Rediscovering the Latent Dimensions of Personality with Large Language Models as Trait Descriptors

Joseph Suh* **Suhong Moon*** **Minwoo Kang*** **David M. Chan**

University of California, Berkeley

{josephsuh, suhong.moon, minwoo_kang, davidchan}@berkeley.edu

Abstract

Assessing personality traits using large language models (LLMs) has emerged as an interesting and challenging area of research. While previous methods employ explicit questionnaires, often derived from the Big Five model of personality, we hypothesize that LLMs implicitly encode notions of personality when modeling next-token responses. To demonstrate this, we introduce a novel approach that uncovers latent personality dimensions in LLMs by applying singular value decomposition (SVD) to the log-probabilities of trait-descriptive adjectives. Our experiments show that LLMs "rediscover" core personality traits such as extraversion, agreeableness, conscientiousness, neuroticism, and openness without relying on direct questionnaire inputs, with the top-5 factors corresponding to Big Five traits explaining 74.3% of the variance in the latent space. Moreover, we can use the derived principal components to assess personality along the Big Five dimensions, and achieve improvements in average personality prediction accuracy of up to 5% over fine-tuned models, and up to 21% over direct LLM-based scoring techniques.

1 Introduction

Over the past decades, researchers in personality psychology have investigated the viability of a general, descriptive model that yields systematic categorization of individual differences in personalities. These efforts culminated in one of the most widely-accepted taxonomies—the *Big Five* model [John et al., 2008, Goldberg, 2013] with extraversion, agreeableness, conscientiousness, neuroticism, and openness as five principal dimensions of personality traits—that has been consistently cross-validated in various studies with different human populations and assessment methods [McCrae and John, 1992, John et al., 1999, Costa and McCrae, 1999, Gosling et al., 2003]. The core underlying principle behind the Big Five is the *lexical hypothesis*, which asserts that the fundamental characteristics of personality are likely encoded in natural language, often as single words [Allport, 1937, Goldberg, 1981, Raad et al., 1998, John et al., 2008]. The five factors were derived from analyses of terms that people frequently use to describe personalities of themselves and others, resulting in lists of trait descriptive adjectives (TDAs) [Norman, 1967, Goldberg, 1992, 2013] that form the basis of psycholexical analyses and have influenced questionnaire-based assessments employed to-date, such as the Big Five Inventory (BFI) [Soto and John, 2017] and NEO-FFI [McCrae and John, 1992, Costa and McCrae, 1999].

More recently, advancements in large language models (LLMs) have sparked investigations of whether these models are capable of recognizing and expressing personality traits as humans [Jiang et al., 2023, Amin et al., 2023, Serapio-García et al., 2023, Peters and Matz, 2024], with potential applications in automatic text-based personality recognition [Mairesse et al., 2007, Kazameini et al., 2020, Zhao et al., 2022, Cao and Kosinski, 2024, Wen et al., 2024]. As LLMs are trained on increasing amounts

*Equal contribution

NeurIPS 2024 Workshop on Behavioral Machine Learning.

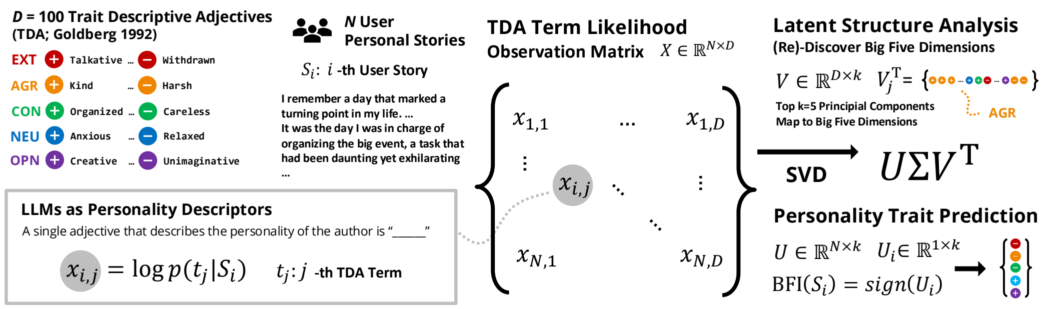

Figure 1: Overview of our approach. Given a set of personal stories and a list of TDAs, a language model computes the likelihood of each trait term describing the author of the story, through which we construct an observation matrix of such log-probabilities. Singular Value Decomposition (SVD) is then applied to this matrix, which yields (1) the loading matrix V that captures the latent dimension structure and (2) the factor matrix U that explains each story author's personality spectrum in the projected low-dimensional space. Results are compared against findings from psychometric literature and binary Big Five labels of authors, respectively.

of text corpora, the expectation is that these models could correlate linguistic features reflected in samples of text to the personality dimensions of the authors [Perez et al., 2023, Moon et al., 2024]. Prior work explores direct prompting, *i.e.* asking to directly predict the Big Five trait scales, or prompting to complete questionnaires [Jiang et al., 2023, Serapio-García et al., 2023]. While these work report that personality prediction may benefit from the use of LLMs, the dependence on direct prompts and questionnaires has potential risks of unreliable and prompt-sensitive results [Sühr et al., 2023, Gupta et al., 2024, Frisch and Giulianelli, 2024].

In this work, we first revisit the lexical hypothesis and the analyses that led to the discovery of the Big Five model: does LLMs' description of personality encode latent structures that align with the Big Five? We examine whether the same principle dimensions reported in human studies emerge when LLMs are prompted to describe personalities without imposing pre-defined taxonomies, thereby allowing natural, unrestricted depiction of personality traits. Specifically, we condition language models so as to measure the likelihoods of each term in the Goldberg's 100-TDA [Goldberg, 1992] in describing the author, over a set of synthetic personal stories. These likelihoods, computed as log-probabilities, are analyzed using singular value decomposition (SVD) to uncover underlying personality factors.

The uncovered factors are then used to predict the polarity of Big Five traits of the authors of unseen stories. Overall, our results demonstrate not only that LLMs encode a latent structure similar to the Big Five model, with the five principal factors explaining 74.3% of total variance, but also our approach to personality trait predictions outperform previous methods, by up to 5% over models fine-tuned on Big Five trait classification and by up to 21% over direct LLM-scoring approaches.

