# Do Assessment Instruments Measure the Same Thing for Humans and LLMs? A Latent Structure Analysis

**Anonymous ACL submission**

## Abstract

The rapid development and growing deployment of large language models (LLMs) have made it increasingly important to understand their capabilities. A common approach is to evaluate LLMs using assessment instruments originally designed to measure skills and competencies in humans, such as standardized exams, and to use performance on these instruments as evidence for generalizable claims about LLMs' underlying abilities on the same skills the assessments are intended to measure in humans. However, from a validity perspective, such inferences require that the relationship between observed performance and underlying constructs established for humans also holds for LLMs. In particular, a necessary condition for transferring score interpretations is similarity in the latent structure of responses to the assessment. In this study, we examine whether this condition holds in two educational contexts: high-school chemistry and a quantitative reasoning section of a university entrance exam. Using a case study design, we compare human response data with responses generated by six multimodal LLMs. Our analytical approach combines exploratory factor analysis, factor congruence, and resampling to assess latent structure similarity across human learners and LLMs. Across both instruments, we find systematic differences between human and LLM factor structures, showing evidence that the analyzed assessments may not measure the same constructs for humans and LLMs. These findings call into question the validity of evaluation practices that use educational assessments to make claims about AI capabilities.

## 1 Introduction

The rapid development and growing deployment of large language models (LLMs) have made the systematic evaluation of their capabilities increasingly important (Raiaan et al., 2024; Laskar et al., 2024). Evaluating LLM performance across knowledge domains serves both as a benchmark of their progress and as a means of understanding how their capabilities compare to those of humans (Hendrycks et al., 2020; Zhu et al., 2025). Established assessment instruments, particularly standardized exams, are commonly used as evaluation frameworks (Achiam et al., 2023; Zhong et al., 2024). The rationale is that using these instruments enables making generalizable claims about LLMs' domain knowledge, skills, or competencies that the instruments were designed to measure (Sánchez Salido et al., 2025; Katz et al., 2024; Jimenez et al., 2023; Borges et al., 2024; Yacobson et al., 2026; Chlapanis et al., 2025).

This inference implicitly assumes that the link between test performance and underlying ability, validated in human populations, extends unchanged to AI systems. However, inferences from test performance to underlying skills depend on the validity of the assessment *in context*, and validity established for a given human population does not imply that the same validity argument holds for LLMs (Mitchell, 2026; Wallach et al., 2025).

A key observation about validity is that it is not a property of the instrument, but of the interpretations and uses of scores, that is, the inferences drawn from observed performance on the instrument to the constructs it is intended to measure (Messick, 1994), which should be observed as probabilistic claims (Mislevy et al., 2003; Liu et al., 2024; Xiao et al., 2023). A necessary (although not sufficient) condition for transfer of an assessment instrument between populations for making the same interpretation is a similar underlying latent structure across the two populations, meaning that the instrument measures the same constructs (Meredith, 1993). Otherwise, the instrument simply measures different things. The goal of the present study is to examine whether this condition holds in science education and quantitative reasoning contexts. It seeks to answer the following research question (RQ):

***RQ*:** To what extent do assessment instruments in science education and mathematical reasoning exhibit similar latent structures when administered to humans and large language models? Our primary analytical approach is based on exploratory factor analysis (EFA). EFA is the standard approach to analyze latent structure of assessment instruments, and similarity in factor structure can be interpreted as a preliminary indicator of construct equivalence across populations (Vandenberg and Lance, 2000; Meredith, 1993). Accordingly, substantial differences in the factor structures obtained from human and LLM responses would suggest that the instrument may not support the same construct-level interpretation across these populations. Here, we use it to examine whether similar latent structures emerge in human and LLMs response data on the same assessment instruments.

We adopt a case study design, applying this analytical approach to two assessment instruments drawn from different STEM contexts: (1) a high-school chemistry diagnostic test completed by several hundred Grade 11–12 students, and (2) a standalone quantitative reasoning section from a high-stakes university entrance examination. The human datasets were augmented with responses generated by six multimodal LLMs (OpenAI: GPT-4o, GPT-5.2; Google: Gemini 1.5 Pro, Gemini 3 Pro; Anthropic: Claude 3.5 Sonnet, Claude 4.5).

The analysis proceeded in two steps. First, we applied EFA separately to human and LLM response data for each instrument, estimating the latent structure for each type of examinee and comparing the resulting structures qualitatively. Second, we compared the resulting structures using factor congruence, optimal factor matching, and repeated resampling to obtain a quantitative measure of structural similarity between LLMs and human responses. The results showed evidence that, across datasets and choices of the number of factors, LLM–human factor structures differ.

The contribution of this work is both methodological and empirical: it introduces a validity-oriented framework for comparing the latent structure of assessments across humans and LLMs, and demonstrates, through two case studies, that established assessment instruments may capture substantially different constructs in humans and LLMs.

