A Large Foundation Model for Assessing Spatially Distributed Personality Traits

Avi Bleiweiss
BShalem Research, CA, USA

Abstract

We explored emulating textually encoded personality information in a large language model. Given its predominant empirical validation, we chose the five-factor model of personality compiled for a broad range of natural languages. Our study assessed personality traits from a multicultural viewpoint over a diverse set of thirty universal contexts. Thus, contributing to the wider comprehension of generalizing relationships among personality traits across cultures. We administered psychometric tests to the language model, examined links between location and personality, and cross validated measures at various levels of trait hierarchy.

Keywords: Five-Factor Model, Personality Traits, Large Language Model, Textual Entailment

1. Introduction

Assessment of individual differences in personality traits is perceived as one of the hallmarks of psychological research (McCrae and Costa, 1999). Personality is most commonly measured using the five-factor model (FFM; McCrae, 2010), a structured concept of traits that represent regularities of thoughts, feelings, and behaviors in humans with descriptive phrases. Expressed in five broad trait disciplines— agreeableness (A), conscientiousness (C), extraversion (E), neuroticism (N), and openness (O)— FFM forms the basis of a personality assessment system. The International Personality Item Pool (IPIP) is a public domain collection of items for use in personality tests. [1] Over the years, items from the IPIP have been consistently transcribed from English to a wide variety of more than 25 other languages (Goldberg et al., 2006).

In our work, we used one of the more comprehensive public-domain representation of the FFM personality inventories, IPIP-NEO-120 (Johnson, 2014; Kajonius and Johnson, 2019). IPIP-NEO-120 with 120 personality items is structurally robust for the benefit of research and practice in personality assessment. The IPIP-NEO-120 is a five-choice answer questionnaire that renders a top-down approach to a hierarchical personality assessment and as such, it gains a broader informative list of traits that recurred in personality measures. To this extent, IPIP-NEO-120 includes six lower-order facet traits equally distributed in each of the factor domains– A, C, E, N, and O— with a total of thirty facet traits. Domains are thus multifaceted collections of specific behaviors that might be grouped in many different ways.

The English edition of the five-way answer choices in IPIP-NEO-120 questionnaire includes: (i) `Very Inaccurate`, (ii) `Moderately Inaccurate`, (iii) `Neither Accurate Nor`

1. https://ipip.ori.org

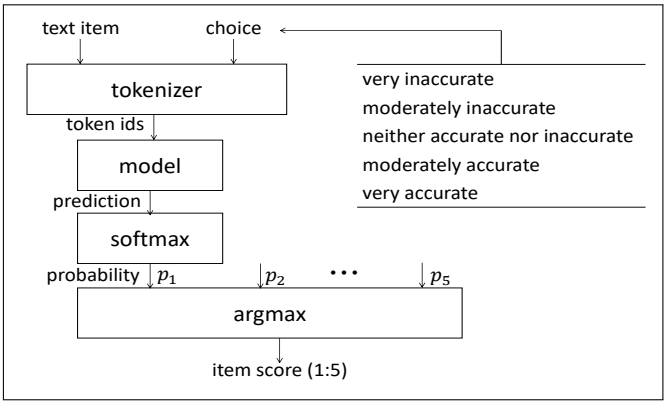


Figure 1: Overview of our textual entailment framework. We pair a text item, the hypothesis, with each of the five answer choices, the premise, and derive pair probabilities $p_{1:5}$ to assess an item score. The text item and answer choices are language matching.

`Inaccurate`, (iv) `Moderately Accurate`, and (v) `Very Accurate`. Answers are scored on a five-point scale with values either ascending from 1 to 5 or descending from 5 to 1 based on the item `keyed` attribute assignment, plus or minus, respectively.

We considered conducting personality assessment on a large language model (LLM), but rather than a text generative model we cast the task as a zero-shot multiple-choice question answering problem. Text generation would require translating dynamic prompts to tens of languages and likely render our study impractical. Following Yin et al. (2019), we used a textual entailment paradigm providing the LLM with a premise-hypothesis reasoning clause and performed multi-label sequence-pair classification (Figure 1). A sample of a text item from each of the IPIP-NEO-120 factor domains along with a facet identifier and the keyed scoring direction are shown in Table 1. Whereas a personality trait is assessed based on the highest entailment probability over five answer choices. In this study, we experimented with an open source LLM and avoided commercial models such as GPT4 (OpenAI, 2024) due to cost constraint and API access that might hinder the option to finetune or alter the model. Although we expected GPT4 to perform well in the multilingual settings we offer.

