Bachelor's Thesis

A Statistical Evaluation of LLM-Generated Data for Psychometrics

Department of Statistics
Ludwig-Maximilians-Universität München

Cosima Fröhner

Munich, June 26th, 2025

Submitted in partial fulfillment of the requirements for the degree of B. Sc.
Supervised by Prof. Dr. Christoph Kern

Abstract

While social sciences increasingly use open big data sources for faster and cheaper insights, many research questions still rely on tailored assessments. A key limitation remains the cost of test development, driven by large samples needed across design iterations. Recent advances in synthetic data generation, especially in large language models (LLMs), may offer a cost-effective alternative by generating synthetic responses from personas. This study investigates LLM-based synthetic data generation for pretesting psychological test items. A synthetic sample of $N = 150$ respondents was generated with GPT-4o for two novel personality item sets and evaluated for fidelity compared to human response distributions, using established metrics (Kullback-Leibler divergence, Wasserstein distance, propensity mean squared error). Utility was assessed by selecting items based on synthetic data and testing model fit with human data according to the train-on-synthetic, test-on-real paradigm. Results show good fidelity on certain items and traits, but limited utility. With high factor loadings and low cross-loadings on all items compared to human data, only 2 items were excluded based on the synthetic sample versus 26 in the human sample. Adjusting the selection algorithm improved model fit, with a non-significant chi-square test of exact fit ($p > .05$). Implications and future directions are discussed.

Keywords: Synthetic Data Generation, Large Language Models, Psychometrics, Automated Test Development, Pretesting

Contents

List of Figures

List of Tables

List of Abbreviations

CFA confirmatory factor analysis

CFI comparative fit index

df degrees of freedom

EFA exploratory factor analysis

GAN generative adversarial networks

IPIP International Personality Item Pool

JS divergence Jensen-Shannon divergence

KL divergence Kullback-Leibler divergence

LLM large language model

MSE mean squared error

OLS ordinary least squares

PCA principal component analysis

pMSE propensity score mean squared error

RMSEA root mean squared error of approximation

SRMR standardized root mean square residual

TLI Tucker-Lewis index

VAE variational autoencoders

1 Introduction

When aiming to measure latent constructs such as honesty-humility or intelligence, a central challenge lies in identifying observable indicators that capture the underlying trait. Direct self-report questions such as "How honest and humble are you?" are unlikely to yield valid measurements. Instead, psychometric instruments are developed, consisting of multiple questions or tasks, referred to as *items*, designed to indirectly assess the target construct. The development and refinement of such instruments to quantify human traits lies at the core of *psychometrics* (Rust et al., 2020). Ensuring that a test possesses desirable psychometric properties, such as effectively measuring the intended psychological construct, typically involves several iterative stages. Beginning with a broad item pool, often informed by literature reviews or expert input, researchers narrow this pool through successive rounds of empirical evaluation, until arriving at a final set of items that forms the measurement scale. However, this process is both time-consuming and costly, often requiring large and representative samples (Liu et al., 2025). These constraints may, at times, even contribute to compromises that ultimately undermine the validity of the final instrument (e.g., see Barrick and Mount, 1991, Berry et al., 2007, Poropat, 2009, Roberts and Bogg, 2004).

Recent advances in LLMs may offer a tool to shift such trade-offs by substituting human samples, thereby making the test development procedure more cost-effective. A growing field of research has begun generating synthetic responses to psychological tests and evaluating them psychometrically with the goal of exploring psychological constructs, such as personality or morals, in LLMs (Ye et al., 2025). However, only limited research has so far explored the potential of LLMs for psychometrics, especially for synthetic data generation (Liu et al., 2025).

Psychometric tests are widely used in high-stakes decision-making contexts, including educational, clinical, and personnel selection settings. Unlike LLM assessments, test development has immediate consequences for real-world outcomes and can shape individuals' life and career trajectories. As such, the potential benefits of using LLM-generated response data to create potentially better tests faster and at lower costs are substantial. However, the standards for quality are equally high.

To this end, this study investigates the potential of LLM generated samples as a scalable, cost-effective alternative to expensive human samples, focusing on their ability to substitute human-response data for psychometric item selection in the pretesting phase. We first generate a novel dataset using *GPT-4o* by prompting the model with a newly developed set of personality items designed to assess work-specific Conscientiousness and Honesty-Humility. We then evaluate the resulting LLM-generated dataset in two ways:

(a) by assessing its similarity to a corresponding human dataset using a set of similarity metrics, and (b) by examining its utility for psychometric item selection. For the latter, we perform item selection independently on both an LLM-generated and a human-generated dataset, and then validate both item selections against a new, held-out human dataset, comparing their respective model fits.

2 Related Work

Synthetic data generation addresses challenges related to data availability and quality across various domains, with similar challenges also arising in psychological test development. Recent advancements in LLMs have increased the relevance of synthetic data generation in this context. This section provides a brief overview of the core concepts of synthetic data generation and emphasizes the specific role of LLMs in generating human-like response data. Subsequently, this study is situated within the emerging field of LLM psychometrics, as well as within the context of automated test development.

2.1 Synthetic Data Generation with LLMs

The central objective in synthetic data generation is the "generation of artificial data with the aim of reproducing the statistical properties of an original dataset" (van der Schaar and Maxfield, 2020). Generative models for this purpose aim to "learn a representation of an intractable probability distribution $\mathcal{X}$ defined over $\mathbb{R}^n$, where n typically is large, and the distribution is complicated" (Ruthotto and Haber, 2021, p. 2). To this end, one usually relies on a finite, often large set of independent and identically distributed (i.i.d.) samples drawn from $\mathcal{X}$ as training data. The goal is to learn a generator function

$$g : \mathbb{R}^q \to \mathbb{R}^n, \tag{1}$$

that maps samples from a simpler, tractable distribution $\mathcal{Z}$, defined over $\mathbb{R}^q$, to points in $\mathbb{R}^n$ that resemble those drawn from $\mathcal{X}$. Specifically, it is assumed that for each sample $\mathbf{x} \sim \mathcal{X}$ there exists at least one $\mathbf{z} \sim \mathcal{Z}$ such that $g(\mathbf{z}) \approx \mathbf{x}$ (Ruthotto and Haber, 2021).
Traditionally, synthetic data generation has been based on statistical models such as copulas (e.g., Li et al., 2020) and Bayesian networks (e.g., Zhang et al., 2017). Recent advances in generative modeling with neural networks have led to the development and application of various methods such as variational autoencoders (VAE) (e.g., Darabi and Elor, 2021, Liu et al., 2023, Ma et al., 2020, Vardhan and Kok, 2020), generative adversarial networks (GAN) (e.g., Baowaly et al., 2019, Choi et al., 2017, Liu et al., 2023, Park

et al., 2018, Xu et al., 2019), and diffusion models (e.g., Kotelnikov et al., 2023, Xu et al., 2023, Yang et al., 2024, Zhang et al., 2023).

These synthetic data generators have emerged as a promising solution to challenges encountered in domains, such as healthcare, finance, and cybersecurity, where collecting large, high-quality datasets is often expensive, time-consuming, or constrained by privacy concerns (Borisov et al., 2022, Dastile et al., 2020, Shwartz-Ziv and Armon, 2022). In these settings, synthetic data can be used to mitigate common training data limitations such as sparsity, class imbalance, and bias against underrepresented groups (Van Breugel and Van Der Schaar, 2024). Consequently, applications include data augmentation, missing value imputation, and rebalancing of class distributions (Jolicoeur-Martineau et al., 2024, Onishi and Meguro, 2023, Sauber-Cole and Khoshgoftaar, 2022).

While similar data challenges arise during the pretesting phase of test development, synthetic data generation has become especially relevant in this context with the emergence of LLMs (e.g., Li et al., 2023, Long et al., 2024).[1] Particularly with respect to modeling human-like response behavior in pretesting studies, LLMs exhibit important properties. First, synthetic responses should resemble those of plausible individuals. Language is often used as a proxy for underlying cognitive and social processes (e.g., Pennebaker and King, 1999, Schwartz et al., 2013). For example, the *lexical hypothesis* in personality psychology posits that personality traits are reflected in language use (Cattell, 1943). Therefore, LLMs may be capable of simulating diverse, human-like response patterns through text. Second, effective pretesting requires that responses reflect the influence of subtle differences in item phrasing on responses. Trained on large-scale corpora via autoregressive objectives (Brown et al., 2020, Vaswani et al., 2017), LLMs learn conditional dependencies that capture syntactic, semantic, and pragmatic structure, making them well-suited to detect and reproduce fine-grained linguistic variation.

2.2 LLM Psychometrics

An emerging research field makes use of psychometric methods to evaluate, validate, and enhance LLMs with respect to their personality, values, morality, and attitudes or opinions (Ye et al., 2025), for instance to justify their use in social science studies (Petrov et al., 2024, Serapio-García et al., 2023, Wang et al., 2024), agent-based modeling (Liu et al., 2025), or chatbots (Wang et al., 2025). Summarized under the term *LLM psychometrics*, Ye et al. (2025) define it as "the interdisciplinary field dedicated to evaluating, understanding, and enhancing LLMs through the application and integration of psycho-

[1] For details on deep learning algorithms and the relevant model architectures underlying LLMs, see `https://slds-lmu.github.io/dl4nlp/` (Aßenbacher et al., 2025).

metric instruments, theories, and principles" (p. 4). The authors give an extensive literature overview of this field, differentiating between the core ideas of construct-oriented psychometric evaluation and task-specific AI-benchmarking, introducing the psychometric measurement framework with its relevant constructs (personality and cognitive constructs), different test formats in psychometrics and their translation to LLM administered prompts, as well as the psychometric quality criteria of the results.[2]

While this overview highlights the extensive research on using psychometrics to evaluate LLMs, there is comparatively limited work on the reverse – using LLMs *for* psychometrics. Nonetheless, LLM psychometrics offers valuable insights for our work, especially in the area of data generation. Particularly relevant is existing research in the third area of LLM psychometrics defined by Ye et al. (2025), named *enhancement*. Research focused on this has explored approaches to manipulate traits to produce personalized and controlled LLM behaviors, facilitating "the controlled simulation or alignment of synthetic personas" (Ye et al., 2025, p. 34). By actively shaping LLMs along desired dimensions such as the Big Five personality traits through prompt-engineering, researchers have shown that outputs can be aligned with human-like personality profiles to certain degrees (e.g., H. Jiang et al., 2023, Serapio-García et al. 2023, Wang et al. 2025). Most recently, Wang et al. (2025) found that in psychometric evaluations of LLM-generated data, the models exhibited higher internal consistency and more clearly defined factor structures than those observed in the human personality profiles they were intended to replicate.

One limitation of studies from LLM psychometrics is a focus on trait-level evaluation, offering little insight into item-level distributions, crucial for test development (Wang et al., 2024). Few studies report item-level results. For instance, Petrov et al. (2024) and Wang et al. (2024) assessed LLM-generated personality data with more focus on item-level, using various scales and LLMs. Both studies concluded from psychometric analyses and item-level distributions, that LLM responses are not yet capable of simulating personality at a human level.

However, the conclusions from these results are limited for test development for several reasons. For instance, with respect to response generation, the researchers rely on the Persona-Chat dataset Zhang et al. (2018), which includes persona statements, such as "My favorite food is mushroom ravioli", to increase variability in the results in a controlled manner. The Persona Chat dataset was designed to support "more engaging chit-chat dialogue agents" (Zhang et al., 2018, p. 1) which may distort responses in the direction of personality traits such as agreeableness or openness. Additionally, Wang et al. (2024) used a simplified nine-level trait scale and randomly paired these with the Persona Chat

[2] This work uses psychometric terminology related to reliability and validity. For an introduction, see Serapio-García et al. (2023) or Ye et al. (2025).

statements, potentially yielding unrealistic personas with limited representativeness and comparability to human data. Finally, most studies in LLM psychometrics draw from established psychometric-inventories, such as International Personality Item Pool (IPIP)-NEO (e.g., Serapio-García et al., 2023) and HEXACO (e.g., Barua et al., 2024). LLMs may have seen the items used and respective response data during training, limiting their value for evaluating unknown pretest items. Additionally, these established scales include only validated items, which is in contrast to the item selection task in pretesting.
Against this background, this study takes the angle of LLM-generated data *for* psychometrics and addresses limitations of existing studies relevant to the use case of item selection for pretesting.

2.3 Automated Test Development

Recent research has begun exploring the use of LLMs in the broader test development process, aiming to improve validity of psychometric instruments while reducing related efforts at various steps in the test development process. Beginning with automated item generation, research has explored the potential of LLMs to create personality items, aiming to reduce the time and effort typically required for literature reviews and expert consultations (e.g., Götz et al., 2023, Hommel et al., 2022, Lee et al., 2023, von Davier, 2018). As another example, Hommel and Arslan (2024) used vector space representations, or embeddings, of item semantics to predict psychometric properties such as correlations between items, reliability of scales, and inter-scale correlations to reduce costly data collection to some degree. To the best of our knowledge, only one study to date by Liu et al. (2025) in the field of educational measurement has explicitly investigated the application of LLMs to psychometrics for the purpose of item selection. In this study, LLM-augmented data was used to recover Item Response Theory parameters of test items. While data generated solely by LLMs lacked sufficient variability, the combination of LLM-generated and human-generated data yielded promising results. While their objectives align with ours, knowledge-based assessments differ from non-cognitive pretests in that they require objectively correct or incorrect answers, which may account for the lower variance observed in their results.

3 This Work

While psychometrics has been used in LLM psychometrics to evaluate LLMs, as described in Section 2.2, limited research has focused on the potential of LLM-generated data for psychometric evaluations, particularly for item selection at the pretesting stage in test development. This study takes this new angle, drawing from the existing literature on synthetic data generation and LLM psychometrics for data generation and evaluation, and refining it for the use-case.

Tackling the risk of overfitting, the personality scale used and related human data have been unpublished at the time of data generation and were thus unknown to the model. The item set itself was generated by GPT-3.5, contributing to research on LLM-assisted item development and the automated test development pipeline that was described in Section 2.3.

Furthermore, our evaluation integrates metrics from synthetic data generation, where evaluation has always been central, and psychometrics – two domains that have largely remained separate to date. Such integration is also critical as LLM psychometrics and AI benchmarking research continue to gain relevance.

Finally, we extend existing evaluations for the target use case of pretesting by conducting item selection on the LLM-generated data. This allows us to illustrate both the practical utility and potential pitfalls when LLMs are prompted for quick diagnostic checks, especially with prompts resembling realistic user behavior.

4 Data Generation

Building on this background, the present study leverages an LLM to generate synthetic data for the use case of item selection during the pretesting stage. The data generation setup includes the choice of model, the pretest items, and a prompt design used to present the items to the model.

4.1 Model Choice

This study employs OpenAI's GPT-4o model with an estimated 200 billion parameters (Abacha et al., 2024), released on May 13, 2024. GPT-4o was selected due to the demonstrated success of OpenAI's models in comparable research contexts (Wang et al., 2025), its state-of-the-art performance to date across a wide range of tasks (OpenAI, 2024), and open accessibility, which supports transparency and reproducibility, aligning with recent recommendations for the use of LLMs in psychological research (Abdurahman et al., 2024). The model was accessed via the OpenAI API without fine-tuning.

4.2 Test Items

The focus of this study is on pretesting structured tests with predefined questions and response formats. Specifically, the Likert-scale response format ("How much do you agree with this statement") is used, which is also often used in evaluating LLM personality traits (Ye et al., 2025). The items used in this study are designed to assess the traits of Conscientiousness and Honesty-Humility in the work context, comprising 20 descriptive rating statements for each construct. The full set of items is provided in Appendix A.1.1 for Conscientiousness at Work and Appendix A.1.2 for Honesty-Humility at Work. For better readability, the item abbreviations also listed in the appendices are predominantly used in the following analyses. These items were selected for their novelty to the model and because analyses on a human dataset prior to this study identified them as including both well- and poorly-performing items. This allows for an investigation into whether the LLM-generated data can help distinguish between more or less suitable items and in this way help form psychometrically sound scales. Additional background on the item set, including definitions of the target constructs, is given in Appendix A.1.3.

4.3 Prompt Design

To generate responses, a role-playing prompt was employed. Similar approaches have been established in LLM psychometrics (e.g., G. Jiang et al., 2023, Serapio-García et al. 2023, Wang et al. 2025). The created personas were based on real-world participant data, incorporating the demographic variables *age*, *sex*, *nationality*, *income*, and *relationship status*. Each persona also included sum scores for general Conscientiousness and Honesty-Humility derived from assessments in the human sample using IPIP-NEO-60 (Maples-Keller et al., 2019) and HEXACO-60 (Ashton and Lee, 2008), respectively. This enabled personality steering while reflecting realistic test development scenarios, in which related personality data may be available, but scores for the target trait scores typically are not. Full details on the underlying human dataset are provided in Appendix A.1.4 (*Human Base Dataset*). Each prompt was submitted as a separate API call using the completion endpoint of the model, generating a full response set per persona. All tests used zero-shot prompting and deterministic settings (`temperature = 0`) to support reproducibility and enable comparison with human data. A prompt excerpt is shown in Appendix A.1.5 and the full version has been made available on GitHub for reuse.[3]

[3] `https://github.com/CFroehner/SynthData-Psych.git`

5 Evaluation Procedure

The use of LLM-generated data for item selection during pretesting poses two main requirements on the synthetic data. First, the LLM-generated data must demonstrate response patterns similar to human response behavior not only at the aggregated trait-score level but also at the item-level. While most research in LLM psychometrics has focused on trait scores, reflecting its interest in the overall personality of LLMs, effective item selection depends on meaningful responses at the item level. Second, the data must reflect not only marginal distributions similar to those found in human data at both the trait and item levels, but also human-like covariance structures among items, which form the basis for item selection. To evaluate these properties in the LLM-generated data, we use two types of metrics: *statistical fidelity* and *data utility.* These criteria have also been suggested in a synthetic data evaluation benchmark by Hansen et al. (2023).

