

Interpretable Humans, Alien LLMs: Expert Analysis of Latent Structures in Assessment Responses

Alona Strugatski*, Licol Zeinfeld*, Jason Cooper,
Shelley Rap, Gil Schwartz, Giora Alexandron
Weizmann Institute of Science

*Equal contribution

{alona.faktor, licol.zeinfeld, jason.cooper, shelley.rap, gil.schwartz, giora.alexandron@weizmann.ac.il}

Abstract

The evaluation of large language models (LLMs) relies heavily on human-designed assessments, implicitly assuming that AI and humans employ similar underlying cognitive constructs. Challenging this assumption, we investigate whether the latent factors governing LLM performance carry the same substantive, human-interpretable meaning as the cognitive constructs governing human learners. Using responses from humans and six LLMs across quantitative reasoning and chemistry assessments, we conducted Exploratory Factor Analysis (EFA) separately for both groups. Subject-Matter Experts (SMEs) then blindly evaluated the resulting factor graphs to ascribe pedagogical meaning to the emerged constructs. SMEs successfully interpreted *most* of the human-derived factors. Conversely, they could not ascribe meaning to any LLM-derived factors in quantitative reasoning and interpreted only half of the LLM factors in chemistry. By combining data-driven EFA with blind expert interpretation, this framework shows that LLMs frequently operate on statistically opaque mechanisms distinct from human reasoning.

1 Introduction

As large language models (LLMs) are increasingly applied to a wide range of cognitive tasks, the systematic evaluation of their capabilities and limitations has become increasingly important (Raihan et al., 2024; Laskar et al., 2024). A common approach to LLM evaluation is using established assessment instruments, particularly standardized exams that were developed for measuring human learners (Achiam et al., 2023; Zhong et al., 2024). Moreover, such evaluations typically follow the rationale that the assessment that was used to evaluate human skills in a certain domain can be used to make arguments about LLM skills within that domain (Katz et al., 2024; Sánchez Salido et al., 2025; Jimenez et al., 2023; Borges et al., 2024; Yacobson

et al., 2026; Chlapanis et al., 2025). However, this seems LLMs' to make an implicit assumption that LLMs performance on those tests can be attributed to the same skills that humans apply to solve the test items, and that the items-to-skill mapping is similar for humans and LLMs.

Recent works start to challenge these LLM evaluation practices from educational measurement perspectives, showing, for example, that the structure of factors differs significantly for LLMs and human learners (Strugatski et al., 2026). This connects to content-driven analyses that reported that LLM and humans are impacted differently from task dimensions (Zeinfeld et al., 2026), and IRT-driven analyses showing that LLM response patterns on educational assessments deviate from the norm and appear as aberrant, compared to human learners (Strugatski and Alexandron, 2025; Sorenson and Hanson, 2024). This raises questions on what are the factors that govern LLM performance.

For human respondents, factors are often interpreted in relation to established theories, domain knowledge, and expected response processes (Flora and Flake, 2017; Leighton and Gierl, 2007). For LLMs, however, this issue remains under-examined in the literature and is of interest both for understanding LLM capabilities and for designing assessments for contexts in which respondents may use AI (legitimately or not). Recent work has analyzed LLM response datasets to identify factors that emerge from their response patterns (Maimon et al., 2025). However, the identified factors were not directly compared with the factors that explain human performance on the same items.

The present paper addresses this gap by analyzing the latent factors that explain human learners' and LLM performance on the same instruments. It is guided by the following research question (RQ): **RQ:** How do subject-matter experts interpret the factors derived from human and LLM responses to assessment instruments?

