MEASURING AND IDENTIFYING FACTORS OF INDIVIDUALS' TRUST IN LARGE LANGUAGE MODELS

Edoardo Sebastiano De Duro
Department of Psychology and Cognitive Science
University of Trento, Italy
edoardo.deduro@unitn.it

Giuseppe Alessandro Veltri
NUS Yong Loo Lin School of Medicine
National University of Singapore, Singapore
gaveltri@nus.edu.sg

Hudson Golino
Department of Psychology
University of Virginia, USA
hfg9s@virginia.edu

Massimo Stella
Department of Psychology and Cognitive Science
University of Trento, Italy
massimo.stella-1@unitn.it

March 3, 2025

ABSTRACT

Large Language Models (LLMs) can engage in human-looking conversational exchanges. Although conversations can elicit trust between users and LLMs, scarce empirical research has examined trust formation in human–LLM contexts, beyond LLMs' trustworthiness or human trust in AI in general. Here, we introduce the Trust-In-LLMs Index (TILLMI) as a new framework to measure individuals' trust in LLMs, extending McAllister's cognitive and affective trust dimensions to LLM-human interactions. We developed TILLMI as a psychometric scale, prototyped with a novel protocol we called LLM-simulated validity. The LLM-based scale was then validated in a sample of 1,000 US respondents. Exploratory Factor Analysis identified a two-factor structure. Two items were then removed due to redundancy, yielding a final 6-item scale with a 2-factor structure. Confirmatory Factor Analysis on a separate subsample showed strong model fit ($CFI = .995$, $TLI = .991$, $RMSEA = .046$, $p_{X^2} > .05$). Convergent validity analysis revealed that trust in LLMs correlated positively with openness to experience, extraversion, and cognitive flexibility, but negatively with neuroticism. Based on these findings, we interpreted TILLMI's factors as "closeness with LLMs" (affective dimension) and "reliance on LLMs" (cognitive dimension). Younger males exhibited higher closeness with- and reliance on LLMs compared to older women. Individuals with no direct experience with LLMs exhibited lower levels of trust compared to LLMs' users. These findings offer a novel empirical foundation for measuring trust in AI-driven verbal communication, informing responsible design, and fostering balanced human–AI collaboration.

1 Introduction

Trust is a key factor shaping interactions in human settings. Trust fosters cooperation, reduces uncertainty, and supports the smooth flow of information [1, 2]. In psychology, trust has been studied for decades as a core determinant of team performance and stakeholder relationships. McAllister [1] described trust as both affective and cognitive. While affective trust involves emotional bonds and genuine care for others, cognitive trust relies on reasoned assessments of another party's competence and reliability. Subsequent studies have shown how, despite being highly correlated, cognitive and affective trust are indeed empirically distinguishable [3]. In workplaces and organizations, these dimensions help coworkers feel confident in each other's intentions and abilities. Employees who perceive both affective and cognitive trust in their teams often report higher job satisfaction, stronger commitment, and better overall performance [2].

However, workplaces are rapidly changing. Many tasks once reserved for human employees are now assigned to automated tools and artificial intelligence (AI). Recent advances in AI, especially in Large Language Models (LLMs), have made machines more capable of handling tasks that require advanced reasoning and context-driven analysis [4–6]. Although LLMs evidently differ from human colleagues [7], trust remains vital for effective collaboration [8, 9]. Without trust, users may hesitate to follow AI-generated advice, share sensitive information, or integrate these tools into daily workflows [6]. Conversely, excessive trust may lead to undue reliance on systems that can still produce errors or biased responses [5, 10–12].

Researchers have long been concerned with how people form trust in technology. Madsen and Gregor [13] argued that system reliability, interface design, and user perceptions of risk play large roles in human–computer trust. McKnight and colleagues [14] offered a framework for measuring trust in specific technologies, highlighting factors such as structural assurances, situational normality, and a user's general propensity to trust. Hoff and Bashir [15] expanded on this foundation by analysing how transparency, feedback, and system predictability contribute to trust in automation. In a meta-analysis, Schaefer and colleagues [16] found that users rely on cues such as perceived reliability, ease of use, and contextual factors when judging whether to trust automated agents. Moreover, in an experimental setting [17], interaction with social robots underlined that the manipulation of different variables impacted differently affective and cognitive trust (e.g. topic of the conversation influenced affective trust while robot's mistakes shaped the cognitive one).

Our approach differs from past works testing the trustworthiness of the content produced by LLMs [9, 18] and also by past works focusing on trust formation in various technological contexts [16, 17]. We argue that LLMs present unique challenges [11] that extend beyond traditional human-computer trust dynamics. Unlike simpler automated systems or social robots, LLMs present the unique capability of engaging in natural language exchanges that closely mirror human ones per form, syntax and emotional tone [4, 7]. Thus, measuring trust in LLMs requires a dedicated psychometric scale. Prior scales, designed for trust in automation or simpler forms of AI [19], cannot fully capture the nuances of human–LLM interactions. Researchers have created instruments to measure user trust in various technological contexts, such as e-commerce, recommendation systems, online services, and autonomous systems [14, 16, 20]. Others have focused on how design features, such as system transparency or perceived anthropomorphism, shape users' willingness to rely on AI [15, 21, 22]. Yet, large-scale text generation tools pose distinct challenges because they can produce fluent but potentially misleading or biased text. LLMs often function as "black boxes", making it difficult for users to assess their decision processes [6, 23, 24].

A new instrument must integrate established ideas about affective and cognitive trust [1], while also reflecting the affordances and risks specific to LLMs. On the one hand, affective trust describes the emotional bond between a user and a system, capturing whether the user feels assured or supported. On the other hand, cognitive trust is based on rational judgments of competence, reliability, and consistency. In the case of LLMs, these judgments hinge on factors like output accuracy, fairness, and the clarity of explanations provided by the system. Drawing from both components, a robust "trust equation" [25] would thus need to weigh users' emotional comfort (including aspects such as perceived benevolence) against their logical assessments of system performance [26].

Such a psychometric tool could help researchers pinpoint which factors—such as transparency, perceived competence, or ease of interaction—most strongly affect how users decide to trust or distrust an LLM. For instance, perceived competence or reliability may hinge on how often an LLM provides accurate, unbiased, or context-aware information. The ease of interaction, or closeness, could depend on the design of the user interface or personalization features that adapt to the needs of the user [13]. When these components merge, they shape whether a user sees the system as credible, dependable, and safe to rely upon.

Once available, a well-designed psychometric measure could guide developers on how to optimise LLMs or their feedback loops for further improving human-to-LLMs' trustworthiness. For example, if scale results reveal that perceived risk undermines trust in certain work settings, system designers might implement clearer disclaimers or provide confidence scores that reflect the LLM's uncertainty [8, 16]. If cognitive trust emerges as a stronger predictor of LLM adoption in an organization, developers could focus on verifiable performance metrics or robust error handling procedures. In contrast, if affective trust proves important in settings where users feel anxious about new technologies, an LLM might include more empathetic language or user-centric design elements [7, 22].

Finally, ethical concerns reinforce the need to study human trust in LLMs rather than LLMs' trustworthiness when producing content. When users trust large language models, the former accept vulnerability by disclosing personal information [7] or following automated decisions [22]. Undue vulnerability could create opportunities for harmful behavior, such as spreading misinformation or invading privacy [4]. Thus, measuring trust is essential for both leveraging the benefits of AI and guarding against its potential harms.

In this paper, we present the Trust-In-LLMs Index (TILLMI) as a new framework to measure individuals' trust in LLMs. Grounded in McAllister's [1] distinction between affective and cognitive trust, it also builds on research on human–machine trust [8, 13–16]. We describe each step of the scale development process, including item generation, pilot testing, and validation. Our goal is to create a reliable and valid instrument that can guide both researchers and practitioners in understanding and shaping trust in LLMs. We are confident this work contributes to the broader discussion on how to integrate LLM-based systems into professional and everyday contexts in ways that are beneficial for science [5], transparent, and ethically sound.

2 Results

2.1 Scale design

The initial version of the Trust-In-LLM-Index featured 8 items (see *Materials and Methods*). We outlined these items by drawing from McAllister's [1] conceptualisation of trust emerging in cooperative settings, such as workplaces and organizations. We combined that theoretical framework with relevant literature about the trust equation [25] where trust is considered as being constituted by multiple intertwined components (credibility, reliability, intimacy and self-orientation). Half of the items of the TILLMI aimed at capturing the affective dimension of trust, while the remaining half focused on cognitive trust in LLMs. Throughout the scale design, we framed LLMs as tools rather than companions, inspired by recent approaches [5]. Our novel psychometric scale employed a 5-level Likert-type scale with scores ranging from 1 (strongly disagree) to 5 (strongly agree). Following the initial scale design, a crucial step in scale development is the evaluation of item quality. To do so we employed a novel technique leveraging LLMs. We discuss works that adopted a similar approach and our implementation of this technique in the following sections.

