PATCH - Psychometrics-AssisTed benCHmarking of Large Language Models: A Case Study of Mathematics Proficiency

Anonymous authors
Paper under double-blind review

Abstract

Many existing benchmarks of large (multimodal) language models (LLMs) focus on measuring LLMs' academic proficiency, often with also an interest in comparing model performance with human test takers. While these benchmarks have proven key to the development of LLMs, they suffer from several limitations, including questionable measurement quality (e.g., Do they measure what they are supposed to in a reliable way?), lack of quality assessment on the item level (e.g., Are some items more important or difficult than others?) and unclear human population reference (e.g., To whom can the model be compared?). In response to these challenges, we propose leveraging knowledge from psychometrics - a field dedicated to the measurement of latent variables like academic proficiency - into LLM benchmarking. We make three primary contributions. First, we introduce PATCH: a novel framework for **P**sychometrics-**A**ssis**T**ed ben**CH**marking of LLMs. PATCH addresses the aforementioned limitations, presenting a new direction for LLM benchmark research. Second, we implement PATCH by measuring GPT-4 and Gemini-Pro-Vision's proficiency in 8th grade mathematics against 56 human populations. We show that adopting a psychometrics-based approach yields evaluation outcomes that diverge from those based on existing benchmarking practices. Third, we release 4 datasets to support measuring and comparing LLM proficiency in grade school mathematics and science against human populations.

1 Introduction

Large language models (LLMs), including their multimodal variants like vision language models, have witnessed significant advancements in recent years. These models are typically evaluated on established benchmarks that assess their performance across a diverse set of tasks, including commonsense reasoning (e.g., HellaSwag by Zellers et al. (2019), WinoGrande by Sakaguchi et al. (2019)), coding (HumanEval by Chen et al. (2021), Natural2Code by Google (2023), and academic proficiency. Academic proficiency, in particular, has become a crucial part of LLM evaluation, as evidenced by the large number of related benchmarks (e.g., MMLU by Hendrycks et al. (2021), ARC by Clark et al. (2018), GSM8K by Cobbe et al. (2021), DROP by Dua et al. (2019), MATH by Hendrycks et al. (2021)), and recent model technical reports' focus on them (e.g., OpenAI, 2023; Google, 2023). In these benchmarks, LLM performance is also often contrasted with human performance.

Despite the success of existing benchmarks in advancing LLM research, they are not without limitations. The *first* concern is measurement quality: Do these benchmarks measure what they are supposed to in a reliable way? Many benchmarks are created via crowd-sourced knowledge, by asking a convenience group of individuals (e.g., crowd workers, paper authors) to create new test items (e.g., GSM8K, DROP) or collecting them from (often undocumented) sources (e.g., websites, textbooks, school exams) (e.g., MATH, MMLU, ARC). Without domain expert input and rigorous testing of item quality, this approach often leads to undesirable outcomes, such as mismatch between a benchmark and its claimed measurement goal, missing information in a question, wrong answer keys, and low data annotation agreement (Nie et al., 2020).[1]

Second, current benchmarks do not account for differences across test items, such as item discriminativeness[2] and difficulty (see § 3.1). For example, consider three items A (easy), B (hard) and C (hard). While answering correctly to A and B would result in the same score as answering correctly to B and C, the latter (i.e., answering correctly to more difficult items) would imply higher proficiency. Furthermore, a benchmark that consists of only easy and hard items will fail to differentiate models with medium proficiency (i.e., low discriminativeness). Thus, without accounting for item differences, benchmarking results, especially model rankings, can be misleading.

Third, while many benchmarks tried to compare LLMs against humans, the human population to be compared is unclear. For instance, human performance in MATH is based on the paper's authors; in MMLU, crowd workers; in MATH, 6 university students in MATH. Using such convenience samples (with none to little information about sample characteristics), the resulting human performance is local to that specific sample and cannot be generalized to other human samples or specific populations.

To address these challenges, we propose integrating insights from psychometrics — a field dedicated to the measurement of latent variables like cognitive abilities and academic proficiency — into the benchmarking process of LLMs. In particular, we draw on two research areas in psychometrics: *item response theory* (see § 3.1) and *test development* (see § 3.2 and 3.3). The former can help to estimate academic proficiency more accurately than common practice in LLM benchmarks (e.g., means, percentages, total scores). It can also provide diagnostic information about the quality of each test item. The latter, test development knowledge from psychometrics, can help to build high quality LLM benchmarks where comparison to specific human populations can be made.

