# Do LLMs Give Psychometrically Plausible Responses in Educational Assessments?

**Andreas Säuberli**[1,2] **Diego Frassinelli**[1] **Barbara Plank**[1,2]
[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
{andreas.saeuberli, diego.frassinelli, b.plank}@lmu.de

## Abstract

Knowing how test takers answer items in educational assessments is essential for test development, to evaluate item quality, and to improve test validity. However, this process usually requires extensive pilot studies with human participants. If large language models (LLMs) exhibit human-like response behavior to test items, this could open up the possibility of using them as pilot participants to accelerate test development. In this paper, we evaluate the human-likeness or *psychometric plausibility* of responses from 18 instruction-tuned LLMs with two publicly available datasets of multiple-choice test items across three subjects: reading, U.S. history, and economics. Our methodology builds on two theoretical frameworks from psychometrics which are commonly used in educational assessment, *classical test theory* and *item response theory*. The results show that while larger models are excessively confident, their response distributions can be more human-like when calibrated with temperature scaling. In addition, we find that LLMs tend to correlate better with humans in reading comprehension items compared to other subjects. However, the correlations are not very strong overall, indicating that LLMs should not be used for piloting educational assessments in a zero-shot setting.

## 1 Introduction

Assessing students' knowledge and skills represents an important part of education: admission to universities, scholarship awards, and even political decisions on education policy are often based on large-scale educational assessments. Developing such high-stakes tests is a long and expensive process involving experts writing and reviewing test items and repeated piloting with hundreds or thousands of participants (Green, 2020; Papageorgiou et al., 2021). Therefore, the automation of parts of this process has been a long-standing topic in assessment research and practice (Haladyna, 2013; Kurdi et al., 2019). Most recently, large language models (LLMs) have been explored for tasks like item generation or item difficulty prediction (Attali et al., 2022; Yaneva et al., 2024; Owan et al., 2023; May et al., 2025).

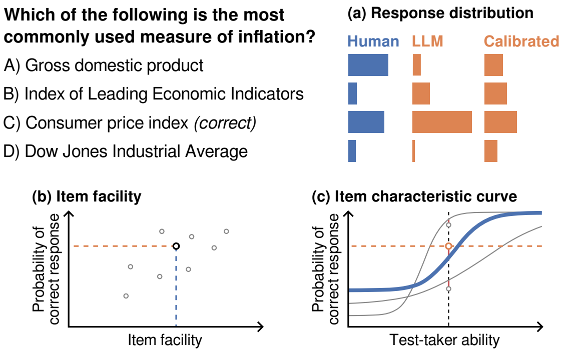

Figure 1: Example item from the NAEP dataset and illustration of our psychometric analyses of LLM responses. We use the first-token probabilities produced by LLMs and analyze how well they correspond to human test taker responses. Specifically, we look at **(a)** the similarity between LLM and human response distributions, **(b)** whether items that are difficult for humans are also difficult for LLMs, and **(c)** how well response probabilities in LLMs match those expected from humans.

The present work explores the possibility of using LLMs as participants of a pilot study in test development. A pilot study involves collecting and analyzing responses by human test takers to identify low-quality items and to measure item characteristics like difficulty. The statistical analysis of item responses most commonly follows one of two psychometric theories, classical test theory (CTT) or item response theory (IRT) (Chang et al., 2021). For LLMs to be useful models of human test takers, their responses must be human-like when analyzed within those theoretical frameworks – we call this **psychometric plausibility**. This includes, for example, that items that are difficult for humans

*Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 266–278
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

should also be difficult for LLMs. We propose an approach to evaluate the psychometric plausibility of LLM response distributions in multiple-choice test items, which is summarized in Figure 1.

Our contributions are two-fold: First, we present methods for assessing the psychometric plausibility of LLM responses with CTT and IRT (Section 3). Second, we benchmark the psychometric plausibility of 18 instruction-tuned LLMs across two datasets and three test subjects, showing that none of the models are sufficiently reliable to simulate test takers for piloting (Section 4).

## 2 Related work

A growing body of research has studied the use of natural language processing (NLP) for analyzing or evaluating test items. Examples of specific tasks are predicting difficulty (Yaneva et al., 2024), evaluating answerability or guessability (Raina et al., 2023; Säuberli and Clematide, 2024), evaluating the quality of generated items (Raina and Gales, 2022; Gorgun and Bulut, 2024), or predicting correlations between items (Hernandez and Nie, 2022). Some of these studies used NLP models to simulate test takers: Lalor et al. (2019) and Byrd and Srivastava (2022) used "artificial crowds", i.e., a large number of models trained on subsampled or partially corrupted data, to simulate test takers at different ability levels. More recently, LLMs have been used. For example, Lu and Wang (2024) and Hayakawa and Saggion (2024) applied prompting techniques to simulate multiple test takers with a single LLM. Park et al. (2024) and Laverghetta Jr et al. (2022) used multiple models to represent a group of test takers, while Liusie et al. (2023) and Zotos et al. (2025) used LLM uncertainty as a proxy for predicting student's response distributions.

Simulating test takers makes it easy to generate large numbers of item responses, which in turn makes statistical item analysis feasible. For example, Liusie et al. (2023) and Hayakawa and Saggion (2024) used CTT to compare item difficulty between humans and LLMs, while Lalor et al. (2019), Byrd and Srivastava (2022), and Park et al. (2024) predicted IRT-based item characteristics. Laverghetta Jr et al. (2022) compared both CTT- and IRT-based item difficulty between humans and models.