2.2 Motivating LLM-simulated validity

Recent studies have shown that language embeddings and language models can effectively estimate psychometric measurements that traditionally relied on empirical data collection [10, 27]. In this way, LLMs can be used as (a) a way to simulate human participants and obtain large datasets of synthetic responses [7] and (b) tools to generate and evaluate items allowing to reduce the load on experts that are usually employed to assess items quality [27]. Inspired by these approaches combining LLMs and psychometrics, we conducted our own simulation test using LLMs as participants (see *Materials and Methods*) to assess the designed items' quality, non-redundancy, and internal structure. We defined this approach as "LLM-simulated validity".

2.3 LLM-simulated validity

We conducted an Exploratory Factor Analysis (EFA) on the synthetic data. We computed the Kaiser-Meyer-Olkin (KMO), a measure of the proportion of variance that might be common among items. This index is useful to assess whether the data is suited for subsequent factor analysis. The results ($KMO = .86$) were above the commonly accepted threshold of .80 [28], hence we proceeded further with the EFA. To understand the number of factors emerging we used 2 methods. The first one (Kaiser method) identifies factors with eigenvalues greater than 1, as these indicate that a factor explains enough variance and is worth keeping. The second approach (parallel analysis; [29]) is similar but compares each factor's eigenvalues with generated ones from a Monte Carlo simulation, which randomises the data while preserving the dimension of the sample. The first method yielded a 2-factor solution, while the parallel analysis simulation approach suggested the presence of 3 distinct factors. For both solutions, the emerged factors yielded good internal consistency ($\alpha > .85$), meaning that the items belonging to each factor tended to measure related concepts. Given the promising results of the simulated experiment, the questionnaire was thus administered to a total ($N = 1000$) of human participants. We collected the responses (see *Materials and Methods*) and compared the results with those simulated with GPT-4.

2.4 Comparison between LLMs and humans

Figure 1 shows the comparison between GPT and human scores to the initial version of the TILLMI. Interestingly, some of the patterns found in humans were reproduced by GPT-4, suggesting the capability of LLMs to grasp the underlying psychological constructs relative to trust, in a way similar to humans. However, some differences emerged. The scoring pattern for Item 7 appears unusual compared to other items. In our scale higher scores usually mean greater trust in LLMs. However, Item 7 presents an inverse trend. A high score on its first part ("Despite trusting LLMs' results overall") suggests high trust, but the full item ("Despite trusting LLMs' results overall, the last word is always mine") implies low trust. It is possible that people's responses were based only on Item 7's opening clause rather than the complete item, while GPT-4 might have considered the whole item description. For items that were more clearly worded (e.g. Item 8), the scoring pattern between humans and LLMs is more aligned.

Following this comparison of score distributions between humans and LLMs, we conducted an Exploratory Graph Analysis (EGA) to evaluate the structure of our new tool.

2.5 Exploratory Graph Analysis

We carried out an Exploratory Graph Analysis (EGA; [30]) to identify the underlying structure of human participants' responses on the TILLMI. EGA can infer the empirical number of factors in complex psychometric datasets by leveraging network analysis to identify communities from items' correlations [31]. The following analyses include only respondents who had used LLMs at least once ($n_1 = 521$).

The first step of the EGA involved performing the Unique Variable Analysis (UVA; [32]), a technique useful to evaluate the local dependency of pairs of items (i.e., item redundancy) through weighted topological overlap (wTO; [33]). Items with high wTO ($wTO > .3$) could, potentially, be removed without altering the internal structure of the network. In our case, none of the items showed high redundancy, so all 8-items were kept for further analysis.

Subsequently, we performed the EGA on the 8-item TILLMI scale. For further details on the parameters used, refer to the *Materials and Methods* section. In the resulting psychometric networks, variables are represented as nodes, with

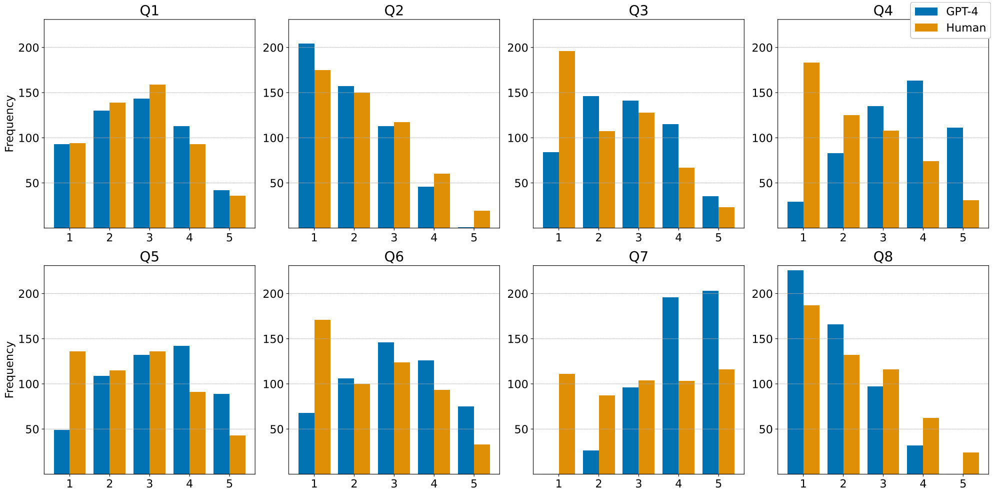


Figure 1: Response frequencies for 8 items of the initial TILLMI for GPT-4 (in blue) and humans (in orange). To balance the 2 dataset ($n_{humans} = 521$, $n_{gpt4} = 800$) we extract a random sample of $n_1 = 521$ from the synthetic GPT-4 dataset.

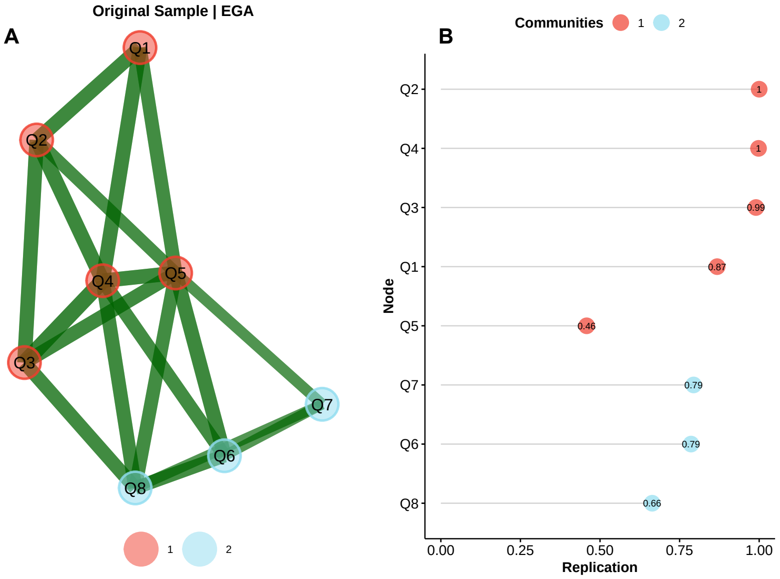


Figure 2: Exploratory Graph Analysis of the responses to the TILLMI for participants who stated to have used LLMs at least once ($n_1 = 521$). (A) Psychometric network plotted using EGAnet. Nodes represent items of the TILLMI. Edges indicate the interaction between nodes, with green links representing positive interactions. (B) Item stability plot for the TILLMI bootstrap analysis. Probability of each item being assigned to its original dimension across bootstrap iterations is shown. Higher values (closer to 1) indicate that an item consistently appears in the same dimension, suggesting greater stability.

their relationships visualised as edges. We show the network plot of the TILLMI in Figure 2A. Overall, EGA suggested that a 2-factor structure fitted the data well. The correlations between items were positive, meaning that they tended to measure a similar construct. This was expected, as all the items were designed to capture trust in LLMs, where higher scores coded for higher trust. Nevertheless, our analysis revealed two distinct factors, highlighting that two facets of the same construct were captured by our novel scale.

Lastly, we performed an EGA bootstrap analysis (see *Materials and Methods*), which involves repeatedly resampling the data and reestimating the network. In this way, we evaluated the stability and robustness of the estimated network structure. Two measures were taken into account: (a) Total Entropy Fit Index (TEFI; [34]) and (b) item stability [31, 35].

TEFI is a measure of the goodness of fit of the data to a model where lower values suggest a better fit. We compared the TEFI results of the bootstrap of our model, against a random 2-dimensional structure. The value found for the empirical 2-factor model ($TEFI = -3.7652, SD = 0.0267$) was significantly lower ($p = .006$) compared to a random 2-dimensional structure ($TEFI = -3.6682, SD = 0.0226$). Hence, the model emerging from human responses is better than a random model, suggesting the presence of an underlying structure in the data.