Our paper makes three primary contributions. *First*, we present **PATCH**: a novel framework for **P**sychometrics-**A**ssis**T**ed ben**CH**marking of LLMs, which addresses the aforementioned limitations of existing benchmarks. *Second*, we demonstrate the implementation of our framework by testing GPT-4 and Gemini's proficiency in 8th grade mathematics using the released test items and data from Trends in International Mathematics and Science Study (TIMSS) 2011. We show empirically how adopting a psychometrics-based approach can lead to evaluation outcomes that diverge from those obtained through conventional benchmarking practices and are more informative, underscoring the potential of psychometrics to reshape the LLM benchmarking landscape. Third, we make our benchmark dataset and evaluation code based on TIMSS 2011 available to future researchers, along with three other math and science datasets based on TIMSS 2011 and 2008. We will encrypt them with a public key to avoid future data contamination, following the advice in Jacovi et al. (2023).

2 Related Work

We are not the first to think of leveraging psychometrics for LLM and more general NLP research. For instance, psychometric scales have been used to examine the psychological profiles of LLMs such as personality traits and motivations (Huang et al., 2024; Pellert et al., 2023; Dillion et al., 2023). The text in these scales can also be used to improve encoding and prediction of social science constructs like personality traits (Kreuter et al., 2022; Vu et al., 2020; Yang et al., 2021; Fang et al., 2023a). Psychometrics-based reliability and validity tests have also been proposed or/and used to assess the quality of model bias measures (Du et al., 2021; van der Wal et al., 2024), text embeddings (Fang et al., 2022), political stance detection (Sen et al., 2020), annotations (Amidei et al., 2020), user representations (Fang et al., 2023b), and general social science constructs (Birkenmaier et al., 2023).

[1]We avoid calling out specific datasets here, but a quick Internet search would reveal many blogs reporting large percentages of errors in existing LLM benchmarks.

[2]In psychometrics, the term "item discrimination" is used. However, given the ambiguity and negative connotation of "discrimination", we adopt "discriminativeness" instead.

The most closely related work to our paper is the use of IRT models in NLP for data splitting (Lalor et al., 2016), comparison of existing evaluation datasets and instances (e.g., difficulty, discriminativeness) (Sedoc & Ungar, 2020; Vania et al., 2021; Rodriguez et al., 2021; Lalor et al., 2018; Rodriguez et al., 2022), as well as identification of difficult instances from training dynamics (Lalor & Yu, 2020; Lalor et al., 2019). Our work distinguishes itself from these papers in two aspects. First, we do not apply IRT to *existing* LLM datasets/benchmarks. Instead, we introduce a framework for benchmarking LLMs by leveraging not only IRT but also test development knowledge from psychometrics. The goal of this framework is to generate new, high-quality benchmarks for LLMs that warrant valid comparison with human populations. Second, we demonstrate our framework with a mathematics proficiency test validated on human populations, and compare LLM performance with human performance. To the best our knowledge, we are the first to apply psychometrically validated (mathematics) proficiency tests to LLMs and make valid model/human comparisons.

3 Preliminaries

3.1 Item Response Theory

IRT refers to a family of mathematical models that describe the functional relationship between responses to a test item, the test item's characteristics (e.g., item difficulty and discriminativeness) and test taker's standing on the latent construct being measured (e.g., proficiency) (AERA et al., 2014). Unlike classical test theory and current LLM benchmarks, which focus on the total or mean score of a test, IRT models takes into account the characteristics of both the items and the individuals being assessed, offering advantages like more accurate estimation of test takers' proficiency, and item quality diagnostics. As such, IRT models have gained widespread adoption in various fields, including education, psychology, and healthcare, where precise measurement and assessment are crucial.

We describe below three fundamental IRT models suitable for different types of test items: the 3-parameter logistic (3PL) model for multiple choice items scored as either incorrect or correct, the 2-parameter logistic (2PL) model for open-ended response items scored as either incorrect or correct, as well as the generalized partial credit (GPC) model for open-ended response items scored as either incorrect, partially correct, or correct.