## 2 Related Work

### 2.1 LLM Evaluation using Educational Assessments

Assessment-based evaluation of LLMs has rapidly become common practice across domains, often by reusing instruments originally developed to measure human knowledge or reasoning (Kasagga et al., 2025; Balunović et al., 2025; Zhong et al., 2024; Du et al., 2025). In medicine, models are evaluated on medical exam-style benchmarks and clinical reasoning datasets (Kasagga et al., 2025; Alaa et al., 2025). In mathematics and science, they are tested on human-targeted problem-solving benchmarks such as AIME 2025 and domain-specific assessments in chemistry and materials science (Balunović et al., 2025; Zaki et al., 2024; Arora et al., 2023; Yacobson et al., 2026; Chen et al., 2025; Guo et al., 2023; Alampara et al., 2024). A similar pattern appears in law and general academic reasoning, where benchmarks such as AGIEval and SuperGPQA draw on standardized exams including the SAT, LSAT, law-related assessments, and graduate-level disciplinary questions (Zhong et al., 2024; Du et al., 2025). This growing practice reflects an implicit assumption that performance on human-designed assessments can be interpreted as evidence of LLM ability in the corresponding domain. This extends across benchmarks that differ in language, task modality, and task domain (such as sequential reasoning, higher-order cognitive tasks), including evaluations specifically constructed to probe current limitations of LLMs (Ahuja et al., 2024; Chang et al., 2024; Balunović et al., 2025; Chlapanis et al., 2025).

While evaluating LLMs via human-targeted assessments has become common practice, recent literature reveals limitations in this approach both generally and across specific domains (Liu et al., 2024; Xiao et al., 2023). For example, in medicine, recent critiques highlight profound gaps in the validity of applying clinical exams to LLMs, noting that high scores on the MedQA benchmark fail to translate to actual clinical decisions relying on the same clinical knowledge (Kasagga et al., 2025). Similarly, in the chemical sciences, established benchmarks like MoleculeNet (Wu et al., 2018) are increasingly criticized for their narrow scope and their inability to provide insights into how models compare relative to human experts, particularly when dealing with specialized molecular structures or equations (Mirza et al., 2025).

These domain-specific failures are often symptoms of deeper technical issues, such as *data contamination* (Kapoor and Narayanan, 2023; Xu et al., 2025; Sainz et al., 2023) or the *brittleness* of model reasoning under minor task perturbations (Ullman, 2023; Inger et al., 2025)

Beyond these domain-specific failures in scientific contexts, broader methodological concerns persist regarding current benchmarking practices. First, the field's heavy reliance on aggregate metrics often obscures how a system will perform in a specific situation, making it difficult to predict real-world performance. This issue is compounded by a lack of transparency; the instance-by-instance evaluation results necessary to "unpack" these aggregate scores are rarely made available for independent audit (Burnell et al., 2023). More fundamentally, there is a growing recognition that while the capabilities of language models are advancing rapidly, the theoretical foundations and norms for their rigorous evaluation lag behind (Balepur et al., 2025), suggesting that established theories of educational measurement may provide a foundation for evaluating the cognitive abilities of artificial 'intelligence' as well (Salaudeen et al., 2025; Mitchell, 2026).Recently, both theoretical interest and the empirical application of these ideas have been rapidly growing in the literature. Aligning with this expanding body of work, we empirically demonstrate evidence highlighting the limitations of making the same inferences about LLMs based on assessments developed for human learners, while also demonstrating the usefulness of established educational measurement workflows for developing methodologically grounded approaches to evaluating LLM capabilities.

### 2.2 Human & LLM Divergences on Educational Assessments

There is increasing evidence that LLMs do not behave like human examinees on educational assessments (Yacobson et al., 2026), and their responses often fail to exhibit consistent psychometrically plausible profiles (Petrov et al., 2024). For example, (Säuberli et al., 2025) found that some models can be made somewhat more human-like through calibration, but the overall correspondence between LLM and human responses remains limited. Similarly, (Liu et al., 2025) shows that no single LLM adequately mimics human respondents, largely because model response distributions are too narrow, even when some psychometric properties can be approximated. In a related multiple-choice setting, (Sorenson and Hanson, 2024) demonstrated systematic differences between human and GenAI response patterns on the same assessments. In STEM education, it was reported that ChatGPT was limited in its ability to solve engineering questions involving figures or diagrams (Borges et al., 2024), and struggled when required to make assumptions about the real world or solve under-specified problems in physics (Wang et al., 2024). Similar observations that LLMs were impacted differently by task dimensions were also made in chemistry and biology (Watts et al., 2023) education contexts. Together, these studies show divergence between human and LLM respondents in educational settings and motivate closer examination of assessment validity.