Table 1: A handful of item queries in English from the IPIP-NEO-120 questionnaire. Showing one query from each of the five-factor domains along with the item facet and the keyed scoring direction.

Query	Domain	Facet	Keyed
Believe that I am better than others	A	5	Minus
Jump into things without thinking	C	6	Minus
Feel comfortable around people	E	1	Plus
Find it difficult to approach others	N	4	Plus
Experience my emotions intensely	O	3	Plus

In this paper, we administered the personality questionnaire to BART (Lewis et al., 2020), a sequence-to-sequence language model trained as a denoising autoencoder for zero-shot learning. Our framework extended the personality test to a broad range of thirty

natural languages listed in Table 2 along with names and two-character codes (ISO-639-1). The availability of multilingual IPIP-NEO-120 versions was facilitated by a multi-decade collaborative research of validated transcription over countries and continents. [2] In theory, given a stable LLM platform with frozen parameters, we would have expected an identical personality score for iterating the psychometric questionnaire across thirty ordinary languages. Thus, an LLM that is emotionally stateless unlike a human should give a single answer to a personality test, regardless of the language in which it was prompted. However, this goal is still elusive as Natural Language Inference (NLI) LLMs pretrained on non-English languages are scarce and the assessments of certain traits are likely to differ.

To the extent of our knowledge, contrastive evaluation of personality using the five-factor model in a multicultural LLM context has not been explored in prior work. Our paper contributes to the increased interest in the roles of culture and location learning for understanding personality.

2. Background

The topic of personalizing LLMs has recently received widened attention by the research community. Work to probe psychological personality in LLMs use FFM, also known as the Big Five model, one of the most prevalent and consistent questionnaire to assess the quality of LLM outputs. A common practice on generative LLMs injects knowledge about the personality trait in the prompt and guides the LLM toward matching target traits. However, prompts are sensitive to model changes and even small alterations may yield considerably different outputs.

We briefly survey recent work on personality emulation in LLMs, highlighting a subset of used models and the prompting strategy when relevant. Serapio-García et al. (2023) introduced an end-to-end method to both administer and validate personality assessment for architectures from the (PaLM; Google AI, 2023) family. Their method leverages the LLM ability to score completions of the provided prompt. In another study, Jiang et al. (2023) draw FFM scores from BART and GPT3.5, and contrast performance between opaque and instruction-guided LLMs. Whereas V Ganesan et al. (2023) investigate zero-shot learning on GPT3 to estimate personality traits from social media posts. They observed comparable performance to a pretrained lexical model when infusing prompt knowledge about the trait. Sorokovikova et al. (2024) extended Big Five assessments to GPT4 and demonstrated differing tendencies for traits when introducing small variation of prompt text and generation parameters.

3. Experiments

In our experiments, we used the BART large checkpoint finetuned on the multi-genre natural language inference (MNLI) dataset (Williams et al., 2018). [3] We used softmax to compute the probability of entailment for every personality item in IPIP-NEO-120 paired with each of the five answer choices. The score $[1.00, 5.00]$ of the answer with the highest probability is assigned to the item. Iterating over thirty languages, we ran inference locally and entirely

2. https://github.com/Alheimsins/b5-johnson-120-ipip-neo-pi-r/tree/main/data
3. https://huggingface.co/facebook/bart-large-mnli

Table 2: Language-specific personality mean and standard deviation across the five IPIP-NEO-120 domains— A, C, E, N, O.