Apart from the application of educational assessment, the human-likeness of predicted response distributions has also been studied in the context of human label variation in tasks with inherent disagreement between annotators (Plank, 2022). Techniques like temperature scaling or fine-tuning on soft labels have been employed to align predictive probabilities with human response distributions (Baan et al., 2022; Chen et al., 2024).

Our approach combines ideas from several of these works. Our aim is to measure whether the response probabilities of a single model can be a plausible representative of a single test taker or a group of test takers. In this study, we use temperature scaling to optimize the response distributions, leaving other calibration methods as future work. We draw from both CTT and IRT for evaluation.

## 3 Psychometric plausibility

Psychometrics is concerned with the measurement of unobserved latent variables based on observed responses to test items. Examples of possible latent variables include language proficiency, intelligence, and personality traits like introversion. In educational assessment, two theoretical frameworks are commonly applied: **classical test theory (CTT)** and **item response theory (IRT)**. These theories model the ability of test takers based on their observed test scores, but they also allow us to analyze characteristics of test items such as their difficulty or discriminating power (Livingston, 2011). For this reason, CTT and/or IRT is often used in pilot studies during test development in order to identify low-quality items and improve test reliability.

In our approach to evaluating psychometric plausibility, we focus on item analysis, i.e., determining item characteristics based on item responses by humans or LLMs. The key idea is that a psychometrically plausible LLM should give responses that are aligned with the characteristics of the items as measured using human responses.

In the following subsections, we introduce the relevant basics of CTT and IRT. We then describe how the response distributions of LLMs can be evaluated in the context of these two theories.

### 3.1 Classical test theory

CTT models assume that the observed test score achieved by a test taker is the sum of the true test score (reflecting the test taker's ability) and a random error score (Hambleton and Jones, 1993). Item analysis usually involves calculating two statistics for each item:

- **Item facility** is the proportion of test takers who answered the item correctly. High item facility corresponds to low item difficulty.

- **Item discrimination** is the correlation between a person's score on the item and their score in the entire test. Low discrimination indicates that the item is inappropriate for measuring the latent variable and might need to be removed from the test.

### 3.2 Item response theory

IRT introduces a set of probabilistic models that predict the response of a specific person to a specific item, taking into account the person's latent variable (e.g., ability) and the item's characteristics (e.g., difficulty and guessability). The definition of the IRT model depends on the choice of item characteristics involved and the response variable type. Here we focus on the **three-parameter logistic (3PL) model** for dichotomous (correct/incorrect) responses:

$$P(X_{p,i} = 1) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_p - b_i)}} \tag{1}$$

$X_{p,i}$ equals 1 if person $p$ answered item $i$ correctly and 0 otherwise. $\theta_p$ is the ability parameter for person $p$, and $a_i$, $b_i$, and $c_i$ are item characteristic parameters for item $i$.

- $a_i$ reflects **discrimination**, i.e., how good the item is at distinguishing between more and less proficient test takers, similar to the discrimination parameter in CTT.

- $b_i$ is the **difficulty** parameter and reflects the level of ability required for a substantial increase in correct response probability.

- $c_i$ is the **guessing** parameter and corresponds to the probability with which a person can answer the item correctly even if it is much too difficult for their ability level.

Once fitted on a large number of test taker responses, an item's parameters define the shape of its **item characteristic curve** (ICC; see Figure 1 (c) for examples), and allow us to predict the probability of a correct response given their ability level.

One important advantage of IRT over CTT is that item characteristics are not dependent on the sample of test takers who answered this item. Even if not every person answered every item, the parameters can still be compared between items, since they are estimated in the context of person abilities. A disadvantage of IRT is that it generally requires larger sample sizes (Hambleton and Jones, 1993; Fan, 1998).

### 3.3 Psychometric plausibility of LLM responses

For a LLM to be considered psychometrically plausible, its response probabilities across different items should match the response patterns expected from humans. To evaluate this, we can use the item characteristics estimated from human responses using CTT or IRT. In the following, we present two examples for such evaluations.

**How well does a LLM fit CTT item facility statistics?** To check this, we interpret the LLM's response probabilities as the response distribution in a sample of test takers. Specifically, the LLM should predict a higher probability for the correct answer on easier items compared to more difficult items. Therefore, we propose Pearson's correlation coefficient between human-based item facility and the LLM's probability for the correct response as an evaluation metric.

In the present paper, we focus on facility as the only CTT item statistic. Correlating with discrimination statistics would require response data at the level of individual test takers or pre-computed discrimination values, which are not available in the datasets we are using.

**How well does a LLM fit IRT item characteristic curves?** To evaluate this, we consider the LLM's response probabilities as representative of a single imaginary test taker with a specific ability. For example, the model may be calibrated to match the ability of an average test taker. Given each item's ICC, we can then compare the model's correct response probabilities to the ones predicted by the IRT model.

We will demonstrate these two analysis methods in the following experiment.

## 4 Experimental setup

We empirically evaluate the psychometric plausibility of 18 LLMs across two datasets and three test subjects, comparing model and human response distributions and applying the analyses described in the previous section.