The 3PL model gives the probability that a test taker, whose proficiency is characterized by the latent variable θ, will respond correctly to item i:

$$P\left(x_i = 1 \mid \theta, a_i, b_i, c_i\right) = c_i + \frac{1 - c_i}{1 + \exp\left(-1.7 \cdot a_i \cdot (\theta - b_i)\right)} \equiv P_{i,1}\left(\theta\right) \tag{1}$$

where x_i is the scored response to item i (1 if correct and 0 if incorrect); θ is the proficiency of the test taker, where higher proficiency has a greater probability of responding correctly; a_i is the slope parameter of item i, characterizing its discriminativeness (i.e., how well the item can tell test takers with higher θ from those with lower θ)[3]; b_i is the location parameter of item i, characterizing its difficulty; c_i is the lower asymptote parameter of item i, reflecting the chances of test takers with very low proficiency selecting the correct answer (i.e., guessing). Correspondingly, the probability of an incorrect response to item i is: $P_{i,0} = P\left(x_i = 0 \mid \theta_k, a_i, b_i, c_i\right) = 1 - P_{i,1}\left(\theta_k\right)$. The 2PL model has the same form as the 3PL model (Equation 1), except that the c_i parameter is fixed at zero (i.e., no guessing).

The GPC model Muraki (1992) gives the probability that a test taker with proficiency θ will have, for the ith item, a response x_i that is scored in the lth of m_i ordered score categories:

[3]The number 1.7 is a scaling parameter to preserve historical interpretation of parameter a_i on the normal ogive scale (Camilli, 1994). Also applies to 2PL and GPC models.

$$P\left(x_i = l \mid \theta, a_i, b_i, d_{i,1}, \cdots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7 \cdot a_i \cdot (\theta - b_i + d_{i,v})\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i \cdot (\theta - b_i + d_{i,v})\right)} \equiv P_{i,l}(\theta) \tag{2}$$

where m_i is the number of response score categories for item i, usually 3; x_i is the scored response to item i, ranging between 0 and $m_i - 1$ (i.e., 0, 1 and 2, for incorrect, partially correct, and correct responses); θ, a_i, b_i have the same interpretations as in the 3PL and 2PL models; $d_{i,1}$ is the category l threshold parameter. Setting $d_{i,0} = 0$ and $\sum_{j=1}^{m_i-1} d_{i,j} = 0$ resolves the indeterminacy of the model parameters.

Assuming conditional independence, the joint probability of a particular response pattern x across a set of n items is given by:

$$P\left(x \mid \theta, \text{ item parameters }\right) = \prod_{i=1}^{n} \prod_{l=0}^{m_i-1} P_{i,l}(\theta)^{u_{i,l}} \tag{3}$$

where $P_{i,l}(\theta)$ is of the form appropriate to the type of item (i.e., 3PL, 2PL or GPC), m_i is equal to 2 for dichotomously scored items and 3 for polytomously scored items, and $u_{i,l}$ is an indicator variable defined as: $u_{i,l} = \begin{cases} 1 \text{ if response } x_i \text{ is in category } l \\ 0 \text{ otherwise} \end{cases}$. This function can be viewed as a likelihood function to be maximized by the item parameters. With the estimated item parameters, θ can then be estimated (Reise & Revicki, 2014).

3.2 Test Development in Psychometrics

Psychometrics	LLM Benchmarking
1. Construct and test need specification.	1. Test need (& construct) specification.
2. Overall planning.	2. Overall planning.
3. Item development.	3. Dataset development.
a. Construct refinement.	a. Existing item collection *OR*
b. Item generation.	- Quality control.
c. Item review.	b. Item creation and/or annotation.
d. Piloting of items.	- Instructions.
e. Psychometric quality analysis.	- (Pilot) study.
4. Test construction and specification.	- Agreement analysis.
5. Implementation and testing.	- Error analysis.
6. Psychometric quality analysis.	4. Dataset construction.
7. Test scoring and norming.	5. Model selection and evaluation.
8. Technical Manual.	6. Benchmark release.

Table 1: Contrasting test development between psychometrics and LLM benchmarking.