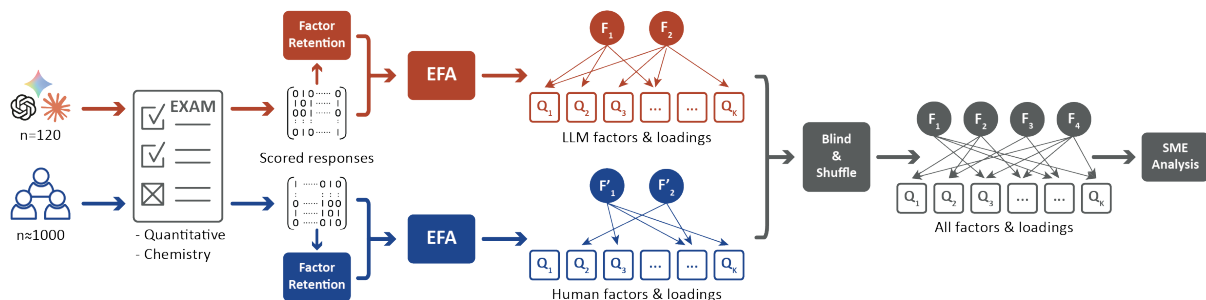


Figure 1: **Methodological Flowchart.** Human and LLM response data for two assessments are processed to extract the underlying factor structure. The blinded and shuffled factor structure results are passed on to subject matter experts for analysis.

To address this question, we used two established assessment instruments in chemistry and quantitative reasoning, with between a few hundred and a few thousands human responders to each. We then ‘administered’ these instruments to six LLM versions, and computed the factor structure for each respondent group (humans and LLMs) using Exploratory factor analysis (EFA). The result of the EFA was a loading graph for the factors computed for humans and LLMs. We then requested Subject-matter experts (SMEs) to interpret the factors identified for each group and attach pedagogical meaning to them. This was conducted blindly, meaning the SMEs did not know if they are interpreting a factor identified for humans or for LLMs. The research flow is illustrated in Figure 1, and the methodology is detailed in Section 3.

Previewing the findings, we found that SMEs were able to ascribe meaning to *all* factors revealed for human learners in both quantitative reasoning and chemistry. In contrast, for LLMs, SMEs were unable to ascribe meaning to *any* of the factors revealed in quantitative reasoning, while in chemistry they were able to meaningfully describe only *two of the four* factors. This raises fundamental questions regarding what explains LLM performance, whether their factors are statistical artifacts or simply ones that cannot be explained by humans, and more.

The key contribution of this study is providing a measurement-oriented perspective on LLM evaluation by moving beyond accuracy-based comparisons to examine the substantive meaning of latent structures derived from assessment responses, and showing that, for the instruments examined, these latent structures not only differed from those of

humans but were also largely uninterpretable to human experts.

2 Related Work

2.1 Evaluating LLMs with Human Assessment Instruments

A common practice in LLM development and evaluation is to reuse instruments originally designed to measure human knowledge or reasoning across domains (e.g., (Kasagga et al., 2025; Balunović et al., 2025; Zhong et al., 2024; Du et al., 2025)). This practice is partly driven by the wide availability of such assessments, many of which use closed-form items that are well suited for automatic evaluation. The use of established assessments also provides a common basis for comparing performance across LLMs and, in some cases, against humans (Kasagga et al., 2025; Yacobson et al., 2026). These comparisons are often treated as evidence of model capabilities in the corresponding domain and are used to expose current capabilities of LLMs. Human-designed assessments are now commonly used in domains such as medical licensing, mathematics, science, legal reasoning, programming, and general academic reasoning, and have been extended across various languages (Balunović et al., 2025; Zaki et al., 2024; Arora et al., 2023; Chen et al., 2025; Guo et al., 2023; Alampara et al., 2024; Zhong et al., 2024; Du et al., 2025).

Despite their practical value, benchmarking LLMs with assessments designed for human learners has several limitations. First, such evaluation may be affected by *contamination*, as publicly available assessments may be included in the LLMs’ training data (Kapoor and Narayanan, 2023; Xu et al., 2025; Sainz et al., 2023). Contamina-

tion may result in models performance being the result of memorization rather than reflecting true ability in the domain. Second, the evaluations typically report aggregate accuracy scores, which often mask more detailed context-specific variations in performance and undermine the prediction of performance in a real-world setting. This issue is exacerbated by the limited availability of item-level evaluation results, which are necessary for detailed independent scrutiny of model performance (Burnell et al., 2023). More fundamentally, equivalent human and LLM assessment performance *does not necessarily imply* that identical underlying constructs are measured, raising a validity concern from educational measurement perspective (Mitchell, 2026; Wallach et al., 2025).