Statistical fidelity evaluates the similarity between synthetic and original data based on statistical properties (Snoke et al., 2018). Measures of statistical fidelity often capture only one aspect of the overall data structure (e.g., Hansen et al., 2023). That is, generative models may achieve high statistical fidelity, yet the synthetic data may fail to capture subtle patterns present in real data (Hansen et al., 2023). This highlights the need to additionally evaluate data utility.

Data utility refers to “how well the synthetic data can be used in place of the real data for a given task” (Hansen et al., 2023, p. 3). In this study, the task is effective item selection in pretesting. To this end, utility is assessed by conducting psychometric item selection on the synthetically generated data and evaluating the resulting model on a new, human dataset. In synthetic data generation literature, this approach is often referred to as the *train-on-synthetic, test-on-real* paradigm (e.g., Hansen et al., 2023).

5.1 Fidelity

To evaluate fidelity, different statistical metrics have been proposed in the synthetic data generation literature, for which Wolf et al. (2024) gives an overview. These metrics include the *Kullback-Leibler divergence (KL divergence)* (e.g., Hansen et al., 2023, Qian et al., 2023) and the *Wasserstein distance* (e.g., Yin et al., 2022), which have been adopted in this study. In addition to these, this study introduces two random forest-based measures of fidelity, namely *propensity score mean squared error (pMSE)* and feature importance. The following section introduces each of these fidelity measures in detail.

5.1.1 Kullback–Leibler Divergence

One commonly used metric is the KL divergence. Given observed values from LLM-generated data as a probability distribution $Q(\cdot)$ and from human-generated data as a probability distribution $P(\cdot)$, the KL divergence compares the two distributions by measuring their relative entropy.

The KL divergence of P from Q is defined as (Kauermann, 2023)

$$KL(P, Q) = \int \log\left(\frac{p(y)}{q(y)}\right) dP(y), \tag{2}$$

where $p(y)$ and $q(y)$ are the densities or probability functions of the distributions P and Q, respectively.

If the distributions are continuous, this becomes

$$KL(P, Q) = \int \log\left(\frac{p(y)}{q(y)}\right) p(y)\, dy. \tag{3}$$

For discrete distributions, the integral is replaced by a sum

$$KL(P, Q) = \sum_{y} p(y) \log\left(\frac{p(y)}{q(y)}\right). \tag{4}$$

The KL divergence is always non-negative and equals zero if and only if $P(\cdot) \equiv Q(\cdot)$. Smaller values indicate greater similarity between the two distributions. Importantly, KL divergence is not symmetric, meaning in general $KL(P, Q) \neq KL(Q, P)$. Therefore, it is not a metric or distance measure, as symmetry is a required property of any metric by definition (Kauermann, 2023). Furthermore, KL divergence becomes infinite when comparing a distribution P to a distribution Q that assigns zero probability to outcomes where P assigns non-zero probability. This is addressed by the *Jensen-Shannon divergence (JS divergence)* (Lin, 2002), which will be used alongside KL divergence to assess robustness. The JS divergence can be interpreted as the total KL divergence to a mixture distribution of the two distributions to be compared (Nielsen, 2020). The formula and a more thorough introduction is for instance given by Nielsen (2020). The `ddjensen` function from the R package `dad` (Boumaza et al., 2021) was used for analysis.

A step-by-step illustration of the KL divergence calculation is provided in Figure 1 based on an example by Kauermann (2023). In this adaptation, the continuous distributions of work-related Honesty-Humility trait scores are compared for human and LLM-generated data. In the top plot, the densities $p(y)$ (human data) and $q(y)$ (LLM data) are shown, which differ by a certain amount. The second plot shows the log ratio $\log(p(y)/q(y))$,

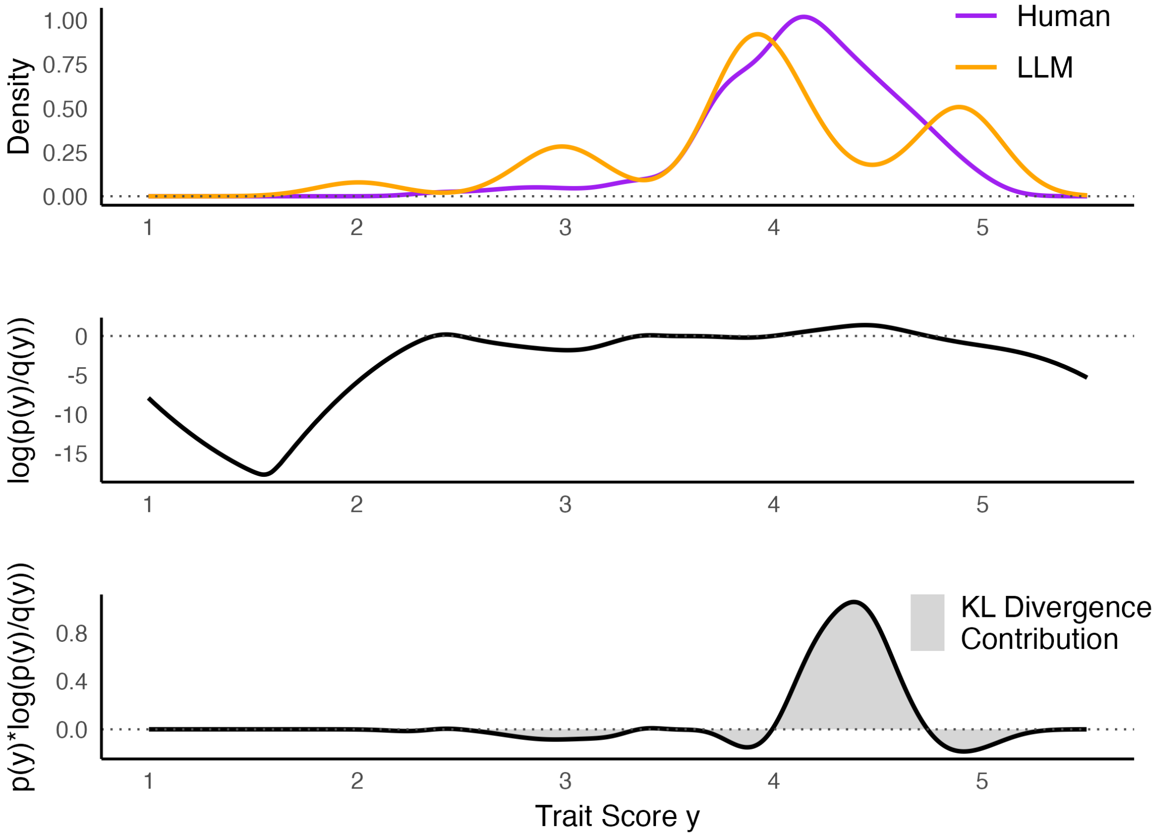


Figure 1: Illustration adapted from Kauermann (2023). Visualization of the Kullback-Leibler divergence calculation for the example of Honesty-Humility at Work trait scores. Top row shows the densities $p(\cdot)$, based on human data, and $q(\cdot)$, based on LLM data. The middle row gives the log ratio $\log(p(\cdot)/q(\cdot))$ and the bottom row shows the ratio weighted by $p(\cdot)$. Integrating the curve in the bottom plot yields the Kullback-Leibler divergence $KL(p(\cdot), q(\cdot))$.

which takes large absolute values especially in the lower range of y. However, the density $p(y)$ is small in these regions. By weighting the log ratio with the corresponding density $p(y)$, the third plot is obtained. As a result, discrepancies in the tails of the distribution, especially at the lower end, are downweighted, while smaller differences near the center of $p(y)$ become more dominant. Integrating this function, as sketched in grey in the bottom plot, yields the KL divergence.

5.1.2 Wasserstein Distance

While the KL divergence is a widely used measure of dissimilarity between probability distributions, it does not account for the geometry of the underlying space $\mathcal{X}$. The p–Wasserstein distance W_p provides an alternative that accounts for this. Rooted in the idea of optimal transport, Wasserstein distance quantifies the "minimal effort required to reconfigure the probability mass of one distribution in order to recover the other distribution" (Panaretos and Zemel, 2019, p. 1).

For a first understanding, Appendix A.2 illustrates the difference between the KL divergence and Wasserstein distance with a concrete example. The plot shows a reference distribution that is compared to two alternative distributions that differ in how far their probability mass lies from that of the reference distribution, which yields the same KL divergence but different Wasserstein distances.

The p-Wasserstein distance between probability measures μ and ν is defined as (Panaretos and Zemel, 2019)

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p \, \mathrm{d}\gamma(x, y) \right)^{1/p}, \quad p \geq 1, \tag{5}$$

where $\Gamma(\mu, \nu)$ denotes the set of probability measures γ on $\mathcal{X} \times \mathcal{X}$ such that $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$ for all Borel subsets $A, B \subseteq \mathcal{X}$. Elements $\gamma \in \Gamma(\mu, \nu)$ are called *couplings* of μ and ν, that is joint distributions on $\mathcal{X} \times \mathcal{X}$ with prescribed marginals μ and ν along each axis. Definition 5 has an intuitive interpretation in the discrete case (Panaretos and Zemel, 2019): given a coupling $\gamma \in \Gamma(\mu, \nu)$ and a pair of locations (x, y), the value of $\gamma(x, y)$ represents the amount of mass from μ located at x that is to be transported to y, in order to transform μ into ν. If the cost of moving a unit of mass from x to y is given by $\|x - y\|^p$, then $W_p(\mu, \nu)$ can be interpreted as the minimal cost required to reconfigure the mass distribution of μ into that of ν.

In the univariate case, the Wasserstein distance no longer requires optimization over different couplings and a closed-form expression is available (Panaretos and Zemel, 2019)

$$W_p(X, Y) = \left(\int_0^1 \left| F_X^{-1}(\alpha) - F_Y^{-1}(\alpha) \right|^p d\alpha \right)^{1/p}, \quad \mathbf{t}_X^Y = F_Y^{-1} \circ F_X, \tag{6}$$

where X and Y are random variables marginally distributed as μ and ν. F_X and F_X^{-1} denote the cumulative distribution function and quantile function of X, respectively, and $\mathbf{t}_X^Y$ is optimal when X is a continuous random variable (Panaretos and Zemel, 2019). When the vectors are of equal length and $p = 1$ is chosen, the `wasserstein1d` function from the R package `transport` (Schuhmacher et al., 2024) used in this study, approximates the Wasserstein distance from Formula 6.[4]

Importantly, computing the Wasserstein distance generally requires the specification of a distance measure between outcomes, often the Euclidean distance, but allowing for any

[4] The results closely matched an implementation of the formula for the case of $p = 1$ provided by Panaretos and Zemel (2019), which defines $W_1(X, Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| \, dt$.

norm (Candelieri et al., 2023). In this study, the implicit assumption of equidistant response categories may not hold. Still, the Wasserstein distance may provide valuable insights beyond KL divergence by respecting the ordinal structure of responses, that is, acknowledging that, for instance, *Strongly Agree* is closer to *Agree* than to *Strongly Disagree*, while keeping this limitation in mind.

5.1.3 Random Forest Based Fidelity

Two additional fidelity metrics were introduced based on a random forest model trained to predict the source of an observation in a combined dataset of human- and LLM-generated responses, namely pMSE and feature importance.[5] The pMSE is frequently defined in the machine learning literature as a utility metric (e.g., Snoke et al., 2018), as it reflects a model-based assessment that is typically related to the task for which the synthetic data is generated. In this study, since utility is defined more narrowly in the sense of test development and its associated modeling approaches, pMSE is grouped under fidelity, although it could reasonably be understood as a utility measure as well. The two fidelity measures are introduced in detail in the following.

Propensity Score Mean Squared Error The pMSE can be used to quantify how well a classifier distinguishes between data from two sources, such as synthetic and human generation. To compute it, a classifier, such as a random forest, is trained on a combined dataset of synthetic and human data, with an added target variable indicating the data source. The pMSE is calculated as the mean squared deviation between each estimated propensity score $\hat{p}_i$, in this study denoting the estimated probability that the observation i is synthetic, and c, the overall proportion of synthetic observations in the combined dataset. In the absence of informative features, c would be the best guess for the data source. The pMSE is defined as (Snoke et al., 2018)

$$\text{pMSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - c)^2, \tag{7}$$

where N denotes the total number of observations, in this study combining both LLM- and human-generated data, $\hat{p}_i$ the estimated propensity score for observation i, and c the proportion of LLM-generated data. In the random forest used in this study, $\hat{p}_i$ corresponds to the proportion of trees that classify observation i as synthetic. A low pMSE indicates poor separability between LLM- and human-generated data, suggesting high similarity.

[5] It is assumed that the reader is familiar with the general principles of random forests. For an introduction, see `https://slds-lmu.github.io/i2ml/chapters/07_forests/` (Bischl et al., 2022).

Conversely, a high pMSE implies that the classifier can better distinguish between the two sources. The random forest algorithm was implemented using the `randomForest` function from the R package `randomForest` (Liaw and Wiener, 2002).

Unlike KL divergence and Wasserstein distance, the pMSE provides a model-based evaluation. Combined with a random forest as the classifier that uses full item responses per trait as input features, the pMSE can capture differences that may not be reflected in KL divergence and Wasserstein distance as implemented in this study but arise through feature interactions. In this way, pMSE complements the previous fidelity criteria by providing first insights beyond marginal distributions, which will be examined more closely through the utility evaluations described in Section 5.2.

Feature Importance To make a statement about similarity at item-level, using the idea of a trained classifier, we can examine feature (item) importance derived from the random forest classifier.

To quantify feature importance, the *Mean Decrease in Impurity* is used. Let $\mathcal{F}$ be an ensemble of N_T decision trees. Then the importance of a feature $X_j \in \{X_1, \ldots, X_m\}$ is defined as (in adherence to Louppe et al., 2013)

$$\mathrm{Imp}(X_j) = \frac{1}{N_T} \sum_{T \in \mathcal{F}} \sum_{\substack{t \in T \\ v(s_t) = X_j}} \pi(t) \Delta i(s_t, t), \tag{8}$$

where $v(s_t)$ denotes the feature used at the splitting criterion s_t at node t, $\pi(t)$ is the proportion of training samples that reach node t, and $\Delta i(s_t, t)$ is the decrease in impurity caused at node t due to s_t. The inner sum ranges over all nodes t in tree T where feature X_j is used for splitting, and the outer sum takes the mean over all trees $T \in \mathcal{F}$.

The impurity decrease $\Delta i(s_t, t)$ is defined as (Louppe et al., 2013)

$$\Delta i(s_t, t) = i(t) - \pi(t_L) i(t_L) - \pi(t_R) i(t_R), \tag{9}$$

where $i(t)$ is the impurity of the parent node t, $i(t_L)$ and $i(t_R)$ are the impurities of the left and right child nodes, respectively. The terms $\pi(t_L)$ and $\pi(t_R)$ denote the proportions of samples that are sent to the left and right child nodes from node t.

In this study, impurity $i(t)$ is measured using the Gini index, which quantifies the probability that a randomly chosen sample would be misclassified if it were assigned a label based on the class distribution at that node. The Gini impurity at node t is given by (Daniya et al., 2020)

$$i(t) = 1 - \sum_{k=1}^{K} \pi_k(t)^2, \tag{10}$$

where K is the number of classes, which in this study is $K = 2$ representing different data sources, and $\pi_k(t)$ is the proportion of samples belonging to class k at node t.

5.2 Utility

To evaluate the utility of LLM-generated data in the pretesting phase of test development, it must effectively support item selection by capturing the relevant underlying covariance structures. To assess this, an item selection algorithm is applied to the LLM-generated dataset which includes exploratory factor analysis (EFA) and internal consistency analysis. This selection process is followed by confirmatory factor analysis (CFA) on a held-out human-generated dataset, following the train-on-synthetic, test-on-real paradigm. For comparison, the same procedure is also applied to a human-generated dataset. The next section introduces the idea behind EFA, internal consistency, and CFA, with a focus on the model fit indices used in the evaluation. A detailed overview of the full item selection algorithm is provided in Appendix A.3.

5.2.1 Exploratory Factor Analysis

Following the introduction by Cudeck (2000), EFA can be understood in two key components. The first is that each observed variable can be expressed as a linear combination of latent factors. In this study these latent factors are assumed to be Conscientiousness and Honesty-Humility at Work.

Formally, let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im})^\top$ denote the observed score vector for individual i across m standardized variables. Each observed variable y_{ij} with $j = 1, \ldots, m$ is modeled as a linear function of l latent factors (Cudeck, 2000)

$$y_{ij} = \sum_{k=1}^{l} f_{jk} z_{ik} + e_{ij}, \tag{11}$$

where the regression coefficient f_{jk} is called *factor loading* of variable j on factor k, z_{ik} is the unobserved score of individual i on factor k and e_{ij} is the residual term. The residuals are assumed to be uncorrelated with the common factors and with each other.

The second component of the model involves explaining the associations among observed variables. Specifically, the association between any two observed variables, $j \neq j'$, is determined by the product of their respective factor loadings and the correlations among the latent factors themselves (Cudeck, 2000).

We define $\mathbf{F} \in \mathbb{R}^{m \times l}$ as the matrix of factor loadings (or regression coefficients), $\mathbf{R}_z \in \mathbb{R}^{l \times l}$ as the factor correlation matrix (matrix of correlations among latent factors), and $\mathbf{U}^2 \in \mathbb{R}^{m \times m}$ as the diagonal matrix of unique variances or residual variance with the assumption that the residuals are uncorrelated. Then the correlation matrix of the observed variables, $\mathbf{R}_y \in \mathbb{R}^{m \times m}$, can be expressed as follows (in line with Cudeck (2000), adapted to matrix notation)

$$\mathbf{R}_y = \mathbf{F}\mathbf{R}_z\mathbf{F}^\top + \mathbf{U}^2. \tag{12}$$

To estimate the relevant parameter matrices, the sample correlation matrix is computed, and the matrices $\hat{\mathbf{F}}$, $\hat{\mathbf{R}}_z$, and $\hat{\mathbf{U}}$ are obtained by fitting the model to the data. Before going into the details of this estimation, one commonly reported value alongside factor loadings is *communality*, which is defined as the proportion of variance in y_j explained by the l latent factors (Shrestha, 2021). Communalities will also be used in the course of this study, as part of the item selection algorithm. Formally, assuming all variables are standardized, that is $\mathrm{Var}(y_j) = 1$, the communality of variable j, denoted h_j, is defined as

$$h_j = 1 - [\mathbf{U}^2]_{jj}. \tag{13}$$

Continuing with the estimation, three central decisions must be made in advance: the method of parameter estimation, the number of factors to extract, and the choice of rotation method.