Language	Code	A		C		E		N		O	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Arabic	ar	3.83	1.86	3.17	2.04	2.00	1.77	2.17	1.86	3.00	2.04
Cantonese	cn	2.96	2.01	2.67	1.95	3.67	1.93	3.33	2.01	2.71	1.99
Danish	da	3.83	1.86	3.17	2.04	2.00	1.77	2.17	1.86	3.00	2.04
Deutsch	de	3.42	0.93	3.08	1.02	2.50	0.88	2.58	0.93	3.00	1.02
English	en	3.46	1.10	3.92	0.83	3.46	0.98	2.08	0.41	3.46	1.02
Spanish	es	2.58	0.93	2.92	1.02	3.50	0.88	3.42	0.93	3.00	1.02
Estonian	et	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Finnish	fi	3.83	1.86	3.17	2.04	2.00	1.77	2.17	1.86	3.00	2.04
French	fr	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Hebrew	he	3.83	1.86	3.17	2.04	2.00	1.77	2.17	1.86	3.00	2.04
Hindi	hi	2.17	1.86	2.83	2.04	4.00	1.77	**3.83**	1.86	3.00	2.04
Croatian	hr	2.17	1.86	2.83	2.04	4.00	1.77	**3.83**	1.86	3.00	2.04
Hungarian	hu	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Indonesian	id	2.58	0.93	2.92	1.02	3.50	0.88	3.42	0.93	3.00	1.02
Icelandic	is	3.42	0.93	3.08	1.02	2.50	0.88	2.58	0.93	3.00	1.02
Italian	it	3.42	0.93	3.08	1.02	2.50	0.88	2.58	0.93	3.00	1.02
Japanese	ja	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Korean	ko	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Mandarin	mn	3.00	1.25	3.00	1.53	3.29	1.37	3.58	1.41	3.50	1.14
Dutch	nl	2.17	1.86	2.83	2.04	4.00	1.77	**3.83**	1.86	3.00	2.04
Norwegian	no	3.42	0.93	3.08	1.02	2.50	0.88	2.58	0.93	3.00	1.02
Polish	pl	2.58	1.89	2.88	2.01	**4.04**	1.71	3.79	1.69	**3.67**	1.83
Portuguese	pt	2.83	1.09	3.71	1.20	3.58	1.25	3.04	1.33	3.29	1.16
Romanian	ro	3.08	1.21	3.29	1.57	2.79	1.28	2.88	1.33	3.08	1.06
Russian	ru	3.83	1.86	3.00	2.04	2.00	1.77	2.17	1.86	3.00	2.04
Albanian	sq	3.33	0.96	**3.96**	0.69	3.38	0.97	2.92	1.02	3.08	1.06
Swedish	sv	2.92	0.41	2.79	0.59	3.17	0.82	3.04	0.20	2.96	0.75
Thai	th	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Ukrainian	uk	**3.92**	1.74	3.17	2.04	1.71	1.52	2.21	1.84	3.00	2.04
Urdu	ur	3.08	1.56	2.88	1.65	2.29	1.43	2.83	1.34	3.04	1.43

on the CPU with up to four workers, while not exceeding 2.5GB of system memory. Our running time last four seconds on average for each item in IPIP-NEO-120.

Trait Hierarchy We conducted our experiments at three FFM hierarchical levels, including questionnaire, domain, and facet. In its original form, IPIP-NEO-120 is represented by a single descriptor object. We further split the questionnaire data into five domain entities, each with 24 trait items, and six facet constructs of 20 items each. Thus, each language expresses a culture using twelve descriptors containing both queries and answer

Table 3: External baseline comparison of personality assessment in English NLI-BART.

System	A		C		E		N		O	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Jiang et al. (2023)	2.17	1.82	2.83	1.99	4.00	1.73	3.83	1.82	3.00	1.82
Ours	3.46	1.10	3.92	0.83	3.46	0.98	2.08	0.41	3.46	1.02

Table 4: Statistical distributions of average assessment scores across all languages for each domain.

Domain	min	max	μ	σ
A	2.17	3.92	**3.12**	0.51
C	2.67	3.96	3.09	0.30
E	1.71	4.04	2.95	0.69
N	2.08	3.83	2.91	0.56
O	2.71	3.67	3.06	0.18

choices that we fed our LLM to perform personality assessment. We follow with our domain evaluation and report questionnaire and facet personality assessments in Appendices A and B, respectively. Unless otherwise noted, the language personality rates are the average of FFM item scores across any of the entire IPIP-NEO-120 questionnaire, one of the five domains, or one of the six facets.