## 3 Methodology

### 3.1 Overview

Our primary analytical approach is based on exploratory factor analysis (EFA) in accordance with (Cudeck, 2000). For each instrument, we applied EFA separately to human and LLMs responses. The analysis pipeline comprised four stages: (1) preprocessing the data, (2) estimating the number of factors to retain using two factor-retention criteria, (3) fitting EFA models and extracting factor structures separately for humans and LLMs, and (4) comparing the resulting structures using factor congruence, optimal factor matching, and repeated sampling. The final stage enabled both LLMs–human comparisons and a human–human similarity baseline. The full code, sample data and prompts to reproduce the analyses in the paper can be found in the GitHub repository: (anonymous link)

### 3.2 Instruments & Human Data

For the present study, we used response data from two different assessment settings. The first was a high-school chemistry diagnostic test administered through Moodle as preparation for the matriculation exam; it included 22 multiple-choice items and responses from 931 students ($M = 71.49$, $SD = 16.95$, out of 100). The second was the quantitative reasoning section of a national university entrance exam, consisting of 20 multiple-choice items and responses from over 4,800 examinees ($M = 12.45$, $SD = 3.75$, out of 20). Both instruments were multimodal, including text-only items as well as items with figures, images, or

formulas. Responses in both datasets were coded dichotomously, with correct answers marked as 1 and incorrect answers as 0.

In navigating the inherent trade-off between data quality and reproducibility, we opted to use a non-public dataset. While this choice limits reproducibility, it was necessary to ensure data quality across two dimensions: (1) It ensured high-quality human response data from real learners making genuine effort under authentic conditions. (2) It increased confidence that the instrument and its solutions had not been previously exposed, which is a prerequisite for EFA comparison. An additional advantage, beyond the scope of this paper, is that these context-specific materials enable future domain-expert interpretation of the results.

### 3.3 Collection of LLM Responses

To generate LLM response data, we collected answers from six multimodal LLMs spanning three model families: OpenAI (GPT-4o, GPT-5.2), Google (Gemini 1.5 Pro, Gemini 3 Pro), and Anthropic (Claude 3.5 Sonnet, Claude 4.5). Since the assessments included figures, formulas, and other visual content, multimodal input support was required for all models. For each model and instrument, we collected 20 independent response sets, yielding 120 total responses across the six models per instrument (chemistry: M = 76.14, SD = 15.41 & quantitative reasoning: M = 11.94, SD = 4.52). Responses were collected through the models' online user interfaces, which is the practical setting in which these large proprietary models are typically used. For each run, we initiated a new temporary chat session to reduce possible carryover from prior prompts. For the primary analyses, responses were pooled across the six models to form a single LLMs group. Conceptually, pooling aligned with our research question, which concerns a comparison of latent structures for human respondents and LLMs as a class, rather than any individual model. Statistically, pooling increased variation in LLM response patterns, which is necessary for estimating the item pair correlations underlying EFA. After pooling, the SD of LLM response patterns was comparable to that of the humans'. This pooling approach is consistent with recent NLP evaluation work (Macko et al., 2023). Additionally, pooling across LLM models is consistent with how EFA estimates aggregate response structure rather than assuming homogeneous human respondents.

### 3.4 Prompting Technique & Response Scoring

For each instrument, the full instrument was uploaded as a PDF through the LLMs' web interface together with an instruction asking the models to provide only the final answer choice for each item (complete prompt: (anonymous link)). This was done to simulate a realistic student-exam setting, where students are presented with the full exam. We used a minimal, zero-shot prompting approach, similar to (Münker, 2025), aiming to elicit direct responses and avoid specific outcome optimization. LLM behavior can be sensitive to prompt design (Zhuo et al., 2024), and in our case, that would introduce an uncontrolled source of variation to the responses, which would make it harder to attribute response patterns to the assessment itself. Model outputs were binarized against the answer key, and skipped or invalid responses were scored as incorrect, consistent with the treatment of human respondents.

### 3.5 Experimental Flow

#### 3.5.1 Preprocessing

Given the binary item-response matrices, we used tetrachoric correlations, which estimate item associations by treating observed correct/incorrect responses as binary indicators of underlying continuous response tendencies. These correlation matrices served as input to all factor-retention and EFA procedures. As a preprocessing step, we screened items that could make these matrices unstable, particularly in the LLM response data. Specifically, we removed items with zero variance in either group and inspected the $2 \times 2$ contingency tables used to estimate each pairwise tetrachoric correlation, since zero cells or very small cell counts can lead to unstable estimates. Based on these diagnostics, items contributing to highly sparse pairwise tables were removed until the remaining item set reached an acceptable stability threshold. This preprocessing step resulted in no item removals for the quantitative reasoning item set, whereas seven items were removed from the chemistry item set: Items 1, 8-10, 12, 17, and 20. Analyses were then restricted to the common set of remaining items across humans and LLMs, ensuring both groups were analyzed on the same item set.

#### 3.5.2 Factor-retention

Before fitting EFA models, we estimated the number of factors to retain using two common retention

methods: the Kaiser criterion and Parallel Analysis (Lee et al., 2017; Nazaretsky et al., 2022). Both methods were applied separately to the preprocessed human and LLM tetrachoric correlation matrices for each instrument. This allowed us to examine whether humans and LLMs showed systematically similar or different evidence regarding the number of underlying factors. Differences at the factor-retention stage are already informative, as they suggest that the assessment may not reflect the same underlying structure across human and LLM responders. Comparing across factor-retention methods allowed us to test (1) between-group robustness: whether observed LLM–human differences in latent structure were method-dependent or robust, and (2) within-dataset stability: whether the item-loading pattern within each dataset remained stable across the different factor-retention methods.