2 Using Language Models as Personality Descriptors

In this work, we hypothesize that LLMs encode personality traits implicitly within their latent spaces. More explicitly, we argue that we can analyze and extract personality traits present in personal stories by carefully querying the LLM and analyzing the log-probabilities of trait-descriptive adjectives (TDAs) within the next token prediction. Here we ask: is it possible to recover traditional categories of personality from the principal components of the resulting token likelihood matrix?

We present the primary steps of our approach in Figure 1. First, we measure the likelihood of a set of personality trait terms describing the story reflecting each individual author, *i.e.* calculate the log-probabilities of the 100 adjectives in the TDA cu-

TDA Assessment Prompt

Following is a personal story.
Essay: ***(Personal Story)***

Question: Based on this essay, describe the personality of the author.
Answer with a single adjective.

Answer: A single adjective that describes the personality of the author is "

Figure 2: A question-answer format prompt that asks for a single adjective describing the personality of the author of the given personal story. A quotation mark is adopted at the prompt suffix to guide the model to complete the sentence with an adjective. For each adjective in the TDA, we measure the log-probability of the model completing the prompt with the given adjective.

rated in Goldberg [1992]. In doing so, we prompt LLMs with a personal story as illustrated in Figure 2. Then, we build an observation matrix where each row contains the log-probabilities of TDA for each user. These log-probabilities can capture the usage frequency of adjectives varies in real life, based on the input prompt. For instance, adjectives associated with extraversion are more commonly used than those related to conscientiousness, leading to differences in latent factor levels [Goldberg, 1981]. Therefore, this subtle difference is captured in log-probabilites, allowing the method to recover the five principal axes corresponding to each trait in the Big Five model. Further details on the measurement process can be found in Appendix A.

After building the observation matrix, we perform singular value decomposition (SVD) on the matrix to identify the top-5 factors. Based on the SVD results, we analyze both the loading matrix and the factor matrix (Section 3). We then report the classification results derived from the factor matrix and compare these with the findings from the previous psychometric literature [John et al., 1999, Goldberg, 1981, 1992, 2013].

2.1 Re-Discovering the Big Five Structure

We start with measuring the log-probabilities of the trait terms in describing each of the users and construct the observation matrix $X \in \mathbb{R}^{N \times D}$, where N is the number of users and D is the number of trait adjectives in Figure 1. We observe that the log-probabilities are biased by the frequency of usage of trait adjectives, as shown in Figure 3, and accordingly apply zero-centering. SVD is applied to the matrix X, resulting in $\tilde{X} = U\Sigma V^\top$—in this decomposition, $U \in \mathbb{R}^{N \times k}$ is the **factor matrix** that represents the latent features of personality traits of each story author, $\Sigma \in \mathbb{R}^{k \times k}$ contains the singular values, and $V \in \mathbb{R}^{D \times k}$ is the **loading matrix** that defines how each trait adjective relates to the latent factors. We compare the factor matrix with factor loading identified in Goldberg [1992] to examine whether the principal components uncovered by the factor matrix resembles the Big Five dimensions.

2.2 Personality Trait Prediction

Each element of the factor matrix U_{ij} represents the extent of the *i*-th story's association with the *j*-th latent factor. Specifically, the sign of the element predicts the binary personality trait (e.g., 'extraverted' or 'introverted') of the story author, and for evaluation, we compare these polarities against the actual personality trait labels. Since our approach is based on an *unsupervised* method, the full procedure does not require explicit labels of personality traits. When labels are available, we can further enhance our analysis by incorporating supervised learning. Specifically, we choose Lasso regression using the log-probability features to predict the personality traits since Lasso regression performs feature selection by identifying the most informative trait adjectives.

3 Experiments

We benchmark our methods against baselines on `PersonaLLM` dataset [Jiang et al., 2023]. `PersonaLLM` dataset is generated by `gpt-4-0613` [Achiam et al., 2023], where `gpt-4-0613` is prompted with Big Five traits and asked to generate stories consistent with the prescribed personality. More details about the dataset are in Appendix B. We consider a suite of LLMs including Meta Llama 3 and Llama 3.1 model families [AI@Meta, 2024] and Mixtral 8x22B [Jiang et al., 2024, MistralAI, 2024]. We present the analysis results that uncover the Big Five structure in Section 3.1 and compare trait prediction accuracy with baselines in Section 3.2.

3.1 The Big Five Structure

The sum of the top-5 squared singular values from SVD accounts for 74.3% of the total variance in the log-probability measurements, indicating that it sufficiently captures the variance. The detailed statistics of singular values are provided in Appendix C.1. We compare elements of principal components with factor loadings from the literature [Goldberg, 1992]. Specifically, identifying which adjectives have large elements in each principal component establishes the correspondence

Table 1: Test accuracy of binary personality assessment using various approaches. The first two rows represent baseline methods—PersonaLLM and fine-tuning DeBERTaV3 He et al. [2021]. For PersonaLLM, we use the numbers reported in Jiang et al. [2023]. Predictions obtained using SVD are comparable to that of the baseline methods, while regression performed on a set of trait adjectives surpasses other methods. 'Instruct' denotes instruction fine-tuned, chat models; otherwise, if unspecified, the pre-trained model is used. Bold-faced numbers indicate the highest accuracy for each of the Big Five traits.

