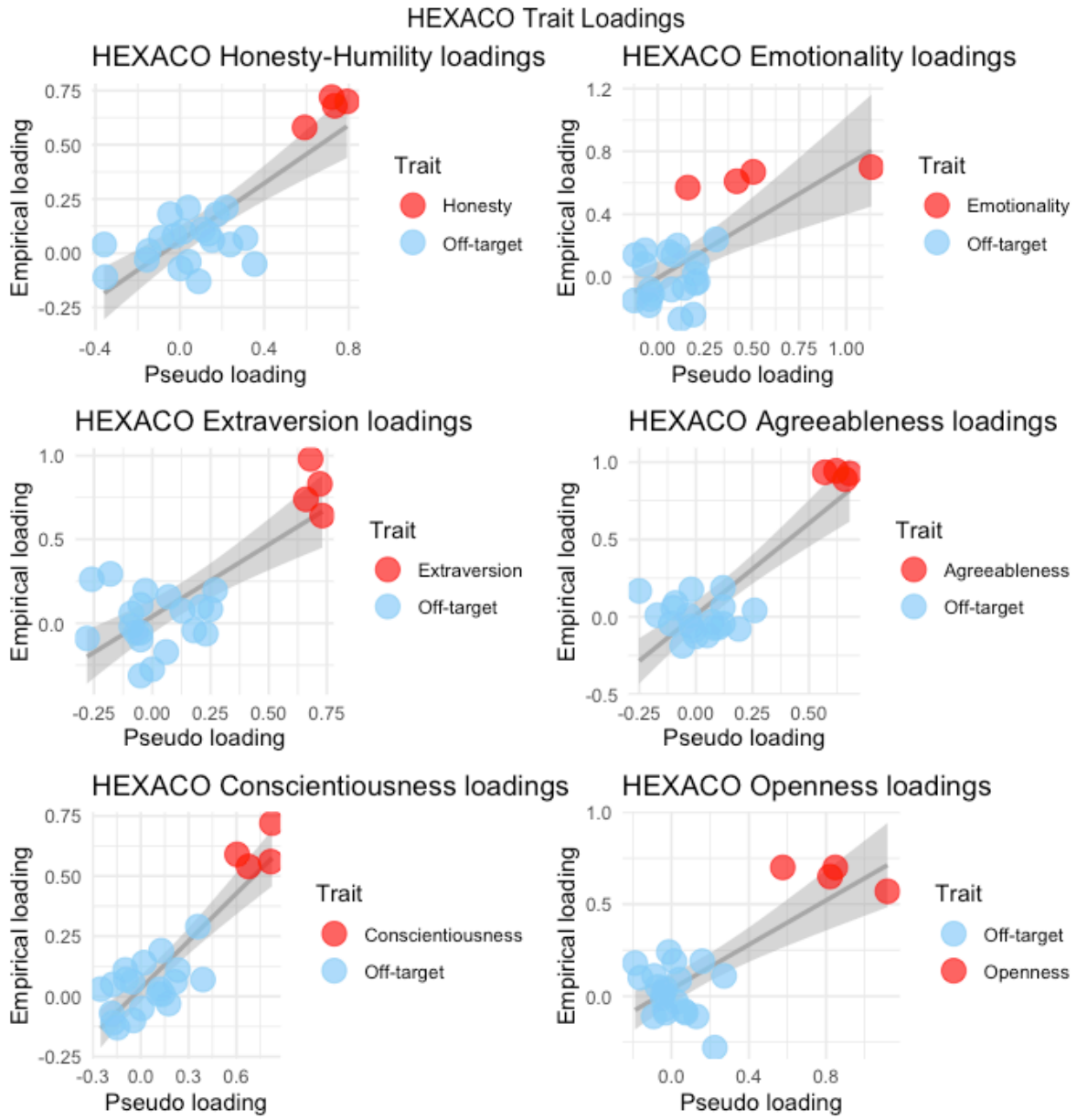

Figure 3. HEXACO Atomic embedding loadings versus empirical loading plots