**Kaiser Criterion** This method retains factors with eigenvalues greater than one (Kaiser, 1974). When factor extraction is based on a correlation matrix, each standardized observed variable contributes one unit of variance. Therefore, an eigenvalue greater than one indicates that the factor explains more variance than a single observed variable. The Kaiser criterion is often used alongside other retention methods, such as parallel analysis.

**Parallel Analysis** We also applied parallel analysis as a factor-retention method (Timmerman and Lorenzo-Seva, 2011). Observed eigenvalues were compared to those obtained from 30 simulated datasets with the same numbers of items and respondents using the `fa.parallel` function from the `psych` package in R (Revelle, 2025). Factors were retained as long as the observed eigenvalues exceeded those expected under random data. Because parallel analysis may yield different factor-retention results across repeated runs, we report the retained number of factors together with its stability, defined as the frequency with which the same number of factors was recovered within each run (see Subsection 4.1).

#### 3.5.3 Factor-extraction

Next, EFA models were fit separately for the human and LLM groups based on the number of factors determined in the factor-retention step, in order to estimate the loading of each item on each factor. We used the default least squares optimization (psych:fa method: `fm = "minres"`) to minimize the difference between the observed and model-reproduced correlation matrices. We also enabled correlation between factor solutions using `rotate = "oblimin"` (Revelle, 2025).

### 3.6 Factor Structure Similarity: LLMs vs. Humans

To quantify the retention of latent factor structure within humans and compare it to LLM-human similarity, we conducted a repeated resampling analysis on the two binary-response datasets – the chemistry and the quantitative reasoning. For each dataset, analyses were run separately with the number of factors fixed to 4, 5, 7 and 8 that were found in 3.5.2. In each iteration, we drew two independent samples of 120 human respondents, denoted $H_1$ and $H_2$, and one independent sample of 120 LLM respondents, denoted $B$. We set the sample size to 120 because this was the number of available LLM response sets, allowing the human and LLM comparisons to be conducted under matched sample-size conditions.

For each resampled subset, we fit EFA models following the procedure in 3.5.1. This yielded loading matrices for $H_1$, $H_2$, and $B$. We first established a human baseline by comparing the factor structures of $H_1$ and $H_2$. We then assessed LLM-human similarity by comparing the factor structure of $B$ to that of $H_2$ from the same iteration.

Let $\mathbf{A}, \mathbf{B} \in R^{p \times k}$ denote two factor-loading matrices defined on the same $p$ items, each with $k$ factors. For factor $r$ in $\mathbf{A}$ and factor $s$ in $\mathbf{B}$, let $\mathbf{a}_r$ and $\mathbf{b}_s$ denote their corresponding loading vectors across items. Factor congruence (Tucker, 1951) was computed as the cosine similarity between these two loading vectors:

$$\phi_{rs} = \frac{\mathbf{a}_r^\top \mathbf{b}_s}{\|\mathbf{a}_r\| \, \|\mathbf{b}_s\|}$$

This yields a congruence matrix $\mathbf{\Phi} \in R^{k \times k}$, where the $(r, s)$ entry is $\phi_{rs}$. Because factor order is arbitrary across EFA solutions, we matched factors using the Hungarian algorithm (Kuhn, 1955), a loss-based linear assignment method that finds the optimal one-to-one correspondence between factors across the two loading matrices. Matching was based on the absolute values of the factor congruence coefficients, so that factors with the same structure but opposite sign were treated as equivalent. We used $1 - |\phi_{rs}|$ as the matching loss, allowing the Hungarian algorithm to maximize absolute congruence through its standard minimization formulation. Each comparison was then summarized by the mean absolute congruence across the matched factor pairs.

This procedure was repeated 100 times for each subset and each factor-number choice, yielding distributions of mean matched congruence scores for the human-human (HH) and LLMs-human (LH) conditions. We compared the HH and LH conditions using one-sided Wilcoxon rank-sum tests across iterations, testing the null hypothesis that HH matched congruence was less than or equal to LH matched congruence against the alternative that it was greater.

## 4 Results

### 4.1 Factor-retention

Table 1 summarizes the factor-retention results for both instruments under the Kaiser criterion and parallel analysis. The Kaiser criterion yielded matching retention results within each instrument – four for the Chemistry, and five for Quantitative reasoning across the two groups. However, parallel analysis yielded different factors between the humans and the LLMs groups in both datasets. In Chemistry, humans consistently retained five factors, whereas LLMs most often retained 4. In Quantitative Reasoning, humans retained 7–8 factors, whereas LLMs consistently retained 5 (same as kaiser retention). With respect to differences in factor retention between groups, the results were mixed and depended on the retention method: the Kaiser criterion yielded the same number of factors for both groups, whereas parallel analysis yielded different numbers.