Test development in psychometrics concerns the process of developing and implementing a test according to psychometric principles (Irwing & Hughes, 2018). Table 1 contrasts psychometric test development (based on Irwing & Hughes (2018)) with common LLM benchmarking procedures (based on Bowman & Dahl (2021); Raji et al. (2021)). What sets psychometric test development apart from typical LLM benchmark development is its focus on ensuring that the test matches a well-defined construct via expert-driven item generation, rigorous pilot testing, use of factor analysis and IRT models for item and test diagnostics, establishment of scoring and normalization standards, and testing on often representative samples of intended test takers. The result of this elaborate process is a high-quality test that can assess the construct of interest for the test takers in a valid and reliable way. Many large-scale assessments, such as PISA, TIMSS and PIRLS, confirm to such a process.

We will use **P**roficiency in **G**rade **S**chool **M**athematics (PGSM) as the construct of interest to further illustrate this process. In Step 1, the construct of interest and the test need are specified. For instance, what do we define PGSM? Is it based on a specific curriculum? What does existing literature say? Which education levels are we interested in? Is the test meant for comparison between students within a school, or between schools within a country? Such questions help us to clarify what we want to measure and how it can be measured.

In Step 2, we make some necessary planning: How many test items? What kind of item format (e.g., multiple choice, short answer questions)? Will the test scores be standardized? How to assess the quality of test items? What are the desired psychometric properties of the test items (e.g., how discriminative and difficult should the items be?) and the test as a whole (e.g., internal consistency)? Will we pilot any test item? Will the test be computer- or paper-based? To sample test takers, what sampling frame and methodology should we use?

In Step 3, we develop test items, which is an iterative procedure involving four steps: (a) construct refinement, where we further clarify the definition of PGSM (e.g., What content domains should be included: number, algebra, probability theory? Is proficiency only about knowing, or also about applying and reasoning?), (b) generate a pool of items with domain experts, (c) review the items for obvious misfit, errors and biases, and (d) pilot the items with an ideally representative sample of test takers. (e) With the responses from the pilot step, we can assess the psychometric properties of the test items with IRT and factor analysis (e.g., item discriminativeness; item difficulty; factor structure[4]). We iterate this procedure until we have a set of test items with acceptable psychometric properties. Then, in Step 4, we construct the PGSM test by specifying, for instance, which items to include (if not all), in which order, how many equivalent test versions, and what scoring instructions to use.

In Step 5, the test gets implemented to the intended test takers, followed by Step 6: another round of quality analysis. If any item displays low quality characteristics (e.g., zero or negative discriminativeness), it will be left out of the final scoring. In Step 7, responses of the test takers are scored for each item, and the resulting item-level scores form the basis for estimating proficiency scores using IRT or simpler procedures like (weighted) sums. Normalizing the proficiency scores are also typical (e.g., a mean of 500 and a standard deviation of 100) to facilitate interpretations and comparisons. Finally, in Step 8, a technical manual is compiled, detailing all the results from Step 1-7, to facilitate correct re-use of the collected data, the test, as well as interpretation of test scores, among other purposes.

3.3 LLM Benchmark Development

Developing LLM benchmarks follows a similar yet different process. Take GSM8K (Cobbe et al., 2021) as an example. According to the GSM8k paper, the authors started by specifying the need for a large, high quality math test at grade school level that is of moderate difficulty for LLMs (Step 1). The implied construct (PGSM) is not linked to any specific curriculum. Then, the overall planning is made (Step 2): The number of items should be in the thousands; The items will be curated by crowd workers; Agreement and error analysis will be used to investigate the quality of the dataset; GPT-3 will be used to benchmark the dataset and verify the difficulty of the dataset; etc. In Step 3, where dataset development takes place, often one of the two strategies is used: *either* collect items from existing datasets and other sources and compile them into a new dataset, *or* (like GSM8K) create own items from scratch possibly with annotations. The latter is usually an iterative procedure consisting of four parts: creating instructions (and possibly a user interface) for item generation and/or annotation; conduct a (pilot) study to collect the items and/or annotations; check annotator agreement; and assess errors associated with the items or annotations. This step is iterated until a sufficient number of items is reached that meets desired quality standards (e.g., high annotator agreement, low error rate). In total, GSM8K collects 8,500 items with solutions, with identified annotator disagreements resolved and a less than 2% error rate. In Step 4, the generated items form the final dataset, which is typically split into training, evaluation and testing partitions. In Step 5, the final selection of LLMs is made and evaluated on the dataset.

[4]The correlational relationship between the test items intended to measure the construct of interest.