Method	Model	Big Five Trait Prediction Accuracy					
		Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Avg.
PersonaLLM	`gpt-4-0613`	0.97	0.69	0.69	0.56	0.59	0.698
Encoder FT	`DeBERTaV3`	**1.000**	0.833	0.786	0.833	0.881	0.867
SVD	`Llama-3.1-70B`	0.905	0.786	0.726	0.774	0.762	0.791
	`Llama-3.1-70B-Instruct`	0.917	0.798	0.536	0.607	0.679	0.707
	`Llama-3-70B`	0.988	0.798	0.667	0.655	0.810	0.783
	`Llama-3-70B-Instruct`	0.905	0.845	0.786	0.607	0.821	0.793
	`Llama-3.1-8B`	0.857	0.702	0.821	0.690	0.667	0.748
	`Llama-3.1-8B-Instruct`	0.702	0.810	0.560	0.738	0.524	0.667
	`Llama-3-8B`	0.869	0.750	0.786	0.726	0.619	0.750
	`Llama-3-8B-Instruct`	0.905	0.679	0.583	0.738	0.548	0.690
	`Mixtral-8x22B`	0.917	0.798	0.536	0.607	0.679	0.707
Lasso	`Llama-3.1-70B`	0.964	**0.964**	0.881	**0.893**	0.857	**0.912**
	`Llama-3.1-70B-Instruct`	0.988	0.845	**0.857**	0.857	**0.905**	0.890
	`Llama-3-70B`	**1.000**	0.893	0.845	0.881	**0.905**	0.905
	`Llama-3-70B-Instruct`	**1.000**	0.833	0.821	0.869	0.845	0.874
	`Llama-3.1-8B`	0.964	0.869	0.821	**0.893**	0.881	0.886
	`Llama-3.1-8B-Instruct`	0.964	0.857	0.786	0.845	0.845	0.860
	`Llama-3-8B`	0.988	0.857	0.798	**0.893**	0.893	0.886
	`Llama-3-8B-Instruct`	0.988	0.857	0.786	0.833	0.833	0.860
	`Mixtral-8x22B`	0.988	0.845	**0.857**	0.857	**0.905**	0.890

between principal components and personality factors. For example, since adjectives that have high loading in the extraversion factor (e.g., 'energetic', 'bashful', 'reserved') also have large elements in the first principal component, the first principal component relates to the extraversion trait. We identify the one-to-one correspondence between principal components and Big Five traits; the top-5 principal components correspond with extraversion, openness, agreeableness, neuroticism, and conscientiousness traits, respectively. Detailed steps for finding the correspondence are given in Appendix C.2.

3.2 Personality Trait Prediction

U matrix of SVD can be interpreted as the personality trait scales of each story author. More specifically, the element of U, U_{ij}, represents the scale of the j-th Big Five trait of the i-th user. By taking the signs of the predicted scales and comparing them with binary personality labels, we measure the accuracy of personality assessment. As a baseline, we employ (1) fine-tuning DeBERTaV3 [He et al., 2021] for the Big Five binary label classification task, and (2) prompting LLMs to evaluate the personality score directly. The details of each baseline method are presented in Appendix D.

The personality prediction accuracy resulting from SVD is comparable to that of the baseline methods. Additionally, we perform Lasso regression using the log-probabilities of trait adjectives as features and binary personality labels as targets. This regression method outperforms other approaches across all personality traits.

4 Further Directions

While the results in latent dimension analysis and personality assessment are promising, several psychometric aspects deserve further exploration.

Convergent Validity: Convergent validity assesses the extent to which different measures that are theoretically related are also experimentally related. To examine this aspect, we need to verify whether a set of widely accepted personality tests, as well as self-perception and external perception tests, consistently reveal a latent structure in personality assessment that aligns with human perception.

Role of Acquaintance: The level of acquaintance is known to influence the accuracy of personality descriptions. Studies indicate that acquaintance between a person and a perceiver generally enhances the assessment accuracy Funder and Colvin [1988], Lee and Ashton [2017]. We may simulate acquaintance with persona steering methods Moon et al. [2024] and perform personality assessments at varying levels of acquaintance to investigate the existence of the same correlation in LLMs.

Test-retest Reliability: Test-retest reliability evaluates the consistency of test results by administering the same test to the same group at different points in time. It could be assessed through behavioral change experiments where language models are in conversational settings. By comparing personality assessments at the beginning and end of the interaction, we may investigate the test-retest reliability.

5 Conclusion

This work introduces a method to probe the latent dimensions of personality using LLMs as trait descriptors. By analyzing key characteristics of principal components, we observe that the personality perception by LLMs encodes a latent structure closely resembling that of human perception. Our method also serves as a robust tool for personality assessment that matches or surpasses the accuracy of previous LLM-based approaches. Building on one of the most intrinsic aspects of human–personality–this framework holds the potential to explore a vast array of real-world values, including ethical and cultural norms, and social behaviors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Gordon Willard Allport. Personality: A psychological interpretation. 1937.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt. *IEEE Intelligent Systems*, 38(2):15–23, 2023.

Xubo Cao and Michal Kosinski. Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14(1):6735, 2024.

PT Costa and RR McCrae. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01):1999, 1999.

Ivar Frisch and Mario Giulianelli. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*, 2024.

David C Funder and C Randall Colvin. Friends and strangers: acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of personality and social psychology*, 55(1):149, 1988.

Lewis R Goldberg. Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology/Sage*, 1981.

Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge, 2013.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality, 2024. URL `https://arxiv.org/abs/2309.08163`.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. corr abs/2111.09543 (2021). *arXiv preprint arXiv:2111.09543*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *ArXiv*, abs/2401.04088, 2024. URL `https://api.semanticscholar.org/CorpusID:266844877`.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*, 2023.

Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 1999.

Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158, 2008.

Amirmohammad Kazameini, Samin Fatehi, Yash Mehta, Sauleh Eetemadi, and Erik Cambria. Personality trait detection using bagged svm over bert word embedding ensembles. *arXiv preprint arXiv:2010.01309*, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

Kibeom Lee and Michael C Ashton. Acquaintanceship and self/observer agreement in personality judgment. *Journal of Research in Personality*, 70:1–5, 2017.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500, 2007.

Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

MistralAI. Mixtral-8x22b, 2024. URL `https://mistral.ai/news/mixtral-8x22b/`.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M Chan. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*, 2024.

Warren T Norman. On estimating psychological relationships: Social desirability and self-report. *Psychological Bulletin*, 67(4):273, 1967.

Ethan Perez et al. Discovering language model behaviors with model-written evaluations. arxiv, dec. 2022, 2023.

Heinrich Peters and Sandra Matz. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, page pgae231, 2024.

Boele De Raad, Marco Perugini, Martina Hrebícková, and Piotr Szarota. Lingua franca of personality: Taxonomies and structures based on the psycholexical approach. *Journal of Cross-Cultural Psychology*, 29(1):212–232, 1998.