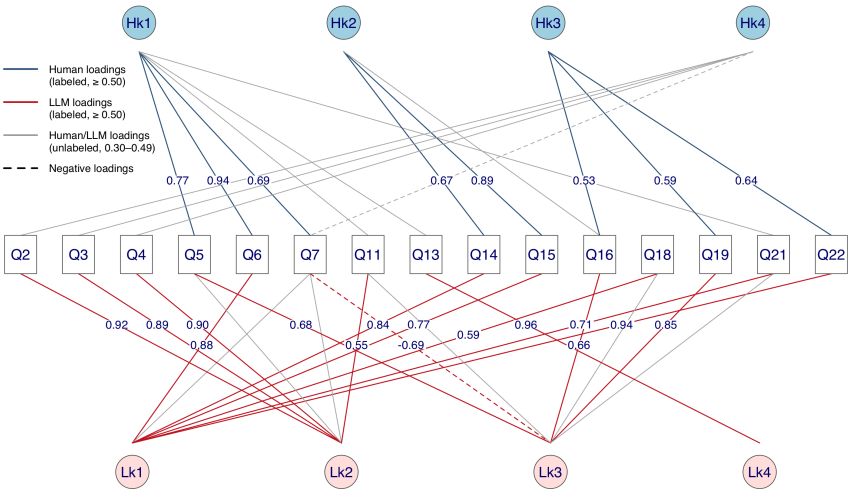
### 4.2 Factor-extraction

Figure 1 shows the EFA factor-loading structures obtained for humans and LLMs for both instruments under the Kaiser-retained factor solution. As seen in the figure, although the retained number of factors was similar for humans and LLMs (4 for Chemistry with human RMSEA = 0.019 & LLM RMSEA = 0.045, 5 for Quantitative Reasoning with human RMSEA = 0.011 & LLM RMSEA = 0.092; for all, confidence  0.90), the loading structure is different. For instance for Chemistry, the third LLM factor, Lk3 (L: LLMs, K: Kaiser, Factor: 3), has loadings larger than 0.5 on Q5, Q16, and Q19. On the human side, these items are loaded on two different factors: Hk1 and Hk2. A similar thing happens with the factor structures of the two groups for the Quantitative Reasoning instrument. One example, among several, is Lk1, which has loadings greater than 0.5 on Q7, Q9, Q17, Q18, and Q20. On the human side, these items are loaded on three different factors: Hk1, Hk2, and Hk3. Analyzing the EFA with parallel analysis-retained factor number also revealed dissimilarities in the item-factor structure and loadings for both instruments (results figure was omitted due to space limit.) To conclude, these factor analyses qualitatively demonstrate differences in the latent structure of the instruments for humans and LLMs. In the next section, we turn to examining these differences quantitatively using factor congruence analysis.

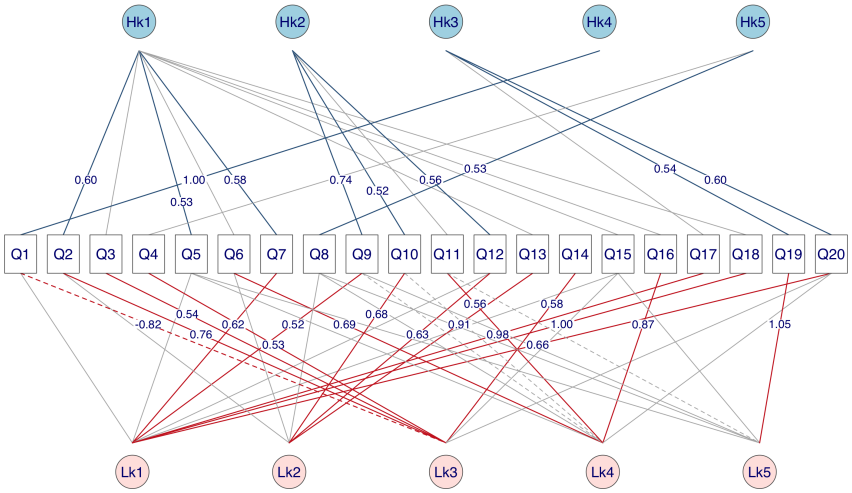
### 4.3 LLM–Human Similarity Analysis

Figure 2 shows the distributions of mean matched factor congruence scores for the HH and LH comparisons across datasets and choices of the number of factors obtained for Kaiser criterion in section 3.5.3. The HH distributions were shifted toward higher values than the LH distributions, indicating greater factor-structure similarity between independently resampled human groups than between LLMs and humans. At the same time, human-human similarity was far from perfect: the distributions were broad rather than concentrated near 1.0, indicating substantial variability across resampling iterations even within the human baseline. For the other conditions, similar results are shown in Table 2. Averaged across all the tested factor solutions (4-8), mean similarity in Quantitative Reasoning was 0.535 for HH comparisons and 0.449 for LH comparisons; in Chemistry, the corresponding means were 0.595 and 0.513. One-sided Wilcoxon rank-sum tests confirmed that this difference was significant ($p < .001$) in all cases.