Finally, in Step 6, the benchmark gets released, which typically consists of the dataset as well as its documentation (often a research paper) and benchmarking results.

Comparison with Psychometrics While sharing similarity with test development in psychometrics, benchmark development for LLMs falls short on four aspects. First, the construct of interest is often under-specified, leading to a mismatch between the intended construct and what the dataset actually measures. Take GSM8K as an example: while the dataset is intended to measure proficiency in grade school mathematics, the target grade level(s) are unclear and it only focuses on one content domain (algebra), missing other relevant ones like geometry and data. This is likely the result of not using established mathematics curriculum and domain experts to develop test items. Second, despite LLM researchers' interest in comparing LLM performance with human test takers (e.g., the GSM8K paper claims that "a bright middle school student should be able to solve every problem"), such comparisons usually cannot be made because the test has not been designed with humans in mind or validated on any (representative samples of) human test taker populations. Third, in addition to agreement and error analysis, LLM benchmarks can benefit from psychometric analysis of test items, (i.e., checking item discriminativeness and difficulty, as well as the factor structure of the items). While this is currently not the norm, there have been some promising attempts (see § 2). Lastly, the released benchmark often does not contain sufficient details about all the steps involved in creating the benchmark. For instance, the GSM8K paper does not include the instructions for item creation and annotation, the results from the pilot study, the agreement numbers, or annotator characteristics, all of which are important for external researchers to independently validate the quality of the benchmark.

4 PATCH: Psychometrics-Assisted benCHmarking of LLMs

Figure 1 illustrates PATCH, our conceptualization of a psychometrics-assisted framework for benchmarking LLMs[5]. In PATCH, the first step is to define the construct of interest (e.g., proficiency in 8th grade mathematics). The second step is to look for an existing validated psychometric test that measures this property; alternatively, a test can be developed from scratch, following the procedures described in § 3.2, which likely requires collaboration with experienced psychometricians. The term "validated" means that the test has been tested on a representative sample of the target population of human test takers and fulfills several psychometric quality requirements (e.g., discriminative items that are well distributed across different difficulty levels; showing high reliability (e.g., high internal consistency) and validity (e.g., the test's factor structure matches the construct definition)).

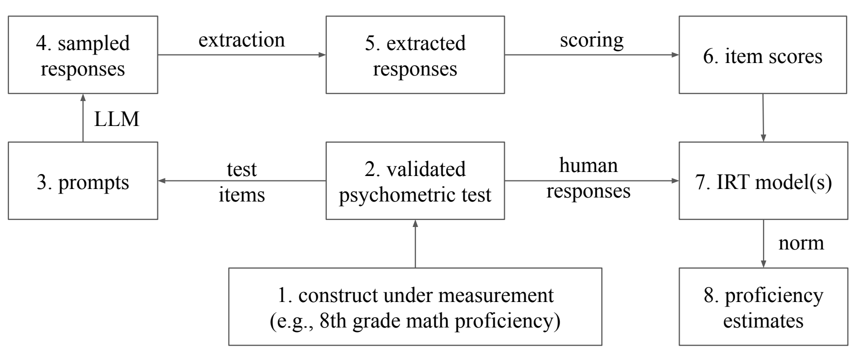


Figure 1: PATCH: A **P**sychometrics-**A**ssis**T**ed framework for ben**CH**marking LLMs.

Next (Step 3→4), we use the items in the validated psychometric test to construct prompts for the LLMs under evaluation, and then sample responses. A response typically consists of a task description, an explanation and an answer (key). Therefore, in Step 4→5, we extract the answer (key) for each item's response, then grade it to obtain item scores (Step 5→6).

[5]Partly inspired by the Hexagon Framework of scientific measurements in Mari et al. (2023).

For Step 2→7, the responses of human test takers can be used to estimate IRT item parameters and subsequently the latent proficiency scores for each test taker (human or LLM) with uncertainty estimates. Multiple IRT models are often used because of the use of different types of test items. These latent scores are typically standardized z-scores (i.e., mean of 0 and SD of 1), which can optionally go through further normalization (e.g., re-scaling to a mean of 500 and a SD of 100) (Step 6→7). These final proficiency scores can be used for comparison with other models and populations.