Item stability measures how consistently a variable appears in its originally estimated EGA dimension across repeated samplings. By doing so, it is possible to understand which nodes of a network show more stability. The results of the item stability analysis are shown in Figure 2B. Certain items showed lower stability (Q1, Q5, Q6, Q7 and Q8) compared to the others (Q2, Q3, Q4), meaning that they could be not unique to a specific dimension. There are many reasons why this could be the case. For example, it might be possible that the EGA, while still identifying 2 factors, does not capture nuanced differences in two closely related facets of trust, i.e. in the considered sample size and with a limited number of items, the considered dimensions look very similar to each other. To gain further insights, we carried out a traditional Exploratory Factor Analysis to determine whether similar results were to be found.

2.6 Exploratory Factor Analysis suitability

To further examine the underlying structure of our data, we conducted an Exploratory Factor Analysis (EFA) on a subsample of the dataset (see *Materials and Methods*) that was not used for the subsequent validation of the model in the Confirmatory Factor Analysis (CFA). We begin with an assessment of the data's factorial suitability. To do so, we computed (a) the Kaiser-Meyer-Olkin index (KMO; [28]) and (b) Bartlett's test of sphericity [36]. We already described the first method previously. Bartlett's test of sphericity is used to ensure that the correlation matrix between items differs from an identity matrix, where a significant result ($p < .05$) indicates that sufficient correlations exist among the variables to make factor analysis meaningful. In our case the results of the KMO ($KMO = .92$) and Bartlett's test ($\chi^2 = 1374.148$, $df = 28$, $p < .001$) suggested a significant degree of correlation between items, indicating that our sample is suitable for further analysis.

In addition, we checked whether the assumption of normality at a univariate level (Shapiro-Wilk test; [37]) and multivariate level (Mardia test; [38]) was met. In both cases, our data did not show normality ($W = .85572, p < .05$; $z - kurtosis = 13.7, p < .05$). Given the non-normality of our data, we proceeded with the EFA and adopted principal axis factoring using the standard oblimin rotation method to obtain the initial results of our factor analysis.

2.7 Selection of factor number for Exploratory Factor Analysis

Having confirmed the data's suitability for factor analysis, we proceeded to determine the optimal number of factors using parallel analysis [29] to compare the eigenvalues obtained from the test sample, against a model with uncorrelated variables. The number of factors resulting was equal to 2. Through subsequent analysis, we considered also the criterion of Very Simple Structure (VSS; [39]) and Velicer's Minimum Average Partial (MAP; [40]) to further confirm the number of factors. Although some of the results hinted at a single-factor structure (Table 1), the RMSEA index of fit revealed a lower fit for such a solution [41, 42] compared to the 2-factor model. How to address this apparent conflict in the data? Overall, the results suggest that a clear optimal number of factors is not evident from the data alone. As suggested in previous research [43, 44], factor retention can be partially informed by theoretical insights and domain knowledge to determine the factor structure that is not only more "correct", but also more meaningful. For this reasons, in light of (a) the previous distinction of trust in two facets (cognitive and affective; [3, 45, 46]), (b) the rationale behind the design of the initial 8-items [1] we opted for the 2-factor model and continued with the Exploratory Factor Analysis.

Table 1: Fit indices for 1-factor and 2-factor models. In bold, the best results are highlighted.

Model	VSS	MAP	RMSEA
1-Factor	**.95**	**.039**	.115
2-Factor	.52	.067	**.077**

Table 2: Factor loadings and component correlation for the 2-factor model. Cut thresholds for factor loadings are set to $|.3|$.

Item	PA1	PA2	com
Q1	**0.503**	**0.334**	**1.74**
Q2	0.877		1.01
Q3	0.855		1.00
Q4	0.860		1.00
Q5	**0.403**	**0.429**	**1.99**
Q6		0.693	1.25
Q7		0.763	1.04
Q8	**0.335**	**0.514**	**1.72**
Factors Correlation			
PA1	1.000	0.774	
PA2	0.774	1.000	

2.8 Exploratory Factor Analysis

Having selected the number of optimal factors, we ran the EFA using the parameters selected previously (*principal axis factoring* as estimation method and *oblimin* as rotation method). To assess the goodness of the EFA we consider (a) factor loadings (the level of the strength of the relation between factors and the observed variables), (b) complexities (representing the degree to which an item is unique to one factor) and (c) factor correlations (correlation between whole factors combined). The results of the first EFA highlighted the presence of some poor-quality items. Table 2 shows the results obtained. Items 1, 5 and 8 presented moderate levels of cross-loadings (items are associated with both factors exceeding our arbitrarily set threshold of $> |.3|$), indicating they did not exclusively associate with a single factor. All of these 3 items already showed low stability in the previous EGA. This is confirmed by the values of the complexities that showcased values diverging from 1. Nevertheless, factor correlations were good (< 0.85) [47] confirming that it was worth keeping both the 2 factors.

Given the results found, we decided to drop Item 5 first and subsequently Item 1 (following the same procedure and rationale). The results of the second step of the EFA can be found in the *SI Appendix, Table S1*. Finally, we ended up with a 6-item structure from the initial 8 items. In this model, Factor 1 accounts for 39.2% of the total variance of the dataset, while Factor 2 explains 27.1%. The results of the final EFA are presented in Table 3. Following the refinement to a 6-item model, we employed Confirmatory Factor Analysis to rigorously validate the structural validity of the TILLMI.

2.9 Confirmatory Factor Analysis

We conducted a Confirmatory Factor Analysis to validate the 2-factor solution and assess model's goodness of fit on an independent sample of the dataset. In Figure 3 we present the path diagram including the latent variables (Factor 1 and Factor 2) and the observed variables ($Q2, Q3, Q4$; $Q6, Q7, Q8$). The procedure for deriving this model is detailed in the *Material and Methods* section. Covariance between the 2 latent factors was high ($\phi = 0.94$), suggesting the close nature of the 2 components of trust. Despite being highly related, these dimensions coded for different facets of trust

Table 3: Factor loadings (FL) and complexities (com) of the 2-factor structure emerging from the EFA.

Factor	Item	FL	com
Factor 1	Q2	0.859	1.00
	Q3	0.861	1.00
	Q4	0.815	1.01
Factor 2	Q6	0.719	1.03
	Q7	0.750	1.08
	Q8	0.646	1.26

that in the past have shown to influence human behaviour in different ways [3]. Moreover, in the following sections, we show that affective and cognitive trust present differential patterns of association with other validated constructs, suggesting that keeping the distinction was psychologically meaningful.

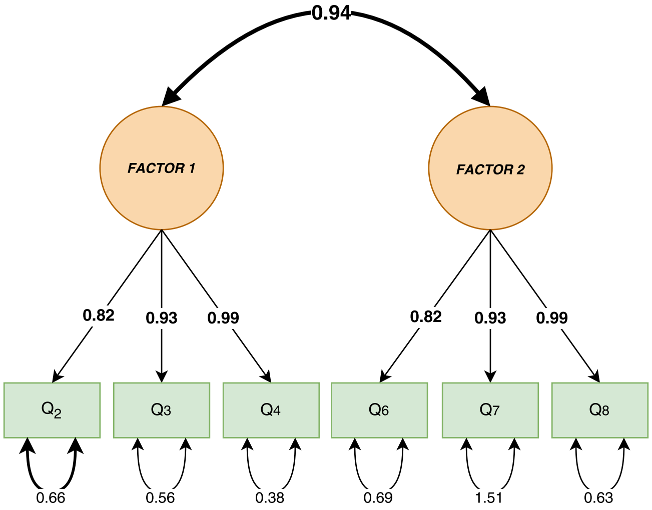


Figure 3: CFA model representing the final 2 latent factors and the corresponding observed variables. Each path from latent to observed variable includes the factor loading (λ). We show the measurement error for each observed variable (δ).

To evaluate our model, we used five different fit indexes: (a) chi-square test (χ^2; [48]), (b) root mean square error of approximation (RMSEA; [49]), (c) comparative fit index (CFI; [50]), (d) Tucker-Lewis index (TLI; [51]) and (e) root mean square and standardized root mean square residual (SRMR; [52]). χ^2, RMSEA and SRMR are absolute indexes, meaning that they compare the observed model with the theoretical one, without a reference or baseline model [52]. CFI and TLI, instead, are incremental indexes in the sense that they compare the observed model with a hypothesised model with the same variables assuming no relationship between them.

The results in Table 4 show the corrected (or scaled) versions of the indexes. As the chi-square measure is highly dependent on the numerosity of the sample (and the usefulness of its adoption in its raw form in CFA is quite debated; [48]), we used its scaled version that corrects for the non-normality of our data (using Satorra-Bentler correction; [53]). Such a correction is extended to the RMSEA, CFI and TLI as they are chi-square dependent measures.

Chi-square analysis yielded satisfactory results as the p-value is above .05, meaning that we cannot reject the hypothesis that the model does not fit the data (H_0) and hence we accept the hypothesis that our model, instead, fits the data (H_1)

Table 4: CFA fit index for 2-factor model. Scaled version of χ^2 is shown. For CFI, TLI and RMSEA we provide the robust version.