First, regarding parameter estimation, one approach is ordinary least squares (OLS) which will be used in this study. It minimizes the squared differences between the sample correlation matrix and the reproduced correlation matrix derived from the estimated parameters (Cudeck, 2000). A widely used alternative is maximum likelihood estimation, which assumes that the variables are normally distributed or at least approximately symmetric (Cudeck, 2000). Additional estimation methods exist and are provided in relevant R packages such as `psych` (Revelle, 2025).

Second, when selecting an appropriate number of factors, various approaches can be employed. Common methods include the eigenvalue-greater-than-unity rule, the scree test, and Horn's parallel analysis (Horn, 1965). Among these, Horn's parallel analysis is widely regarded as one of the most accurate approaches for determining the number of factors to retain (Glorfeld, 1995, Hayton et al., 2004, Velicer et al., 2000, Zwick and Velicer, 1986). The specific procedure behind Horn's parallel analysis is described in Appendix A.4. In this study, parallel analysis was performed using the `paran` function from the R package `paran` (Dinno, 2025).

Third, a rotation method must be specified. Since different sets of factor loadings can produce the same correlation matrix, factor analysis relies on rotation to impose a simple structure, resolving this indeterminacy and yielding interpretable solutions (Cudeck, 2000). Numerous rotation methods are available. Following the recommendation of Cudeck (2000), this study employs the oblique rotation method *oblimin*, as described in Revelle (2025), which allows for correlations among factors by estimating the off-diagonal elements of the factor correlation matrix $\mathbf{R}_z$ from the data. To assess the robustness of the findings, EFA is also conducted using *varimax* rotation (Kaiser, 1958), an orthogonal method that assumes the off-diagonal elements of $\mathbf{R}_z$ are zero. EFA was carried out using the `fa` function from the R package `psych` (Revelle, 2025).

After conducting EFA, we can conduct item selection based on the factor loadings. Based on previous research, the following selection algorithm is applied. Building on established item selection procedures with human data, all items with primary factor loadings below 0.40 (e.g., Boateng et al., 2018, Clark and Watson, 2019, Ford et al., 1986a, Götz et al., 2023, Rosellini and Brown, 2021) or items with cross-loadings above 0.30 (Boateng et al., 2018, Costello and Osborne, 2005, Götz et al., 2023) are removed. This elimination process is repeated iteratively, removing items from the pool until no items remain that meet the exclusion criteria (see Step 1, named *Iterative EFA Filtering*, in the item selection algorithm, Appendix A.3). Second, all items with communalities below 0.30 are excluded (Carpenter, 2018, Fabrigar et al., 1999, Götz et al., 2023). This criterion is outlined as Step 2 of the item selection algorithm, titled *Communality Thresholding*. As the final selection step, referred to as *Top-Loading Selection* in the item selection algorithm, the top $k = 4$ items with the highest factor loadings are retained, following an *Internal Consistency Check* described in the next section. The top loading selection allows for better comparison between the scales derived from synthetic and human data because of equal length and reflects a practical scenario in which a researcher may aim to develop a short and efficient scale, as opposed to retaining all acceptable items for subsequent CFA.

5.2.2 Internal Consistency

Conducting EFA on LLM-generated data yields a model that can be further refined by evaluating internal consistency and potentially removing items that negatively impact it. In the selection algorithm in Appendix A.3, this step is referred to as the *Internal Consistency Check*. Beyond item selection, internal consistency of the final scales is also assessed during the validation analyses.

A widely established measure of internal consistency is *Cronbach's* α (Cronbach, 1951), which is defined as (Mcneish, 2017)

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} \sigma_j^2}{\sigma^2},\right) \tag{14}$$

where m denotes the number of items, σ_j^2 the variance of item j, and σ^2 the variance of the total scores, calculated as the sum of item scores. Higher values of Cronbach's α indicate greater internal consistency. Item selection in the internal consistency check is based on iteratively dropping one item at a time and comparing the change in Cronbach's α to a predefined threshold for exclusion.

One limitation of Cronbach's α is its assumption of *tau equivalence*, stating that "each item on a scale contributes equally to the total scale score" (Mcneish, 2017, p. 4). When this assumption is violated, *McDonald's* ω (McDonald, 2013) is often preferred, which makes use of a factor analytic approach. To ensure robustness, McDonald's ω_{total} will be evaluated alongside Cronbach's α in validating the final scales.[6] Assuming that the latent construct variance is constrained to 1, and that measurement errors are uncorrelated, ω_{total} is defined as follows (Mcneish, 2017)

$$\omega_{total} = \frac{\left(\sum_{j=1}^{m} f_j\right)^2}{\left(\sum_{j=1}^{m} f_j\right)^2 + \sum_{j=1}^{m} \theta_{jj}}, \tag{15}$$

where f_j denotes the factor loading of item j on a single latent factor, and θ_{jj} denotes the error variance of item j (corresponding to $[\mathbf{U}^2]_{jj}$ in Formula 13). McDonald's ω is conceptually similar to Cronbach's α in that it compares the variability explained by the items with the total variability of the scale. When factor loadings are equal across items, ω_{total} is equivalent to Cronbach's α (McDonald, 2013).

The analysis was conducted using the `alpha` and `omega` functions from the `psych` package (Revelle, 2025). The `omega` function in `psych` performs an EFA, implementing a specific variance decomposition along with a Schmid–Leiman rotation of the factor solution (Schmid and Leiman, 1957). The interested reader is referred to Mcneish (2017), providing a more detailed discussion on the nuanced differences in implementation of McDonald's ω in `psych`.

[6] Compared to McDonald's $\omega_{hierarchical}$, McDonald's ω_{total} assumes unidimensionality and no additional dimensions (Mcneish, 2017). This assumption will be applied per trait in the later analysis.

5.2.3 Confirmatory Factor Analysis and Fit Indices

CFA assesses the extent to which a hypothesized model resulting for instance from EFA aligns with (unseen) empirical data by "testing whether the model-implied covariance structure reproduces the empirical covariance matrix (or resembles it very strongly)" (Goretzko et al., 2024, p. 124). To evaluate this, different methods exist. Firstly, a chi-squared test can be conducted assessing the overall model fit. Secondly, various indices can be used to quantify the goodness of fit or the deviance from the perfect model fit. In the following, four of these metrics are introduced and later evaluated for the models derived from synthetic and from human data: *root mean squared error of approximation (RMSEA)*, *standardized root mean square residual (SRMR)*, *Tucker-Lewis index (TLI)* and *comparative fit index (CFI)*. CFA and the calculation of fit indices were conducted using the `cfa` function from the `lavaan` package (Rosseel, 2012) in R.

Root Mean Squared Error of Approximation The RMSEA (Steiger, 1998) replaces the null hypothesis underlying the global χ^2 test of model fit. Rather than testing the hypothesis of exact fit, that is that the specified model perfectly reproduces the population covariance structure, it assesses the extent of approximate fit (Goretzko et al., 2024). Specifically, the statistic is derived from the likelihood-ratio test comparing the fitted model to a saturated model, which has zero degrees of freedom (df), reproducing the observed variance-covariance matrix. The resulting χ^2 statistic is evaluated relative to its expected value under the null hypothesis, which is equal to the model's df (Goretzko et al., 2024). The RMSEA is given by (Goretzko and Bühner, 2022)

$$\text{RMSEA} = \sqrt{\max\left(\frac{\chi^2 - df}{(N-1)df},\ 0\right)}, \tag{16}$$

where N denotes the sample size. Values of RMSEA closer to zero indicate better model fit.

Standardized Root Mean Square Residual The SRMR (Hu and Bentler, 1998) is another measure of model misfit that "quantifies the average standardized differences between each bivariate empirical correlation and the respective model-implied counterpart" (Goretzko and Bühner, 2022, p. 125). It is defined as follows (Goretzko and Bühner, 2022)

$$\text{SRMR} = \sqrt{\frac{\sum_{j=1}^{m}\sum_{j'=1}^{m}\left(\frac{s_{jj'}-\hat{\sigma}_{jj'}}{s_j s_{j'}}\right)^2}{m(m+1)}}, \tag{17}$$

where m denotes the number of observed variables or items, $s_{jj'}$ is the empirical covariance between the j-th and j'-th items, $\hat{\sigma}_{jj'}$ is the covariance implied by the model, and s_j, $s_{j'}$ are the empirical standard deviations of the j-th and j'-th items, respectively. An SRMR of zero indicates that the model perfectly reproduces the empirical covariance matrix and higher values indicate increasing misfit.

Tucker-Lewis Index The TLI (Tucker and Lewis, 1973) assesses fit of a proposed model relative to a baseline model. This baseline model, also *independence model*, assumes that all observed variables are uncorrelated and that the error variance is zero, representing a diagonal matrix (Goretzko and Bühner, 2022). The TLI is computed as (Goretzko and Bühner, 2022)

$$\mathrm{TLI} = \frac{\frac{\chi^2_{\mathrm{Independence}}}{df_{\mathrm{Independence}}} - \frac{\chi^2}{df}}{\frac{\chi^2_{\mathrm{Independence}}}{df_{\mathrm{Independence}}} - 1}, \tag{18}$$

where $\chi^2_{\mathrm{Independence}}$ and $df_{\mathrm{Independence}}$ denote the chi-square statistic and df for the independence model, and χ^2 and df are the corresponding values for the proposed model, as described previously in the context of RMSEA. The TLI values can exceed one, with higher values indicating better model fit.

Comparative Fit Index The CFI (Bentler, 1990) makes use of the independence model as the TLI and is defined as (Goretzko and Bühner, 2022)

$$\mathrm{CFI} = 1 - \frac{\max(\chi^2 - df, 0)}{\max\left(\chi^2_{\mathrm{Independence}} - df_{\mathrm{Independence}},\ \chi^2 - df,\ 0\right)}. \tag{19}$$

CFA and the evaluation of fit indices were conducted using the `cfa` function from the `lavaan` package (Rosseel, 2012) in R.

One challenge in interpreting fit indices is determining which results are sufficient to justify proceeding with a model. This study will use established guidelines as broad reference points to aid interpretation. Specifically, RMSEA and SRMR values below 0.10 are considered acceptable and below 0.06 excellent, while CFI and TLI values above 0.90 are deemed acceptable and above 0.95 excellent (Finch and West, 1997, Hu and Bentler, 1999). While these guidelines are widely used, they can be misleading, as their suitability depends on aspects such as model misspecification (e.g., Bentler, 1990) and sample size (e.g., Ainur et al., 2017, Fan et al., 1999). Recent work proposes dynamic fit index cutoffs tailored to specific data contexts (e.g., see McNeish, 2023, McNeish and Wolf, 2023). However, as this study focuses on comparing model fits, conventional guidelines are used only as an additional reference, while acknowledging their limitations.

6 Results

In the following, the relevant three datasets are introduced for analysis, with main focus on the LLM-generated, synthetic dataset. Subsequently, results for fidelity analyses based on KL divergence, Wasserstein distance, pMSE and random forest feature importance are reported, both on trait and item level. Finally, utility of the data in the context of pretesting is reported.

6.1 Synthetic, Base and Test Set

The following results draw from three datasets: (1) a primary human dataset (*human base dataset*), which serves three purposes: (a) to provide aggregated, generic personality scores and demographic variables used in the prompt design from Section 4.3 for generating synthetic data, (b) to provide the corresponding item-level data for fidelity analysis, and (c) to support utility analyses and illustrate the results of a standard test development procedure using real human data; (2) the synthetic dataset, generated by the LLM (*synthetic* or *LLM-generated dataset*); and (3) a held-out human dataset composed of different individuals (*human comparison dataset / human test set*). This third dataset does not replicate the original sample but serves two purposes: (a) to support fidelity analysis by illustrating expected sampling variability, and (b) to conduct CFA on both synthetic- and human-based models using unseen data. Each of the three datasets contains $N = 150$ observations. Details on the human samples and their data collection are provided in Appendix A.1.4.

As an initial validation of the data generation procedure, trait- and item-level distributions were compared visually, together with mean and standard deviation per item. Furthermore, bias and mean squared error (MSE) were evaluated, to compare the assumed ground truth human response values with their predicted, LLM-generated response values. Details on the calculation of bias and MSE are provided in Appendix A.5.

With respect to trait-level distributions, Figure 2 shows the kernel density estimates of trait scores, that is the mean response score per individual, for Conscientiousness at Work and Honesty-Humility at Work across the three different datasets. The two human-derived distributions appear similar for both traits, particularly for Conscientiousness at Work, and both exhibit slight left-skewness. In contrast, the synthetic trait score distribution for Conscientiousness at Work exhibits higher kurtosis, while the distribution of Honesty-Humility at Work trait scores appears multimodal, rather than unimodal. Overall, this suggests that the LLM-generated distributions deviate more substantially from the human distributions.

On the item-level, Figures 3 and 4 show the response count distributions for each item in the Conscientiousness and Honesty-Humility at Work item pools, broken down by response category and data source (human base or synthetic). Many items across both traits exhibit left-skewed response patterns in both human and synthetic data. For Conscientiousness at Work, synthetic responses generally show lower variance compared to human responses. In contrast, for Honesty-Humility at Work, the variance in synthetic responses is often comparable to or exceeds that of human responses.

Sample mean and standard deviations per item, displayed in Appendix A.6 support these patterns, reporting relatively high item means across both human and LLM responses. For Conscientiousness at Work, the standard deviation ratios ($\mathrm{SD}_{\mathrm{LLM}}/\mathrm{SD}_{\mathrm{Human}}$) are consistently below 1, ranging from 0.33 to 0.66, indicating lower dispersion in synthetic responses. For Honesty-Humility at Work, several items show ratios above 1, ranging from 0.63 to 1.28, suggesting more variability in the synthetic data for those items.

With respect to bias, that is the mean prediction error per item, synthetic responses tend to systematically underestimate the human response level for most items across both traits (see Appendix 8a). The item with the largest positive bias is "I would never tell a lie to get a promotion", indicating that synthetic responses tend to overestimate agreement for this item, leaning more toward "strongly agree" than the human responses.

Looking at the ranking of items by their MSE in Appendix 8b, it can be observed that rankings differ between bias and MSE. This suggests that some items exhibit relatively high bias but low MSE (e.g., item `HH_W_doublecheck`), meaning that synthetic responses may systematically over- or underestimate the true response, but do so with relatively low variability or high consistency. Conversely, other items show low bias but high MSE (e.g., item `C_W_othersresponsibility`), indicating that while there is little systematic over- or underestimation, deviations are more variable or inconsistent.

Based on this initial validation, differences between human- and synthetic responses appear to exist, though they vary across traits and items. These differences will be further quantified in the following fidelity analysis, and their implications for the utility of LLM-generated data in pretesting contexts will be explored.

6.2 Fidelity

To evaluate fidelity, datasets are compared at both the item and trait level. The main focus lies on the comparison between the human base dataset and the LLM-generated dataset, as introduced in the previous section. The human comparison dataset is included where

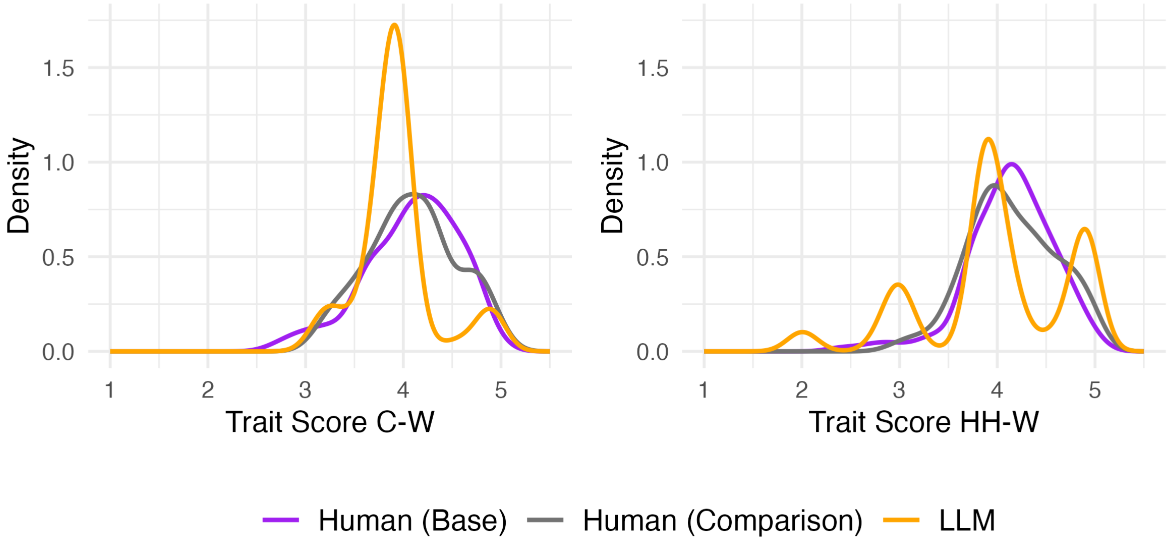


Figure 2: Kernel density estimates of trait scores for Conscientiousness at Work (C-W) and Honesty-Humility at Work (HH-W), based on three sources: the human base dataset, the human comparison dataset, and the LLM-generated dataset ($N = 150$ observations per source). A Gaussian kernel with a bandwidth of approximately 0.15 was used for all density estimations.

additional context on the natural variability between two human samples seems helpful. Trait-level fidelity is assessed first, using KL divergence, Wasserstein distance, and pMSE. The results are summarized in Table 1. KL divergence between the human base and LLM-generated distributions is relatively high for Conscientiousness at Work (2.28) compared to Honesty-Humility at Work (0.32). This observation is consistent with the kernel density estimates from Figure 2 and the theoretical properties of KL divergence discussed in Section 5.1.1. The higher KL divergence for Conscientiousness at Work can likely be attributed to high kurtosis in the synthetic distribution in regions where human density is also high, resulting in higher penalization. In contrast, for Honesty-Humility at Work, higher density in the synthetic data occurs in regions of low human density, which contributes less to KL divergence. In both traits, the divergence between the two human datasets is considerably lower. Computing the JS divergence, as a more robust alternative to KL divergence, yielded values of $D_{JS} = 0.23$ for Conscientiousness at Work and $D_{JS} = 0.08$ for Honesty-Humility at Work. This indicates that the qualitative pattern observed with KL divergence was preserved, although the difference in divergence between the two traits was smaller.