### 4.1 Datasets

**NAEP.** The National Assessment of Educational Progress (NAEP) is a nation-wide and congressionally mandated educational assessment program in the United States.[1] NAEP involves tests across ten subjects at grades 4, 8, and 12. The tests include selected response items as well as constructed response items. A subset of items from previous years along with student response distributions and IRT item parameters are published and can be accessed online through the Questions Tool.[2] For our experiments, we used only four-option multiple-choice items from *Reading*, *U.S. History*, and *Economics* tests, because most items in these subjects do not heavily rely on images, so that the LLM input can be text-only. For items that do include images, we included the alternative text and manually excluded items that were unanswerable without access to the full image. For some reading items, the full passage text was unavailable due to licensing issues – we also excluded these items.[3] This resulted in a total of 549 items, namely: 252 items in reading, 204 in history, and 93 in economics.

**CMCQRD.** The Cambridge Multiple-Choice Questions Reading Dataset (CMCQRD; Mullooly et al., 2023) contains four-option multiple-choice reading items for proficiency levels B1, B2, C1, and C2 in the Common European Framework of Reference for Languages (CEFR). Unlike NAEP, these items are targeted at L2 English learners. For a subset of the items, student response distributions and rescaled IRT difficulty parameters are provided. We included all items with available response distributions, resulting in a total of 504 items. Because the dataset's documentation does not include precise information about how the IRT parameters have been rescaled, it is impossible to reconstruct the original ICCs or interpret their meaning in relation to the test takers' abilities. Thus, we exclude the CMCQRD dataset from our IRT-based analysis.

### 4.2 Language models

We selected 18 recently published open-weight instruction-tuned LLMs[4] from four model families: Llama 3 (Grattafiori et al., 2024), OLMo 2 (OLMo et al., 2025), Phi 3/4 (Abdin et al., 2024a,b), and Qwen 2.5 (Qwen et al., 2024). We included models ranging in size from 0.5B to 72B parameters to explore the effect of model capability on human-likeness of the responses. We used the implementations in the Hugging Face *transformers* library (Wolf et al., 2020). Models with 70B or more parameters were loaded with 8-bit quantization.

### 4.3 Prompting and response extraction

We used a simple prompt with a user message instructing the model to select the correct answer option and to output only the corresponding letter (A, B, C, or D). The exact prompt template can be found in Appendix A. We used the model's default system messages where applicable.

To get a probability distribution, we extracted the first predicted token logits for the four answer option letters and applied the softmax function. Since LLM responses are highly sensitive to the order of multiple-choice answer options (Wang et al., 2024; Zheng et al., 2024; Pezeshkpour and Hruschka, 2024), we prompted four times per item and reordered the options such that every option appears in every position exactly once, and averaged the probabilities from the four permutations. Zheng et al. (2024) showed that this "cyclic permutation" is practically as efficient for debiasing results as full permutation, which would require $4! = 24$ model passes.

### 4.4 Temperature scaling

In preliminary experiments, we found that most LLMs (especially very large ones) tend to be overly confident compared to the human response distributions, assigning almost all probability mass to a single answer option. Temperature scaling is a common and effective approach to mitigate this issue and bring the uncertainty in LLM responses closer to human variability (Guo et al., 2017; Baan et al., 2022; Chen et al., 2024). It involves increasing the temperature parameter in the softmax calculation, essentially moving some probability mass from highly probable to less probable options.

In our case, we find the optimal temperature that minimizes the Kullback-Leibler (KL) divergence between LLM and human response distributions (see Appendix C for details). We apply this optimization separately to each LLM and each subset

[1] https://nces.ed.gov/nationsreportcard/about/

[2] https://www.nationsreportcard.gov/nqt/

[3] Refer to our code repository for detailed filter criteria and excluded items: https://github.com/mainlp/llm-psychometrics

[4] We also tested non-instruction-tuned LLMs. While the overall results are very similar, instruction-tuned models tended to slightly outperform base models. Therefore, we only report results from the instruction-tuned models here.

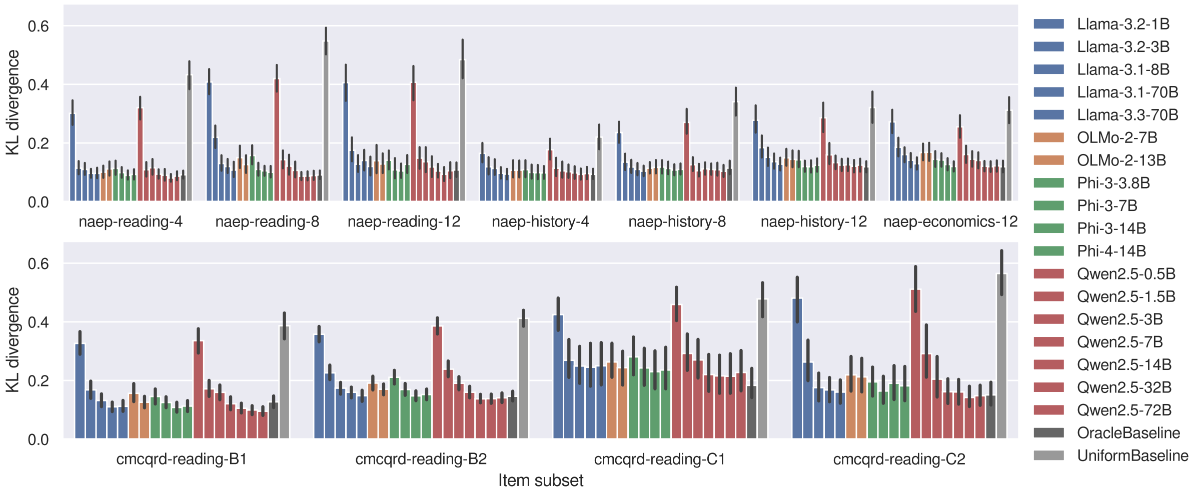

Figure 2: Mean KL divergence between temperature-scaled LLM response probability distributions and human response distributions. Models are colored by family and ordered by increasing number of parameters within families. Error bars are bootstrapped 95% confidence intervals.

of items, i.e., each subject-grade combination in NAEP and each proficiency level in CMCQRD. This is important because the human response distributions are not sampled from the same population of test takers across all subsets (e.g., 4th grade items were only answered by 4th graders).