Taken together, these limitations suggest that evaluating LLMs with human-designed assessments requires moving beyond aggregate performance comparisons towards finer analyses of response patterns across humans and LLMs.

2.2 Divergence of Human Learners and LLMs on Educational Assessment

Increasing evidence indicates that LLMs diverge in multiple aspects from human examinees on educational assessments (Yacobson et al., 2026), that their response patterns appear as aberrant compared to human learners (Sorenson and Hanson, 2024; Strugatski and Alexandron, 2025, 2026), and that their response distributions remain narrow, consistently failing to yield psychometrically plausible, human-like trait profiles even after targeted calibration (Petrov et al., 2024; Säuberli et al., 2025; Liu et al., 2025). Additionally, evidence regarding task dimensions shows that specific item features, such as multimodal inputs or under-specified real-world assumptions in STEM contexts, systematically disadvantage models and induce differential item functioning (DIF) between the two populations (Borges et al., 2024; Wang et al., 2024; Watts et al., 2023; Zeinfeld et al., 2026).

2.3 Educational Measurement Perspectives on LLM Evaluation

The core of assessment is construct validity, which is about establishing a connection between performance on a test and inference on a latent skill it is meant to measure (Mislevy et al., 2003; Mesnick, 1994). In educational assessment, the assumption is that the cognitive modeling is known, since developers carefully structure tasks so that a

human's performance provides direct evidence of their knowledge. However, evaluating LLMs using assessments designed for humans undermines this basic assumption, since they compute answers using mechanisms that are entirely different from human cognition. Thus, there is a growing need to design evaluation frameworks specifically built for artificial agents (Liu et al., 2024; Maimon et al., 2025; Balepur et al., 2025). This has led to increasing interest in whether principles from educational and psychological measurement can provide a stronger basis for LLM evaluation (Salaudeen et al., 2025; Mitchell, 2026).

One possible approach based on educational measurement is to use a data-driven latent structure discovery technique, such as exploratory multidimensional IRT (Reckase, 2009), latent class analysis (Collins and Lanza, 2010) that is useful to identify discrete cognitive processing profiles, and network-based approaches (Golino and Epskamp, 2017), which are being adapted to uncover hidden response patterns (Münker, 2025). An additional very common approach is EFA (Vandenberg and Lance, 2000; Meredith, 1993). EFA is based on estimating factors from patterns of item covariation. In human testing, these statistical tools analyze how responses group together to reveal the underlying cognitive skills driving test performance. EFA is becoming increasingly useful in AI evaluation (Maimon et al., 2025; Strugatski et al., 2026). However, it does not explain *what* these factors are, and this poses a significant interpretive challenge (Flora and Flake, 2017). Experts can interpret these constructs by analyzing item content while also drawing on their understanding of the domain, curricular structures, and learning progressions, and by considering principles of cognitive diagnostic assessment (Leighton and Gierl, 2007).

3 Methodology

3.1 Overview

Our methodology combines exploratory factor analysis (EFA) with blind subject-matter expert (SME) interpretation. The goal of our procedure was to identify latent factors underlying human and LLM response patterns and examine how SMEs characterize those factors based on the items that loaded on them. As shown in Figure 1, the analysis pipeline consisted of six steps. (1) **Data collection:** Human and LLM item responses were collected for two instruments, yielding separate response