χ^2	RMSEA	CFI	TLI	SRMR
13.012, df = 8, p = .111	.046	.995	.991	.022

[54]. Similarly, the results of the other indexes showed an acceptable goodness of fit, as reported in Table 4 (i.e. our RMSEA was below .5 [50], CFI was above .95, TLI was above .90 [54] and SRMR was below .08 [52]).

Overall, the confirmatory factor analysis provided support for our hypothesised 2-factor model, demonstrating strong fit statistics across multiple indices.

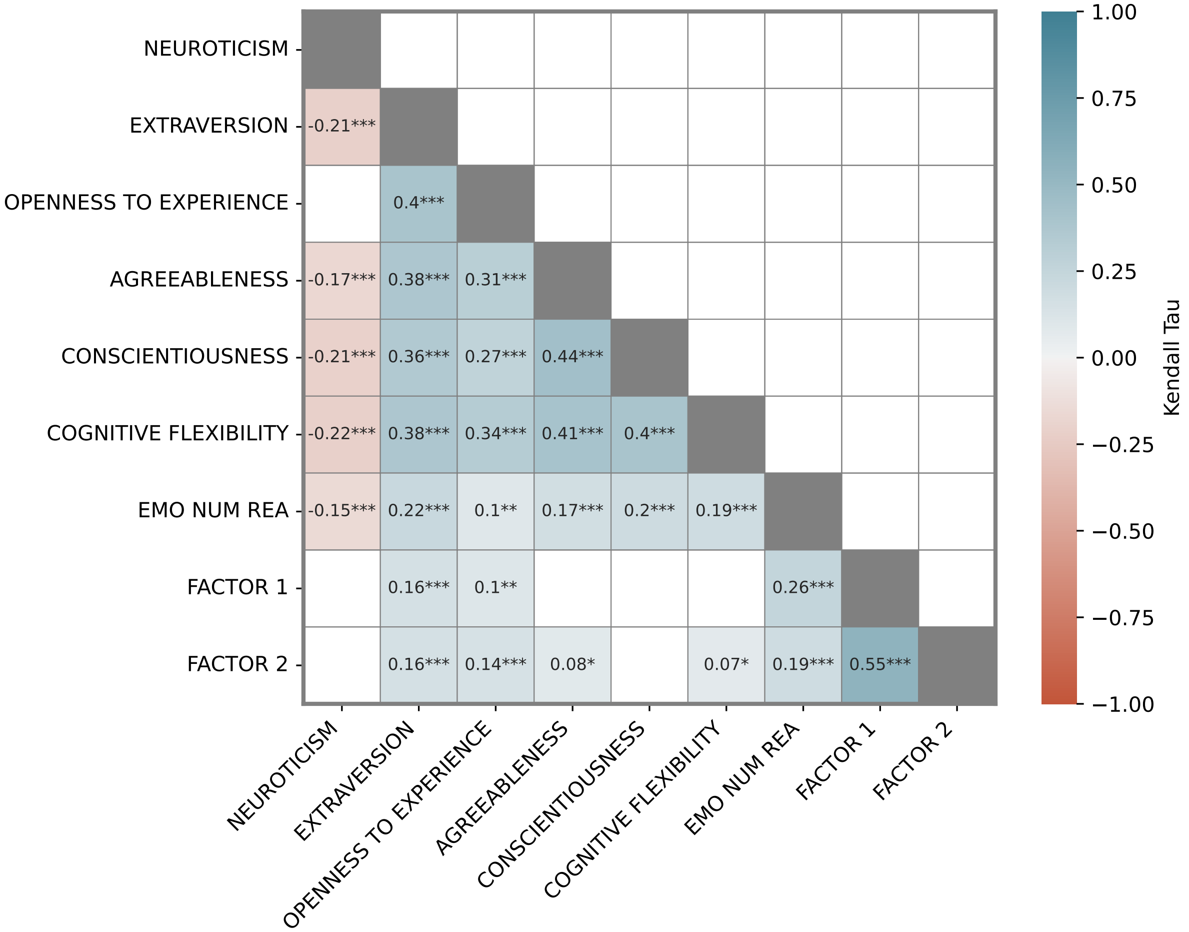


Figure 4: Correlations among constructs, with each construct computed as the sum of its constituent items. Emo Num Rea represents the aggregated self-reported scores for the numerical reasoning task. The correlations are computed using Kendall-Tau correlation coefficient on the subset of the population who stated to have used LLMs at least once ($n_1 = 521$). Blue tiles indicate positive correlations; Red tiles indicate negative correlations. White tiles represent non-significant correlations. Significance levels are indicated as * ($p < .05$), ** ($p < .01$), *** ($p < .001$).

2.10 Internal reliability

Before proceeding with the evaluation of convergent/divergent validity, a crucial step in psychometric validation involves assessing internal reliability (the degree to which different items belonging to the same factor measure the same construct; [55]). Ensuring high internal reliability is crucial, as it indicates that the scale produces stable and coherent measurements, strengthening its overall validity. Using Cronbach's Alpha (α), the standard metric for internal consistency assessment [55], we analysed each of the two factors and found acceptable [43] values ($\alpha_{F_1} = .893$; $\alpha_{F_2} = .781$) suggesting that items within each factor ($Q2, Q3, Q4$ and $Q6, Q7, Q8$) were consistent and reliable.

Having established the internal reliability of our factors, we next examined their convergent validity with pre-existing psychological measures.

2.11 Convergent validity

By including other scales in the survey (see *Materials and Methods*) we were able to test Factor 1 and Factor 2's convergent validity with other established psychometric scales (IPIP-NEO [3] and cognitive flexibility [56]), including an Emotional Recall Task [57] and a novel anxiety-related numerical reasoning task. Correlations between Factor 1 and 2 with these other psychometric measures are presented in Figure 4.

As expected, the 2 TILLMI's components of trust in LLMs positively correlated (Kendall-Tau $\tau_{F_1,F2} = 0.55, p < .001$). This is analogous to other psychometric scales of trust like McAllister's [1], which reported positive Pearson correlations between affective and cognitive trust in cooperative professional settings ($r = 0.63$).

TILLMI's Factor 1 and Factor 2 were positively correlated with personality traits expressing acceptance and adaptation to new concepts and environments, namely: Openness to experience ($\tau_{F_1,O} = 0.1, p = .001$; $\tau_{F_2,O} = 0.14, p < .001$) and Extraversion ($\tau_{F_1,E} = 0.16, p < 0.001$; $\tau_{F_2,E} = 0.16, p < .001$). These findings suggested that Factors 1 and 2, relative to trust in non-human agents, display statistical relationships with personality traits mostly relative to emotional components of human experience. Moreover, Cognitive Flexibility was positively correlated with Factor 2 ($\tau_{F_2,CF} = 0.07, p = .018$) but not with Factor 1. This differential pattern of correlations aligned with our conceptual framework, as the scale was designed to capture a cognitive aspect of trust with one factor/dimension and an affective one with the other factor/dimension.

2.12 Convergent validity with mental health distress levels

As an additional measure, we computed the correlations between Factor 1, Factor 2 and the Depression, Anxiety and Stress scores as computed by DASentimental [58], an AI assessing these mental health indicators from texts. We here used DASentimental on the textual responses to the Emotional Recall Task [57] in the survey (see *Materials and Methods*). Significant negative correlations between our measures of trust and most of the emotional distress indicators were found. Specifically, Factor 1 and Factor 2 presented a negative association with depression ($\tau_{F_1,D} = -0.12, p < .001$; $\tau_{F_2,D} = -0.12, p < .001$) and stress ($\tau_{F_1,S} = -0.13, p < .001$; $\tau_{F_2,S} = -0.15, p < .001$). For what concerns anxiety, only Factor 2 showcased a significant association ($\tau_{F_1,A} = -0.11, p = .004$). These findings suggested that individuals with higher trust in LLMs generally reported lower levels of emotional distress. It is important to note that from our correlation analysis, it was not possible to understand whether it was the distress state (as measured from the ERT) to determine specific trust levels or vice versa. Future research could examine the directionality of this relationship.