Gerard Saucier. Effects of variable selection on the factor structure of person descriptors. *Journal of personality and social psychology*, 73(6):1296, 1997.

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. *arXiv preprint arXiv:2307.00184*, 2023.

Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. Challenging the validity of personality tests for large language models. *arXiv e-prints*, pages arXiv–2311, 2023.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Ruosong Yang, and Shuaiqi Liu. Affective-nli: Towards accurate and interpretable personality recognition in conversation. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 184–193. IEEE, 2024.

Xiaoming Zhao, Zhiwei Tang, and Shiqing Zhang. Deep personality trait recognition: a survey. *Frontiers in Psychology*, 13:839619, 2022.

A Log-probability Measurement Details

A.1 Why are TDA Log-probabilities Meaningful?

We argue that log-probabilities of personality trait descriptive adjectives computed by LLMs are indicative measures of how accurately those adjectives describe a person. While these log-probabilities encode likelihood that the model assigns to describing the person with a specific adjective, the prediction of such likelihoods are influenced not only by the explicit or implicit references to behavioral characteristics reflected in the story, but also by the patterns of general language use. Figure 3 illustrates that: the mean log-probability for the most probable adjective ('introverted') is -5.7, whereas it drops to -20.0 for the least probable adjective ('unenvious'). Despite this difference, variances in log-probabilities between stories for a specific trait adjective convey meaningful signals. For example, the probability of using 'unenvious' for trait description is generally low but it is even lower for some stories while relatively higher for others. We plot the correlation between log-probabilities for different adjectives (Figure 4) to analyze the variance. The correlation matrix demonstrates the consistency of log-probabilities measured by LLMs. In more detail, we observe that a strong positive correlation among log-probabilities of trait adjectives shares the same personality trait and pole, and a strong negative correlation among those of opposite poles. This evidence supports our assertion that log-probability differences effectively capture the accuracy of an adjective being a description of the person.

A.2 Trait Descriptive Adjectives

There are multiple sets of trait descriptive adjectives in a literature Goldberg [1992, 2013], Saucier [1997] with varying numbers of adjectives and different ways of presentation (e.g. unipolar v.s. bipolar). We adopt a widely-used trait adjective sets, the 100 unipolar adjectives set Goldberg [1992]. 100 trait adjectives are listed in Table 2.

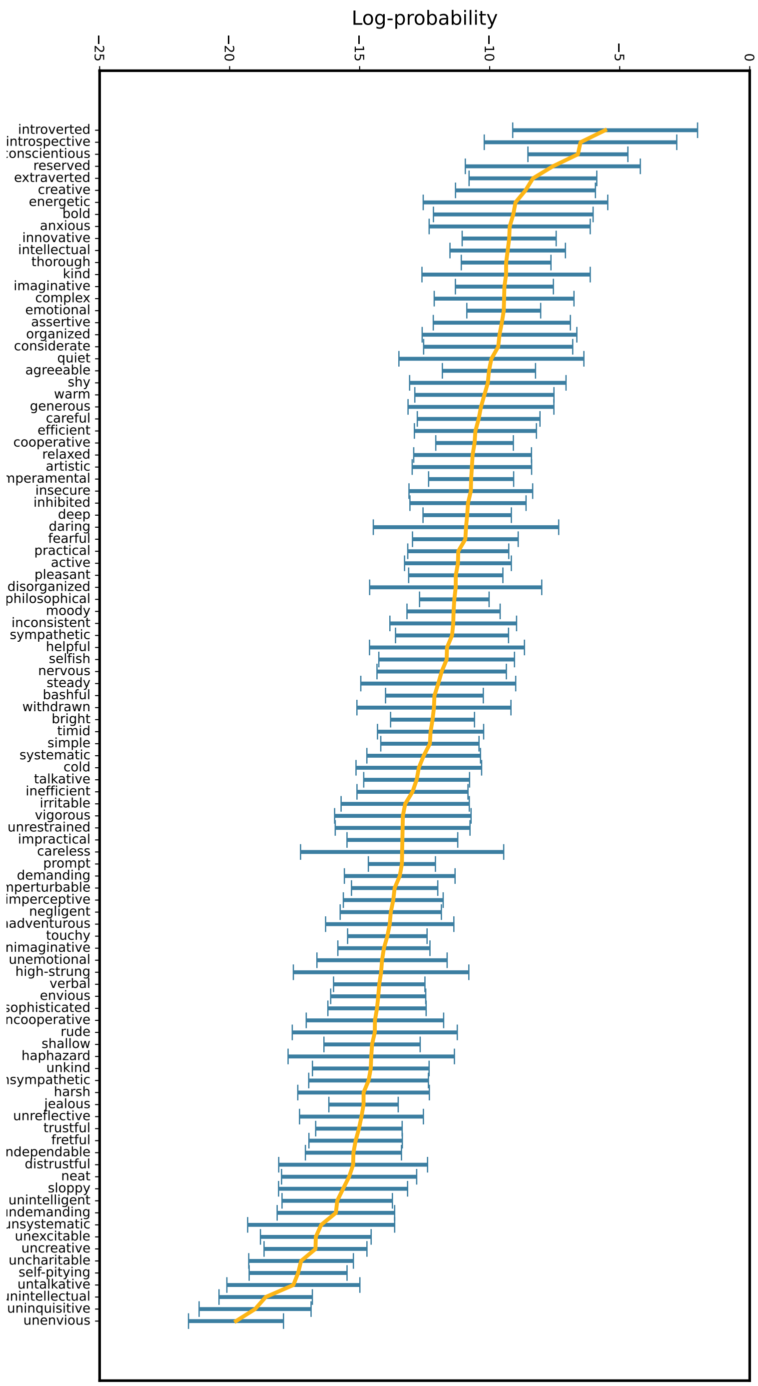

Figure 3: Mean (gold) and standard deviation (blue) of log-probabilities for 100 trait adjectives, sorted by the mean. Log-probabilities are measured from the `PersonaLLM` dataset (Appendix B) with pretrained `Llama-3.1-70B`, decoding temperature $T = 1$.