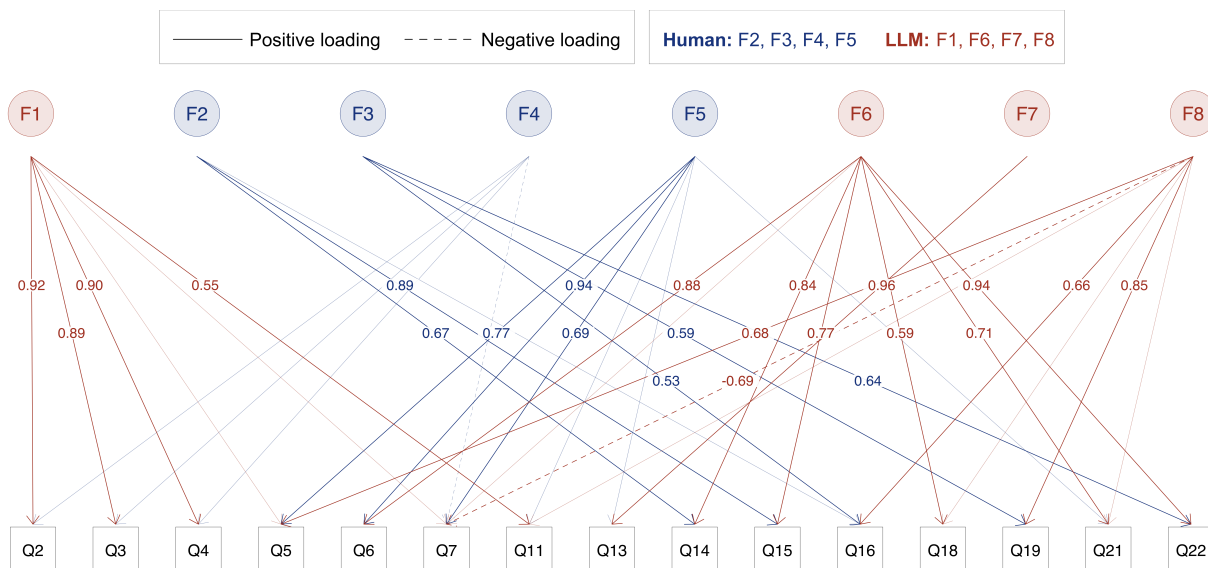


Figure 2: **EFA item-loading patterns for the chemistry assessment.** Human and LLM factors and loadings are shown in blue and red for visual clarity; however, factor structures were originally extracted independently for human and LLM respondents before being combined, shuffled and blinded into a monochromatic version of this figure in preparation for SME evaluation. $|\text{loadings}| \geq 0.5$ are labeled; $0.3 \leq |\text{loadings}| < 0.5$ are unlabeled; $|\text{loadings}| < 0.3$ not shown.

datasets for each group. (2) **Data preparation:** Responses in both datasets were scored as correct (1) or incorrect (0), converted into binary matrices, and preprocessed. (3) **Applying factor retention criteria:** The number of factors for each response group was determined based on the preprocessed matrices. (4) **Fitting EFA:** EFA models were fitted separately for humans and LLMs, producing a factor-loading structure for each group based on the factor-retention result. (5) **Blind & shuffle:** The EFA results from both groups were grouped, blinded and shuffled in preparation for SME analysis. (6) **SME interpretation:** SMEs examined the set of items that loaded on to each factor. The full code, sample data, and prompts to reproduce the analyses in the paper can be found here: ([link](#)).

3.2 Data

Instruments & Human Data In this study, we analyzed responses from two human-designed assessment instruments. (1) A 22-item high-school *chemistry* diagnostic assessment administered through Moodle to students preparing for the matriculation examination. The dataset included responses from 931 students, with an average score of ($M=71.49$, $SD=16.95$, out of 100). (2) A 20-item *quantitative reasoning* section from a national university entrance exam, with access to more than 4,800 examinee responses ($M=12.45$, $SD=3.75$, out of 20);

for the analysis 979 were used. Both assessments included multiple-choice items and contained a mixture of textual and multimodal content, including figures, images, and formulas.

While the data is not publicly available, it provides an advantage of combining high-quality human responses with domain-specific assessments designed and administered in real educational settings. This combination is essential for SMEs, who depend on evaluating real-world tasks with disciplinary content, and representations.

Collection & Pooling of LLM Responses LLM response data were generated using six multimodal models from three major proprietary families: Anthropic’s Claude 3.5 Sonnet and Claude 4.5, OpenAI’s GPT-4o and GPT-5.2, and Google’s Gemini 1.5 Pro and Gemini 3 Pro. Since both assessments contained visual and symbolically rich material, including figures, formulas and disciplinary representations, we restricted data collection to models capable of processing multimodal input. For each model–instrument combination, we obtained 20 independent response sets, resulting in 120 LLM response sets per instrument. Across these pooled responses, LLM performance was ($M=76.14$, $SD=15.41$) on the chemistry instrument and ($M=11.94$, $SD=4.52$) on the quantitative reasoning instrument.