Domain Evaluation In Table 2, we show language specific scores of personality assessment at the domain level. The two-character language codes and domain initials are sorted alphabetically for clarity. We provide both mean μ and standard deviation σ metrics for each language-domain pair. Polish had the all-around top personality score in the E and O domains with 4.04 and 3.67, respectively. While Ukrainian leads the E domain with 3.92, Albanian 3.96 in C, and Croatian, Dutch, and Hindi are tied at 3.83 in N. Although not at the highest rank, English signifies the most domain-stable with a 3.46 score in three out-of-five domains. Table 3 provides baseline performance comparison to Jiang et al. (2023) English personality scores in BART. Their 3.17 domain average is fairly consistent with our 3.28 measure. In Table 4, we present complementary statistical distributions of average assessment measures for each domain and across all languages. We anticipated varied personality scores for different cultures, but surprisingly domain-to-domain average rates appear mostly concurring with a tiny deviation of 0.09. Additionally, we observed several occurrences of zero standard deviation owing to a relatively modest trait sample in a domain.

Human and LLM One of the largest cross-cultural research on personality assessment (Schmitt et al., 2007) used the Big Five model and administered a self-report survey to thousands of individuals representing 56 nations and 30 languages. In a similar study, Kajonius and Giolla (2017) applied the IPIP-NEO-120 questionnaire to assess personality characteristics from a large sample of individuals residing in twenty two countries. Overall, the countries surveyed are a matching subset to the origins of our thirty languages. In Table 5, we provide cross-country and cross-lingual mean scores for the five trait domain

factors comparing human to BART measures. The Pearson correlation coefficients when pairing BART with Schmitt et al. (2007) and Kajonius and Giolla (2017) are -0.79 and -0.49, respectively. Albeit the inverse correlation, these figures suggest a sufficiently strong human-LLM relationship. As expected, both human and BART scores are fairly balanced across domains with a sample deviation of under five percent.

Table 5: Comparing cross-country and cross-lingual average scores for domain-level personality trait assessment administered to humans and BART, respectively.

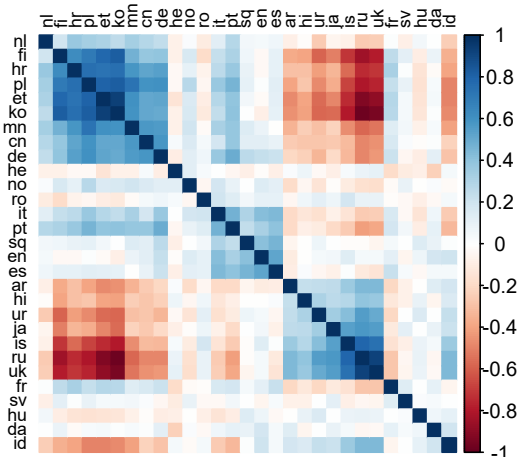
Model	Study	A	C	E	N	O
Human	Schmitt et al. (2007)	2.39	2.38	2.43	2.52	2.46
	Kajonius and Giolla (2017)	4.25	4.50	4.45	4.50	4.05
BART	Ours	3.12	3.09	2.95	2.91	3.06

Language Correlation Using hierarchical clustering, we show in Figure 2(a) all language-language correlations on personality scores at the questionnaire level. We note that all-pair correlations are rendered among twenty nine languages, excluding the extremely low resourced Thai language that yielded an identical personality score of 3.0 for all the questionnaire queries, hence leading to an inconsistent zero variance. Most prominent are two distinct cultural clusters of the highest positive relationships that are drawn symmetrically along the correlogram diagonal (dark blue): $C_1 = (\text{Finnish}, \text{Croatian}, \text{Polish}, \text{Estonian}, \text{Korean})$, with a Pearson correlation that ranges from 0.63 for Finnish-Croatian to 0.94 for Estonian-Korean, and $C_2 = (\text{Icelandic}, \text{Russian}, \text{Ukrainian})$ with 0.75 for Icelandic-Ukrainian to 0.93 for Russian-Ukrainian. On the other hand, the most negative relationships (dark red) are interpreted asymmetrically off-diagonal for the same clusters and span from -0.56 for Icelandic-Croatian to -0.98 for Russian-Korean.