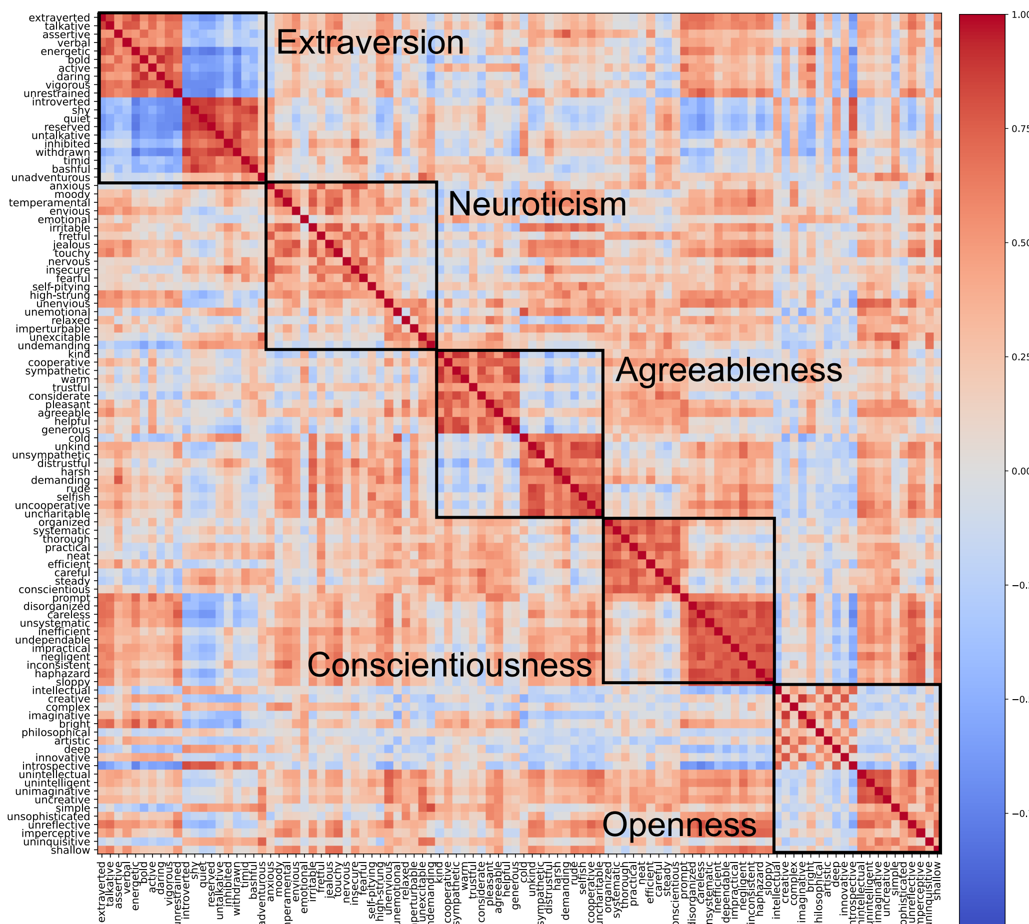

Figure 4: Visualization of correlation between trait adjectives. Datasets, models, and decoding parameters are identical to those of Figure 3. Five boxes with black edges indicate personality traits that adjectives belong to, drawn for visual aid. Trait adjectives that share Big Five trait show strong correlation, either positive or negative. We note that correlations between adjectives of different Big Five trait also show moderate level of correlation (e.g., 'introverted' and 'introspective'). This may imply that an adjective is related to several latent factors (instead of a single latent factdor) or that Big Five personality traits are not orthogonal.

Table 2: 100 unipolar trait descriptive adjectives grouped by Big Five traits and the pole. Pole refers to one end of a personality trait's continuum. Each of the five traits is considered a spectrum between two opposite extremes, or poles.

Trait	**Extraversion**		**Neuroticism**		**Agreeableness**	
Pole	(+)	(-)	(+)	(-)	(+)	(-)
Adjectives	extraverted talkative assertive verbal energetic bold active daring vigorous unrestrained	introverted shy quiet reserved untalkative inhibited withdrawn timid bashful unadventurous	anxious moody temperamental envious emotional irritable fretful jealous touchy nervous insecure fearful self-pitying high-strung	unenvious unemotional relaxed imperturbable unexcitable undemanding	kind cooperative sympathetic warm trustful considerate pleasant agreeable helpful generous	cold unkind unsympathetic distrustful harsh demanding rude selfish uncooperative uncharitable

Trait	**Conscientiousness**		**Openness**	
Pole	(+)	(-)	(+)	(-)
Adjectives	organized systematic thorough practical neat efficient careful steady conscientious prompt	disorganized careless unsystematic inefficient undependable impractical negligent inconsistent haphazard sloppy	intellectual creative complex imaginative bright philosophical artistic deep innovative introspective	unintellectual unintelligent unimaginative uncreative simple unsophisticated unreflective imperceptive uninquisitive shallow

A.3 Tokenization of Adjectives

As LLMs model next-token responses, a log-probability of an adjective is a sum of log-probabilities of tokens that consist the adjective. A single adjective is tokenized into multiple tokens. For example, an adjective 'sophisticated' is tokenized into four tokens with `Llama3` tokenizer: 's', 'oph', 'istic', and 'ated'. Therefore, to compute the log-probability for 'sophisticated' with the prompt (Figure 2), we compute four instances of log-probability: (1) log-probability of 's' given the prompt, (2) log-probability of 'oph' given the prompt appended with 's', (3) log-probability of 'istic' given the the prompt appended with 'soph', and (4) log-probability of 'ated' given the prompt appended with 'sophistic'. Four instances of log-probability are added to measure the log-probability for 'sophisticated'. Similar process is repeated for each trait adjectives.

A.4 Decoding Parameter: Temperature

Result in the main text is obtained with sampling temperature $T = 1.0$. We can compute the log-probability at different sampling temperatures with a single round of computation at $T = 1.0$, if we have access to log-probabilities for a complete set of vocabularies. This can be done with an efficient LLM serving system Kwon et al. [2023]. The process below describes the computation process at a token level.