The generated responses were then pooled across

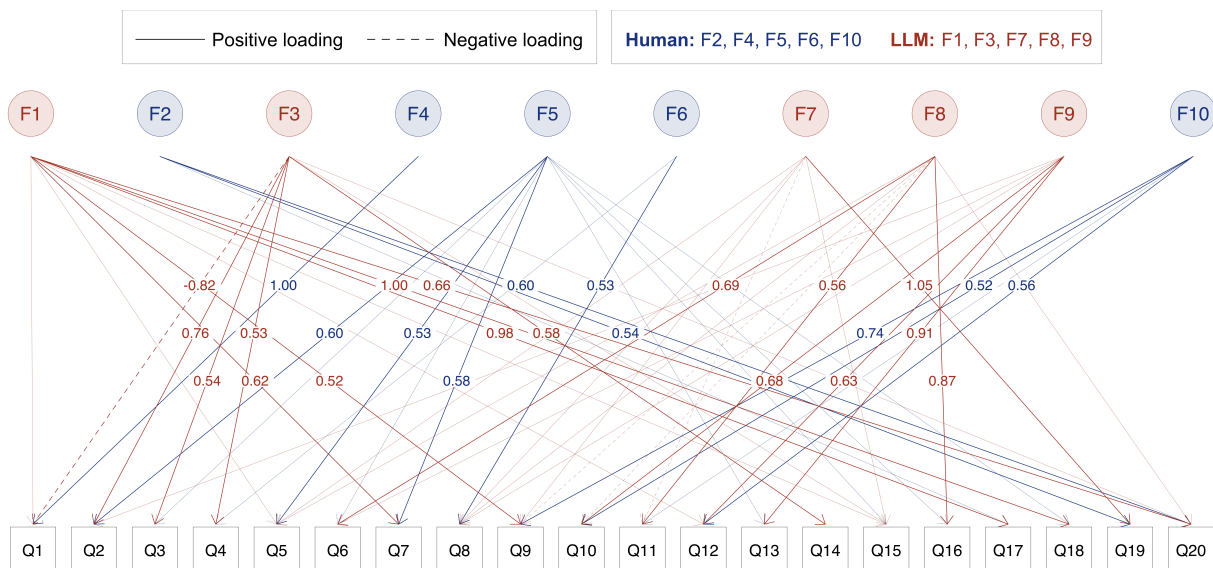


Figure 3: **EFA item-loading patterns for the quantitative reasoning assessment.** Factor structures independently derived for each group. Visualization and blinding procedures for SME evaluation follow the identical monochromatic protocol utilized for chemistry assessment.

the six models into a single LLM respondent group, consistent with recent NLP evaluation work that analyzes performance across multiple models (Macko et al., 2023). This decision follows from the goal of the study to compare the latent response structures that emerge for human respondents and for LLMs broadly, rather than to characterize model-specific factor structures. Additionally, pooling increased variation in LLM response patterns, consistent with the human data variation, allowing for stable EFA fit results. Conceptually, pooling introduces heterogeneity into the data, aligning with EFA’s assumption that human respondents exhibit heterogeneous response patterns.

LLM Prompting Technique To prevent context leakage, we collected responses via the models’ web interfaces using independent, temporary chat sessions. Mirroring authentic human test administration, we provided the complete instrument as a single PDF rather than item-by-item, instructing the models to output only their final selected answers. Following (Münker, 2025), we adopted a minimal, zero-shot prompting strategy to elicit direct LLM responses. Since LLM outputs can vary substantially with prompt engineering (Zhuo et al., 2024), we intentionally avoided prompt engineering that would, in our case, introduce an additional source of difficult-to-control variation.

3.3 Exploratory Factor Analysis

Data preparation

(1) Response Scoring Finally, model responses were scored against the answer key and converted to binary outputs. As with the human response data, omitted, malformed, or otherwise invalid answers were treated as incorrect. **(2) Preprocessing** Since factor-retention and EFA procedures require continuous distributions for item-responses, we used the binary item-response data to estimate these distributions. As a preprocessing step, we screened the item sets and removed items that could produce unstable estimates. We first excluded any items with zero variance in either human or LLM group. Then, we inspected the 2×2 contingency tables underlying each item-pairwise correlation and removed items contributing to empty or highly sparse cells. As a result, no items were excluded from the quantitative reasoning item set. Seven items were excluded from the chemistry item set: 1, 8–10, 12, 17, and 20. All subsequent analyses were conducted on the common retained item set for humans and LLMs.