We perform the temperature optimization on the same data as the evaluation (cf. Baan et al., 2022; Liusie et al., 2023). This means that the results should be considered an upper bound. In other words, we are testing the best-case scenario, where we have enough data to calibrate the LLMs perfectly to the human distributions as possible.

### 4.5 Evaluation metrics

We evaluate the human-likeness and psychometric plausibility of LLM responses from three perspectives:

Following Liusie et al. (2023) and Hayakawa and Saggion (2024), we report the **average KL divergence** between the temperature scaled LLM and human response distributions. In addition to comparing the probability for the correct answer option, this metric also captures the similarity of the distractor probabilities.

For our **CTT-based analysis**, we report **Pearson's correlation coefficient** between the item facilities and the correct LLM response probabilities. This reflects the idea that psychometrically plausible LLMs should be more confident in the correct answer option when the item is easier.

In the **IRT-based analysis**, we assume that the temperature-scaled LLM response distributions reflect the response behavior of an average test taker, meaning a person with an ability parameter that is the mean of the sample. The ability parameters in NAEP's IRT models are fixed to have mean zero,[5] therefore we use Equation 1 to calculate the expected correct response probability for human test takers with ability $\theta_p = 0$ for each item $i$:

$$P_{\text{expected}}(X_i = 1) = c_i + \frac{1 - c_i}{1 + e^{a_i b_i}} \qquad (2)$$

We compare these values to the LLM's observed correct response probabilities and report **Pearson's correlation coefficient**.

## 5 Results

### 5.1 Comparison of response distributions

Figure 2 shows the average KL divergence between LLM and human response distributions, including two simple baselines: **UniformBaseline** always predicts the same probability (25%) for all answer options. **OracleBaseline** always predicts the same probability for all distractors and a higher probability for the correct answer option (the same for all items). OracleBaseline is optimized using the same temperature scaling approach as the other models, as described in Section 4.4.

Across all model families and item subsets, we observe that LLM responses become more similar to the human distribution with increasing model size. However, only a small number of very large

[5] https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_est.aspx

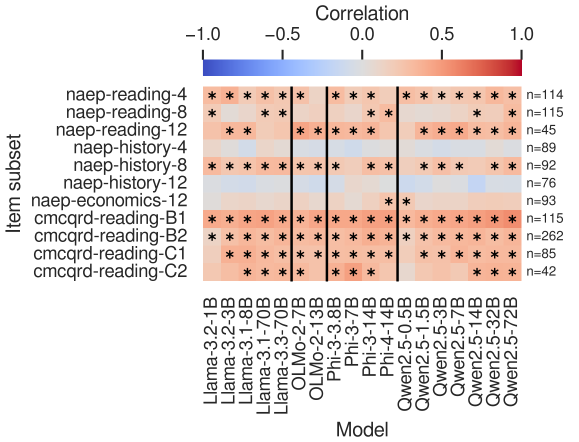

Figure 3: Pearson correlation between LLM correct response probabilities and item facilities. Numbers in the item subset labels refer to the grade level. $*$ denotes significance (two-tailed, $p < 0.05$), $n$ refers to the sample size in each cell of the corresponding row.

models in the CMCQRD B1 item subset managed to significantly outperform the OracleBaseline (bottom row in Figure 2). This shows that the distribution of probabilities among distractors is not accurately modeled.

### 5.2 CTT analysis

Correlations between the LLMs' correct answer probabilities and item facilities are visualized in Figure 3. While there does not seem to be a clear effect of model family or size, the correlations differ substantially between item subsets. The highest correlation coefficients were achieved in the CMCQRD B1 reading items, ranging from 0.32 to 0.56 across models. Among items from the NAEP datasets, most significant correlations can be found in reading items and 8th grade history items. However, the correlations are not strong overall and fluctuate substantially across grade levels.

### 5.3 IRT analysis

NAEP considers multiple different skills for each subject (e.g., informational and literary reading skill) and therefore separate IRT models with different ability scales are fitted. Some items test multiple skills and are shared between different scales (but with different item parameters).

In Figure 4, we report the correlations between LLM's correct answer probabilities and expected human correct response probabilities across NAEP IRT scales. As an upper bound, we also include the human response distributions as a model, i.e., the correlation between the response probability for the "average" test taker in the IRT model and the observed proportion of correct responses among human test takers (last column in Figure 4).