While we are mindful of commonly used heuristic thresholds for factor congruence (e.g., .95 for near-equivalence, .90-.94 for very high similarity, and .85-.89 for moderate similarity), the present findings are best interpreted relative to the HH baseline, because similarity within humans was itself variable across resamples. Accordingly, the main result is not perfect reproducibility in the human condition, but the consistent gap between HH and LH matching across instrument data and number HH condition is used as an empirical baseline for expected structural agreement under repeated resampling, with the LH condition falling reliably below it.




(a) EFA structure of the Chemistry instrument for human responses (blue) and LLMs responses (red).




(b) EFA structure of the Quantitative Reasoning instrument for human responses (blue) and LLMs responses (red).

Figure 1: **Human vs. LLM EFA Structures Across Instruments.** Factor retention based on Kaiser criterion.

## 5 Discussion

The main finding of this study is that both qualitative visual analysis of the factor graphs and quantitative analysis using factor congruence showed considerable differences between the latent factor structures computed through EFA for both instruments. The quantitative analysis used within-human repeated sampling as the baseline distribution and showed that LLM-human similarity remained considerably lower than human-human similarity, measured through mean matched factor congruence scores. This result was consistent across instruments and choices of the number of factors.

Taken together, these findings answer our RQ by demonstrating that the latent structures of these evaluated case studies – assessment instruments in science education and mathematical reasoning – exhibit limited similarity when administered to human learners and LLMs.

Viewed through an assessment lens, these findings point to a validity-of-inference problem. The issue is not only whether LLMs obtain high scores, but whether those scores support the same construct-level interpretation that the assessment was designed to support for human respondents. In this sense, the present results broaden the validity perspective on LLM benchmarking: benchmark success may indicate that a model can produce correct answers under a given evaluation format, without establishing that the same underlying abilities are being measured. This helps explain why LLMs may perform well on a benchmark yet respond in unexpected or inconsistent ways on related tasks that appear to require the same benchmarked capabilities (Kasagga et al., 2025; Ullman, 2023). More

| Instrument | Kaiser Criterion | | Parallel Analysis | |
|---|---|---|---|---|
| | Humans | LLMs | Humans | LLMs |
| Chemistry | 4 | 4 | **5: 100%** | **4: 62.4%**, 3: 16%, 5: 16%, 2: 5.2%, 7: 0.4% |
| Quantitative Reasoning | 5 | 5 | **8: 66%**, 7: 34% | **5: 100%** |

Table 1: Factor-retention results for the Chemistry and Quantitative Reasoning instruments under the Kaiser criterion and parallel analysis. In the parallel analysis, the percentages refer to the proportion of runs that yielded each result.

| Inst. | $k$ | L-H sim. | H-H sim. | Cohen's $d$ | $W$ |
|---|---|---|---|---|---|
| Quant. | 4 | 0.438 | 0.543 | $1.80^{***}$ | 8987.0 |
| | 5 | 0.432 | 0.533 | $1.93^{***}$ | 9060.5 |
| | 7 | 0.456 | 0.531 | $1.67^{***}$ | 8930.5 |
| | 8 | 0.470 | 0.531 | $1.39^{***}$ | 8370.0 |
| Chem. | 4 | 0.476 | 0.620 | $2.14^{***}$ | 9441.5 |
| | 5 | 0.500 | 0.587 | $1.45^{***}$ | 8499.0 |
| | 7 | 0.530 | 0.582 | $1.08^{***}$ | 7648.5 |
| | 8 | 0.547 | 0.591 | $0.87^{***}$ | 7276.0 |

Table 2: Mean matched factor congruence scores across instruments and factor choices. H denotes human and L denotes LLMs, and Cohen's $d$ denotes the standardized difference between the two distributions. Statistical significance is denoted by $^{***}$, for $p < .001$.

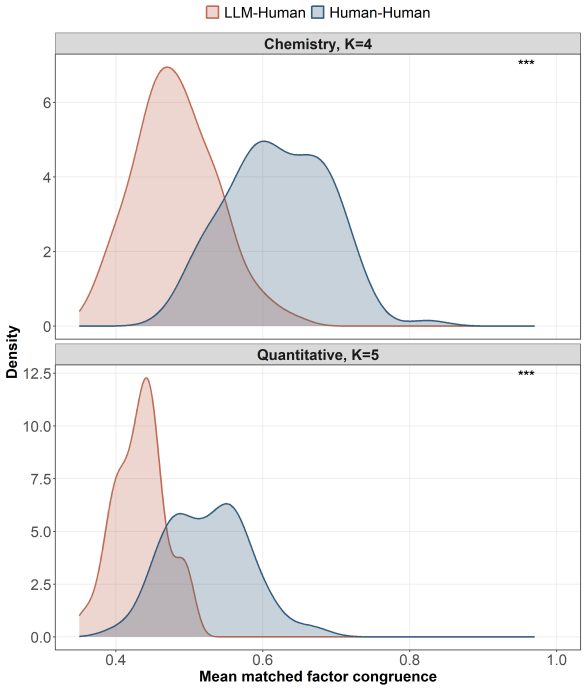



Figure 2: **Distribution of mean matched factor congruence scores**. Computed with Hungarian algorithm, for Human-Human and LLM-Human across datasets and Kaiser factor choice, on 100 resampled repetitions.

generally, our findings suggest that the common practice of evaluating LLMs with human-designed assessments carries a real risk of invalid interpretation if score's meaning is assumed to transfer without re-establishing validity in the new context.