Factor-retention We specified the dimensionality of the EFA models by applying the Kaiser criterion (eigenvalues > 1) (Lee et al., 2017; Nazaret-sky et al., 2022; Kaiser, 1974) independently to the human and LLM tetrachoric correlation matrices. This retention method yielded an identical number of factors for both populations within each instrument: four for chemistry and five for quantitative reasoning, establishing the structural baseline needed to examine their respective item-loading

patterns.

Factor-extraction Using the retained number of factors, we fit EFA models separately for the human and LLM response groups. Factor extraction was conducted with minimum residual estimation in the psych package, using fa with fm = "minres", which seeks a solution that closely reproduces the observed correlation matrix. Because the extracted factors were not assumed to be independent, we applied an oblique rotation using rotate = "oblimin" (Revelle, 2025). Figures 2 and 3 present the EFA loading structure results for the human and LLM response groups for both instruments using the Kaiser-retained factor solutions and table 1 the quality of fit matrices for human (all RMSEA < 0.02) and LLM (all RMSEA < 0.1) with confidence (> 0.90) for both, representing a good model fit for humans and marginal for LLMs.

While an identical number of factors were used for both groups (table 1), the underlying item-loading structures diverged (?). For example, items Q5, Q16, and Q19 loaded heavily (> 0.5) onto a single LLM factor (F8), yet split across two distinct human factors (F2 and F5). This structural misalignment also permeated quantitative reasoning assessments (see 3). Next, we examine how SMEs interpreted these factor structures.

3.4 SME Analysis

Experts in each of the fields – mathematics and chemistry – were asked to try to explain the factors that emerged from EFA (Sireci, 1998). To avoid bias, the factors were blinded in the sense that they were randomly assigned numbers from 1 to 10, obscuring which reflect human responses and which reflect LLM responses. The instruction was: What do the items that loaded highly on each factor have in common, and in what ways are they different from other items? The experts were instructed to consider three kinds of characteristics of the items: epistemic (e.g. curricular content), cognitive (e.g. order of thinking required to solve) and psychometric (e.g. effectiveness of elimination strategies). Experts ordered the items in each factor and asked themselves what the first two have in common, what the third has in common with them, and so on. If and when common characteristics were found, the items outside the factor were analyzed with respect to these characteristics. The quality of an explaining characteristic was measured in terms of the number of items that load strongly on the factor and hold the characteristic and the number of items

outside the factor that do not.

4 SME Analysis Results

4.1 Quantitative Reasoning Factors

In the quantitative reasoning instrument, whose factor structure is shown in Figure 3, none of the LLM factors could be explained by the experts in a satisfactory manner. Three human factors were explained, as follows: Factor F10 (epistemic): Reading data from a graph. Items Q9-12 all refer to data presented graphically, and they all loaded on this factor, and only on it (Q9, Q10 and Q12 strongly, Q11 weakly). Factor F5 (psychometric): It is possible to check the correctness of each offered answer without solving the problem constructively. This is a strategy that is often advocated in preparing for the psychometric test. In each of the items that loaded strongly on this factor (Q2, Q5, Q7) it is helpful to substitute the offered answers for the unknown and to check for each if it is a correct answer. This is also the case for some of the items that loaded weakly. For example, in item Q3, substituting the smallest answer (0) and recognizing that this is a possible value for $(x - y)^2$ solves the problem. Factor F2 (cognitive/psychometric): It is possible to solve the general problem by working with an imagined example. This is a well known psychometrics strategy, which is sometimes called autonomous concrete instantiation (as opposed to substitution of the given answers) (Koedinger and Nathan, 2004). In item Q20, this involves arbitrarily selecting three prime numbers and working with their product and, in item Q19, imagining the vertices of an arbitrary equilateral triangle and trying to imagine a fourth point for which the condition holds. Item Q17, which loaded weakly on this factor, does not seem to have this characteristic. Item Q13, which did not load on this factor, may seem to have this characteristic (work with arbitrary A, B, and C for which the given holds); however, there the question is which claim holds necessarily, hence an arbitrarily selected example is not generic and will generally not lead to the correct answer.