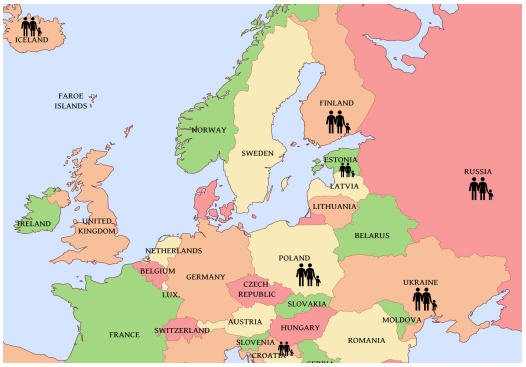
4. Discussion

Synthesizing personality traits in LLM outputs using FFM-based psychometric tests have proven empirically reliable. However, LLMs in prior monolingual work (Serapio-García et al., 2023) were assessed exclusively with English-language psychometric tests. Our study seeks to lay the foundation for better understanding multilingual interaction between LLMs. The Big Five model already possess cross-cultural generalization and each human language represents a distinct personality profile identified by an apparent culture resemblance. Non-Western languages thus express additional culture-specific dimensions for measuring personality in LLMs. It is a common practice to guide the LLM and generate optimized outputs consistent with an instruction. However, in a multicultural environment like ours, a unique prompting configuration for each ordinary language is required to ensure equivalence of the personality test.

Personality and Location Our language correlation experiments highlighted two cultural clusters ($C1, C2$) of enhanced personality similarity. The clusters extend countrywide and circumscribe seven nations of the Eastern-Europe region. Evidently two distinct subgroups of population are observed: (a) Poland, Ukraine, and Russia; and (b) Finland, Russia, and Estonia; each in close geographical proximity with a shared border (Figure 2(b)),



(*a*) All language-pair correlogram at the questionnaire hierarchy level (red and blue colors render negative and positive relationships, respectively).



(*b*) Personality traits differ by location in the Eastern-Europe region. Notable subgroups are (Poland, Ukraine, Russia) and (Finland, Russia, Estonia).

Figure 2: Language relationships (a) leading to location based personality (b).

thus suggesting regional differences in the distribution of personality traits and social characteristics.

Contextual Bias Our study assessed the monolingual performance of personality traits spanning a broad spectrum of typologically diverse thirty languages, while running all psychometric tests within the confines of a single monolingual language model. Using BART with a large size of pretraining English data, we anticipated uneven language performance in our downstream multilingual task. To this extent, this behavior might identify a cultural personality bias potentially remedied with an instruction-tuned LLM. While a prompting design such as in Sorokovikova et al. (2024) may optimize LLM outputs and improve performance of some languages, the labor cost of manual prompt translation of items across all languages is nonetheless significant.

Score Consistency To validate consistency of our assessed personality scores across IPIP-NEO-120 domains, we conducted in addition cross-hierarchy matching to both questionnaire (Appendix A) and facet (Appendix B) levels. Distinctly the Polish language remains leading on both questionnaire and facet_1 tiers with a 3.39 (Table 6) and 4.45 (Table 8) scores, respectively. We note that facet-to-facet personality assessment (Table 9) renders mostly concurring rates that only deviate by an inconsequential 0.09 measure.

5. Conclusion

In this study, we explored the uncharted dimension of location and its impact on individual psychology when administering FFM personality tests in an LLM. Using a broad range of thirty cultural contexts, we observed distinct language inclination for specific traits, as well

as cluster formation of related cultures. Future research includes configuring a prompt to differentiate personality traits based on geo-political communities.

6. Limitations

The FFM devises a lower facet scale to partition a domain of trait items. However, this results in facets with a small four-trait samples. To ensure a stable and more robust analysis we divided the entire IPIP-NEO-120 questionnaire into six facet clusters with twenty traits each. Our language correlation study, while pointing to cultural relationships on the merit of geographical proximity and social form, the generalizing of observations made in this paper requires FFM and LLM improvements. Choosing NLI BART as our central language model has seemingly hampered exploring instruction-tuned LLMs to their full extent for assessing personality traits. However, manually curating prompts to achieve desired LLM outputs is time consuming and difficult to manage for tens of languages. In contrast, barring low-resource languages, machine translated prompts is a more effective approach to consider.

7. Ethical Statement

We honor and support the NeurIPS Code of Ethics. Our questionnaire data was scraped from the internet and we ensured its translation validity. The IPIP-NEO-120 data has an MIT license we abide by and use it for research-only. Although we restructured the personality test data to improve LLM efficacy, this paper does not release new assets. Our study refrains from crowdsourcing or research with human subjects. The sole exception is in comparing our LLM personality measures to an average profile assessment of a large sample of humans cross-country, conducted by an external study (Table 5). Our findings of a plausible link between personality and physical location suggest a cultural relationships across national borders, however, we believe this has no negative societal impact.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

Lewis R. Goldberg, John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006. doi: https://doi.org/10.1016/j.jrp.2005.08.007.