It can be said that the heart of PATCH is a validated psychometric test, which not only provides the basis for accurate measurement of a model's capability of interest but also facilitates comparison between LLMs and human test takers. Unfortunately, developing such a test can be a long and expensive process; utilizing existing tests can be a shortcut, which, however, need to satisfy three requirements: clear human population reference; test items available (released); human responses and/or item parameter estimates available. The second and third requirement are in practice difficult to meet, as many test institutes do not make their test items public due to commercial interests (e.g., SAT) or the need to measure trends over time (e.g., PISA). Collaboration with test institutes would be ideal.

To the best of our knowledge, when it comes to academic proficiency tests, only TIMSS and PIRLS tests from certain years can be readily used for PATCH-based LLM benchmarking. TIMSS measures proficiency in grade school mathematics and science (4th grade, 8th grade, and final year of secondary school), while PIRLS assesses reading comprehension in 9-10 year olds. Both TIMSS and PIRLS are administered in a large number of countries and regions with representative student samples, enabling country/region-level comparisons. In the following section, we demonstrate PATCH by measuring GPT-4 and Gemini's proficiency in 8th grade mathematics, using the latest available data from TIMSS: TIMSS 2011.

5 Demonstration: Measuring LLM Proficiency in 8th Grade Mathematics

5.1 Data: TIMSS 2011 8th Grade Mathematics

56 countries/regions participated in TIMSS 2011, with typically a random sample of about 150 schools in each country/region and a random sample of about 4,000 students from these schools. These sample sizes are determined on the basis of a $\leq .035$ standard error for each country's mean proficiency estimate. The use of random sampling makes unbiased proficiency estimates possible at the population level. TIMSS 2011 has released a publicly available database[6], of which three components are relevant to our study:

Test Items The TIMSS 2011 study has released 88 mathematics test items, 48 of which are multiple choice, 30 open-ended items scored as either incorrect or correct, and 10 open-ended items scored as either incorrect, partially correct, or correct. These items assess four content domains representative of 8th grade mathematics curriculum (agreed upon by experts from participating countries/regions): number, algebra, geometry, data and chance. Within each domain, items are designed to cover various subtopics (e.g., decimals, functions, patterns) and three cognitive domains: knowing, applying and reasoning. These test items are only available in a PDF file that can be downloaded from the NCES website, which includes also scoring instructions.[7] To extracting them into a format compatible with LLMs, we used OCR tools to extract as much textual information as possible, converted mathematical objects (e.g., numbers, symbols, equations, tables) into LaTeX format (following earlier benchmarks like MATH (Hendrycks et al., 2021)) and figures into JPEG format. See Appendix A.1 for examples. We will release this LLM-compatible version of test items, as well as a mathematics dataset from TIMSS 2008 and two science datasets from TIMSS 2011 and 2008.

IRT and Item Parameters The second part of the dataset concerns the specific IRT model used for each test item and the estimated item parameters (e.g., discriminativeness, difficulty), which can be used to reconstruct the IRT formulas for estimating proficiency scores.

[6] https://timssandpirls.bc.edu/timss2011/international-database.html
[7] https://nces.ed.gov/timss/pdf/TIMSS2011_G8_Math.pdf

Student Responses and Proficiency Estimates Lastly, the students' responses to each test item and their proficiency estimates are available, allowing us to construct proficiency score distributions for each country and region.

5.2 LLMs: GPT-4 with Vision and Gemini-Pro-Vision

Considering that more than 1/3 of the test items contain visual elements, we chose two vision language models: GPT-4 with Vision (GPT-4V) and Gemini-Pro-Vision, using the respective APIs. We are aware of other LLMs with vision capabilities. However, our goal is to showcase PATCH instead of benchmarking all relevant LLMs.

A major concern in using these closed-source LLMs is data contamination, which is difficulty to check due to inaccessible training data. However, as our focus is on demonstrating the PATCH framework, data contamination is less worrying. Furthermore, data contamination is still unlikely for four reasons. First, these test items are copyrighted, forbidding commercial use. Second, the test items are hard to extract from the PDF file. Third, to the best of our knowledge, these test items do not exist in current LLM mathematics benchmarks. Fourth, we request GPT-4V and Gemini-Pro-Vision to explain or provide solutions to the test items' IDs (available in the PDF file). Both failed to recognize these specific test IDs.