Let N be a vocabulary size of a language model, W a list of conditioning tokens, and w_{target} a next token of which we want to compute log-probability. Our goal is to compute the log-probability $\text{LP}(w_{\text{target}}|W;T_{\text{t}})$ at a target temperature T_{t} given a full access to log-probabilities at an original temperature T_{o}, i.e. we know $\text{LP}(w_i|W;T_{\text{o}})$ for $i \in [1, N]$. We can use the following relation

$$\text{LP}(w_{\text{target}}|W;T_{\text{t}}) = \log\left(\frac{\exp\left(\frac{l_{\text{target}}}{T_{\text{t}}}\right)}{\sum_{i=1}^{N}\exp\left(\frac{l_i}{T_{\text{t}}}\right)}\right) = \log\left(\frac{1}{\sum_{i=1}^{N}\exp\left(\frac{l_i - l_{\text{target}}}{T_{\text{t}}}\right)}\right)$$

$$= \log\left(\frac{1}{\sum_{i=1}^{N}\exp\left(\frac{l_i - l_{\text{target}}}{T_{\text{o}}}\cdot\frac{T_{\text{o}}}{T_{\text{t}}}\right)}\right) = \log\left(\frac{1}{\sum_{i=1}^{N}\exp\left((\text{LP}(w_i|W;T_{\text{o}}) - \text{LP}(w_{\text{target}}|W;T_{\text{o}}))\cdot\frac{T_{\text{o}}}{T_{\text{t}}}\right)}\right)$$

where l_i represents a logit for the i-th token. Computation at an adjective level is straightforward by using token-level computations repeatedly for each token consisting an adjective (A.3).

B Dataset

Synthetic stories are adopted from the `PersonaLLM` dataset Jiang et al. [2023]. These stories were generated using `GPT-4-0613` by prompting with prescribed Big-Five personality traits (e.g., "You are a character who is extroverted, agreeable, conscientious, emotionally stable, and closed to experience.") along with an instruction "Please share a personal story in 800 words. Do not explicitly mention your personality traits in the story." The authors generated 10 stories for each of 32 combinations of binary Big-Five personality traits, and filtered out stories that explicitly contained trait-related lexicons. This process resulted in a total of 208 synthetic stories. We use these 208 stories as an input for log-probability measurement, and employ their binary Big-Five personality labels for the prediction accuracy measurement.

C Additional Experiment Results

C.1 Singular Values of SVD

We present the statistics of the singular values in Figure 5. The gold-colored bars represent the singular values sorted by magnitude, and the blue line illustrates the cumulative explained variance ratio. The explained variance ratio measures the proportion of the total variance in the original dataset that is accounted for by each principal component; it is calculated as the ratio of a principal component's eigenvalue to the sum of the eigenvalues of all principal components. We observe a significant drop between the fifth and sixth singular values and the cumulative explained variance ratio at the fifth principal component is 0.743. This fact indicates that top-5 components sufficiently explain the variance of the latent dimensions.

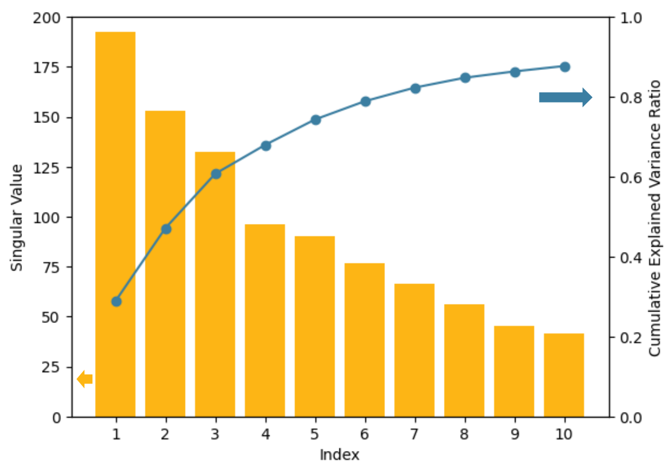

Figure 5: Singular values corresponding to each principal component (gold) and cumulative explained variance ratio (blue). At the fifth principal component, cumulative explained variance ratio is 0.743.

C.2 Analyzing Elements of Singular Vectors

In the SVD $\tilde{X} = U\Sigma V^\top$, the relation between the *i*-th original dimension and the *j*-th principal component is quantified by the matrix element V_{ij}. In terms of our approach, it is equivalent to the explainability of likelihood of the *i*-th trait adjective (e.g., 'introverted') by the *j*-th factor (e.g., extraversion). Therefore, by analyzing trait adjectives that have the largest or smallest elements in one principal component, we can infer the alignment of that principal component with a Big Five personality trait. For example, if adjectives 'assertive' and 'talkative' have the largest elements while 'shy' and 'introverted' have the smallest elements in the first principal component, the first principal component is aligned with the extraversion factor.

Table 3 shows 10 adjectives with the largest elements and 10 adjectives with the smallest elements for each principal component. Although each dimension does not have a perfect correspondence with a specific Big Five personality trait, we are able to observe tendency. For example, the adjectives having the largest and smallest elements in the first principal component are mostly related with extraversion factor, including 'energetic' (0.232), 'daring' (0.180), and 'quiet' (-0.216). Repeating the analysis for other principal components, we report that the dimensions 1 through 5 are related with Big Five extraversion, openness, agreeableness, neuroticism, and conscientiousness factors, respectively.

We further investigate the relation of Big Five traits with principal components by examining the sign of elements in Table 4. From our previous observation in Table 3, we focus on the signs of elements of extraversion, openness, agreeableness, neuroticism, and conscientiousness adjectives in principal components

1 through 5, respectively. Adjectives of the same pole have same sign with few exceptions. For example, in the first principal component, adjectives associated with the (+) pole of extraversion ('energetic', . . . , 'verbal') have positive elements, while adjectives associated with the (-) pole of extraversion ('unadventurous', . . . , 'quiet') have negative elements. Based on these findings we claim that principal components from SVD and Big Five personality traits have a one-to-one correspondence.