Google AI. PaLM 2 technical report. Technical report, Google, 2023. https://arxiv.org/abs/2305.10403.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=I9xE1Jsjfx.

John A. Johnson. 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, 2014. doi: https://doi.org/10.1016/j.jrp.2014.05.003.

Petri Kajonius and E M Giolla. Personality traits across countries: Support for similarities rather than differences. *PLoS ONE*, 12(6), 2017. doi: https://doi.org/10.1371/journal.pone.0179646.

Petri J. Kajonius and John A. Johnson. Assessing the structure of the five factor model of personality (ipip-neo-120) in the public domain. *Europe's Journal of Psychology*, 15(2): 260–275, June 2019. doi: 10.5964/ejop.v15i2.1671.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703.

R. R. McCrae and Jr. Costa, P. T. A five-factor theory of personality. In L. A. Pervin and O. P. John, editors, *Handbook of personality: Theory and research*, pages 139–153. Guilford Press, New York, NY, 1999.

Robert R. McCrae. The place of the ffm in personality psychology. *Psychological Inquiry*, 21 (1):57–64, 2010. doi: 10.1080/10478401003648773.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2024. https://arxiv.org/abs/2303.08774.

David P. Schmitt, Jüri Allik, Robert R. McCrae, and Verónica Benet-Martínez. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, 38(2):173–212, 2007. doi: 10.1177/0022022106297299.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.

Aleksandra Sorokovikova, Sharwin Rezagholi, Natalia Fedorova, and Ivan Yamshchikov. LLMs simulate big5 personality traits: Further evidence. In Ameet Deshpande, EunJeong Hwang, Vishvak Murahari, Joon Sung Park, Diyi Yang, Ashish Sabharwal, Karthik Narasimhan, and Ashwin Kalyan, editors, *Personalization of Generative AI Systems (PERSONALIZE)*, pages 83–87, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.personalize-1.7.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. Systematic evaluation of GPT-3 for zero-shot personality estimation. In Jeremy Barnes, Orphée De Clercq, and Roman Klinger, editors, *Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wassa-1.34.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101.

Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Empirical Methods in Natural Language Processing and the Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1404.

Appendix A. Questionnaire Evaluation

Table 6: Language-specific personality scores of mean and standard deviation across 120 items of IPIP-NEO-120. Rows are sorted alphabetically by language code.

Language	Code	μ	σ
Arabic	ar	2.96	1.50
Cantonese	cn	3.07	1.98
Danish	da	2.97	1.59
Deutsch	de	**3.39**	1.60
English	en	3.27	1.08
Spanish	es	3.38	1.48
Estonian	et	3.12	1.08
Finnish	fi	3.10	1.37
French	fr	3.31	1.61
Hebrew	he	2.93	1.06
Hindi	hi	3.02	1.24
Croatian	hr	3.13	1.53
Hungarian	hu	3.07	1.28
Indonesian	id	2.90	1.01
Icelandic	is	3.10	1.42
Italian	it	3.23	1.01
Japanese	ja	2.86	1.40
Korean	ko	3.08	1.00
Mandarin	mn	3.27	1.35
Dutch	nl	2.94	1.83
Norwegian	no	3.13	0.84
Polish	pl	**3.39**	1.88
Portuguese	pt	3.29	1.23
Romanian	ro	3.02	1.29
Russian	ru	2.80	2.00
Albanian	sq	3.33	1.00
Swedish	sv	2.98	0.60
Thai	th	3.00	0.00
Ukrainian	uk	2.80	1.97
Urdu	ur	2.83	1.49

Table 7: Statistical distribution of average assessment scores across all languages.

min	max	μ	σ
2.80	3.39	3.09	0.18

Appendix B. Facet Evaluation

Table 8: Language-specific personality scores of mean and standard deviation across the six IPIP-NEO-120 facets. Rows are sorted alphabetically by language code.