The differences observed in KL divergence diminish when considering the Wasserstein distance and the ordering between traits even reverses. Specifically, the Wasserstein

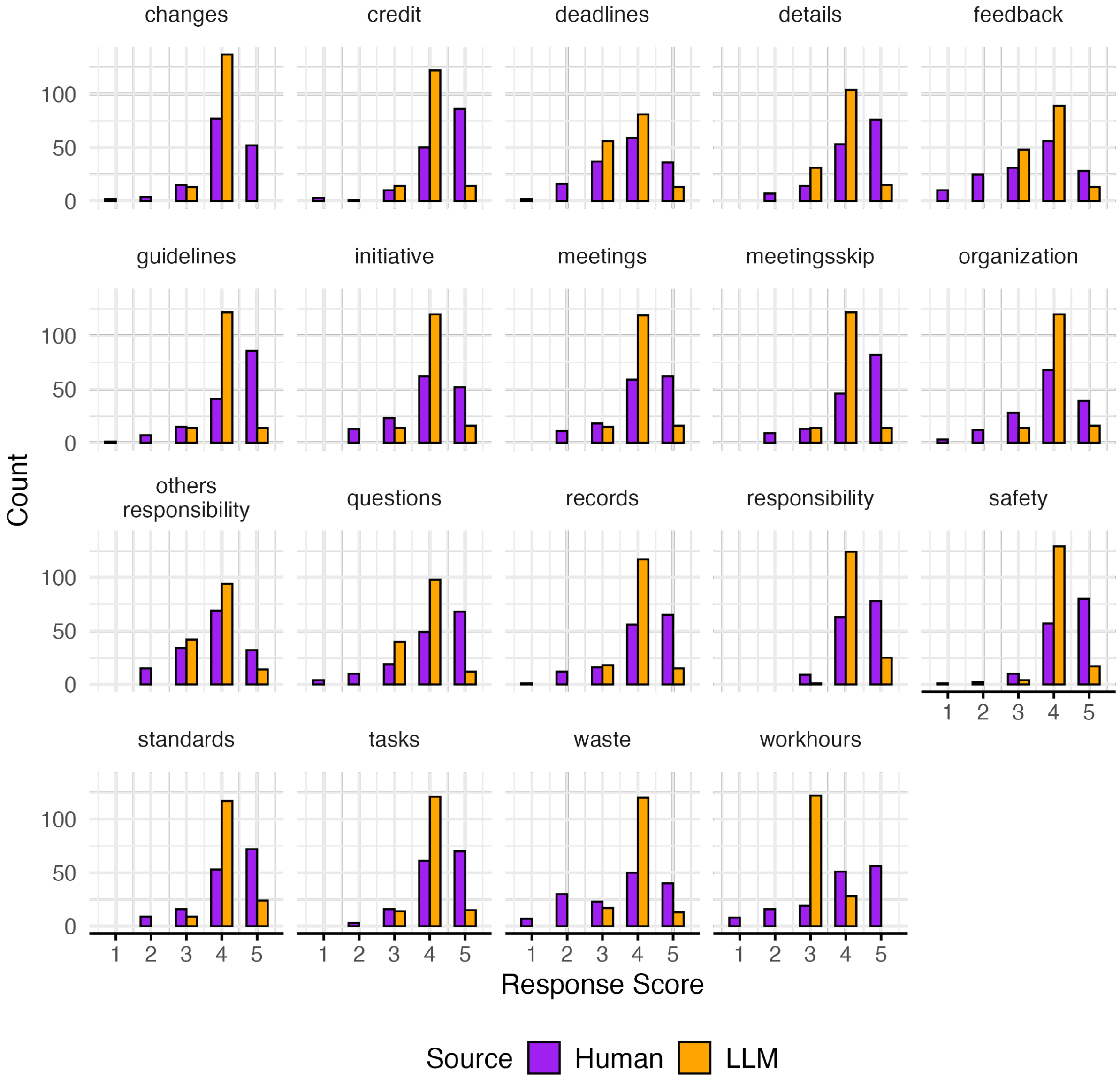


Figure 3: Response counts for each item in the Conscientiousness at Work item pool, broken down by response category (1 = *Disagree* to 5 = *Agree*) and data source (human base dataset, LLM-generated dataset). Full item texts are provided in Appendix A.1.1, with each item labeled using the prefix "C_W_" followed by its title. Notably, no data was available in the human dataset for the Conscientiousness at Work item "I double check my work before submitting it." (`C_W_doublecheck`). Consequently, this item was removed from both the human and LLM-generated datasets during preprocessing.

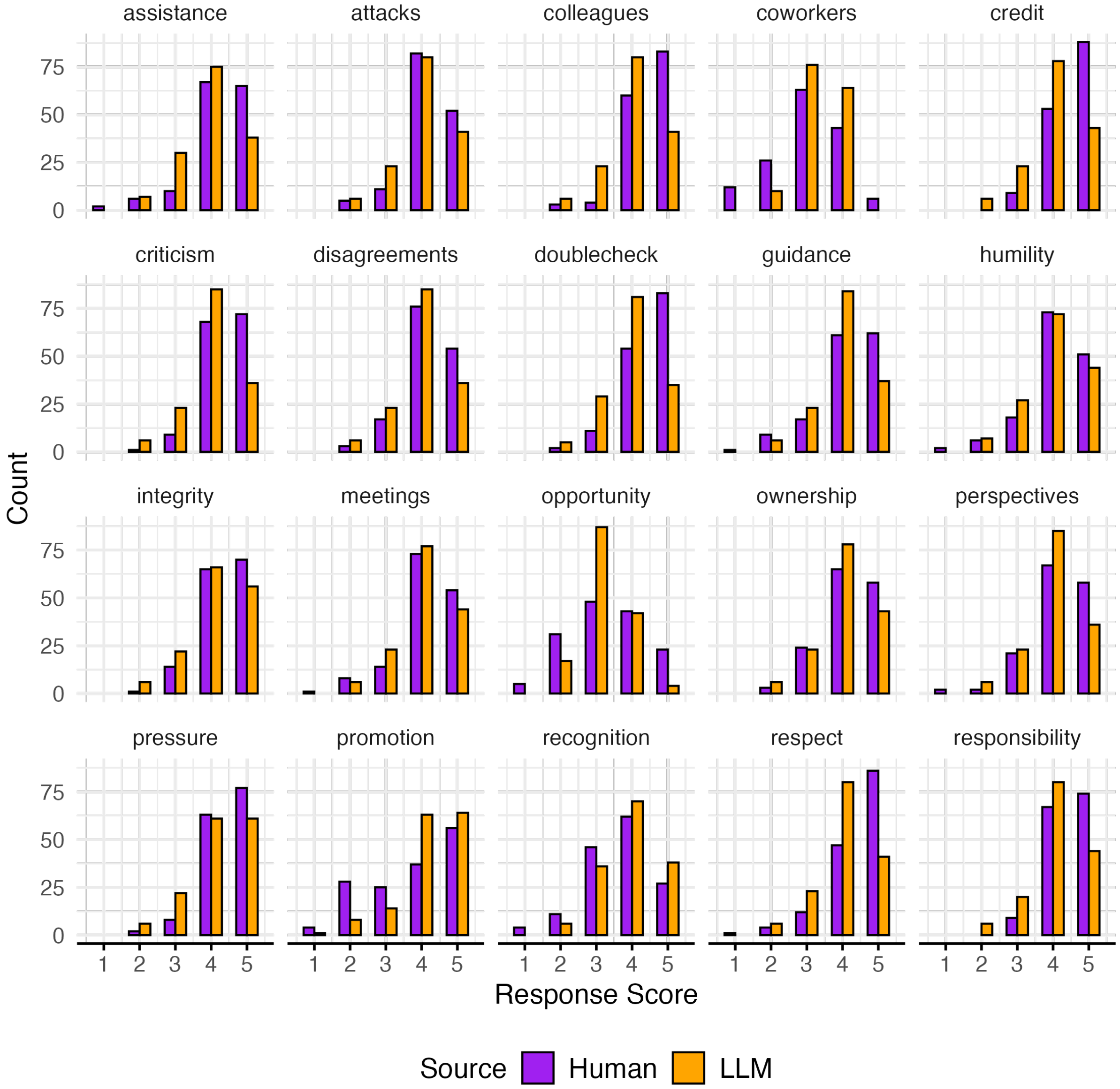


Figure 4: Response counts for each item in the Honesty-Humility at Work item pool, broken down by response category (1 = *Disagree* to 5 = *Agree*) and data source (human base dataset, LLM-generated dataset). Full item texts are provided in Appendix A.1.2, with each item labeled using the prefix “HH_W_” followed by its title.

distance between human and LLM distributions is smaller for Conscientiousness at Work than for Honesty-Humility at Work ($0.23 < 0.30$). This suggests that when accounting for the distance between probability masses, as done by the Wasserstein distance, the LLM-generated scores for Conscientiousness at Work are overall closer to the corresponding human distribution than for Honesty-Humility at Work. This is under the assumption that the response scale is equidistant. As with KL divergence, the Wasserstein distances between the two human datasets remain lower for both traits.[7]

Finally, pMSE shows comparable values for Conscientiousness and Honesty-Humility at Work. Results suggest that the random forest model assigns propensity scores higher than chance, indicating that synthetic and human responses are generally distinguishable. In contrast, the human-to-human comparison yields propensity scores that are close to chance. Notably, unlike KL divergence and Wasserstein distance, pMSE was calculated based on individual item responses rather than aggregated trait scores. These individual item responses will be examined in more detail through the following item-level fidelity assessment.

Table 1: Kullback-Leibler (KL) divergence, Wasserstein distance, and propensity mean squared error (pMSE) between the human base (Hum.) and LLM-generated dataset, and between the human base and human comparison (Comp.) datasets, for the traits Conscientiousness at Work (C-W) and Honesty-Humility at Work (HH-W).

Trait	KL Divergence		Wasserstein Distance		pMSE	
	Hum./LLM	Hum./Comp.	Hum./LLM	Hum./Comp.	Hum./LLM	Hum./Comp.
C-W	2.28	0.09	0.23	0.06	0.20	0.02
HH-W	0.32	0.10	0.30	0.07	0.19	0.02

For item-level fidelity analyses, KL divergence, Wasserstein distance and feature importance were computed, alongside JS divergence for robustness. The results are presented in Figure 5 (KL divergence, Wasserstein distance, JS divergence) and Figure 6 (feature importance), showing the Conscientiousness and Honesty-Humility at Work items ranked according to each respective measure.

As an initial observation, the results of the three fidelity metrics at the item level are highly correlated. Specifically, the KL divergence and mean decrease in Gini yield a Pearson correlation of $r_{\text{KL,Gini}} = 0.97$, and Wasserstein distance correlates with both the mean decrease in Gini and the KL divergence at $r_{\text{WS,Gini}} = r_{\text{WS,KL}} = 0.90$. This suggests that, while capturing different aspects of fidelity, the general item-level trends seem to be consistent across metrics. Notably, the Conscientiousness at Work item "I am flexible with my work hours" (`C_W_workhours`) ranks among the highest across all three metrics.

[7] Using $p = 2$ instead of $p = 1$ in the Wasserstein distance calculation preserved the qualitative pattern of results.

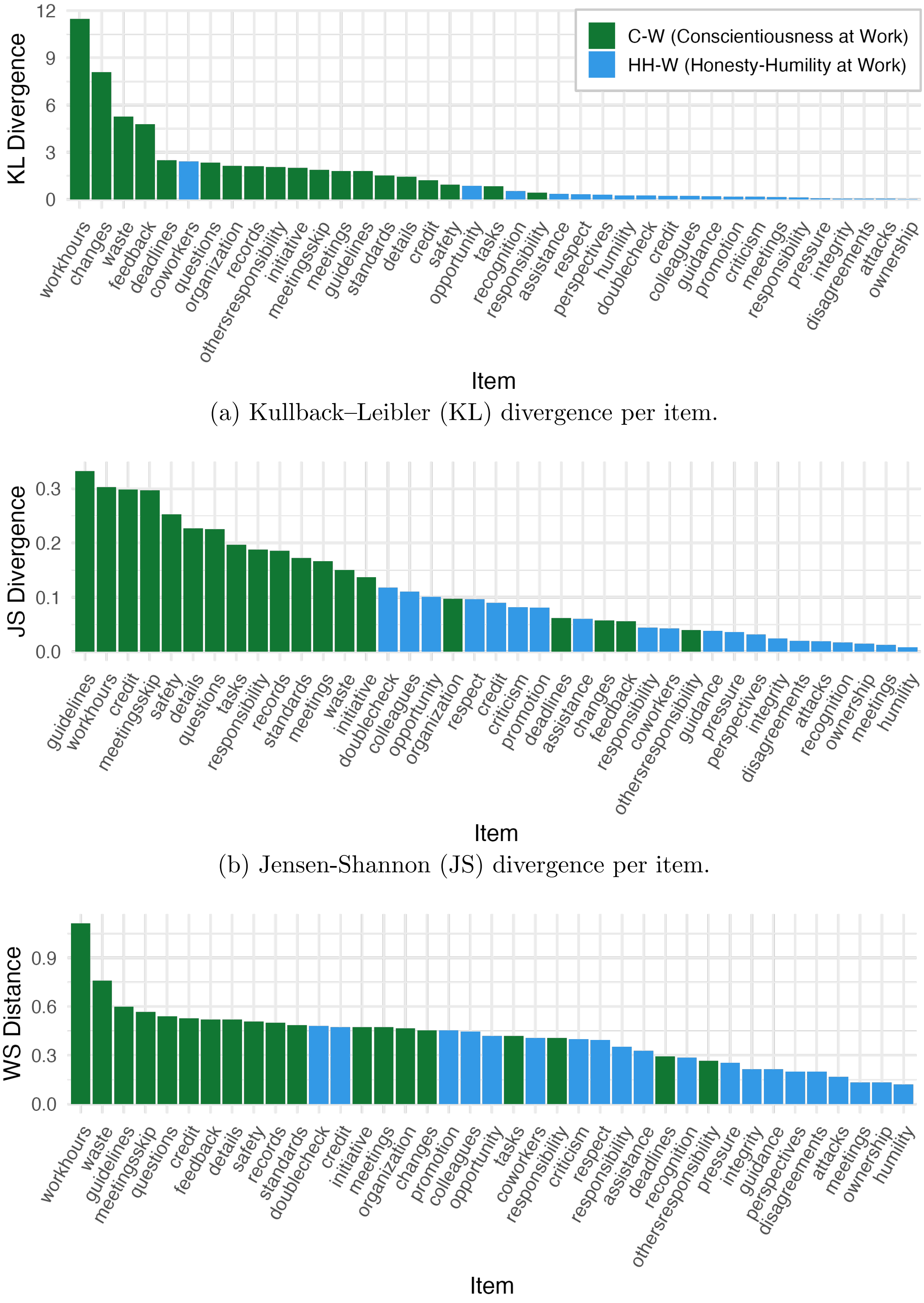
Going back to Figure 3, the distribution of responses for this item shows very low variance of LLM-generated responses compared to human respones, concentrated at the neutral category.

Regarding KL divergence per item (Figure 5a), the overall ranking shows that the item `C_W_ workhours` exhibits the highest divergence, followed by the items `C_W_changes,` `C_W_waste`, and `C_W_feedback`. Beyond these, differences in KL divergence across items are smaller, relatively consistently. In contrast, the more robust JS divergence displays greater variability across items (Figure 5b). Furthermore, the item-level KL divergence results are consistent with the previous trait level observations. Synthetic responses to Conscientiousness at Work items exhibit higher divergence than those for Honesty-Humility at Work items. One exception is the Honesty-Humility at Work item “I put the needs of my coworkers before my own,” which shows higher divergence compared to other items within the trait. However, this exception is not evident when using the JS divergence, for which Conscientiousness at Work items still tend to show lower fidelity than Honesty–Humility at Work items, but the difference is less pronounced.

For Wasserstein distance at the item level (Figure 5c), the items `C_W_workhours` and “I waste time at work” (`C_W_waste`) exhibit the greatest distance between synthetic and human responses. Looking at Figure 3 for the item `C_W_waste`, the LLM response distribution again shows small variance compared to the human responses, concentrated in the *Rather Agree* category. Subsequent items show relatively similar Wasserstein distances, suggesting less variation in fidelity. Only the final Honesty-Humility at Work items display comparably low Wasserstein distances, such as “At all times, I aim to maintain an attitude of humility” (`HH_W_humility`) and “When in a leadership position, I take ownership of successes and failures” (`HH_W_ownership`).

Additionally, the feature importance of a random forest model was assessed, with the results displayed in Figure 6. Feature importance was measured by the mean decrease in Gini impurity from a random forest trained on the full item pool of Conscientiousness at Work and Honesty-Humility at Work items. The results are broadly consistent with previous findings. Notably, every Conscientiousness at Work item consistently shows higher importance that Honesty-Humility at Work items, indicating that responses to these items contribute more strongly to changes in node impurity during prediction.

Based on the fidelity analysis, the results indicate higher trait fidelity for Honesty-Humility at Work than Conscientiosuness at Work based on KL divergence and JS divergence, comparable results based on pMSE, and higher fidelity for Conscientiousness at Work based on Wasserstein distance. As an initial check, we computed additional Pearson correlations to examine whether trait-specific differences in fidelity might be influenced by differences in the strength of association between generic trait scores – used to prompt the model –



(a) Kullback–Leibler (KL) divergence per item.