If benchmark scores do not support valid construct-level inferences, then the issue extends beyond the interpretation of current results to the design of benchmarks themselves. It raises the question of how LLM benchmarks should be constructed so that the capabilities they claim to measure are clearly specified and validated on generalization that is not already available in the training data (Inger et al., 2025). From this perspective, benchmark design must address not only task difficulty, but also the constructs being measured and whether those constructs remain interpretable across changing model architectures (Maimon et al., 2025). This also has important implications for educational applications built with LLMs. If students and LLMs do not share the same latent structure on a task, LLMs may be unreliable as tutors, ineffective at simulating students for various purposes, or unsuitable for assessment development.

## 6 Conclusions & Future Work

Assessment instruments show evidence of measuring different constructs for humans and LLMs. Thus, applying assessments designed for human test-takers under the assumptions that (1) these assessments measure the same skills in LLMs, and (2) LLMs' results on such assessments can be used to make inferences about those skills, raises validity concerns and does not adhere to educational measurement standards.

This work advances the emerging field of applying approaches from psychometrics and educational measurement to the evaluation of LLMs. Future research should broaden the scope of assessment instruments, expand the range of analytical methods employed, and incorporate human expert judgment in the interpretation of results.

## Ethics statement

We acknowledge that this work complies with the ACL Code of Ethics. The research and its data collection procedures were approved by the Institutional Review Board and the Ministry of Education.

## Limitations

We acknowledge several limitations in our data analysis that should be considered when interpreting our findings. First, using non-public datasets constrains absolute reproducibility (sec. 3.2). We mitigate this by open-sourcing our codebase and making the complete testing instruments available to researchers upon request. Second, a key limitation regarding external validity is that the results are based on a small number of instruments and specific GenAI tools. Third, systematic biases may arise from variations data generation procedure from LLMs such as prompting strategies, generation parameters, and User interface use. While we standardized these configurations across all LLMs, different model architectures can process identical prompts uniquely, altering response patterns. To support reproducibility and critical assessment, we have publicly released our complete codebase, prompts, and analysis scripts.

In terms of internal validity, the LLMs dataset is relatively small compared to the human sample (120 responses). We also group different LLMs together, assuming based on literature (Maimon et al., 2025; Münker, 2025), that they can be treated as a single population, despite differences in their underlying architectures, which are not publicly disclosed (though they are generally assumed to be autoregressive, decoder-style transformer models).

A broader methodological limitation concerns the use of proprietary LLM systems for evaluation with non-public assessment materials. Although the evaluation instrument was private, uploading it to proprietary tools introduces a potential risk of data leakage or future contamination. Therefore, we cannot fully rule out the possibility that the instrument may become accessible to, or influence, subsequent versions of these systems.

Additionally, this study addresses one aspect of validity, namely internal structure – other aspects, such as generalization to related tasks designed to measure the same underlying abilities, are not examined here.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. 2024. Probing the limitations of multimodal language models for chemistry and materials research. *Preprint*, arXiv:2411.16955. Introduces MaCBench; also published as an AI4Mat @ NeurIPS 2024 workshop paper on OpenReview.

Daman Arora, Himanshu Gaurav Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543. Association for Computational Linguistics.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with llms? a) forced b) flawed c) fixable d) all of the above. *Preprint*, arXiv:2502.14127.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*.

Beatriz Borges, Negar Foroutan, Deniz Bayazit, Anna Sotnikova, Syrielle Montariol, Tanya Nazaretsky, Mohammadreza Banaei, Alireza Sakhaeirad, Philippe Servant, Seyed Parsa Neshaei, and 1 others. 2024. Could chatgpt get an engineering degree? evaluating higher education vulnerability to ai assistants. *Proceedings of the National Academy of Sciences*, 121(49):e2414955121.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, and Jose Hernandez-Orallo. 2023. Rethinking "human-level" performance: A framework for AI evaluation. *arXiv preprint arXiv:2301.05051*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Xiuying Chen, Tairan Wang, Taicheng Guo, Kehan Guo, Juexiao Zhou, Haoyang Li, Mingchen Zhuge, Jürgen Schmidhuber, Xin Gao, and Xiangliang Zhang. 2025. Unveiling the power of language models in chemical research question answering. *Communications Chemistry*, 8(1):4.

Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. GreekBarBench: A challenging benchmark for free-text legal reasoning and citations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119, Suzhou, China. Association for Computational Linguistics.