Table 3: 10 adjectives with the largest and smallest loadings in each principal component, from dimension 1 to 5. Extraversion adjectives predominantly have extreme elements in the first principal component, agreeableness adjectives for dimension 3, neuroticism adjectives for dimension 4, and conscientious adjectives for dimension 5. This tendency reveals that each principal component corresponds to a specific Big Five trait.

Dimension 1				**Dimension 2**			
Adjective	Factor	Pole	Loading	Adjective	Factor	Pole	Loading
careless	CON	-	0.268	anxious	NEU	+	0.202
energetic	EXT	+	0.232	introverted	EXT	-	0.190
disorganized	CON	-	0.200	withdrawn	EXT	-	0.187
haphazard	CON	-	0.188	distrustful	AGR	-	0.187
daring	EXT	+	0.180	inhibited	EXT	-	0.180
vigorous	EXT	+	0.176	reserved	EXT	-	0.176
extraverted	EXT	+	0.174	insecure	NEU	+	0.176
unrestrained	EXT	+	0.170	cold	AGR	-	0.172
high-strung	NEU	+	0.168	unemotional	NEU	-	0.167
unsystematic	CON	-	0.156	irritable	NEU	+	0.159
intellectual	OPN	+	-0.081	active	EXT	+	-0.046
steady	CON	+	-0.087	innovative	OPN	+	-0.049
bashful	EXT	-	-0.092	generous	AGR	+	-0.054
untalkative	EXT	-	-0.106	vigorous	EXT	+	-0.067
withdrawn	EXT	-	-0.137	artistic	OPN	+	-0.075
shy	EXT	-	-0.171	bold	EXT	+	-0.081
introverted	EXT	-	-0.188	imaginative	OPN	+	-0.087
reserved	EXT	-	-0.200	creative	OPN	+	-0.114
quiet	EXT	-	-0.216	energetic	EXT	+	-0.115
introspective	OPN	+	-0.238	daring	EXT	+	-0.147

Dimension 3				**Dimension 4**				**Dimension 5**			
Adjective	Factor	Pole	Loading	Adjective	Factor	Pole	Loading	Adjective	Factor	Pole	Loading
kind	AGR	+	0.309	nervous	NEU	+	0.304	haphazard	CON	-	0.257
helpful	AGR	+	0.267	anxious	NEU	+	0.272	shy	EXT	-	0.173
generous	AGR	+	0.262	high-strung	NEU	+	0.251	disorganized	CON	-	0.167
considerate	AGR	+	0.251	complex	OPN	+	0.233	relaxed	NEU	-	0.162
warm	AGR	+	0.232	creative	OPN	+	0.192	sloppy	CON	-	0.162
steady	CON	+	0.190	fretful	NEU	+	0.167	careless	CON	-	0.150
organized	CON	+	0.189	fearful	NEU	+	0.167	introverted	EXT	-	0.147
sympathetic	AGR	+	0.189	imaginative	OPN	+	0.158	inconsistent	CON	-	0.144
pleasant	AGR	+	0.180	organized	CON	+	0.154	unadventurous	EXT	-	0.144
neat	CON	+	0.173	daring	EXT	+	0.149	bashful	EXT	-	0.143
introverted	EXT	-	-0.080	unexcitable	NEU	-	-0.095	conscientious	CON	+	-0.128
harsh	AGR	-	-0.081	uncooperative	AGR	-	-0.100	cold	AGR	-	-0.130
insecure	NEU	+	-0.084	uninquisitive	OPN	-	-0.109	thorough	CON	+	-0.138
cold	AGR	-	-0.090	selfish	AGR	-	-0.126	organized	CON	+	-0.167
moody	NEU	+	-0.091	unsophisticated	OPN	-	-0.132	systematic	CON	+	-0.188
irritable	NEU	+	-0.099	unkind	AGR	-	-0.135	demanding	AGR	-	-0.199
withdrawn	EXT	-	-0.102	uncharitable	AGR	-	-0.139	unemotional	NEU	-	-0.204
complex	OPN	+	-0.104	undemanding	NEU	-	-0.145	harsh	AGR	-	-0.204
distrustful	AGR	-	-0.157	unsympathetic	AGR	-	-0.162	efficient	CON	+	-0.225
rude	AGR	-	-0.165	rude	AGR	-	-0.169	assertive	EXT	+	-0.230

Table 4: Elements (loadings) of extraversion, openness, agreeableness, neuroticism, and conscientiousness adjectives for principal components 1 to 5, respectively. Adjectives belonging to the same pole (either (+) or (-)) maintain same polarity, except for few exceptions: 'neat' in conscientiousness trait, whose element is close to 0; 'complex', 'introspective', 'deep', 'intellectual' in openness trait. Combining with the results from Table 3, we conclude that latent factors from SVD resemble the Big Five factors observed in human personality assessment Goldberg [1992].

Extraversion Adjectives in Dimension 1						**Openness Adjectives in Dimension 2**					
Positive Loadings			**Negative Loadings**			**Positive Loadings**			**Negative Loadings**		
Adjective	Pole	Loading	Adjective	Pole	Loading	Adjective	Pole	Loading	Adjective	Pole	Loading
energetic	+	0.233	unadventurous	-	-0.010	uncreative	-	0.099	philosophical	+	0.000
daring	+	0.181	inhibited	-	-0.041	uninquisitive	-	0.098	bright	+	-0.028
vigorous	+	0.177	timid	-	-0.077	unintelligent	-	0.095	innovative	+	-0.049
extraverted	+	0.175	bashful	-	-0.092	unintellectual	-	0.094	artistic	+	-0.075
unrestrained	+	0.171	untalkative	-	-0.106	imperceptive	-	0.091	imaginative	+	-0.087
bold	+	0.154	withdrawn	-	-0.137	simple	-	0.090	creative	+	-0.114
assertive	+	0.117	shy	-	-0.171	shallow	-	0.087			
talkative	+	0.117	introverted	-	-0.188	unreflective	-	0.083			
active	+	0.102	reserved	-	-0.200	unimaginative	-	0.082			
verbal	+	0.088	quiet	-	-0.216	complex	+	0.071			
						introspective	+	0.063			
						unsophisticated	-	0.027			
						deep	+	0.019			
						intellectual	+	0.011			