(b) Jensen-Shannon (JS) divergence per item.

(c) Wasserstein distance per item.

Figure 5: Item-level fidelity between LLM-generated and human responses across Conscientiousness at Work (C-W) and Honesty-Humility at Work (HH-W) items. Note that certain item suffixes appear twice per plot. Combined with their respective trait prefix (C-W, HH-W) they are unique and represent different items, as listed in Appendices A.1.1 and A.1.2.

and contextualized trait scores – used for fidelity evaluation – in the human data. The results revealed a stronger association between generic and contextualized Conscientiousness scores than between the corresponding Honesty-Humility scores. Item level analyses provide more detailed insights, showing variability in fidelity across items within both pools. In particular, the Conscientiousness at Work pool includes items with notably large divergences, such as `C_W_workhours` but also items with divergence comparable to Honesty-Humility at Work items, depending on the fidelity measure considered. The implications of this observed limited fidelity in certain items when using the LLM-generated data for item selection are examined in the subsequent utility analysis.

6.3 Utility

To assess the implications of partially limited fidelity in item selection within a pretesting context, item selection was conducted on the LLM-generated dataset and validated using the human test dataset. Before applying the selection algorithm, a parallel analysis was performed to determine the number of factors, as described in Section 5.2.1. This analysis supported the retention of two components, aligning with the hypothesized structure in the literature, which distinguishes Conscientiousness and Honesty-Humility as separate traits (e.g., Howard and Van Zandt, 2020, Oh et al., 2014). The EFA was then conducted with a prespecified two-factor solution. Appendix A.8.1 presents the initial factor loadings before item removal, using oblimin rotation. Notably, the correlation matrix was singular. As a result, the pseudo-inverse was applied by the `psych` package (Revelle, 2025). Excluding the two item exceptions `C_W_workhours` and `C_W_changes`, the initial factor loadings showed high loadings on the intended trait (ranging from 0.89 to 0.97) and small loadings on the other trait (ranging from 0 to 0.14).

In the first step of the selection algorithm, the iterative EFA filtering, two items (`C_W_workhours` and `C_W_changes`) were excluded during the first iteration. No further items were removed in subsequent iterations. In addition, neither the communality thresholding step nor the internal consistency check led to further exclusions. During the internal consistency check, the correlation matrix was not positive definite, and therefore smoothing was applied by the `psych` package (Revelle, 2025). The factor loadings after item selection are summarized in Appendix A.8.2. With a small number of exclusions, the final loadings are similar to the initial ones.

The same procedure was applied to the human base dataset. As with the LLM-generated data, the parallel analysis suggested retention of two components. The initial factor loadings are shown in Appendix A.8.3. Compared to the LLM-generated data, the human base data exhibited overall lower factor loadings, including negative loadings (between

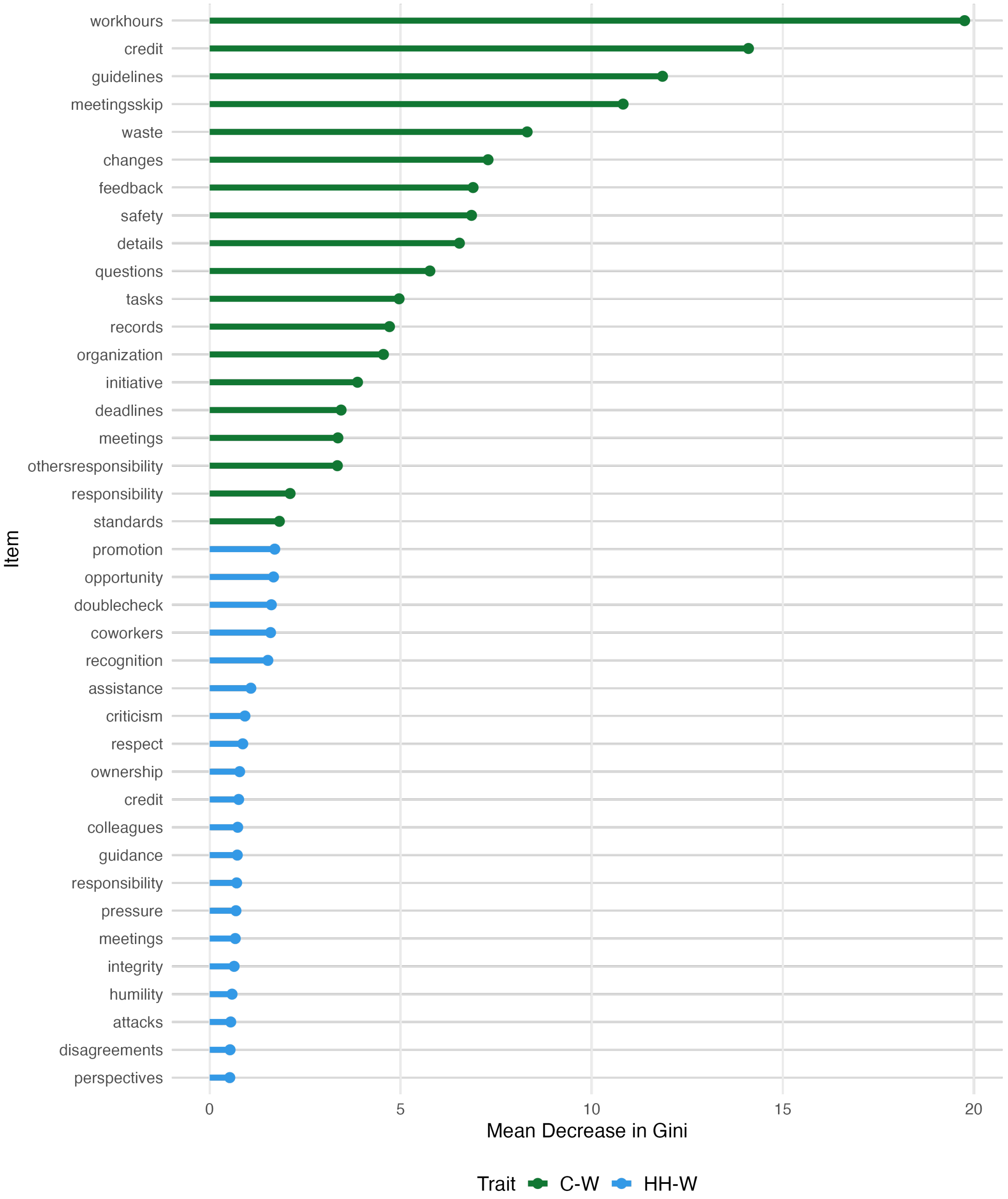


Figure 6: Feature importance based on the mean decrease in Gini impurity from a random forest (`ntree` = 500). The model was trained to predict the source of each observation (LLM-generated vs. human base data) using all responses to Conscientiousness at Work (C-W) and Honesty-Humility at Work (HH-W) items. The dataset included $N = 300$ observations (150 per source). Note that certain item suffixes appear twice per plot. Combined with their respective trait prefix (C-W, HH-W) they are unique and represent different items, as listed in Appendices A.1.1 and A.1.2.

-0.02 and 0.78), and more frequent cross-loadings. When the selection algorithm was applied to the human base data, 12 items were excluded in the first EFA filtering iteration and an additional 11 items were removed during the communality thresholding step. No further items were excluded during the internal consistency check. Two additional items, `HH_W_integrity` and `HH_W_doublecheck`, were excluded because they showed primary loadings on the Conscientiousness at Work factor, contrary to their intended alignment with the Honesty-Humility at Work construct. A third item, `HH_W_assistance`, was excluded because its factor loading fell below the 0.40 threshold following the model adjustments after communality-based exclusions. Comparing the selection outcomes between the LLM-generated and human base datasets, the two items during the EFA filtering in the LLM-generated dataset were also excluded when applying the selection algorithm on the human base dataset. In the human base data, one of these items was removed during iterative EFA filtering and the other during the communality thresholding step. The final factor loadings from the human base data, after item selection, are presented in Appendix A.8.4. Despite item refinement, the human base data continued to exhibit greater variability in factor loadings compared to the LLM-generated dataset, with loadings ranging from 0.47 to 0.83.

To validate the item selection based on LLM-generated data and compare it to the human-based selection, a CFA was conducted on both models using the unseen human comparison dataset. The results for each fit index of the model derived from LLM-generated data (LLM model) and the model derived from human base data (human model), are presented in Table 2.

The results indicate that the LLM model shows poorer fit than the human model in terms of the CFI and TLI, while fit is comparable between the models for the RMSEA and SRMR. Neither model fully meets the established fit criteria that were introduced in Section 5.2.3. Specifically, the LLM model falls below the recommended thresholds for CFI ($0.71 < 0.90$) and TLI ($0.69 < 0.90$). The RMSEA ($0.08 < 0.10$) and SRMR ($0.08 < 0.10$) can be deemed acceptable. The human model also falls just short of the recommended cut-off for CFI, yielding a value of exactly 0.90, and fails to meet an acceptable value for TLI ($0.88 < 0.90$). The RMSEA and SRMR values are slightly worse than those of the LLM model but can be deemed acceptable. Finally, we observe a significant result in the Chi-squared goodness-of-fit test for both models with $\chi^2_{\text{LLM}}(628) = 1235.43$, $p < .05$; $\chi^2_{\text{human}}(53) = 107.62$, $p < .05$. These results indicate a lack of fit for both models.

Table 2: Fit indices from confirmatory factor analysis for models estimated from LLM-generated data and human base data: comparative fit index (CFI), Tucker–Lewis index (TLI), root mean squared error of approximation (RMSEA) with confidence interval (CI), and standardized root mean square residual (SRMR). Indices were calculated on the unseen human comparison dataset. Note that the LLM model includes 37 items, and the human model 12 items.

Fit Measure	**LLM Model**	**Human Model**
CFI	0.71	0.90
TLI	0.69	0.88
RMSEA	0.08	0.08
90% CI	[0.07, 0.09]	[0.06, 0.10]
SRMR	0.08	0.07

Furthermore, internal consistency was assessed for each model. Across both traits, internal consistency was higher for the LLM than the human model. Specifically, the Conscientiousness at Work scale yielded a Cronbach's alpha of $\alpha_{LLM} = 0.85$ for the LLM-based version and $\alpha_{human} = 0.78$ for the human-based version. For the Honesty-Humility at Work scale, the corresponding values were $\alpha_{LLM} = 0.90$ and $\alpha_{human} = 0.82$. The McDonald's omega coefficients differed from the corresponding alpha values by no more than 0.01. It is important to note that the number of items differed between models. The LLM scale included 37 items, while the human scale included 12. With respect to internal consistency, a larger number of items can artificially inflate Cronbach's alpha, regardless of the actual coherence among items (e.g., Cortina, 1993). The following analyses, which use top-k selected items, enable more directly comparable results.

Selecting the top-k items per trait from the final loadings table yielded the respective short scales. A value of $k = 4$ was chosen. The selected items are marked in bold in the full item tables in Appendices A.1.1 and A.1.2. Of the eight items selected based on the LLM-generated data, three are also found in the item set selected from the human base dataset. The remaining five differ, and three of them were explicitly excluded during the human-based selection process before top-k selection.

The fit indices of the new short-version models are presented in Table 3. Both models demonstrated excellent model fit across all indices after selecting the top four items. The LLM-based model exhibited slightly better fit compared to the human model. Furthermore, the Chi-squared goodness-of-fit tests for both models were non-significant, with $\chi^2_{\text{LLM}}(19) = 17.16$, $p > .05$ and $\chi^2_{\text{human}}(19) = 25.57$, $p > .05$, indicating good model fit.

Table 3: Fit indices from confirmatory factor analysis for models estimated from LLM-generated data and human base data using top-loading item selection ($k = 4$): comparative fit index (CFI), Tucker–Lewis index (TLI), root mean squared error of approximation (RMSEA) with confidence interval (CI), and standardized root mean square residual (SRMR). Indices were calculated on the unseen human comparison dataset.

Fit Measure	**LLM Model**	**Human Model**
CFI	1.00	0.98
TLI	1.01	0.97
RMSEA	0.00	0.05
90% CI	[0.00, 0.06]	[0.00, 0.09]
SRMR	0.04	0.05

With respect to internal consistencies, the short Conscientiousness at Work scale yielded a Cronbach's alpha of $\alpha_{LLM} = 0.69$ for the LLM-based version and $\alpha_{human} = 0.71$ for the human-based version. For the short Honesty-Humility at Work scale, the corresponding values were $\alpha_{LLM} = 0.72$ and $\alpha_{human} = 0.74$. The McDonald's omega coefficients differed from the respective alpha values by no more than 0.02.

To assess the robustness of these results with respect to the rotation method, the full utility procedure was also conducted using varimax rotation instead of oblimin. For the LLM-generated data, this resulted in the same item exclusions by the item selection algorithm and the same selection of top-loading items for the short scales per trait, along with identical fit indices for the respective models. In the human data, the change in rotation method led to four additional item exclusions during the first iteration of the iterative EFA filtering step. The fit indices for the long-form human model using varimax were very similar to those obtained with oblimin, with the only notable difference being a slightly higher CFI under varimax rotation. For the short-form model, using varimax instead of oblimin, the fit indices improved for the human-based model. In this setting, the small differences in fit indices previously observed under oblimin rotation between the short-form LLM- and human-based models disappeared. Full results for varimax rotation can be reproduced using the provided R code.

7 Discussion

In this study, synthetic data was generated using GPT-4o via a prompt-based approach, and its potential for item selection in pretesting was evaluated. This novel use-case posed two key requirements on the generated data. First, the LLM-generated data needed to demonstrate fidelity not only at the trait-score level but also at the item-level. While research in LLM psychometrics has primarily focused on trait scores, reflecting its interest in overall personality of LLMs, effective item selection requires meaningful responses at the item level. To assess this, three fidelity criteria derived from synthetic data generation literature were applied. Second, the data needed to reflect not only comparable marginal distributions at both the trait and item levels but also human-like covariance structures among items. These inter-item covariances form the basis for EFA and internal consistency measures, both of which are used to select the most psychometrically sound items to construct a final scale. To assess this, EFA and internal consistency evaluations were conducted on the LLM-generated data, with results validated on human response data. Additionally, the inclusion of the pMSE as a fidelity metric partially accounted for both item-level similarity and the preservation of inter-item covariance structures.

With respect to the first requirement of item-level fidelity, the results varied across both items and fidelity metrics. The most notable difference appeared between items measuring different traits. Overall, items from the Honesty-Humility at Work pool demonstrated higher fidelity. However, certain Conscientiousness at Work items also performed well with respect to fidelity, despite this trait showing some of the most pronounced divergences, for example in the item "I am flexible with my work hours".

The observed distributional discrepancies, combined with the observed bias and high MSE in certain items, point towards open potential in prompt design to enhance alignment and fidelity. At the same time, variation in fidelity within the item set implies that fidelity may be influenced by the specific item content and target traits being measured. This observation aligns with prior findings from the LLM psychometrics literature, which also report mixed outcomes depending on the constructs and metrics evaluated (Ye et al., 2025). The finding that the association between generic Conscientiousness and Conscientiousness at Work trait scores in the human data is stronger than the corresponding association for Honesty-Humility supports this point. While alignment between the prompted and generated trait scores likely contributes to fidelity outcomes, it does not seem to be the only determining factor for fidelity.

The partially divergent results across fidelity metrics highlight the importance of employing a diverse set of evaluation metrics when assessing prompts for LLM-generated pretesting data but also more broadly in LLM psychometrics research. For instance, results based on KL divergence deviated from those based on Wasserstein distance and pMSE at trait-level. Especially, KL divergence is sensitive to low-probability events, which can lead to instability when certain response categories are rarely observed. Such cases are not uncommon in LLM-generated data or for weakly functioning items, making KL divergence potentially unsuitable as a standalone metric for evaluating item-level fidelity in future pretesting settings.

With respect to the second main requirement of covariance alignment, most items in the

LLM-generated dataset exhibited high primary factor loadings and low cross-loadings. This finding is consistent with previous studies (e.g., Wang et al., 2025) that have observed comparable loading structures in LLM-generated data. However, this led to substantially lower selectivity. Only two items were excluded based on the LLM-generated dataset, compared to 26 items in the human-generated dataset. Furthermore, the LLM-generated data produced a singular correlation matrix at the point of EFA, pointing towards additional challenges with the data. The resulting model fit for the LLM-based solution was slightly worse than that derived from the human data. Notably, also the human-derived model did not consistently achieve acceptable fit across all fit indices.

Additionally, we evaluated model fit after applying top-k selection. The results showed improved fit for both the LLM and human-derived models, both meeting the established guidelines on all fit indices. Notably, the LLM-derived model achieved slightly better fit than the human-derived model under oblimin rotation, with both models satisfying established guidelines across all fit indices. Internal consistencies were slightly lower for the LLM-derived scale but remained acceptable given the small number of items included. This finding suggests that LLM-generated responses may indeed be sensitive to item quality. However, this sensitivity may operate on different thresholds than those typically used in human-based psychometrics and selection criteria may need to be recalibrated when applied to synthetic data.

In conclusion, the LLM-generated data exhibited relevant limitations in supporting item selection. The observed inconsistencies in fidelity between items as well as utility, depending on whether top-k selection is applied, raise concerns about the reliability of LLM-generated data as a tool for streamlining the pretesting process at this stage. Rather than offering a basis for early item-filtering, the results suggest a risk of substantially overestimating item quality and factor structure when relying on LLM responses alone. One could argue that even eliminating a small number of items, such as two, may still offer practical benefits in terms of time and cost savings, especially in large samples. However, this must also be weighed against the current costs of data generation with LLMs via API, especially when larger sample sizes are desired. Nonetheless, the fact that items with the highest factor loadings in the LLM-generated data did yield a coherent scale points to some promise for this approach. Realizing the full potential of this idea, however, may require rethinking the prompting strategy or adapting the item selection procedure to better align with the properties of LLM-generated responses.