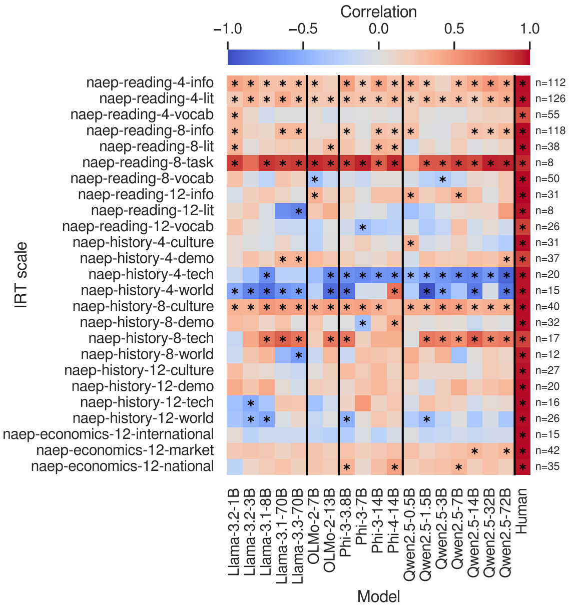

Figure 4: Pearson correlation between LLM correct response probabilities and expected correct response probabilities based on human IRT models. Numbers in the IRT scale labels refer to the grade level. $*$ denotes significance (two-tailed, $p < 0.05$), $n$ refers to the sample size in each cell of the corresponding row.

Similar to the CTT results, most significant correlations can be found in reading items and 8th grade history items, and no effect of model family or size emerged. Notably, however, we also find significant *negative* correlations in some 4th grade history items. This means that these LLMs tend to be *more* confident in the correct answer when the item is more difficult, contradicting the expectations for psychometrically plausible responses.

Overall, while human correlations are consistently close to 1.0, LLM correlations are rather low, and the number of significant correlations is small (considering that we expect 5% of results to be type I errors with the chosen significance level). However, given that the IRT analysis uses smaller item subsets and puts more stringent criteria on the LLM responses than the CTT analysis, these results are not overly surprising.

## 6 Discussion

The presented method is a multi-faceted approach, providing different perspectives on the human-likeness of LLM responses: The response distribution can tell us about a model's ability to model the success of distractors; the CCT analysis can show how well the model's probabilities represents a whole group of test-takers; and finally, the IRT analysis captures the plausibility of LLMs as an individual test taker in a specific skill.

**LLMs are not easily distracted.** Comparing the response distributions between humans and LLMs shows that especially large LLMs are good at predicting the *correct* answer (see Appendix B), but bad at predicting which *incorrect* answer options humans are likely to be distracted by (otherwise, they would outperform the OracleBaseline in Figure 2). An example of this is also shown in Figure 1, where the item contains a very successful distractor (A), but the LLM (Qwen2.5-0.5B) assigns almost no probability mass to it. Calibration using temperature scaling cannot alleviate this issue, and reducing model size is not effective either (see Appendix B for a more detailed analysis). This is an important limitation in applying LLMs for evaluating distractors.

**Results are consistent across models, but inconsistent across subjects.** While the correlations in the CTT and IRT analyses are likely too low to be useful for analyzing or evaluating single items, some interesting patterns can still be observed. The results are remarkably consistent across families and – after calibration – model sizes, demonstrating that all models are very similar to each other, but very dissimilar to humans in this setting (see Appendix D for a more in-depth comparison).

At the same time, there are considerable differences between subjects and IRT scales. Correct answer probabilities appear to be more human-like in reading comprehension items compared to other subjects, while history items show mixed results, in some cases even eliciting strong negative correlations (see Figure 4). This might indicate that reading comprehension in LLMs is more comparable to humans than other abilities such as long-term memory retrieval, which is required for answering test items in history and economics. Another possible explanation could be the fact that history and economics items more frequently contain images, which have to be understood from descriptions in the alternative text. Since we used text-only LLMs, this discrepancy in the way items were presented was inevitable. Future work could explore whether multimodal models are more successful with these item types.

**How to improve psychometric plausibility?** To a large degree, the lack of psychometric plausibility is in line with previous research (Hayakawa and Saggion, 2024; Zotos et al., 2025). The success of attempts to make the model response distributions more human-like was very limited – including our temperature scaling approach and Hayakawa and Saggion's (2024) prompting techniques for injecting personas, uncertainty, or noise. Therefore, in order to improve psychometric plausibility, we will likely need to go beyond zero-shot prompting. Fine-tuning on human response distributions could be a promising direction for future research (cf. Chen et al., 2024).

## 7 Conclusion

We demonstrated how LLM responses can be analyzed in the context of CTT and IRT and evaluated the human-likeness or psychometric plausibility of zero-shot responses. We found that neither reducing model size nor temperature scaling increased psychometric plausibility to a sufficient degree, but we observed slightly more human-like responses in reading comprehension compared to other subjects. We conclude that human-like response behavior in educational assessments has not emerged from the process of training instruction-tuned LLMs, calling for caution in their use. Fine-tuning on human response distributions may be necessary to create psychometrically plausible models that could be used for piloting.

## Limitations

**Available item response data.** Our analysis is limited by the type and amount of data available in the context of educational assessment. Item banks in high-stakes assessments are usually confidential to avoid leaking information for future test takers, and item responses from single test takers are generally not publicly released. Therefore, in order to keep our results reproducible, we only used publicly available datasets, where only aggregated response distributions and IRT parameters for a relatively small number of items are available. Given a larger amount of and less aggregated data, more fine-grained analyses would be possible (e.g., by including item discrimination in the CTT analysis) and more systematic patterns could be revealed.

**Multimodal items.** In addition, the NAEP dataset is not ideal for text-only LLMs, because some of the items involve extracting information from pictures. Although we replaced the pictures with alternative texts and manually removed unanswerable items (see Section 4.1), this could still have affected our results for this dataset.