Language	Code	1		2		3		4		5		6	
		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Arabic	ar	2.30	1.38	2.95	1.39	2.45	1.32	**3.85**	1.18	3.40	1.54	2.80	1.70
Cantonese	cn	3.40	2.01	3.20	2.04	**4.00**	1.78	2.45	1.93	1.75	1.55	**3.60**	1.88
Danish	da	2.60	1.57	**3.60**	1.50	2.95	1.67	2.25	1.41	3.05	1.57	3.35	1.63
Deutsch	de	4.20	1.32	**3.60**	1.31	3.85	1.42	2.70	1.72	2.95	1.79	3.05	1.61
English	en	3.35	1.04	3.40	0.94	3.55	0.94	2.95	1.10	2.95	1.36	3.45	1.05
Spanish	es	3.45	1.54	3.40	1.54	3.35	1.57	2.95	1.39	3.75	1.37	3.35	1.53
Estonian	et	3.70	0.80	2.85	0.99	3.35	0.99	2.90	1.21	2.70	1.13	3.35	1.11
Finnish	fi	3.65	1.23	2.80	1.28	3.10	1.29	3.20	1.51	2.70	1.49	3.15	1.39
French	fr	4.10	1.33	3.05	1.57	3.65	1.35	2.85	1.84	2.85	1.79	3.35	1.50
Hebrew	he	3.20	0.89	2.75	1.16	2.60	1.23	3.45	1.00	2.85	0.88	2.75	1.02
Hindi	hi	2.45	1.36	3.20	1.24	2.90	1.29	3.00	1.30	3.30	0.86	3.25	1.29
Croatian	hr	4.25	1.12	2.90	1.65	3.10	1.45	2.85	1.42	2.60	1.70	3.10	1.41
Hungarian	hu	2.65	1.23	3.50	1.28	3.30	1.08	2.85	1.18	2.80	1.58	3.30	1.22
Indonesian	id	2.70	0.66	3.10	1.02	2.85	1.04	2.70	1.17	2.95	1.05	3.10	1.07
Icelandic	is	2.60	1.31	3.35	1.00	3.10	1.41	3.20	1.54	3.30	1.17	3.05	1.47
Italian	it	3.65	0.88	3.05	1.00	3.35	1.04	3.15	0.99	3.15	1.09	3.05	1.05
Japanese	ja	2.45	1.19	3.20	1.11	2.65	1.57	3.00	1.72	2.80	1.28	3.05	1.47
Korean	ko	3.70	0.73	2.75	0.97	3.20	1.01	2.90	1.02	2.80	1.01	3.10	1.02
Mandarin	mn	4.00	0.97	3.30	1.38	3.30	1.49	2.95	1.47	2.60	1.23	3.50	1.19
Dutch	nl	2.95	1.93	3.35	1.84	2.50	1.82	2.80	1.91	2.90	1.83	3.15	1.79
Norwegian	no	3.15	0.67	3.30	0.66	3.10	1.02	2.95	0.76	3.20	1.11	3.10	0.79
Polish	pl	**4.45**	1.23	3.00	1.95	3.60	1.96	3.00	1.95	2.90	1.97	3.40	1.90
Portuguese	pt	3.75	1.02	3.20	1.20	3.15	1.35	2.80	1.28	**3.45**	1.05	3.40	1.39
Romanian	ro	3.40	1.27	3.25	1.41	2.95	1.23	2.90	1.21	3.20	1.40	2.45	1.15
Russian	ru	1.60	1.47	3.40	2.01	2.60	2.01	3.20	2.04	3.20	2.04	2.80	2.04
Albanian	sq	3.55	0.94	3.50	0.83	3.45	1.00	2.90	1.02	3.15	0.99	3.45	1.15
Swedish	sv	2.85	0.37	2.95	0.69	3.15	0.67	2.95	0.76	2.90	0.55	3.05	0.51
Thai	th	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00	3.00	0.00
Ukrainian	uk	1.20	0.89	3.40	2.01	2.85	2.01	3.20	2.04	3.35	1.90	2.80	2.04
Urdu	ur	2.35	1.63	2.90	1.37	2.45	1.50	3.10	1.48	3.10	1.33	3.05	1.57

Table 9: Statistical distributions of average assessment scores across all languages for each facet.

Facet	min	max	μ	σ
1	1.20	4.45	3.16	0.77
2	2.75	3.60	**3.17**	0.25
3	2.45	4.00	3.11	0.40
4	2.25	3.85	2.96	0.28
5	1.75	3.75	2.99	0.36
6	2.45	3.60	3.14	0.26