8 Limitations and Future Research

As with any study, this research has limitations that should be acknowledged. A primary limitation lies in the use of a relatively minimalist prompt. While this approach establishes a baseline and facilitates direct comparison with underlying human data, selectivity might be enhanced through more refined prompting strategies. Future research could consider explicitly stating the goal of item discrimination in the prompt, randomizing the item order, or enriching the prompt with additional persona information, including internal and external circumstances in which the test is completed. Such modifications could yield a more human-like factor structure that facilitates clearer differentiation between items. If a realistic human base dataset is available, prompt tuning could be approached as a hyperparameter optimization task, where parameters such as temperature, prompt length, persona specificity, and even model choice are systematically varied. Optimizing these settings could help align the synthetic data more closely with human-like response distributions.

Furthermore, there are certain limitations regarding the item selection algorithm. Psychometric item selection typically involves analyzing item response theory parameters to estimate item characteristics, such as difficulty and discrimination (Embretson and Reise, 2013). Future studies with larger samples should consider incorporating these analyses to potentially enhance item selectivity. Additionally, the item selection algorithm used in this study is based entirely on established selectivity criteria from human pretesting studies. The findings from the top-k selection, however, suggest a need to adapt these criteria. Existing thresholds may not account for the generally higher and less diffuse factor loadings observed in synthetic data. For example, combining established item selection methods with novel LLM-based approaches, such as reliability estimation using vector representations proposed by Hommel and Arslan (2024), may be necessary to fully leverage the potential of LLM data generation.

Moreover, the items used for this study should be considered in relation to further adjacent psychological constructs. This study did not assess how the synthetic data aligns with the covariance structures of the Big Five or HEXACO traits, nor with external variables. As a proof of concept, the current research lays the groundwork for future studies to expand this work by generating data across a broader array of traits. In this context, it may also be valuable to incorporate a qualitative analysis of item performance, for example by examining whether differences arise between items that align with or conflict with the ethical guidelines of the LLM.

Finally, this study evaluated only the fully LLM-generated datasets for its utility in item selection. Prior work by Liu et al. (2025) demonstrated the benefits of data augmentation strategies that combine human and LLM-generated data, showing improved item selectivity in knowledge-based assessments compared to using smaller, purely human-based datasets. This approach may also hold promise for pretesting in non-cognitive contexts. Furthermore, pretesting could benefit from augmenting samples with LLM-generated data to enhance diversity and support data generation for target populations that are difficult or impossible to reach. Collectively, these applications highlight the potential to streamline and scale test development processes, even when LLM-generated datasets are used only in part rather than exclusively.

9 Outlook

Data collection for pretesting using LLMs represents an initial step in automated test development. In the future, this automation may advance even further. With the help of machine learning, it could eventually become feasible to generate complete, ready-to-use assessments with minimal human input. The implications of such a development for both social science research and industrial applications would be profound. While the current study makes clear that this level of automation has not yet been achieved, and that careful oversight and clear guidelines remain essential, it offers an early demonstration of the potential of machine learning, particularly LLMs, in supporting the test development process.

References

Abacha, A. B., Yim, W.-w., Fu, Y., Sun, Z., Yetisgen, M., Xia, F. and Lin, T. (2024). Medec: A benchmark for medical error detection and correction in clinical notes, *arXiv preprint arXiv:2412.19260* .

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A. and Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research, *PNAS nexus* **3**(7).

Ainur, A., Sayang, M., Jannoo, Z. and Yap, B. (2017). Sample size and non-normality effects on goodness of fit measures in structural equation models., *Pertanika Journal of Science & Technology* **25**(2).

Ashton, M. C. and Lee, K. (2007). Empirical, theoretical, and practical advantages of the hexaco model of personality structure, *Personality and social psychology review* **11**(2): 150–166.

Ashton, M. C. and Lee, K. (2008). The prediction of honesty–humility-related criteria by the hexaco and five-factor models of personality, *Journal of Research in Personality* **42**(5): 1216–1228.

Aßenbacher, M., Heumann, C., Roth, B., Schütze, H., Stephan, A., Weißweiler, L. and Sawitzki, M. (2025). Deep learning for natural language processing (dl4nlp), `https://slds-lmu.github.io/dl4nlp/`. Accessed: 2025-06-22.

Baowaly, M. K., Lin, C.-C., Liu, C.-L. and Chen, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks, *Journal of the American Medical Informatics Association* **26**(3): 228–241.

Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis, *Personnel psychology* **44**(1): 1–26.

Barua, A., Brase, G., Dong, K., Hitzler, P. and Vasserman, E. (2024). On the psychology of gpt-4: Moderately anxious, slightly masculine, honest, and humble, *arXiv preprint arXiv:2402.01777* .

Bentler, P. M. (1990). Comparative fit indexes in structural models., *Psychological bulletin* **107**(2): 238.

Berry, C. M., Ones, D. S. and Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis., *Journal of applied psychology* **92**(2): 410.

Bing, M. N., Davison, H. K. and Smothers, J. (2014). Item-level frame-of-reference effects in personality testing: An investigation of incremental validity in an organizational setting, *International Journal of Selection and Assessment* **22**(2): 165–178.

Bischl, B., Bothmann, L., Scheipl, F., Pielok, T., Wimmer, L., Li, Y., Kolb, C., Schalk, D., Seibold, H., Molnar, C. and Richter, J. (2022). Introduction to Machine Learning (I2ML), `https://slds-lmu.github.io/i2ml/`. Accessed: 2025-05-20.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R. and Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer, *Frontiers in public health* **6**: 149.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M. and Kasneci, G. (2022). Language models are realistic tabular data generators, *arXiv preprint arXiv:2210.06280* .

Boumaza, R., Santagostini, P., Yousfi, S. and Demotes-Mainard, S. (2021). dad: an r package for visualisation, classification and discrimination of multivariate groups modelled by their densities, *The R Journal* **13**(2): 386.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners, *Advances in neural information processing systems* **33**: 1877–1901.

Candelieri, A., Ponti, A., Giordani, I. and Archetti, F. (2023). On the use of wasserstein distance in the distributional analysis of human decision making under uncertainty, *Annals of Mathematics and Artificial Intelligence* **91**(2): 217–238.

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers, *Communication methods and measures* **12**(1): 25–44.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters., *The journal of abnormal and social psychology* **38**(4): 476.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F. and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks, *Machine learning for healthcare conference*, PMLR, pp. 286–305.

Clark, L. A. and Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments., *Psychological assessment* **31**(12): 1412.

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications., *Journal of applied psychology* **78**(1): 98.

Costello, A. B. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis, *Practical assessment, research, and evaluation* **10**(1).

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D. and Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors, *Educational and Psychological Measurement* **70**(6): 885–901.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *psychometrika* **16**(3): 297–334.

Cudeck, R. (2000). Exploratory factor analysis, *Handbook of applied multivariate statistics and mathematical modeling*, Elsevier, pp. 265–296.

Dan, Y., Ahmed, A. A. A., Chupradit, S., Chupradit, P. W., Nassani, A. A. and Haffar, M. (2021). The nexus between the big five personality traits model of the digital economy and blockchain technology influencing organization psychology, *Frontiers in psychology* **12**: 780527.

Daniya, T., Geetha, M. and Kumar, K. S. (2020). Classification and regression trees with gini index, *Advances in Mathematics: Scientific Journal* **9**(10): 8237–8247.

Darabi, S. and Elor, Y. (2021). Synthesising multi-modal minority samples for tabular data, *arXiv preprint arXiv:2105.08204* .

Dastile, X., Celik, T. and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey, *Applied Soft Computing* **91**: 106263.

Dinno, A. (2014). Gently clarifying the application of horn's parallel analysis to principal component analysis versus factor analysis.

Dinno, A. (2025). *paran: Horn's Test of Principal Components/Factors*. R package version 1.5.4.
URL: *https://CRAN.R-project.org/package=paran*

Embretson, S. E. and Reise, S. P. (2013). *Item response theory for psychologists*, Psychology Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research., *Psychological methods* **4**(3): 272.

Fan, X., Thompson, B. and Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes, *Structural equation modeling: a multidisciplinary journal* **6**(1): 56–83.

Finch, J. F. and West, S. G. (1997). The investigation of personality structure: Statistical models, *Journal of Research in personality* **31**(4): 439–485.

Ford, J. K., MacCallum, R. C. and Tait, M. (1986a). The application of exploratory factor analysis in applied psychology: A critical review and analysis, *Personnel psychology* **39**(2): 291–314.

Ford, J., MacCallum, R. and Tait, M. (1986b). The application of exploratory factor analysis in applied psychology: A critical review and analysis, *Personnel Psychology* **39**(2): 291–314.

Glorfeld, L. W. (1995). An improvement on horn's parallel analysis methodology for selecting the correct number of factors to retain, *Educational and psychological measurement* **55**(3): 377–393.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures, *Journal of Research in personality* **40**(1): 84–96.

Goretzko, D. and Bühner, M. (2022). Factor retention using machine learning with ordinal data, *Applied Psychological Measurement* **46**(5): 406–421.

Goretzko, D., Siemund, K. and Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis, *Educational and Psychological Measurement* **84**(1): 123–144.

Götz, F. M., Maertens, R., Loomba, S. and van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development., *Psychological Methods* .

Green, S. B., Levy, R., Thompson, M. S., Lu, M. and Lo, W.-J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis, *Educational and Psychological Measurement* **72**(3): 357–374.

Hansen, L., Seedat, N., van der Schaar, M. and Petrovic, A. (2023). Reimagining synthetic tabular data generation through data-centric ai: A comprehensive benchmark, *Advances in neural information processing systems* **36**: 33781–33823.

Hayton, J. C., Allen, D. G. and Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis, *Organizational research methods* **7**(2): 191–205.

Hoffmann, S. (2022). Concepts of statistical modeling: Principal component analysis, Lecture, Institute of Statistics, LMU Munich.

Holtrop, D., Born, M. P., de Vries, A. and de Vries, R. E. (2014). A matter of context: A comparison of two types of contextualized personality measures, *Personality and Individual Differences* **68**: 234–240.

Hommel, B. E. and Arslan, R. C. (2024). Language models accurately infer correlations between psychological items and scales from text alone, *Preprint at PsyArXiv https://doi.org/10.31234/osf.io/kjuce* .

Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H. and Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation, *Psychometrika* **87**(2): 749–772.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis, *Psychometrika* **30**(2): 179–185.

Howard, M. C. and Van Zandt, E. C. (2020). The discriminant validity of honesty-humility: A meta-analysis of the hexaco, big five, and dark triad, *Journal of Research in Personality* **87**: 103982.

Hu, L.-t. and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification., *Psychological methods* **3**(4): 424.

Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural equation modeling: a multidisciplinary journal* **6**(1): 1–55.

Humphreys, L. G. and Montanelli Jr, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors, *Multivariate Behavioral Research* **10**(2): 193–205.

Hunthausen, J. M., Truxillo, D. M., Bauer, T. N. and Hammer, L. B. (2003). A field study of frame-of-reference effects on personality test validity., *Journal of Applied Psychology* **88**(3): 545.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C. and Zhu, Y. (2023). Evaluating and inducing personality in pre-trained language models, *Advances in Neural Information Processing Systems* **36**: 10622–10643.

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D. and Kabbara, J. (2023). Personallm: Investigating the ability of large language models to express personality traits, *arXiv preprint arXiv:2305.02547* .

John, O. P., Srivastava, S. et al. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Jolicoeur-Martineau, A., Fatras, K. and Kachman, T. (2024). Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees, *International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1288–1296.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**(3): 187–200.

Kauermann, G. (2023). Vorlesungsunterlagen zu inferenzstatistik 1 und 2: Statistics – coping with uncertainty. Lecture notes, September 28, 2023.

Kotelnikov, A., Baranchuk, D., Rubachev, I. and Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models, *International Conference on Machine Learning*, PMLR, pp. 17564–17579.

Lee, P., Fyffe, S., Son, M., Jia, Z. and Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing, *Journal of Business and Psychology* **38**(1): 163–190.

Li, Z., Zhao, Y. and Fu, J. (2020). Sync: A copula based framework for generating synthetic data from aggregated sources, *2020 International Conference on Data Mining Workshops (ICDMW)*, IEEE, pp. 571–578.

Li, Z., Zhu, H., Lu, Z. and Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations, *arXiv preprint arXiv:2310.07849* .

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest, *R News* **2**(3): 18–22.
URL: *https://CRAN.R-project.org/doc/Rnews/*

Lin, J. (2002). Divergence measures based on the shannon entropy, *IEEE Transactions on Information theory* **37**(1): 145–151.

Liu, T., Qian, Z., Berrevoets, J. and van der Schaar, M. (2023). Goggle: Generative modelling for tabular data by learning relational structure, *The Eleventh International Conference on Learning Representations*.

Liu, Y., Bhandari, S. and Pardos, Z. A. (2025). Leveraging llm respondents for item evaluation: A psychometric analysis, *British Journal of Educational Technology* **56**(3): 1028–1052.

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G. and Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey, *arXiv preprint arXiv:2406.15126* .

Louppe, G., Wehenkel, L., Sutera, A. and Geurts, P. (2013). Understanding variable importances in forests of randomized trees, *Advances in neural information processing systems* **26**.

Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M. and Zhang, C. (2020). Vaem: a deep generative model for heterogeneous mixed type data, *Advances in Neural Information Processing Systems* **33**: 11237–11247.

Maples-Keller, J. L., Williamson, R. L., Sleep, C. E., Carter, N. T., Campbell, W. K. and Miller, J. D. (2019). Using item response theory to develop a 60-item representation of the neo pi–r using the international personality item pool: Development of the ipip–neo–60, *Journal of personality assessment* **101**(1): 4–15.

McDonald, R. P. (2013). *Test theory: A unified treatment*, psychology press.

Mcneish, D. (2017). Psychological methods thanks coefficient alpha, we'll take it from here thanks coefficient alpha, we'll take it from here, *Psychological Methods, may* **29**.

McNeish, D. (2023). Dynamic fit index cutoffs for categorical factor analysis with likert-type, ordinal, or binary responses., *American Psychologist* **78**(9): 1061.

McNeish, D. and Wolf, M. G. (2023). Dynamic fit index cutoffs for one-factor models, *Behavior Research Methods* **55**(3): 1157–1174.

Nielsen, F. (2020). On a generalization of the jensen–shannon divergence and the jensen–shannon centroid, *Entropy* **22**(2): 221.

Oh, I.-S., Le, H., Whitman, D. S., Kim, K., Yoo, T.-Y., Hwang, J.-O. and Kim, C.-S. (2014). The incremental validity of honesty–humility over cognitive ability and the big five personality traits, *Human Performance* **27**(3): 206–224.

Onishi, S. and Meguro, S. (2023). Rethinking data augmentation for tabular data in deep learning, *arXiv preprint arXiv:2305.10308* .

OpenAI (2024). Gpt-4o, `https://platform.openai.com/docs/models/gpt-4o`. Accessed: 2025-05-16.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances, *Annual review of statistics and its application* **6**(1): 405–431.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H. and Kim, Y. (2018). Data synthesis based on generative adversarial networks, *arXiv preprint arXiv:1806.03384* .

Pennebaker, J. W. and King, L. A. (1999). Linguistic styles: language use as an individual difference., *Journal of personality and social psychology* **77**(6): 1296.

Petrov, N. B., Serapio-García, G. and Rentfrow, J. (2024). Limited ability of llms to simulate human psychological behaviours: a psychometric analysis, *arXiv preprint arXiv:2405.07248* .

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance., *Psychological bulletin* **135**(2): 322.

Qian, Z., Cebere, B.-C. and van der Schaar, M. (2023). Synthcity: facilitating innovative use cases of synthetic data in different data modalities, *arXiv preprint arXiv:2301.07573* .

Revelle, W. (2025). *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois. R package version 2.5.3.
URL: *https://CRAN.R-project.org/package=psych*

Roberts, B. W. and Bogg, T. (2004). A longitudinal study of the relationships between conscientiousness and the social-environmental factors and substance-use behaviors that influence health, *Journal of personality* **72**(2): 325–354.

Rosellini, A. J. and Brown, T. A. (2021). Developing and validating clinical questionnaires, *Annual review of clinical psychology* **17**(1): 55–81.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling, *Journal of Statistical Software* **48**(2): 1–36.

Rust, J., Stillwell, D. and Kosinski, M. (2020). *Psychometrics: From Practice to Theory and Back Again*, Routledge, London. Published on December 24, 2020.

Ruthotto, L. and Haber, E. (2021). An introduction to deep generative modeling, *GAMM-Mitteilungen* **44**(2): e202100008.

Saccenti, E. and Timmerman, M. E. (2017). Considering horn's parallel analysis from a random matrix theory point of view, *Psychometrika* **82**(1): 186–209.

Sauber-Cole, R. and Khoshgoftaar, T. M. (2022). The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey, *Journal of Big Data* **9**(1): 98.

Schmid, J. and Leiman, J. M. (1957). The development of hierarchical factor solutions, *Psychometrika* **22**(1): 53–61.

Schuhmacher, D., Bähre, B., Bonneel, N., Gottschlich, C., Hartmann, V., Heinemann, F., Schmitzer, B. and Schrieber, J. (2024). *transport: Computation of Optimal Transport Plans and Wasserstein Distances*. R package version 0.15-4.
URL: *https://cran.r-project.org/package=transport*

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. and Ungar, L. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach, *PLoS ONE* **8**.
URL: *https://pdfs.semanticscholar.org/cf84/60b19f3ab20d3a2bc52a4205e5698eaaaafa.pdf*

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Abdulhai, M., Faust, A. and Matarić, M. (2023). Personality traits in large language models, *arXiv preprint arXiv:2307.00184v4* .

Shrestha, N. (2021). Factor analysis as a tool for survey analysis, *American journal of Applied Mathematics and statistics* **9**(1): 4–11.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need, *Information Fusion* **81**: 84–90.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. (2018). General and specific utility measures for synthetic data, *Journal of the Royal Statistical Society Series A: Statistics in Society* **181**(3): 663–688.

Steiger, J. H. (1998). A note on multiple sample extensions of the rmsea fit index.