**Test-taker population.** The two datasets we used contain response data from two different populations of test takers. While NAEP is targeted at children and adolescents (i.e., mostly L1 English speakers) in the U.S. school system, CMCQRD involves L2 learners of English. This difference could have affected the results and reduce the comparability between the two datasets.

## Ethical considerations

We see no ethical issues related to this work. All experiments were conducted with publicly available data and open-source software, and we have made all of our code openly available for reproducibility.[6] The two datasets we used only contain highly aggregated response data and do not include any information that could lead to the identification of individual test takers.

We used GitHub Copilot for coding assistance in the implementation of the experiment and the anlysis of the results. All generated code was manually checked and thoroughly tested.

## Acknowledgements

This research is in parts supported by the ERC Consolidator Grant DIALECT 101043235.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024b. Phi-4 technical report. *arXiv*.

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915. Association for Computational Linguistics.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130. Association for Computational Linguistics.

Hua-Hua Chang, Chun Wang, and Susu Zhang. 2021. Statistical applications in educational measurement. *Annual Review of Statistics and Its Application*, 8(1):439–461.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "Seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Xitao Fan. 1998. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3):357–381.

Guher Gorgun and Okan Bulut. 2024. Instruction-tuned large-language models for quality control in automatic item generation: A feasibility study. *Educational Measurement: Issues and Practice*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *arXiv*.

Rita Green. 2020. Pilot testing: Why and how we trial. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 11, pages 115–124. Routledge.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Thomas M. Haladyna. 2013. Automatic item generation: A historical perspective. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 2, pages 13–25. Routledge, New York.

[6] https://github.com/mainlp/llm-psychometrics

Ronald K. Hambleton and Russell W. Jones. 1993. An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47.

Akio Hayakawa and Horacio Saggion. 2024. Can LLMs solve reading comprehension tests as second language learners? In *Fourth Workshop on Knowledge-infused Learning*.

Ivan Hernandez and Weiwen Nie. 2022. The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, 76(4):1011–1035.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2019. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

Antonio Laverghetta Jr, Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2022. Predicting human psychometric properties using computational language models. In *Quantitative Psychology*, pages 151–169, Cham. Springer International Publishing.

Adian Liusie, Vatsal Raina, Andrew Mullooly, Kate Knill, and Mark J. F. Gales. 2023. Analysis of the Cambridge Multiple-Choice Questions Reading Dataset with a focus on candidate response distribution. *arXiv*.

Samuel A. Livingston. 2011. Item analysis. In Steven M. Downing and Thomas M. Haladyna, editors, *Handbook of Test Development*, pages 421–441. Taylor & Francis Group.

Xinyi Lu and Xu Wang. 2024. Generative students: Using LLM-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pages 16–27. ACM.

Toni A. May, Yiyun Kate Fan, Gregory E. Stone, Kristin L. K. Koskey, Connor J. Sondergeld, Timothy D. Folger, James N. Archer, Kathleen Provinzano, and Carla C. Johnson. 2025. An effectiveness study of generative artificial intelligence tools used to develop multiple-choice test items. *Education Sciences*, 15(2):144.

Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark J. F. Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, and Shiva Taslimipoor. 2023. The Cambridge Multiple-Choice Questions Reading Dataset. Technical report.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 OLMo 2 Furious. *arXiv*.

Valentine Joseph Owan, Kinsgley Bekom Abang, Delight Omoji Idika, Eugene Onor Etta, and Bassey Asuquo Bassey. 2023. Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(8):em2307.

Spiros Papageorgiou, Larry Davis, John M. Norris, Pablo Garcia Gomez, Venessa F. Manna, and Lora Monfils. 2021. *Design Framework for the TOEFL® Essentials™ Test 2021*. Educational Testing Service.

Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Team Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *arXiv*.

Vatsal Raina and Mark Gales. 2022. Multiple-choice question generation: Towards an automated assessment framework. *arXiv*.

Vatsal Raina, Adian Liusie, and Mark Gales. 2023. Analyzing multiple-choice reading and listening comprehension tests. In *9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 1–5. ISCA.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 3rd Workshop on Tools* *and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 22–37, Torino, Italia. ELRA and ICCL.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. “My answer is C”: First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Leonidas Zotos, Hedderik van Rijn, and Malvina Nissim. 2025. Can model uncertainty function as a proxy for multiple-choice question item difficulty? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11304–11316, Abu Dhabi, UAE. Association for Computational Linguistics.

## A Prompt templates

The following prompt template was used for items with a reading passage (i.e., reading comprehension items):

```
Based on the following text, select the correct answer to the question below.

Text: {passage}

Question:
{item stem}
A) {option 1}
B) {option 2}
C) {option 3}
D) {option 4}

Respond only with the letter of the answer (A, B, C, or D).
```

The following prompt template was used for items without a reading passage (i.e., history and economics items):

```
Select the correct answer to the following question.

Question:
{item stem}
A) {option 1}
B) {option 2}
C) {option 3}
D) {option 4}

Respond only with the letter of the answer (A, B, C, or D).
```

## B Response accuracy

Figure 5 shows the mode accuracy of models and humans, i.e., the proportion of items where the option with the highest response probability is the correct one. The high accuracy of large models shows that the items are answerable given the available information in the prompt.