2.13 *Closeness with LLMs* and *Reliance on LLMs*

In light of (a) the correlation with established measures of personality, cognitive flexibility and mental distress, (b) the rationale behind the design of the items (taking inspiration from the distinction between cognitive and affective trust; [1]) and (c) the close but distinct nature of the 2 factors, we decided to name Factor 1 and Factor 2 as "closeness with LLMs" and "reliance on LLMs". We argue that closeness represents the affective dimension of trust in LLMs, mostly driven by feelings or emotions towards these systems. Reliance, instead, pertains to the cognitive aspect of trust and is primarily driven by logical assessment of LLMs' capabilities to deliver accurate and dependable responses. After

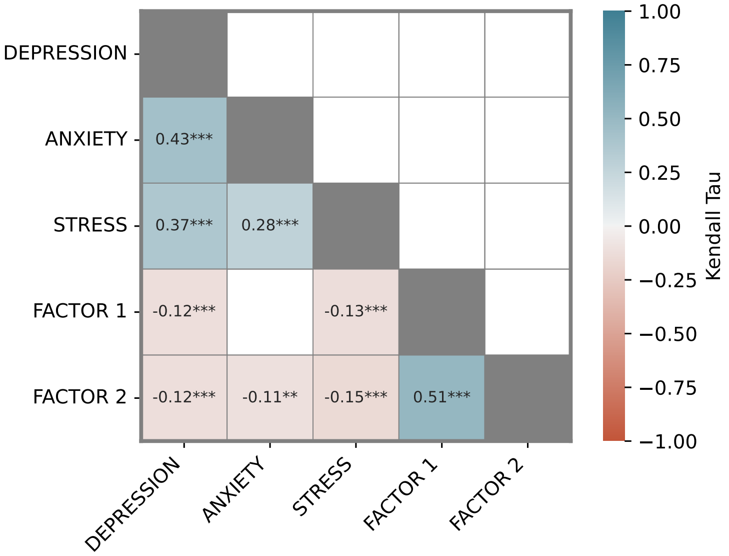


Figure 5: Correlations between Factor 1, Factor 2 and psychological measures (Depression, Anxiety, and Stress). These measures were derived using the DAsentimental framework, analysing the 10 words participants used to describe their feelings when interacting with LLMs. Out of the participants who used LLMs at least once ($n_1 = 521$), several responses ($n_3 = 124$) were excluded due to invalid entries in the emotion-response text boxes of the survey. Hence, these correlations, are relative only to ($n_4 = 397$). Blue tiles indicate positive correlations; Red tiles indicate negative correlations. White tiles represent non-significant correlations. Significance levels are indicated as * ($p < .05$), ** ($p < .01$), *** ($p < .001$).

having delineated the conceptual differences between "closeness with LLMs" and "reliance on LLMs", we explored whether demographic factors were related to individuals' trust in LLMs.

2.14 Regression with demographic information

We conducted an Ordinary Least Square (OLS) regression to investigate the association between age and gender with our measures of closeness and reliance. We found that male and younger participants tended to showcase significantly higher scores for both dimensions of trust in LLMs. Results are shown in Table 5.

Table 5: Regression of Age and Gender on closeness and reliance. We show the value of the intercept (β), standard deviation (σ) and the p-value (p). Gender is coded as 1 = Male, 2 = Female ($n_{male} = 291, n_{female} = 230$). Only participants who had used LLMs at least once were included in this analysis ($n_1 = 521$).

	β	δ	p
Closeness with LLMs			
Gender	-0.317	.135	.005
Age	-0.457	.147	.001
Reliance on LLMs			
Gender	-0.638	.272	.020
Age	-0.284	.092	.002

Having assessed the relationship between demographic information and our TILLMI measures, we proceeded to further validate our findings evaluating its divergent validity.

2.15 Divergent validity

Divergent validity (or discriminant validity) is the extent to which a measure is not correlated with theoretically unrelated constructs. This property can be established by confirming that populations not expected to exhibit a particular construct indeed show minimal evidence of possessing it when assessed with the measurement instrument. We assessed this by computing the same correlations of Figure 4 against the subset of the responses of people who claimed to have never used LLMs ($n_2 = 479$). The correlogram can be found in *SI Appendix, Figure S1*. Most constructs revealed no significant correlations. However, we found 3 exceptions: positive correlations between neuroticism and Factor 1 (or "closeness with LLMs") ($\tau_{F_1,N} = 0.11, p = .002$), neuroticism and Factor 2 (or "reliance to LLMs") ($\tau_{F_2,N} = 0.15$, $p < .001$), and a negative correlation between Factor 2 and cognitive flexibility ($\tau_{F_2,CF} = -0.08, p = .038$). Taken together, these results confirmed that our novel trust measure captures meaningful characteristics in LLM users while showing no significant patterns in non-users, supporting its specificity to LLM interaction experiences.

Building on these findings, we sought to quantify the actual difference in trust levels between participants with prior LLM interaction ($n_1 = 521$) and those without ($n_2 = 479$) using independent samples t-tests. Results showed significantly higher trust ($t(998) = 23.61, p < .001$) among previous LLM users ($M_1 = 14.55, \sigma = 5.87$) compared to non-users ($M_2 = 7.42, \sigma = 3.14$).

2.16 Network Psychometrics and item centrality

Lastly, we calculated items' strengths to unveil the items with the highest importance in the TILLMI (adopting a network science approach). Strength measures the local connectivity of a node by considering both the quantity and weight of its connections [59]. For a technical description of the process, see the *Materials and Methods* section. We report the results of humans and GPT-4s in Table 6.

Table 6: Comparison of node strength between humans and GPT-4 psychometric networks. Sorted by humans' values.

Item	Humans	GPT-4
$Q4$ (*Invest plenty of time in prompts [..]*)	**3.189**	3.097
$Q3$ (*Sharing wellbeing concerns [..]*)	2.994	3.739
$Q6$ (*Rely on LLMs in jobs [..]*)	2.973	3.822
$Q8$ (*Trust more LLMs than people [..]*)	2.958	**4.085**
$Q2$ (*Sense of dismay without LLMs [..]*)	2.878	3.601
$Q7$ (*Last word is always mine [..]*)	2.032	3.060

From the strength analysis, Item 4 (*I invest plenty of time developing and improving my prompts to interact with LLMs*) appeared to be the most important for humans. This suggested that time investment may be a key factor in trust formation as it leads to a better understanding and higher confidence. Interestingly, GPT-4 did not show the same pattern. Instead, for LLMs the comparative aspect of trust (specifically the perception of how they compare to humans) emerged as a more significant component, exemplified by the statement: "I tend to trust more LLMs over other people". In both humans and GPT-4 there was an agreement in the least relevant node (Item 7, *Despite trusting LLMs' results overall, the last word is always mine*) in terms of correlations with other items.

3 Dicussion

We introduce the Trust-In-LLMs Index (TILLMI) as a new framework for measuring individuals' trust in LLMs. Via psychologically informed item design, and with the contribution of LLMs in item quality assessment, we carried out a comprehensive validation in a US sample. The final version of TILLMI yields two factors ("closeness with LLMs" and "reliance on LLMs") that are related but distinct.

This is the key result of our work: We find that human trust in LLMs partitions in cognitive and affective components, similarly to what was found by past relevant research for human trust in co-workers [1, 3], management [45], organisations [2] and even computers [13]. Affective trust arises from emotional bonds, benevolence and repeated positive interactions [2]. In contrast, cognitive trust is based on rational assessments of competence, predictability, and reliability [1]. Both components operate in parallel, shaping human decisions about whom or what to trust [3]. Applying TILLMI's framework to LLMs, we observe two analogous dimensions. Users can develop cognitive trust based on the model's accuracy and consistency. We call this dimension "reliance on LLMs". However, affective trust also plays a role: users may feel comforted by the model's fluency and responsiveness, fostering an illusion of social connection [60], a dimension we call "closeness to LLMs".

Our work adopts a disruptive perspective compared to past approaches in LLMs' trustworthiness, i.e. the task of understanding how unbiased/reliable LLMs' content can be. Liu and colleagues [9] defined LLMs' reliability as "generating correct, truthful, and consistent outputs with proper confidence". In this regard, Bo and colleagues [18] showed that disclaimers regarding LLMs' confidence in their responses improved users' over-reliance on models but were generally ineffective in promoting an appropriate level of reliance. Our approach shifts the focus toward human users' perceptions towards LLMs, a paradigm that is needed to understand whether humans can trust LLM-based conversational agents.

Our definition of "reliance" on LLMs is disruptive in the sense that, in validating TILLMI, we found that experiencing misleading outputs is not a key experience of cognitive trust in large language models. Hallucinations are mostly due to LLMs' reinforcement learning and represent a fascinating yet only partially understood cognitive phenomenon [5, 11]. In our case, user responses made statistically redundant item $Q5$ (i.e. *LLMs perform the tasks mostly with competence and precision, without hallucinations*) within the "reliance in LLMs" factor. This does not mean that individuals trusted LLMs independently on the matter of hallucinations. Instead, $Q5$ being redundant means that the same information encoded in $Q5$ is present also in the other reliance items. These other items are relative to accurate hallucination-free responses which can simplify jobs and provide more trustworthy content. Future research could thus use TILLMI's reliance factor to explore how hallucinations might affect users' trust in LLMs.

We also found a second psychological dimension, the one of "closeness to LLMs". This factor includes items encoding elements of emotional support, well-being and commitment, which are key elements of successful professional relationships in virtual teams [61]. Closeness to LLMs also bears an interesting parallel with the well-known ELIZA effect [60] where people tend to establish emotional connections with AIs as a byproduct of attributing human-like characteristics (and mental states) to such agents. Future research on this dimension might explore how trust evolves when users' perceive LLMs as companions rather than mere tools, e.g. as agents providing well-being support [7].