Swift, V. and Peterson, J. B. (2019). Contextualization as a means to improve the predictive validity of personality models, *Personality and Individual Differences* **144**: 153–163.

Timmerman, M. E. and Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis., *Psychological methods* **16**(2): 209.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis, *Psychometrika* **38**(1): 1–10.

Van Breugel, B. and Van Der Schaar, M. (2024). Why tabular foundation models should be a research priority, *arXiv preprint arXiv:2405.01147* .

van der Schaar, M. and Maxfield, N. (2020). Synthetic data: breaking the data logjam in machine learning for healthcare, `https://www.vanderschaar-lab.com/synthetic-data-breaking-the-data-logjam-in-machine-learning-for-healthcare/`. Accessed: 2025-05-30.

Vardhan, L. V. H. and Kok, S. (2020). Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders, *Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37 th International Conference on Machine Learning*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.

Velicer, W. F., Eaton, C. A. and Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components, *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* pp. 41–71.

von Davier, M. (2018). Automated item generation with recurrent neural networks, *psychometrika* **83**(4): 847–857.

Wang, P., Zou, H., Yan, Z., Guo, F., Sun, T., Xiao, Z. and Zhang, B. (2024). Not yet: Large language models cannot replace human respondents for psychometric research.

Wang, Y., Zhao, J., Ones, D. S., He, L. and Xu, X. (2025). Evaluating the ability of large language models to emulate personality, *Scientific reports* **15**(1): 519.

Wolf, M., Tritscher, J., Landes, D., Hotho, A. and Schlör, D. (2024). Benchmarking of synthetic network data: Reviewing challenges and approaches, *Computers & Security* p. 103993.

Worthington, R. L. and Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices, *The counseling psychologist* **34**(6): 806–838.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan, *Advances in neural information processing systems* **32**.

Xu, W., Hu, W., Wu, F. and Sengamedu, S. (2023). Detime: Diffusion-enhanced topic modeling using encoder-decoder based llm, *arXiv preprint arXiv:2310.15296* .

Yang, Z., Yu, H., Guo, P., Zanna, K., Yang, X. and Sano, A. (2024). Balanced mixed-type tabular data synthesis with diffusion models, *arXiv preprint arXiv:2404.08254* .

Ye, H., Jin, J., Xie, Y., Zhang, X. and Song, G. (2025). Large language model psychometrics: A systematic review of evaluation, validation, and enhancement, *arXiv preprint arXiv:2505.08245* . Project website: `https://llm-psychometrics.com`, GitHub: `https://github.com/ValueByte-AI/Awesome-LLM-Psychometrics`.

Yin, Y., Lin, Z., Jin, M., Fanti, G. and Sekar, V. (2022). Practical gan-based synthetic ip header trace generation using netshare, *Proceedings of the ACM SIGCOMM 2022 Conference*, pp. 458–472.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H. and Karypis, G. (2023). Mixed-type tabular data synthesis with score-based diffusion in latent space, *arXiv preprint arXiv:2310.09656* .

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. and Xiao, X. (2017). Privbayes: Private data release via bayesian networks, *ACM Transactions on Database Systems (TODS)* **42**(4): 1–41.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D. and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?, *arXiv preprint arXiv:1801.07243* .

Zwick, W. R. and Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain, *Multivariate behavioral research* **17**(2): 253–269.

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain., *Psychological bulletin* **99**(3): 432.

A Appendix

A.1 Experimental Set-Up

A.1.1 Item Set for Conscientiousness at Work (C_W)

Table 4: Items included in the final short scales based on top-k selection are shown in bold.

#	**Abbreviation**	**Statement**
1	C_W_details	I am organized and pay attention to details.
2	C_W_tasks	**I finish tasks on time.**
3	C_W_records	I keep accurate records of my work.
4	C_W_responsibility	I take responsibility for my mistakes.
5	C_W_doublecheck	I double check my work before submitting it.
6	C_W_standards	I set high standards for myself.
7	C_W_safety	I follow safety protocols.
8	C_W_workhours	I am flexible with my work hours.
9	C_W_deadlines	I complete assignments ahead of deadlines.
10	C_W_initiative	**I take initiative in learning new skills or procedures.**
11	C_W_feedback	I always seek feedback from my supervisors or peers on how to improve my work performance.
12	C_W_othersresponsibility	I make sure that others are held accountable for their responsibilities.
13	C_W_organization	**I proactively look for ways to contribute to the organization's goals and objectives.**
14	C_W_meetings	**I come prepared and well-informed to meetings.**
15	C_W_changes	I am open to changes in working practices or processes.
16	C_W_waste	I waste time at work.
17	C_W_credit	I take credit for the accomplishments of others.
18	C_W_questions	I rarely ask questions when something is unclear.
19	C_W_guidelines	I ignore guidelines and policies.
20	C_W_meetingsskip	I skip out on important meetings.

A.1.2 Item Set for Honesty-Humility at Work (HH_W)

Table 5: Items included in the final short scales based on top-k selection are shown in bold.

#	Abbreviation	Statement
1	HH_W_promotion	I would never tell a lie to get a promotion.
2	HH_W_coworkers	I put the needs of my coworkers before my own.
3	HH_W_responsibility	I try to recognize when I have made a mistake and take responsibility for it.
4	HH_W_perspectives	I strive to understand different perspectives.
5	HH_W_criticism	I am willing to listen to constructive criticism.
6	HH_W_credit	**When given credit, I make sure to share it with those who deserve it.**
7	HH_W_pressure	Even when there is pressure from others, I remain honest about my work.
8	HH_W_opportunity	If given the opportunity, I will try to benefit myself over others.
9	HH_W_colleagues	**When talking with colleagues, I always aim to stay positive and respectful.**
10	HH_W_disagreements	In disagreements, I strive to be open-minded and accept other opinions.
11	HH_W_respect	**When working in teams, I give everyone equal respect regardless of rank or experience.**
12	HH_W_integrity	Even when facing tough challenges, I maintain my integrity.
13	HH_W_guidance	When working on tasks that are difficult or unfamiliar, I seek guidance rather than staying quiet.
14	HH_W_ownership	When in a leadership position, I take ownership of successes and failures.
15	HH_W_recognition	Rather than seeking recognition or praise for myself, I prefer to focus on team success.
16	HH_W_doublecheck	To ensure accurate results, I double check my work before submitting it.
17	HH_W_meetings	In meetings or presentations, I make sure not to exaggerate facts.
18	HH_W_assistance	If a colleague is struggling with their work load, I offer assistance if possible.
19	HH_W_attacks	**When someone has done something wrong, I treat them fairly and avoid personal attacks.**
20	HH_W_humility	At all times, I aim to maintain an attitude of humility.

A.1.3 Definition of Target Constructs

The test items used in this study stem from a previously developed but at the time of data generation unpublished item set targeting two contextualized personality traits: *Conscientiousness* and *Honesty-Humility at Work*. Contextualizing personality assessments (e.g., for workplace or romantic settings) has been shown to yield incremental validity compared to generic instruments like the IPIP (Goldberg et al., 2006) or HEXACO (Ashton and Lee, 2007) (e.g., Bing et al., 2014, Holtrop et al., 2014, Hunthausen et al., 2003, Swift and Peterson, 2019). Although this approach is supported by research, it remains underused, likely due to the high costs of adaptation and re-validation (e.g., Holtrop et al., 2014). Using a contextualized item set thus illustrates one use case of how synthetic data generation may help balance cost and predictive gain.
Conscientiousness is defined as an individual's tendency to complete tasks reliably and persistently (John et al., 1999). Employees scoring high on this trait often demonstrate "competence, organization, willingness, high achievement, consideration, and self-discipline" (Dan et al., 2021, p. 3).
Honesty-Humility refers to the degree to which individuals are sincere, modest, and uninterested in status or material gain (Ashton and Lee, 2007). It also reflects a tendency to cooperate with others, even in situations where exploitation would have no negative consequences (Ashton and Lee, 2007).

A.1.4 Human Dataset

The original dataset was collected via Prolific (`https://www.prolific.co/`) from a gender-balanced sample of $N = 326$ participants. Eligibility criteria included native English proficiency and completion of at least 50 prior surveys. To ensure relevance to workplace-related questions, participants were required to be currently employed or to have been employed within the past 12 months. Each participant received £5.33 (approximately USD $6.73) for their participation. For this study, the following two unique subsets were sampled from this original dataset.

Human Base Dataset. A sample of $n = 150$ participants was used to generate the synthetic dataset and evaluate fidelity. Demographic variables and trait scores (mean item scores per trait) were used for prompting the LLM; item-level variables were used for fidelity analysis. Ages in the dataset ranged from 18 to 77 years ($M = 42.11$, $SD \approx 12.6$), with 81 identifying as female and 69 as male. Countries of residence were the UK ($n = 128$), Canada ($n = 11$), Ireland ($n = 7$), and other countries. Self-reported annual income fell into predefined brackets with below £10,000 ($n = 9$), £20,000–£29,999 ($n =$

51), £30,000–£39,999 ($n = 30$), and within £90,000–£99,999 ($n = 1$). The majority ($n = 136$) indicated being in a (casual) relationship, while 14 identified as single.

Human Comparison Dataset. A second, new sample of $n = 150$ participants was used to assess variation across human samples during the fidelity analysis and later served as test set for CFA. Ages ranged from 19 to 65 years ($M = 35.43$), with 67 female and 83 male participants. Most were based in the UK ($n = 100$), followed by South Africa ($n = 17$), the US ($n = 7$), Canada ($n = 5$), and others. Reported income was below £10,000 ($n = 25$), £20,000–£29,999 ($n = 38$) and £30,000–£39,999 ($n = 24$), and £100,000–£149,999 ($n = 3$). Relationship status was similar to Subset 1, with 126 participants in (casual) relationships and 24 identifying as single.

A.1.5 Prompt Excerpt

```
prompt = f"""
    Meet this participant, a {Input['Age']}-year-old {Input['Sex']} from
    {Input['Nationality']} with an annual income of {Input['income']}.
    In terms of personal life, the participant is
    {Input['relationship_status_word']}.

    The participant provided responses to a Big Five Personality test.
    The resulting scores are given below. Based on the resulting score,
    infer how the same participant would likely respond to each of the
    new statements, using a scale from 1 (Disagree) to 5 (Agree).
    Evaluate each new item individually based on the trait score.
    Respond with a single number (1 - 5) per new item, separated by
    commas. Do not include explanations - only the numbers.

    Participant score on Conscientiousness {Input['Gen_C_sumscore']}
    (with possible minimum 12 and possible maximum 60)

    New Test Items for Conscientiousness at Work (Rate each 1 - 5):

    1. I am organized and pay attention to details.
    2. I finish tasks on time.

    [...] """
```

A.2 Illustration of KL Divergence vs. Wasserstein Distance

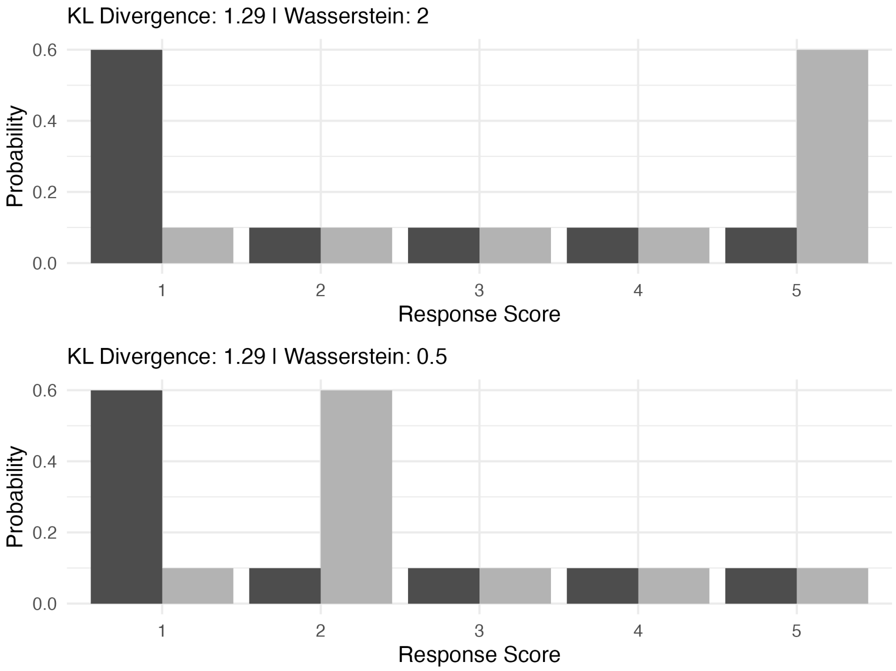


Figure 7: Illustration of a reference distribution (dark grey) compared to two alternatives (light grey) that yield the same Kullback-Leibler (KL) divergence but different Wasserstein distances.

A.3 Item Selection Algorithm

1. **Iterative EFA Filtering.** Let λ_{jk} denote the loading of item j on factor k, where $k \in \{1, \ldots, l\}$. Remove any item j that satisfies either of the following (Boateng et al., 2018, Clark and Watson, 2019, Costello and Osborne, 2005, Ford et al., 1986a, Rosellini and Brown, 2021)

$$\begin{aligned} &\max_k |\lambda_{jk}| < 0.40 \quad \text{(weak primary loading)} \\ \text{or} \quad &\exists\, k' \neq \arg\max_k |\lambda_{jk}| \text{ such that } |\lambda_{jk'}| > 0.30 \quad \text{(cross-loading).} \end{aligned}$$

   Repeat EFA and reapply the criteria iteratively until no remaining items meet the exclusion conditions.

2. **Communality Thresholding.** For each retained item j, let h_j^2 denote its communality. Exclude items where (Carpenter, 2018, Fabrigar et al., 1999, Worthington and Whittaker, 2006)

$$h_j^2 < 0.30.$$

3. **Internal Consistency Check.** Compute the change in Cronbach's α (Cronbach, 1951) when item j is dropped by

$$\Delta\alpha_j = \alpha_{\text{drop},j} - \alpha_{\text{total}}.$$

   Remove items with $\Delta\alpha_j > 0.001$ for both raw and standardized α (Götz et al., 2023).

4. **Top-Loading Selection.** For each factor k, select the top 4 items with the highest absolute loadings, selecting

$$\text{Top-4}\{|\lambda_{jk}|\}.$$

A.4 Parallel Analysis

Horn's parallel analysis (Horn, 1965) provides a method for determining the number of factors to retain in factor analysis. Unlike the Kaiser criterion, parallel analysis accounts for sampling variability by comparing observed eigenvalues to those obtained from randomly generated eigenvalues, rather than relying solely on the observed eigenvalues (Horn, 1965). The procedure is as follows (Dinno, 2014, Hoffmann, 2022):

1. Conduct a principal component analysis (PCA) on the observed data to obtain the eigenvalues λ_j for $j = 1, \dots, m$, where m denotes the number of observed variables.

2. Generate multiple random datasets with the same number of variables m and observations as the observed dataset. For this study 100 random datasets were generated, assuming the variables are independent and normally distributed.

3. For each random dataset, compute the eigenvalues and calculate the mean eigenvalue for each component across the simulated datasets $\bar{\lambda}_j^r$. Optionally, compute the standard deviation for use in certain modified procedures (Glorfeld, 1995).

4. Determine the suggested number of factors by comparing the observed eigenvalue $\lambda_j$to the corresponding mean eigenvalue from the random datasets using the retention criterion (Dinno, 2014)

$$\lambda_j \begin{cases} > \bar{\lambda}_j^r & \text{retain,} \\ \leq \bar{\lambda}_j^r & \text{do not retain (and stop).} \end{cases}$$

Note: As noted by Saccenti and Timmerman (2017), Horn's parallel analysis is often used to determine the number of factors in common factor analysis (the underlying idea of EFA), even though it was originally developed for PCA (see Step 1), raising theoretical concerns about this conceptual mismatch (Ford et al., 1986b, Humphreys and Montanelli Jr, 1975). Empirical comparisons suggest that methods specifically adapted to common factor analysis tend to perform worse (Crawford et al., 2010) or similar (Green et al., 2012) to the original form of Horn's method. Despite its roots in PCA, Horn's parallel analysis appears to provide a reasonable approximation of the number of major common factors, making it a useful heuristic in practice (Timmerman and Lorenzo-Seva, 2011, Zwick and Velicer, 1982).

A.5 Bias and Mean Squared Error

To evaluate synthetic responses compared to human-generated responses, as a first step, bias and MSE per item are computed. Let $\{(y_i^{\text{human}}, y_i^{\text{synthetic}})\}_{i=1}^{n}$ denote the paired observations for respondent $i \in \{1, \ldots, n\}$, where y_i^{human} represents the human-generated response based on a real-world sample and $y_i^{\text{synthetic}}$ the corresponding synthetic, LLM-generated response, based on demographic and trait level information about the same individual.

The bias per item, defined as the mean difference between synthetic and human responses, is given by

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i^{\text{synthetic}} - y_i^{\text{human}} \right) \tag{20}$$

and reflects systematic deviations in the synthetic data. Positive values indicate systematic overestimation and negative values systematic underestimation relative to the ground truth. In addition, the MSE as mean squared differences between synthetic and human responses per item is calculated by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i^{\text{synthetic}} - y_i^{\text{human}} \right)^2 . \tag{21}$$

MSE captures both bias and variance in the synthetic responses. Bias and MSE are defined for continuous data and assume equal intervals between values. For this study, we treat the 5-point Likert scale as interval-scaled to justify their use.

A.6 Means and Standard Deviations

A.6.1 LLM vs. Human Responses for Conscientiousness at Work Items

Table 6: Means and standard deviations (SD) of responses from LLM-generated and human base data for items in the Conscientiousness at Work (C-W) item pool. The final column reports the SD ratio (LLM / Human).