Robert Cudeck. 2000. 10 - exploratory factor analysis. In Howard E.A. Tinsley and Steven D. Brown, editors, *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, pages 265–296. Academic Press, San Diego.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems*, volume 36.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Nurit Cohen Inger, Yehonatan Elisha, Bracha Shapira, Lior Rokach, and Seffi Cohen. 2025. Forget what you know about LLMs evaluations - LLMs are like a chameleon. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21664–21677, Suzhou, China. Association for Computational Linguistics.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues?, 2024. *URL https://arxiv. org/abs/2310.06770*, 7.

Henry F. Kaiser. 1974. An index of factorial simplicity. *Psychometrika*, 39(1):31–36.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9).

Alousious Kasagga, Aayam Sapkota, Gayan Changaramkumarath, J. M. Abucha, M. M. Wollel, N. Somannagari, M. Y. Husami, Kirubel Tesfaye Hailu, and E. Kasagga. 2025. Performance of chatgpt and large language models on medical licensing exams worldwide: A systematic review and network meta-analysis with meta-regression. *Cureus*, 17(10):e94300.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382(2270).

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, and 1 others. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816.

Sunbok Lee, Zhongzhou Chen, David Pritchard, Alex Kimn, and Andrew Paul. 2017. Factor analysis reveals student thinking using the mechanics reasoning inventory. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*, L@S '17, page 197–200, New York, NY, USA. Association for Computing Machinery.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9960–9987, Singapore. Association for Computational Linguistics.

Aviya Maimon, Amir DN Cohen, Gal Vishne, Shauli Ravfogel, and Reut Tsarfaty. 2025. Iq test for llms: An evaluation framework for uncovering core skills in llms. *Preprint*, arXiv:2507.20208.

William Meredith. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

Samuel Messick. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, 23(2):13–23.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Mohammad Elahi, Hani M. Elbeheiry, María Victoria Gil, Christina Glaubitz, Maximilian Greiner, Caroline T. Holick, Tim Hoffmann, and 16 others. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034.

Robert J Mislevy, Linda S Steinberg, and Russell G Almond. 2003. Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1):3–62.

Melanie Mitchell. 2026. On evaluating cognitive capabilities in machines (and other "alien" intelligences).

Simon Münker. 2025. Fingerprinting LLMs through survey item factor correlation: A case study on humor style questionnaire. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 245–258, Suzhou, China. Association for Computational Linguistics.

Tanya Nazaretsky, Mutlu Cukurova, and Giora Alexandron. 2022. An instrument for measuring teachers' trust in ai-based educational technology. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, LAK22, page 56–66, New York, NY, USA. Association for Computing Machinery.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

William Revelle. 2025. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.5.6.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. 2025. Measurement to meaning: A validity-centered framework for ai evaluation. *Preprint*, arXiv:2505.10573.

Eva Sánchez Salido, Roser Morante, Julio Gonzalo, Guillermo Marco, Jorge Carrillo-de Albornoz, Laura Plaza, Enrique Amigo, Andrés Fernandez García, Alejandro Benito-Santos, Adrián Ghajari Espinosa, and Victor Fresno. 2025. Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6184–6200, Abu Dhabi, UAE. Association for Computational Linguistics.

Benjamin Sorenson and Kenneth Hanson. 2024. Identifying generative artificial intelligence chatbot use on multiple-choice, general chemistry exams using Rasch analysis. *Journal of Chemical Education*, 101(8):3216–3223.

Andreas Säuberli, Diego Frassinelli, and Barbara Plank. 2025. Do LLMs give psychometrically plausible responses in educational assessments? In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 266–278.

Marieke E. Timmerman and Urbano Lorenzo-Seva. 2011. Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16(2):209–220.

Ledyard R. Tucker. 1951. A method for synthesis of factor analysis studies.

Tomer D. Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Robert J. Vandenberg and Charles E. Lance. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1):4–70.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating generative AI systems is a social science measurement challenge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82232–82251. PMLR.

Karen D Wang, Eric Burkholder, Carl Wieman, Shima Salehi, and Nick Haber. 2024. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. In *Frontiers in Education*, volume 8, page 1330486. Frontiers Media SA.

Field M Watts, Amber J Dood, Ginger V Shultz, and Jon-Marc G Rodriguez. 2023. Comparing student and generative artificial intelligence chatbot responses to organic chemistry writing-to-learn

assignments. *Journal of Chemical Education*, 100(10):3806–3817.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Cheng Xu, Nan Yan, Shuhao Guan, Changhong Jin, Yuke Mei, Yibing Guo, and Tahar Kechadi. 2025. DCR: Quantifying data contamination in LLMs evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23002–23020, Suzhou, China. Association for Computational Linguistics.

Elad Yacobson, Yael Schleifer, Ziva Bar-Dov, Shelley Rap, Ron Blonder, and Giora Alexandron. 2026. Benchmarking ai on standard chemistry exams: Llms still underperform compared to high school students. *Journal of Science Education and Technology*.

M. Zaki, Jayadeva, Mausam, and N. M. A. Krishnan. 2024. Mascqa: Investigating materials science knowledge of large language models. *Digital Discovery*, 3(2):313–327.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. TableEval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7146, Suzhou, China. Association for Computational Linguistics.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.