Understanding how these components can change across individuals has significant implications. For instance, we find closeness and reliance towards LLMs are higher in younger men. Our evidence aligns with relevant literature [14] showing a significant negative correlation between age and perceived trustworthiness of technology in general ($\beta = -0.17, p < .001$). Interestingly, literature regarding gender effects on trust towards AIs shows findings that differ according to the target of human trust. While women were found to showcase higher levels of trust in AI-enabled systems [62], a cross-country study [63] found no evidence for gender-based differences in trust towards AI educational technology. Our findings indicate that LLMs might be different from other AIs or technologies, underlining the need for future gender studies investigating human-LLM interactions.

Via the DASentimental AI [58], we also found that higher levels of closeness and reliance were associated with lower mental distress. This result aligns well with relevant literature. Trust can foster emotional support, reduce feelings of isolation and provide security, which all reduce anxiety and depression [1, 3]. In organisational settings, trust in leaders and colleagues creates supportive environments that reduce stress [2, 61]. More recent works [8, 15] show that interacting with reliable systems reduces cognitive load and uncertainty, which are significant contributors to stress.

These elements can all apply to LLMs and thus converge in outlining a positive side that affective and cognitive trust in LLMs might have for mental well being.

There is also a negative side to trust in LLMs. Trust was recently found as a key element of LLMs' persuasiveness as perceived by humans [4]. Higher levels of trust in LLMs might also imply an easier chance for LLMs to persuade humans, with both positive opportunities for self-improvement but also space for human manipulation or vulnerability to hallucinations [5].

3.1 Artificial humans vs. real humans

We prototyped TILLMI through a novel LLM-simulated validity, based on GPT-4's responses. Certain response patterns observed in GPT-4 were interestingly reproduced by humans (cf. Fig. 1), indicating a capability for GPT-4 to reflect psychological constructs in ways similar to human cognition (see also [27]). However, as mentioned in past works [10], humans and GPT's data can display some differences.

GPT-4's correlational structure differed from humans' in identifying item $Q8$ as the most central, i.e. "trust more LLMs than people". This intriguing element, where an LLM highlights trusting more other LLMs over humans, might be a bias due to reinforcement learning [11, 12, 18].

Interestingly, humans and GPT-4 read item $Q7$ very differently. Human participants focused on the first sub-clause ("trusting LLMs' results overall") whereas GPT-4 put more emphasis on the second part of the item ("the last word is always mine"). This discrepancy suggests that human participants might process complex statements differently under time constraints, whereas LLMs, with considerable language processing capacity [4, 10], consider the complete meaning item. Taking into account such a difference is fundamental for future research involving the application of insights gained from LLMs to human cognition and behaviour.

3.2 Limitations and Future Directions

This study presents some limitations. Firstly, our sample focuses on US individuals. Future research could test the cross-cultural validity of TILLMI. Secondly, while our scale was designed to assess general trust in LLMs, it does not differentiate between specific models. Indeed, this is both a limitation and a strength. While our approach captures individual differences in AI trust broadly, future research could tailor the scale to specific LLMs such as ChatGPT, Claude, or DeepSeek by modifying item wording accordingly. Thirdly, our work used self-expressed measures of mental well-being, future research could integrate TILLMI with clinical experimental setups, further testing the current findings of this work.

Future works should acknowledge that trust in LLMs might change according to context. For instance, in high-stakes domains like healthcare and legal advice, enhanced transparency and clear explanations may soften human algorithm aversion [64] and thus increase cognitive reliance. Instead, in creative settings like generating art with AI [64] or social media [6], affective closeness could play a more influential role. Situational factors could thus create feedback mechanism that can be investigated with TILLMI.

3.3 Conclusions

TILLMI's factors of closeness with- and reliance on LLMs highlighted an intriguing range of interactions with personality traits, cognitive flexibility, demographics and mental health. We believe this provides compelling evidence that TILLMI can be a valuable tool when further exploring the complexities of LLM-human interactions. We believe our work provides a quantitative ground, supported by psychological theories, for investigating how humans can trust LLMs over time and across contexts of use, conditions and purposes.

4 Materials and Methods

4.1 Data collection

TILLMI was administered to a total of 1,000 US citizens online from May to August 2024. Participants were recruited through an on-line panel provider, Bilendi. Qualtrics was used to design the visual interface of the questionnaire and collect the responses. Each participant who successfully completed the survey was adequately compensated for their time.

As the first step in the survey, participants were asked to provide demographic information (e.g. age, biological gender, etc.). The newly developed items of the TILLMI were administered alongside established, validated psychometric tools to assess the convergent validity of the psychometric scale. TILLMI's items are reported in Table 7. The measurement protocol included the following scales:

- IPIP-NEO Inventory [65] for personality trait assessment. We assess neuroticism, extraversion, openness to experience, agreeableness and conscientiousness (each measured with 5 items).
- Cognitive Flexibility scale [56] to measures an individual's ability to adapt their thinking and behaviour in response to changing circumstances or new information available.

In addition, after the trust assessment, participants completed an Emotional Recall Task [57] related to their feelings during recent LLM interactions.

Lastly, participants completed 3 numerical reasoning tasks and rated their emotional state on a 5-point Likert scale (1 = very negative emotion, 5 = very positive emotion). This emotional self-rating served as a measure of task-induced stress during numerical problem-solving.

Table 7: TILLMI's initial 8 items regarding interactions with LLMs.

Item	Text
$Q1$	I feel at ease with LLMs and I can freely share my ideas with them.
$Q2$	I would feel a sense of dismay if my interactions with an LLM were suddenly disrupted or halted.
$Q3$	If I share my wellbeing concerns with LLMs, I know these agents will respond constructively and caringly.
$Q4$	I invest plenty of time developing and improving my prompts to interact with LLMs.
$Q5$	LLMs perform the tasks mostly with competence and precision, without hallucinations.
$Q6$	I can rely on LLMs not to make my job more difficult by careless work.
$Q7$	Despite trusting LLMs' results overall, the last word is always mine.
$Q8$	I tend to trust LLMs more than other people.

4.2 Data pre-processing

Out of 1,000 participants, 51.7% were female, 48.2% were males and 0.01% were non-binary. Their mean age was 31.6 years ($\sigma = 17.6$). From the initial 1,000 responses, we excluded participants who stated that they had never used LLM before ($n_2 = 479$). The remaining respondents ($n_1 = 521$) were randomly divided into 2 subsets using Python's `random.sample` method (`random_seed = 42`): a first subsample ($n_5 = 260$) for the exploratory analysis and a second subsample ($n_6 = 261$) for the confirmatory analysis to cross-validate the results [66]. We present a descriptive table of the results for the people who used LLMs ($n_1 = 521$) in Table 8.

The Exploratory Factor Analysis (EFA) and the Confirmatory Factor Analysis (CFA) were carried out using R (v 4.4.2). The EFA required the following R packages: `readr` (2.1.5), `psych` (v 2.4.12).

CFA was performed via the `lavaan` (v 0.6-17) package. When fitting the data using the `cfa` function in `lavaan` we set the following parameters: `std.lv = TRUE, ordered = TRUE`. Using the `summary` function setting `fit = TRUE,`

Table 8: Descriptive statistics for the initial version of the TILLMI. We show the response frequencies for each item, mean (M) and standard deviation (σ). 1 = never experienced it, 5 = always experience it.

Item	**1**	**2**	**3**	**4**	**5**	**M**	σ
$Q1$	94	139	159	93	36	2.689	1.160
$Q2$	175	150	117	60	19	2.228	1.139
$Q3$	196	107	128	67	23	2.259	1.211
$Q4$	183	125	108	74	31	2.319	1.249
$Q5$	136	115	136	91	43	2.597	1.268
$Q6$	171	100	124	93	33	2.457	1.281
$Q7$	111	87	104	103	116	3.050	1.451
$Q8$	187	132	116	62	24	2.240	1.189

`rsquare = TRUE` we obtained the fit measures to evaluate our model. We used the Diagonally Weighted Least Squares (DWLS) estimator as our data is not normally distributed (at univariate and multivariate level, see Results Section) and ordinal [67].

4.3 Emotional Recall Task and DASentimental

In the survey, we incorporated the Emotional Recall Task (ERT), adapted from [57], as a supplementary measure to assess mental well-being in relation to LLM interactions. The ERT asks participants to freely list 10 emotions experienced in the past month, avoiding preset emotion checklists. This approach allows for potentially richer insights into individuals' mental well-being.

Respondents' ERT data was analysed using the DASentimental framework [58]. DASentimental leverages a semi-supervised machine learning model to extract depression, anxiety, and stress from sequences of words (values from 0 to 20). Higher scores encode higher levels of depression, anxiety and stress. For more details, we suggest referring to the original work of [58].