Item	**Mean**	**Mean**	**SD**	**SD**	**SD Ratio**
	(LLM)	(Human)	(LLM)	(Human)	
C_W_details	3.89	4.32	0.55	0.83	0.66
C_W_tasks	4.01	4.32	0.44	0.74	0.59
C_W_records	3.98	4.15	0.47	0.95	0.49
C_W_responsibility	4.16	4.46	0.39	0.61	0.64
C_W_standards	4.10	4.25	0.46	0.88	0.52
C_W_safety	4.09	4.42	0.37	0.74	0.50
C_W_workhours	3.19	3.87	0.39	1.18	0.33
C_W_deadlines	3.71	3.74	0.62	0.99	0.63
C_W_initiative	4.01	4.02	0.45	0.92	0.49
C_W_feedback	3.77	3.45	0.60	1.17	0.51
C_W_othersresponsibility	3.81	3.79	0.58	0.89	0.65
C_W_organization	4.01	3.85	0.45	0.97	0.46
C_W_meetings	4.01	4.15	0.46	0.90	0.51
C_W_changes	3.91	4.15	0.28	0.81	0.35
C_W_waste	3.97	3.57	0.45	1.21	0.37
C_W_credit	4.00	4.43	0.43	0.81	0.53
C_W_questions	3.81	4.11	0.56	1.04	0.54
C_W_guidelines	4.00	4.36	0.43	0.89	0.48
C_W_meetingsskip	4.00	4.34	0.43	0.87	0.49

A.6.2 LLM vs. Human Responses for Honesty-Humility at Work Items

Table 7: Means and standard deviations of responses from LLM-generated and human base data for items in the Honesty-Humility at Work (HH-W) item pool. The final column reports the SD ratio (LLM / Human).

Item	Mean (LLM)	Mean (Human)	SD (LLM)	SD (Human)	SD Ratio
HH_W_promotion	4.21	3.75	0.87	1.21	0.72
HH_W_coworkers	3.36	3.03	0.61	0.97	0.63
HH_W_responsibility	4.08	4.43	0.76	0.61	1.25
HH_W_perspectives	4.01	4.18	0.75	0.82	0.91
HH_W_criticism	4.01	4.41	0.75	0.64	1.17
HH_W_credit	4.05	4.53	0.78	0.61	1.28
HH_W_pressure	4.18	4.43	0.83	0.66	1.26
HH_W_opportunity	3.22	3.32	0.67	1.07	0.63
HH_W_colleagues	4.04	4.49	0.77	0.65	1.18
HH_W_disagreements	4.01	4.21	0.75	0.72	1.04
HH_W_respect	4.04	4.42	0.77	0.80	0.96
HH_W_integrity	4.15	4.36	0.81	0.68	1.19
HH_W_guidance	4.01	4.16	0.75	0.90	0.83
HH_W_ownership	4.05	4.19	0.78	0.77	1.01
HH_W_recognition	3.93	3.65	0.81	0.95	0.85
HH_W_doublecheck	3.97	4.45	0.75	0.69	1.09
HH_W_meetings	4.06	4.14	0.78	0.84	0.93
HH_W_assistance	3.96	4.25	0.80	0.85	0.94
HH_W_attacks	4.04	4.21	0.77	0.72	1.07
HH_W_humility	4.02	4.10	0.81	0.86	0.94

A.6.3 Human Comparison vs. Human Base Responses for Conscientiousness at Work Items

Table 8: Means and standard deviations (SD) of responses from human comparison (Human Comp.) and human base (Human) data for items in the Conscientiousness at Work (C-W) item pool. The final column reports the SD ratio (Human Comp. / Human).

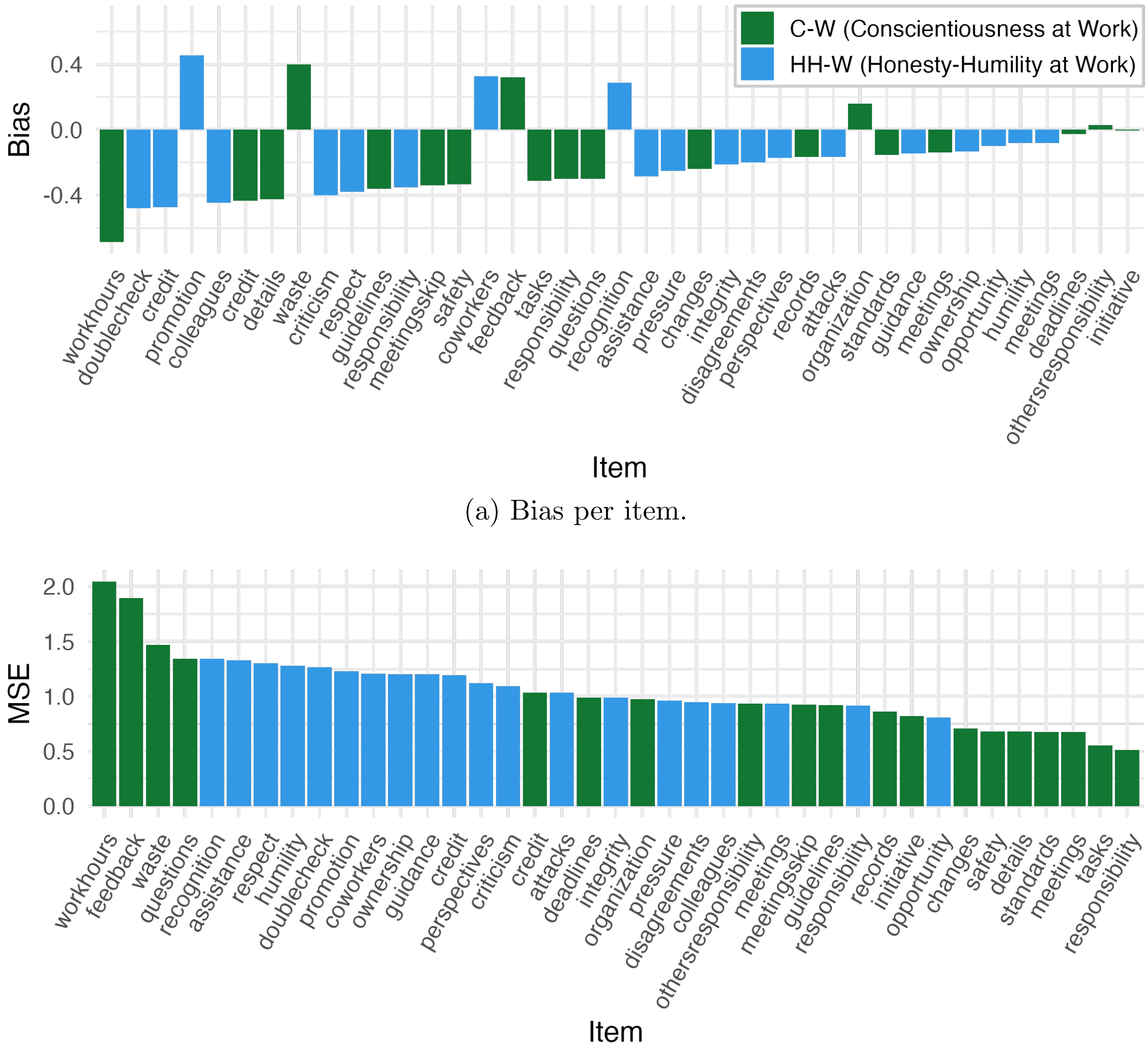
Item	**Mean** (Human Comp.)	**Mean** (Human)	**SD** (Human Comp.)	**SD** (Human)	**SD Ratio**
C_W_details	4.25	4.32	0.77	0.83	0.93
C_W_tasks	4.29	4.32	0.81	0.74	1.09
C_W_records	4.09	4.15	0.98	0.95	1.03
C_W_responsibility	4.51	4.46	0.62	0.61	1.02
C_W_standards	4.27	4.25	0.83	0.88	0.94
C_W_safety	4.45	4.42	0.71	0.74	0.96
C_W_workhours	3.84	3.87	1.16	1.18	0.98
C_W_deadlines	3.70	3.74	1.02	0.99	1.03
C_W_initiative	4.13	4.02	0.81	0.92	0.88
C_W_feedback	3.56	3.45	1.05	1.17	0.90
C_W_othersresponsibility	3.78	3.79	0.90	0.89	1.01
C_W_organization	4.01	3.85	0.81	0.97	0.84
C_W_meetings	4.13	4.15	0.80	0.90	0.89
C_W_changes	4.17	4.15	0.74	0.81	0.91
C_W_waste	3.49	3.57	1.13	1.21	0.93
C_W_credit	4.41	4.43	0.92	0.81	1.14
C_W_questions	4.03	4.11	0.98	1.04	0.94
C_W_guidelines	4.33	4.36	0.93	0.89	1.04
C_W_meetingsskip	4.29	4.34	0.86	0.87	0.99

A.6.4 Human Comparison vs. Human Base Responses for Honesty-Humility at Work Items

Table 9: Means and standard deviations (SD) of responses from human-comparison (Human Comp.) and human base (Human) data for items in the Honesty-Humility at Work (HH-W) item pool. The final column reports the SD ratio (Human Comp. / Human).

Item	**Mean** (Human Comp.)	**Mean** (Human)	**SD** (Human Comp.)	**SD** (Human)	**SD Ratio**
HH_W_promotion	3.84	3.75	1.24	1.21	1.02
HH_W_coworkers	3.06	3.03	1.09	0.97	1.12
HH_W_responsibility	4.47	4.43	0.62	0.61	1.02
HH_W_perspectives	4.24	4.18	0.77	0.82	0.94
HH_W_criticism	4.29	4.41	0.62	0.64	0.97
HH_W_credit	4.57	4.53	0.58	0.61	0.95
HH_W_pressure	4.36	4.43	0.67	0.66	1.02
HH_W_opportunity	3.27	3.32	1.14	1.07	1.07
HH_W_colleagues	4.44	4.49	0.69	0.65	1.06
HH_W_disagreements	4.27	4.21	0.65	0.72	0.90
HH_W_respect	4.35	4.42	0.75	0.80	0.94
HH_W_integrity	4.29	4.36	0.64	0.68	0.94
HH_W_guidance	4.17	4.16	0.78	0.90	0.87
HH_W_ownership	4.27	4.19	0.77	0.77	1.00
HH_W_recognition	3.82	3.65	0.95	0.95	1.00
HH_W_doublecheck	4.35	4.45	0.77	0.69	1.12
HH_W_meetings	4.21	4.14	0.79	0.84	0.94
HH_W_assistance	4.28	4.25	0.76	0.85	0.89
HH_W_attacks	4.23	4.21	0.71	0.72	0.99
HH_W_humility	4.22	4.10	0.79	0.86	0.92

A.7 Item-Level Bias and Mean Squared Error



(a) Bias per item.

(b) Mean Squared Error (MSE) per item.

Figure 8: Item-level bias (mean prediction error) and mean squared error between LLM-generated and human responses across Conscientiousness at Work (C-W) and Honesty-Humility at Work (HH-W) items. Note that certain item suffixes appear twice per plot. Combined with their respective trait prefix (C-W, HH-W) they are unique and represent different items, as listed in Appendices A.1.1 and A.1.2.

A.8 EFA Loadings

A.8.1 Before Item Selection (LLM Data)

Table 10: Exploratory factor loadings (oblimin rotation) based on LLM-generated data for the full set of Conscientiousness (C) and Honesty-Humility (HH) at Work items. Loadings > 0.40 are in bold.

Item	Factor 1	Factor 2
C_W_details	**0.89**	-0.02
C_W_tasks	**0.95**	-0.01
C_W_records	**0.94**	-0.01
C_W_responsibility	**0.69**	0.14
C_W_standards	**0.82**	0.04
C_W_safety	**0.83**	0.03
C_W_workhours	0.16	0.08
C_W_deadlines	**0.79**	-0.05
C_W_initiative	**0.95**	-0.03
C_W_feedback	**0.81**	0.01
C_W_othersresponsibility	**0.84**	-0.01
C_W_organization	**0.95**	-0.03
C_W_meetings	**0.96**	-0.01
C_W_changes	-0.03	0.32
C_W_waste	**0.92**	0.02
C_W_credit	**0.92**	0.02
C_W_questions	**0.81**	-0.05
C_W_guidelines	**0.92**	0.02
C_W_meetingsskip	**0.92**	0.02
HH_W_promotion	0.01	**0.90**
HH_W_coworkers	0.05	**0.83**
HH_W_responsibility	-0.00	**0.97**
HH_W_perspectives	-0.00	**0.98**
HH_W_criticism	-0.00	**0.98**
HH_W_credit	0.00	**0.99**
HH_W_pressure	-0.03	**0.94**
HH_W_opportunity	0.11	**0.86**
HH_W_colleagues	-0.00	**0.99**
HH_W_disagreements	-0.00	**0.98**
HH_W_respect	-0.00	**0.99**
HH_W_integrity	-0.04	**0.95**
HH_W_guidance	0.02	**0.98**
HH_W_ownership	0.00	**0.99**
HH_W_recognition	-0.06	**0.95**
HH_W_doublecheck	0.13	**0.89**
HH_W_meetings	0.00	**0.99**
HH_W_assistance	-0.07	**0.96**
HH_W_attacks	-0.00	**0.99**
HH_W_humility	-0.03	**0.97**

A.8.2 After Item Selection (LLM Data)

Table 11: Exploratory factor loadings (oblimin rotation) after item selection based on LLM-generated data for Conscientiousness (C) and Honesty-Humility (HH) at Work items. Loadings > 0.40 are in bold.

Item	Factor 1	Factor 2
C_W_meetings	**0.96**	-0.01
C_W_initiative	**0.95**	-0.03
C_W_organization	**0.95**	-0.03
C_W_tasks	**0.95**	-0.01
C_W_records	**0.94**	-0.01
C_W_guidelines	**0.92**	0.02
C_W_credit	**0.92**	0.02
C_W_meetingsskip	**0.92**	0.02
C_W_waste	**0.92**	0.02
C_W_details	**0.89**	-0.02
C_W_othersresponsibility	**0.84**	-0.01
C_W_safety	**0.82**	0.03
C_W_standards	**0.82**	0.04
C_W_questions	**0.81**	-0.04
C_W_feedback	**0.81**	0.01
C_W_deadlines	**0.79**	-0.05
C_W_responsibility	**0.69**	0.13
HH_W_respect	-0.00	**0.99**
HH_W_attacks	-0.00	**0.99**
HH_W_colleagues	-0.00	**0.99**
HH_W_credit	0.00	**0.99**
HH_W_ownership	0.00	**0.99**
HH_W_meetings	0.00	**0.99**
HH_W_guidance	0.02	**0.98**
HH_W_perspectives	-0.00	**0.98**
HH_W_disagreements	-0.00	**0.98**
HH_W_criticism	-0.00	**0.98**
HH_W_responsibility	-0.00	**0.97**
HH_W_humility	-0.03	**0.97**
HH_W_assistance	-0.07	**0.96**
HH_W_integrity	-0.03	**0.95**
HH_W_recognition	-0.06	**0.94**
HH_W_pressure	-0.03	**0.94**
HH_W_promotion	0.01	**0.90**
HH_W_doublecheck	0.13	**0.89**
HH_W_opportunity	0.11	**0.86**
HH_W_coworkers	0.05	**0.83**

A.8.3 Before Item Selection (Human Data)

Table 12: Exploratory factor loadings (oblimin rotation) based on human data for the full set of Conscientiousness (C) and Honesty-Humility (HH) at Work items. Loadings > 0.40 are in bold.

Item	Factor 1	Factor 2
C_W_details	**0.78**	0.05
C_W_tasks	**0.64**	-0.01
C_W_records	**0.65**	0.00
C_W_responsibility	0.36	0.39
C_W_standards	**0.59**	0.09
C_W_safety	0.30	0.40
C_W_workhours	0.20	-0.01
C_W_deadlines	**0.62**	-0.11
C_W_initiative	**0.47**	0.12
C_W_feedback	0.12	0.31
C_W_othersresponsibility	**0.48**	-0.07
C_W_organization	0.34	0.34
C_W_meetings	**0.75**	0.02
C_W_changes	-0.10	**0.42**
C_W_waste	0.39	0.15
C_W_credit	-0.02	0.26
C_W_questions	0.20	0.26
C_W_guidelines	0.22	**0.40**
C_W_meetingsskip	0.11	0.34
HH_W_promotion	-0.21	**0.57**
HH_W_coworkers	-0.07	**0.47**
HH_W_responsibility	0.34	**0.46**
HH_W_perspectives	0.12	**0.55**
HH_W_criticism	0.13	**0.43**
HH_W_credit	0.17	**0.51**
HH_W_pressure	0.23	**0.47**
HH_W_opportunity	-0.32	**0.67**
HH_W_colleagues	0.08	**0.62**
HH_W_disagreements	0.06	**0.62**
HH_W_respect	0.04	**0.49**
HH_W_integrity	**0.47**	0.25
HH_W_guidance	0.04	**0.45**
HH_W_ownership	0.35	0.27
HH_W_recognition	0.01	**0.48**
HH_W_doublecheck	**0.55**	-0.02
HH_W_meetings	-0.09	**0.48**
HH_W_assistance	0.21	**0.42**
HH_W_attacks	0.12	**0.47**
HH_W_humility	0.09	**0.44**

A.8.4 After Item Selection (Human Data)

Table 13: Exploratory factor loadings (oblimin rotation) after item selection based on human data for Conscientiousness (C) and Honesty-Humility (HH) at Work items. Loadings > 0.40 are in bold.

Item	Factor 1	Factor 2
C_W_details	**0.83**	0.06
C_W_meetings	**0.71**	0.03
C_W_deadlines	**0.70**	-0.17
C_W_records	**0.64**	0.04
C_W_tasks	**0.63**	0.03
C_W_standards	**0.54**	0.09
HH_W_disagreements	-0.08	**0.80**
HH_W_perspectives	0.01	**0.65**
HH_W_attacks	0.00	**0.63**
HH_W_colleagues	0.09	**0.62**
HH_W_credit	0.16	**0.49**
HH_W_pressure	0.25	**0.47**

B Electronic appendix

All data files, the Python script used for data generation (including the prompt), and the R code for data analysis are available on GitHub: `https://github.com/CFroehner/SynthData-Psych.git`.

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, June 26th, 2025

Cosima Fröhner