4.2 Chemistry Factors

To better understand the chemistry instrument's underlying factor structure, shown in Figure 2, an initial attempt was made to identify commonalities based on chemistry-related content. However, the factor structure could not be adequately explained

Table 1: Factor Retention and EFA Fit for Human and LLM Samples

Instrument	Factors Retained		EFA Fit Results			
	Human	LLM	Human		LLM	
			RMSEA	Confidence	RMSEA	Confidence
Chemistry	4	4	0.019	> 0.90	0.045	> 0.90
Quantitative reasoning	5	5	0.011	> 0.90	0.092	> 0.90

on the basis of content alone. Instead, several similarities were identified in the structural characteristics of the items rather than in their substantive content. One exception to this pattern is discussed separately below, as it remains unclear whether the factor emerged due to the nature of the items themselves or due to their shared content. Factors F1 and F4, which primarily included Items Q2, Q3, and Q4, appeared to group together because these items consisted of subsections in which all correct responses were selected from only two possible answer options. This shared response structure may account for the observed factor loading. Nevertheless, it is important to note that additional multiple-choice and sentence-completion items did not load onto this factor, suggesting that response format alone does not fully explain the pattern. Factor F2 consisted of two items that were sequentially related, with one item serving as a continuation of the other. This close structural and contextual relationship likely contributed to their loading onto the same factor. Factors F3 and F8 included items Q16 and Q19, both of which incorporated a true/false response format. The shared evaluative structure of these items may explain their association within the factor analysis. Factor F5 may be interpreted, at least partially, on the basis of content. The items associated with this factor addressed dissolution processes at the microscopic level. At the same time, items Q6 and Q7 also demonstrated substantial visual and procedural similarity, as they required a highly comparable mode of response. In contrast, no clear explanation could be identified for factor F6. Factor F7 was not ascribed a meaning since only one item loaded on it.

5 Discussion

The results show that, on the instruments examined, not only that LLMs use different ‘skills’ to answer items, but that these skill may not be human interpreted. The fact that in most cases the SMEs did not find any logic in the LLMs’ reasoning indicates a misalignment between how SMEs and

LLMs organize domain knowledge. And if LLM factors are largely opaque for SMEs, it becomes unclear what are the constructs that we are measuring for LLMs anyway. This provides a different explanation to why LLMs sometimes perform well on a benchmark but poor on tasks that SMEs see as requiring the same skill set (Kasagga et al., 2025; Ullman, 2023), and extends recent studies arguing that making inference about LLM skills based on assessments designed for evaluating these skills among human learners may be ungrounded (Wallach et al., 2025; Inger et al., 2025; Ullman, 2023).

Another possibility explaining SMEs’ failure to find meaning in the LLM factors is that their factors represent non-human skills. But a different possibility is that the LLMs behavior is merely the result of a statistical process that cannot be meaningfully explained as resulting from latent traits. Delving into this question may require deeper investigation into how LLMs represent domain knowledge and activate it during problem solving.

This leads to a broader question. A widely held view in neuroscience is that the basic organization of the brain is broadly similar across individuals (Kanwisher, 2010; Yeo et al., 2011), and that differences in abilities arise in large part from environmental factors, such as the learning opportunities individuals encounter throughout their lives (Noble et al., 2015). In contrast, LLMs differ substantially in their underlying architectures, in addition to “environmental” differences related to their “learning opportunities”. Given this variability in underlying mechanisms, an interesting question is whether a unified definition and measurement of abilities – independent of the mechanisms that underlie them – is even feasible.

Limitations. We acknowledge several limitations in our data analysis that should be considered when interpreting our findings. First, using non-public datasets constrains reproducibility, but was important for reducing the risk of contamination, which could have bias the results. In terms of internal validity, the LLMs dataset is relatively small com-

pared to the human sample. We also pooled different LLMs together, assuming that they can be treated as a single population, despite differences in their underlying architectures. This is an accepted, yet debatable practice (Maimon et al., 2025; Munker, 2025). The analysis itself was conducted by a small number of SMEs and did not include an inter-rater agreement process. In terms of external validity, the research is based on a small number of instruments, and was produced in a single language.