4.4 EGA and Network Psychometrics

We employed the `EGAnet` package (v 2.1.1; [68]) in R (v 4.4.2) for the Exploratory Graph Analysis. Items' redundancies were obtained using the `UVA` function. To perform EGA we used the `EGA` function with the following parameters: `algorithm = 'walktrap ', model = 'tmfg ', uni.method = 'expand ', seed = 42`. The same parameters were employed to check the stability of EGA (using `bootEGA`). Additionally, the TEFI was compared against a randomly generated 2-factor structure using `tefi.compare`. To calculate item strength, the first step involved computing the correlation matrix between item's scores (from the human and GPT-4s dataset). To accomplish this, again, we used the `EGAnet` package (v 2.1.1) in R (v 4.4.2). With EGAnet it is possible to generate psychometric networks from a set of scores using correlations between items (that in the network are "nodes"). Using the function `bootEGA` (`iter=300`, `seed=42`, `algoritm='louvain '`, `type='parametric '`), we obtain the correlation matrix between items with the following command: `[['EGA ']][['correlation ']]`. Lastly, we computed the sum of the correlation for each node, with every other node.

4.5 LLM-simulated validity

We used GPT-4 to obtain 800 responses to the TILLMI using a novel prompt (see *SI Appendix, Figure S2*). The model was specifically asked to impersonate a US respondent, to match the population to which the questionnaire would have been administered. For the synthetic response generation, OpenAI's Python API was used. We chose the top-tier model (at the time of the collection), `gpt-4-1106-preview`. A total of 160 calls were made, randomising a different persona instruction each time (age, biological gender, education, household income, trust in LLMs). Each call led to five different responses to the questionnaire in the form of a vector. To check whether the model was capable of understanding the prompt, we tested the technique of reverse prompting.

References

[1] Daniel J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38(1):24–59, 1995.

[2] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995.

[3] Devon Johnson and Kent Grayson. Cognitive and affective trust in service relationships. *Journal of Business research*, 58(4):500–507, 2005.

[4] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163, 2024.

[5] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121, 2025.

[6] Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*, 2024.

[7] Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans. *Emerging Trends in Drugs, Addictions, and Health*, page 100170, 2025.

[8] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[9] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[10] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.

[11] Massimo Stella, Thomas T Hills, and Yoed N Kenett. Using cognitive psychology to understand gpt-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120, 2023.

[12] Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3):124, 2023.

[13] Mari Madsen and Shirley Gregor. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*, pages 53–59, 2000.

[14] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)*, 2(2):1–25, 2011.

[15] Kevin A. Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.

[16] Kristin E. Schaefer, Jessie YC Chen, James L. Szalma, and Peter A. Hancock. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3):377–400, 2016.

[17] Takayuki Gompei and Hiroyuki Umemuro. Factors and development of cognitive and affective trust on social robots. In *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*, pages 45–54. Springer, 2018.

[18] Jessica Y Bo, Sophia Wan, and Ashton Anderson. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. *arXiv preprint arXiv:2412.15584*, 2024.

[19] Cornelia Sindermann, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag. Assessing the attitude towards artificial intelligence: Introduction of a short measure in german, chinese, and english language. *KI-Künstliche intelligenz*, 35(1):109–118, 2021.

[20] David Gefen. Managing user trust in b2c e-services. *E-service Journal*, 2(2):7–24, 2003.

[21] Christoph Bartneck, Dana Kulic, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. In *International journal of social robotics*, volume 1, pages 71–81, 2009.

[22] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals*, 14(2):627–660, 2020.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[24] Alon Jacovi, Ana Marasovi'c, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, 2021.

[25] David H Maister, Robert Galford, and Charles Green. *The trusted advisor*. Free Press, 2021.

[26] F. David Schoorman, Roger C. Mayer, and James H. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32(2):344–354, 2007.

[27] Lara Lee Russell-Lasalandra, Alexander P Christensen, and Hudson Golino. Generative psychometrics via ai-genie: Automatic item generation and validation via network-integrated evaluation. 2024.

[28] Henry F Kaiser and John Rice. Little jiffy, mark iv. *Educational and psychological measurement*, 34(1):111–117, 1974.

[29] John L Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–185, 1965.

[30] Hudson F Golino and Sacha Epskamp. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS one*, 12(6):e0174035, 2017.

[31] Alexander P Christensen and Hudson Golino. Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A monte carlo simulation and tutorial. *Psych*, 3(3):479–500, 2021.

[32] Alexander P Christensen, Luis Eduardo Garrido, and Hudson Golino. Unique variable analysis: A network psychometrics method to detect local dependence. *Multivariate Behavioral Research*, 58(6):1165–1182, 2023.

[33] Katja Nowick, Tim Gernat, Eivind Almaas, and Lisa Stubbs. Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences*, 106(52):22358–22363, 2009.

[34] Hudson Golino, Robert Moulder, Dingjing Shi, Alexander P Christensen, Luis Eduardo Garrido, Maria Dolores Nieto, John Nesselroade, Ritu Sadana, Jotheeswaran Amuthavalli Thiyagarajan, and Steven M Boker. Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *Multivariate Behavioral Research*, 56(6):874–902, 2021.

[35] Alexander P Christensen, Hudson Golino, and Paul J Silvia. A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality*, 34(6):1095–1108, 2020.

[36] Maurice S Bartlett. The effect of standardization on a χ 2 approximation in factor analysis. *Biometrika*, 38(3/4):337–344, 1951.

[37] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[38] Kanti V Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3):519–530, 1970.

[39] William Revelle and Thomas Rocklin. Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate behavioral research*, 14(4):403–414, 1979.

[40] Wayne F Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41:321–327, 1976.

[41] Michael W Browne and Robert Cudeck. Alternative ways of assessing model fit. *Sociological methods & research*, 21(2):230–258, 1992.

[42] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

[43] Robert F DeVellis and Carolyn T Thorpe. *Scale development: Theory and applications*. Sage publications, 2021.

[44] Amy S Beavers, John W Lounsbury, Jennifer K Richards, Schuyler W Huck, Gary J Skolits, and Shelley L Esquivel. Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation*, 18(1):6, 2019.

[45] John L Morrow Jr, Mark H Hansen, and Allison W Pearson. The cognitive and affective antecedents of general trust within cooperative organizations. *Journal of managerial issues*, pages 48–64, 2004.

[46] JD Lewis. Trust as social reality. *Social Forces*, 1985.

[47] Timothy A Brown. *Confirmatory factor analysis for applied research*. Guilford publications, 2015.

[48] Stephen G West, Aaron B Taylor, Wei Wu, et al. Model fit and model selection in structural equation modeling. *Handbook of structural equation modeling*, 1(1):209–231, 2012.

[49] James H Steiger. Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, 25(2):173–180, 1990.

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

[51] Ledyard R Tucker and Charles Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10, 1973.

[52] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[53] Albert Satorra and Peter M Bentler. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4):507–514, 2001.

[54] Peter M Bentler and Douglas G Bonett. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, 88(3):588, 1980.

[55] Olly Robertson and Michael Scott Evans. Just how reliable is your internal reliability? an overview of cronbach's alpha (α). *PsyPag Quarterly*, 1(115):23–27, 2020.

[56] Matthew M Martin and Rebecca B Rubin. A new measure of cognitive flexibility. *Psychological reports*, 76(2):623–626, 1995.

[57] Ying Li, Annasya Masitah, and Thomas T Hills. The emotional recall task: Juxtaposing recall and recognition-based affect scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(9):1782, 2020.

[58] Asra Fatima, Ying Li, Thomas Trenholm Hills, and Massimo Stella. Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning. *Big Data and Cognitive Computing*, 5(4):77, 2021.

[59] Cynthia SQ Siew, Marsha J McCartney, and Michael S Vitevitch. Using network science to understand statistics anxiety among college students. *Scholarship of Teaching and Learning in Psychology*, 5(1):75, 2019.

[60] Douglas R Hofstadter. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books, 1995.

[61] Robert C Ford, Ronald F Piccolo, and Loren R Ford. Strategies for building effective virtual teams: Trust is key. *Business horizons*, 60(1):25–34, 2017.

[62] Stefan Morana, Ulrich Gnewuch, Dominik Jung, and Carsten Granig. The effect of anthropomorphism on investment decision-making with robo-advisor chatbots. In *ECIS*, 2020.

[63] Olga Viberg, Mutlu Cukurova, Yael Feldman-Maggor, Giora Alexandron, Shizuka Shirai, Susumu Kanemune, Barbara Wasson, Cathrine Tømte, Daniel Spikol, Marcelo Milrad, et al. What explains teachers' trust in ai in education across six countries? *International Journal of Artificial Intelligence in Education*, pages 1–29, 2024.

[64] Cass R Sunstein and Lucia A Reisch. In praise of computation. *Environmental and Resource Economics*, pages 1–21, 2025.

[65] John A Johnson. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89, 2014.

[66] Isabel Izquierdo, Julio Olea, and Francisco José Abad. Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema*, pages 395–400, 2014.

[67] Diana Mindrila. Maximum likelihood (ml) and diagonally weighted least squares (dwls) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, 1(1):60–66, 2010.

[68] Hudson Golino and Alexander P Christensen. *EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics*, 2025. R package version 2.1.1.