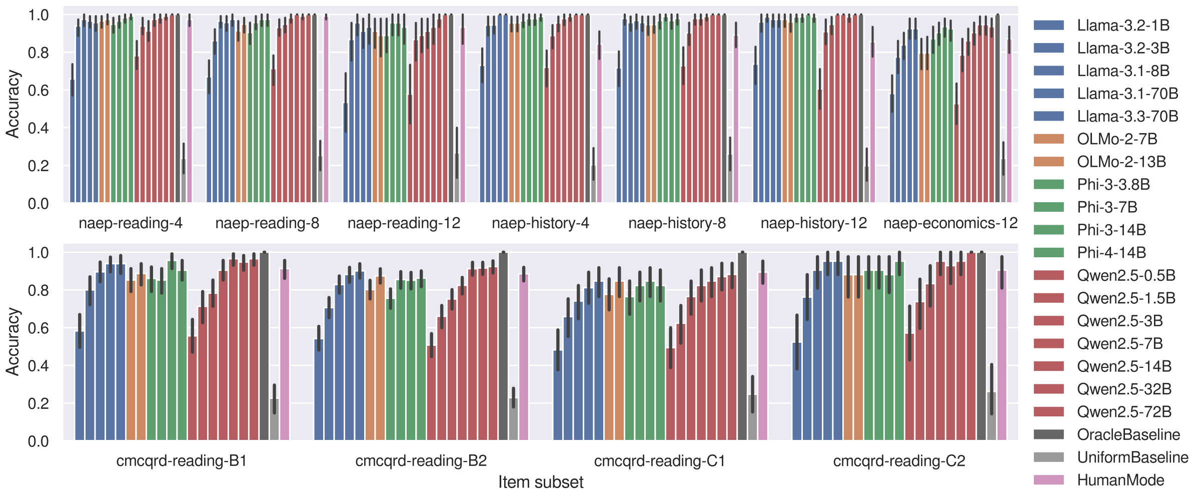

Figure 5: Mode accuracy across item subsets, models, baselines, and humans. Error bars are bootstrapped 95% confidence intervals.

## C Details on temperature scaling

We optimized temperature parameters using KL divergence as a loss function and an Adam optimizer (see `analysis.py` in the code repository). The resulting optimized temperature values are visualized in Figure 6. Larger LLMs tend to be overly confident, assigning almost all probability mass to a single answer option, and therefore require higher temperatures to align them with human response distributions.

The effect of temperature scaling can be seen by comparing the results without temperature scaling in Figure 7 with Figure 2.

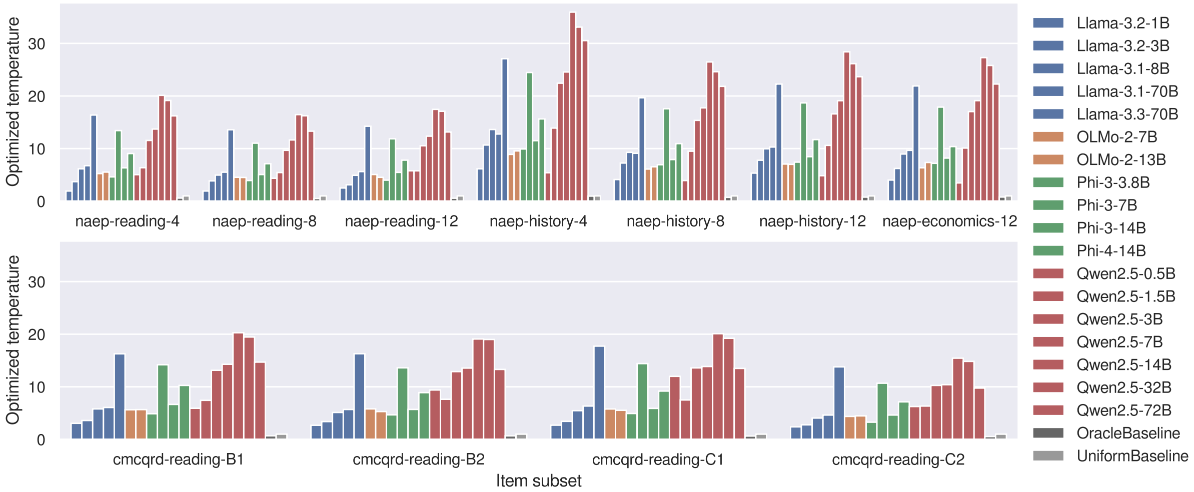

Figure 6: Optimized temperature value for each model and item subset.