Agreeableness Adjectives in Dimension 3						**Neuroticism Adjectives in Dimension 4**					
Positive Loadings			**Negative Loadings**			**Positive Loadings**			**Negative Loadings**		
Adjective	Pole	Loading	Adjective	Pole	Loading	Adjective	Pole	Loading	Adjective	Pole	Loading
kind	+	0.309	uncharitable	-	-0.020	nervous	+	0.304	unenvious	-	-0.045
helpful	+	0.267	demanding	-	-0.034	anxious	+	0.272	imperturbable	-	-0.059
generous	+	0.262	unsympathetic	-	-0.047	high-strung	+	0.251	relaxed	-	-0.067
considerate	+	0.251	selfish	-	-0.049	fretful	+	0.167	unemotional	-	-0.079
warm	+	0.232	unkind	-	-0.058	fearful	+	0.167	unexcitable	-	-0.095
sympathetic	+	0.189	uncooperative	-	-0.063	insecure	+	0.119	undemanding	-	-0.145
pleasant	+	0.180	harsh	-	-0.081	moody	+	0.073			
agreeable	+	0.149	cold	-	-0.090	emotional	+	0.067			
cooperative	+	0.141	distrustful	-	-0.157	temperamental	+	0.061			
trustful	+	0.087	rude	-	-0.165	envious	+	0.045			
						irritable	+	0.042			
						touchy	+	0.029			
						jealous	+	0.027			
						self-pitying	+	0.011			

Conscientiousness Adjectives in Dimension 5					
Positive Loadings			**Negative Loadings**		
Adjective	Pole	Loading	Adjective	Pole	Loading
haphazard	-	0.257	prompt	+	-0.022
disorganized	-	0.167	careful	+	-0.056
sloppy	-	0.162	practical	+	-0.102
careless	-	0.150	steady	+	-0.122
inconsistent	-	0.144	conscientious	+	-0.128
impractical	-	0.123	thorough	+	-0.138
unsystematic	-	0.121	organized	+	-0.167
inefficient	-	0.101	systematic	+	-0.188
undependable	-	0.072	efficient	+	-0.225
negligent	-	0.044			
neat	+	0.003			

C.3 Personality Prediction with SVD

In this section, we elaborate how we can predict the personality traits based on SVD analysis. We compare the sign of a column of the factor matrix $U \in \mathbb{R}^{N \times k}$ with Big Five personality labels $L \in \{0, 1\}^{N \times 5}$, where $k = 5$ in this work. However, since we have not yet determined which principal component corresponds to each personality trait, we need a method to establish this correspondence. To achieve this, we construct the accuracy matrix, P, where each element P_{ij} is the prediction accuracy of j-th personality trait with i-th principal component.

P is in Table 5. An accuracy of 1 is close to a perfect prediction, while an accuracy of 0.5 is close to a random guess. When comparing the first column of U with the Big Five extraversion label, the accuracy is 0.899. Comparing the first column of U with Big Five labels other than extraversion gives accuracy of 0.505, 0.581, 0.591, 0.524, all close to a random guess. Notably, for each principal component, there exist only *one* Big Five label that yields high prediction accuracy. Columns 1 to 5 have high accuracy when compared with extraversion, openness, agreeableness, neuroticism, conscientiousness labels, respectively. This is the same one-to-one correspondence relation obtained in C.2.

Table 5: Accuracy matrix P comparing the signs of the first five principal components (rows) with the Big Five personality traits (columns). Each entry P_{ij} represents the accuracy of predicting the j-th personality trait from the i-th principal component. The matrix is derived from the training dataset using SVD with $k = 5$. The highest accuracy in each row is highlighted in bold, illustrating a one-to-one correspondence between principal components and personality traits.

Index	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
1	**0.899**	0.581	0.591	0.505	0.524
2	0.663	0.586	0.510	0.548	**0.798**
3	0.548	**0.817**	0.596	0.615	0.567
4	0.533	0.514	0.562	**0.726**	0.600
5	0.524	0.572	**0.803**	0.543	0.582

D Baseline Experiment Details

D.1 Fine-tuning Encoder Models

As a baseline, we fine-tune the DeBERTa-V3 Large model He et al. [2021] using `PersonaLLM` dataset. We place a 5-way classifier on top of the output of the [CLS] token from the encoder model. We utilize binary cross-entropy loss for fine-tuning. The AdamW optimizer Loshchilov et al. [2017] is utilized along with a cosine learning rate scheduler and a linear warm-up. The number of warm-up steps is set to 300. We divide `PersonaLLM` dataset into 80% for training, 10% for validation, and 10% for test. Hyperparameter tuning is performed on the validation set and the test set is evaluated only once using the best-performing hyperparameter configuration from the validation set. We perform a hyperparameter sweep over batch size, learning rate, weight decay, and number of training epochs. The hyperparameter configurations we use are: $\{4, 8, 16, 32\}$ for batch size, $\{5e-5, 1e-4, 2e-4, 4e-4\}$ for learning rate, $\{0.01, 0.02\}$ for weight decay, and $\{5, 10, 15, 20\}$ for the number of training epochs.

D.2 Prompting LLMs to Evaluate Personality

Prior works prompt language models to assess the Big Five personality score in a Likert scale, usually on a spectrum of 1 to 5. Various prompts have been designed: to list a few, ‘please rate how accurately this describes you on a scale from 1 to 5 (where 1 = "very inaccurate", 2 = "moderately inaccurate", 3 = "neither accurate nor inaccurate", 4 = "moderately accurate", and 5 = "very accurate")' Serapio-García et al. [2023] 'Rate the personality of the person called "user" on the Big Five personality dimensions. Pay attention to how people's personalities might be reflected in the way they respond to questions and what they share about themselves. Provide your response on a scale from 1 to 5 for the traits Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Provide only the numbers.' Peters and Matz [2024], and 'What is your guess for the big-five personality traits of someone who said "*text*", answer low or high with bullet points for the five traits? It does not have to be fully correct. You do not need to explain the traits. Do not show any warning after.' Amin et al. [2023]. Authors of `PersonaLLM` dataset Jiang et al. [2023] also utilized the prompting method. The published accuracy value from the paper is taken as a baseline method accuracy.