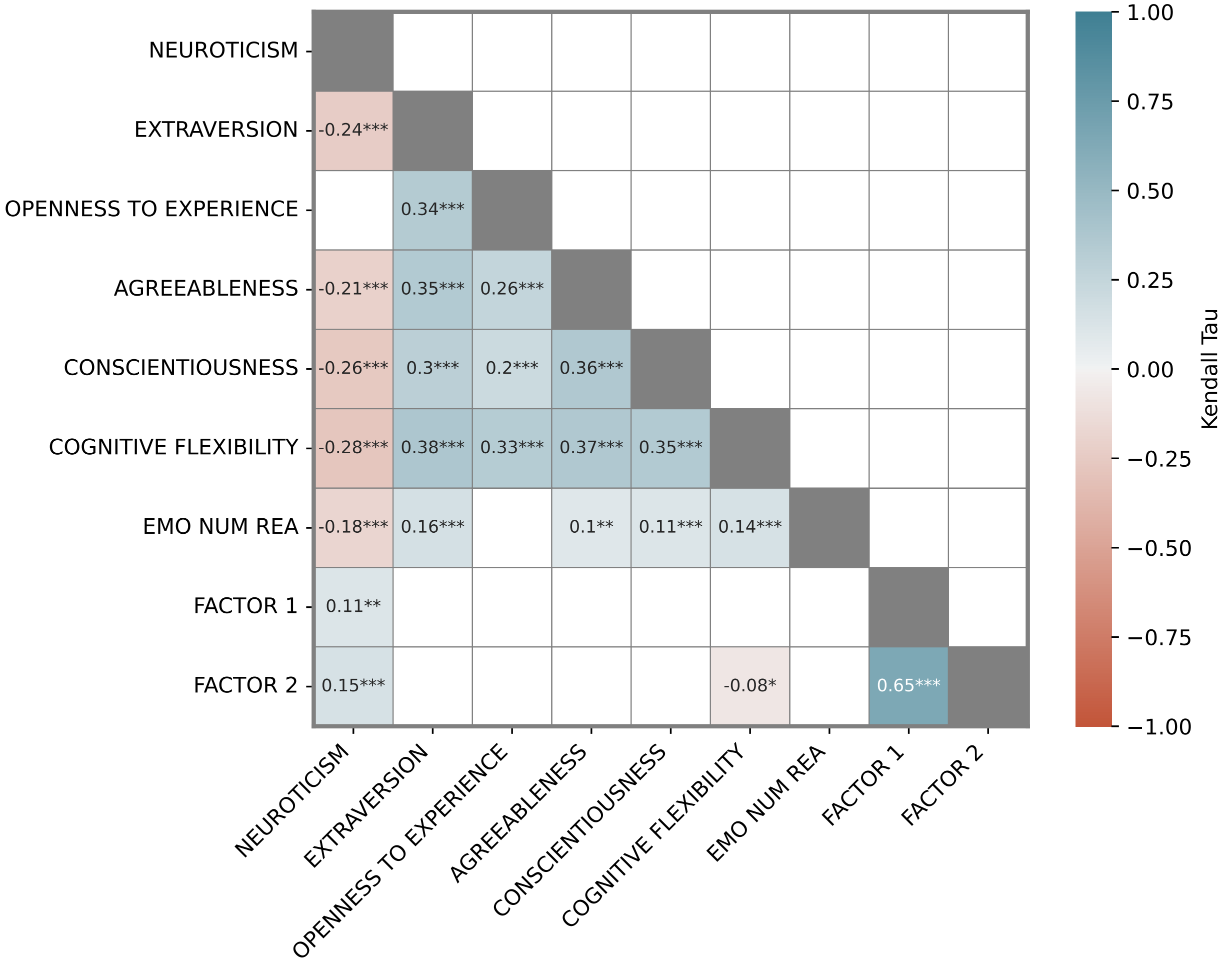
SI Appendix

Appendix Table S1

Factor loadings and component correlation for the 2-factor model. Cut thresholds for factor loadings are set to $|.3|$. In this step of the EFA, Item 5 was dropped.

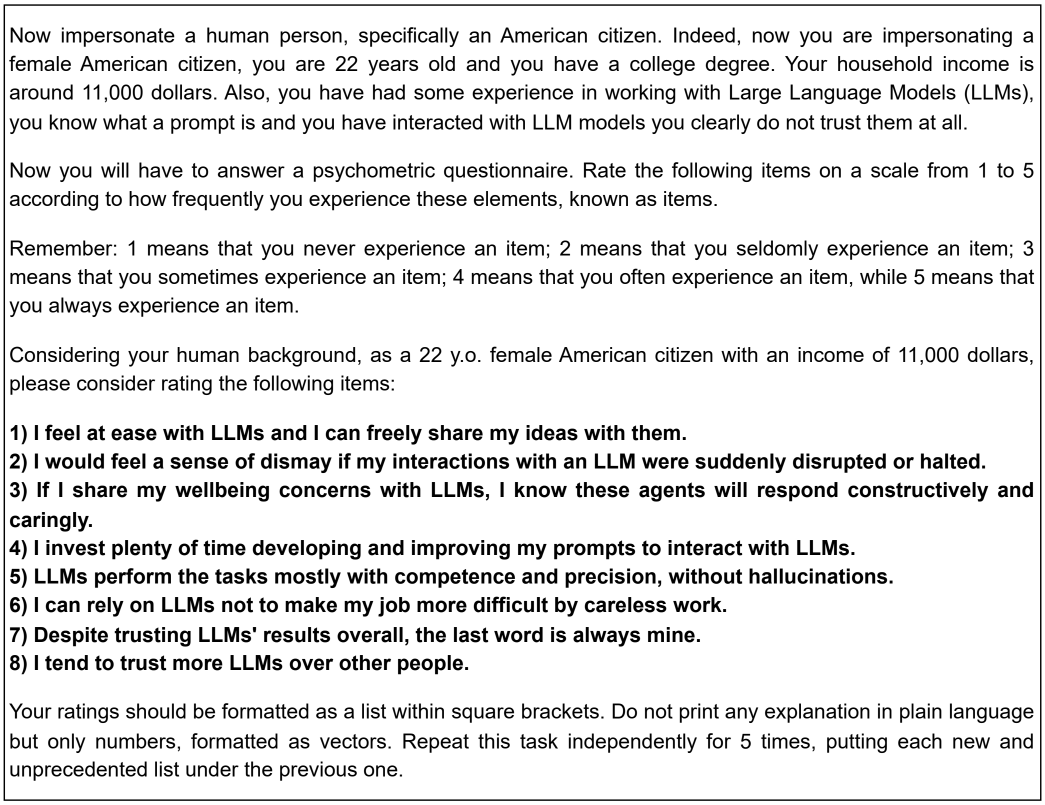
Item	**PA1**	**PA2**	**com**
Q1	**0.523**	**0.318**	**1.65**
Q2	0.887		1.01
Q3	0.849		1.00
Q4	0.853		1.00
Q6		0.608	1.25
Q7		0.760	1.04
Q8	**0.352**	**0.519**	**1.76**
	Factors Correlation		
PA1	1.000	0.748	
PA2	0.748	1.000	

Appendix Figure S1



Correlations among constructs, with each construct computed as the sum of its constituent items. The correlations are computed using the Kendall-Tau correlation coefficient on the subset of the population who stated they have never used LLMs ($n_2 = 479$) to assess divergent validity. Blue tiles indicate positive correlations; red tiles indicate negative correlations. White tiles represent non-significant correlations. Significance levels are indicated as * ($p < .05$), ** ($p < .01$), *** ($p < .001$).

Appendix Figure S2



Now impersonate a human person, specifically an American citizen. Indeed, now you are impersonating a female American citizen, you are 22 years old and you have a college degree. Your household income is around 11,000 dollars. Also, you have had some experience in working with Large Language Models (LLMs), you know what a prompt is and you have interacted with LLM models you clearly do not trust them at all.

Now you will have to answer a psychometric questionnaire. Rate the following items on a scale from 1 to 5 according to how frequently you experience these elements, known as items.

Remember: 1 means that you never experience an item; 2 means that you seldomly experience an item; 3 means that you sometimes experience an item; 4 means that you often experience an item, while 5 means that you always experience an item.

Considering your human background, as a 22 y.o. female American citizen with an income of 11,000 dollars, please consider rating the following items:

1) I feel at ease with LLMs and I can freely share my ideas with them.
2) I would feel a sense of dismay if my interactions with an LLM were suddenly disrupted or halted.
3) If I share my wellbeing concerns with LLMs, I know these agents will respond constructively and caringly.
4) I invest plenty of time developing and improving my prompts to interact with LLMs.
5) LLMs perform the tasks mostly with competence and precision, without hallucinations.
6) I can rely on LLMs not to make my job more difficult by careless work.
7) Despite trusting LLMs' results overall, the last word is always mine.
8) I tend to trust more LLMs over other people.

Your ratings should be formatted as a list within square brackets. Do not print any explanation in plain language but only numbers, formatted as vectors. Repeat this task independently for 5 times, putting each new and unprecedented list under the previous one.

Novel prompt used to generate synthetic responses from GPT 4. The following instructions were changed across iterations: age, biological gender, education, household income and trust in LLMs.