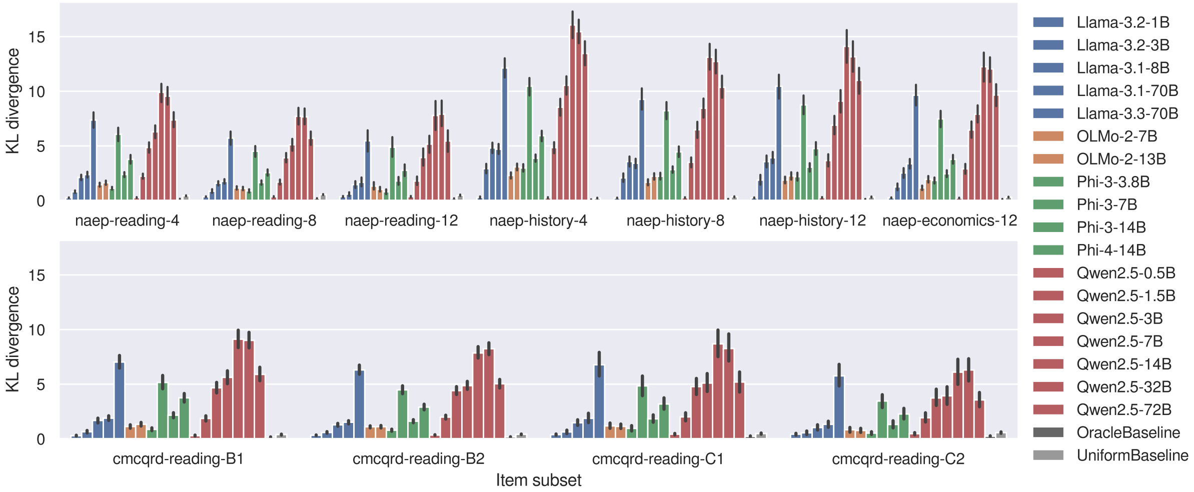

Figure 7: Mean KL divergence between LLM response probability distributions *without* temperature scaling and human response distributions. Error bars are bootstrapped 95% confidence intervals.

## D Additional results for CTT analysis

In addition to the correlations between models and humans in Figure 3, Figure 8 shows the full correlation matrices, including model-model correlations. This confirms that the LLMs are much more similar to each other than to humans. In addition, models of similar sizes (but different model families) tend to be more similar to each other compared to models of different sizes.

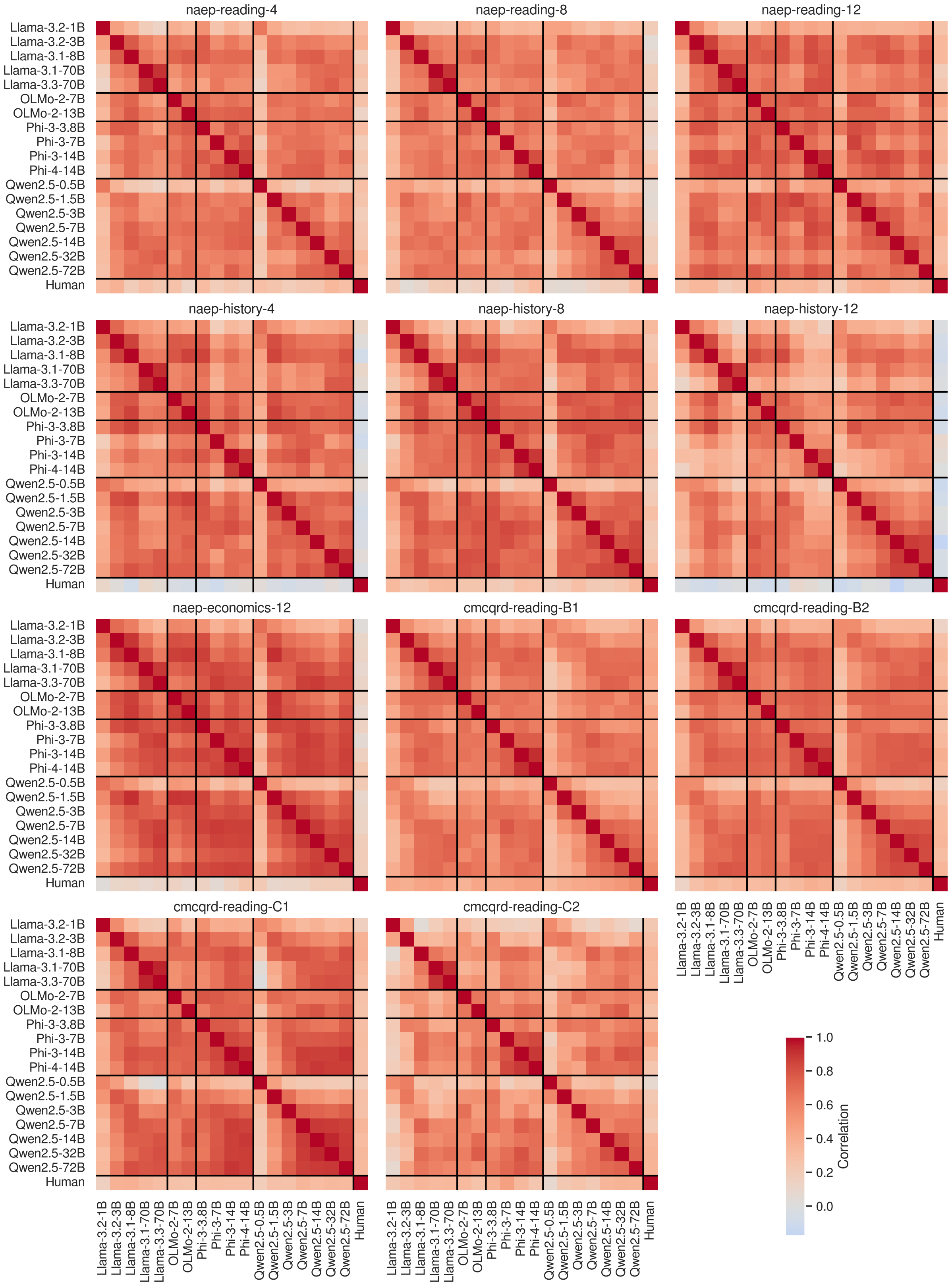

Figure 8: Pearson correlation between all LLM correct response probabilities and human item facilities.