Implications to educational applications. If students and LLMs do not share the same latent structure on a task, their use in educational applications that rely on a common interpretation of the meaning space may hinder their usability as personalized tutors, for assessment design, and for simulating students for purposes such as creating learning pals.

Contribution & Future Work. This work makes two primary contributions. First, we present a methodological framework to examine the substantive meaning of latent structures derived from assessment responses by combining EFA with blind SME interpretation. Second, we contribute empirical findings highlighting the inherent difficulty human experts face when attempting to ascribe substantive meaning to the latent constructs that emerge from LLMs. In future work, we will up-scale the scope of assessment tools to include more modalities and a larger volume of data, including open-source datasets. We also plan to explore the pooling effect by examining the emerged constructs of each model separately.

Ethics Statement

The research and its data collection procedures were approved by the Institutional Review Board No. 3331-2.

Acknowledgments

This work was supported by the Knell Family Institute for Artificial Intelligence, at the Weizmann Institute of Science, Israel. The authors thank the National Institute for Testing and Evaluation for providing access to psychometric exam data.

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*.

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 Nazaret-sky, 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*.

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. *GreekBar-Bench: 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.

Linda M. Collins and Stephanie T. Lanza. 2010. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley Series in Probability and Statistics. John Wiley & Sons.

- 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*.
- David Flora and Jessica Flake. 2017. [The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation](#). *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 49:78–88.
- H. F. Golino and S. Epskamp. 2017. [Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research](#). *PLoS ONE*, 12(6):e0174035.
- 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.
- 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](https://arxiv.org/abs/2310.06770), 7.
- Henry F. Kaiser. 1974. [An index of factorial simplicity](#). *Psychometrika*, 39(1):31–36.
- Nancy Kanwisher. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170.
- 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).
- Kenneth Koedinger and Mitchell Nathan. 2004. [The real story behind story problems: Effects of representations on quantitative reasoning](#). *Journal of The Learning Sciences - J LEARN SCI*, 13:129–164.
- 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.
- Jacqueline P Leighton and Mark J Gierl, editors. 2007. [Cognitive diagnostic assessment for education: Theory and applications](#). Cambridge University Press.
- 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.
- 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 Munker. 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.
- Kimberly G Noble, Suzanne M Houston, Natalie H Brito, Hauke Bartsch, Eric Kan, Joshua M Kuperman, Natacha Akshoomoff, David G Amaral, Cinnamon S Bloss, Ondrej Libiger, and 1 others. 2015. Family income, parental education and brain structure in children and adolescents. *Nature neuroscience*, 18(5):773–778.
- 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.
- Mark D. Reckase. 2009. *Multidimensional Item Response Theory*. Statistics for Social and Behavioral Sciences. Springer.
- 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 Fernández 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.
- Stephen G. Sireci. 1998. [The construct of content validity](#). *Social Indicators Research*, 45(1):83–117.
- 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.
- Alona Strugatski and Giora Alexandron. 2025. [Applying IRT to distinguish between human and generative AI responses to multiple-choice assessments](#). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK ’25, page 817–823, New York, NY, USA. Association for Computing Machinery.
- Alona Strugatski and Giora Alexandron. 2026. [Applying psychometric methods to distinguish between human and generative AI responses to multiple-choice assessments](#). *Computers and Education: Artificial Intelligence*, 11:100668.
- Alona Strugatski, Licol Zeinfeld, and Giora Alexandron. 2026. [Do assessment instruments measure the same thing for humans and llms? A latent structure analysis](#). *Preprint*, arXiv:2608.15630.
- 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.
- 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.
- 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*.
- BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, and 1 others. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*.
- 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.
- Licol Zeinfeld, Alona Strugatski, Ziva Bar-Dov, Ron Blonder, Shelley Rap, and Giora Alexandron. 2026. [Identifying items on which humans and chatbots diverge using differential item functioning](#). In *Proceedings of the 19th International Conference on Educational Data Mining*, pages 479–486, Seoul, Republic of Korea. International Educational Data Mining Society.
- 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.
- 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.