5.3 Prompts and Temperature

We design two separate prompts for each test item: the system message and the user message. We design the system message according to the prompt engineering guide by OpenAI[8], utilizing chain-of-thought and step-by-step instructions on how to respond to the user message (i.e., with a classification of question type, an explanation and an answer (key)). The system message is the same for all test items (see Appendix A.2). Furthermore, to account for LLMs' sensitivity to slight variations in prompts (e.g., Sclar et al., 2023; Loya et al., 2023), we generate 10 additional variants of the system prompt with slight perturbations (e.g., lowercase a heading, vary the order of unordered bullet points).

The user message is item-specific, containing both the item's textual description and the associated image(s) in base 64 encoded format. See Appendix A.1 for examples.[9]

Following OpenAI (2023)'s technical report, we set the temperature parameter at 0.3 for multiple choice items and 0.6 for the others. See Appendix B for example responses.

5.4 Response Scoring and Proficiency Estimation

We manually examine the sampled responses from GPT-4V and Gemini-V and score them following the official scoring guide of TIMSS 2011. Then, for multiple choice items, we apply the 3PL model (Equation 1); for open-ended items, we apply the GPC model (Equation 2) if partially correct response is admissible, otherwise the 2PL model. We use maximum likelihood to obtain unbiased estimates of model proficiency scores (θ) with the mirt package in R (Chalmers, 2012). This results in 11 θ estimates per model due to the use of 11 system message variants. We then use inverse variance weighting (Marín-Martínez & Sánchez-Meca, 2010) to obtain a weighted θ and its 95% confidence interval (CI) for each model.

5.5 Results

Figure 2 shows the proficiency score distribution and ranking of 15 selected participating countries & regions, GPT-4V and Gemini-Pro-Vision.[10] The proficiency scores on the left panel (x-axis) are percentages of correct responses, which is the default approach in current LLM benchmarking; the proficiency estimates on the right panel are based on IRT. We make

[8] https://platform.openai.com/docs/guides/prompt-engineering

[9] While we are aware of other prompt engineering techniques, such as few-shot prompting and self-consistency, we did not experiment with them, as our focus is on demonstrating the use of PATCH.

[10] Only 15 countries are shown here to save space. The complete figures can be found in Appendix C.

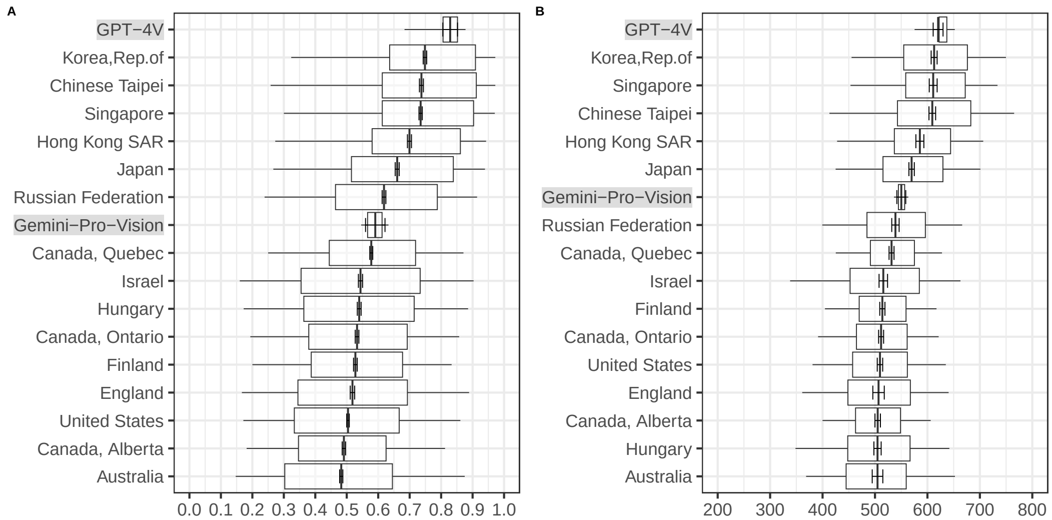


Figure 2: Distribution of proficiency estimates for GPT-4V, Gemini-Vision-Pro and selected participants of the TIMSS 2011 8th grade mathematics test. Left figure (A) shows the proficiency estimates based on the percentages of correct responses. Right figure (B) shows the IRT-based proficiency estimates. The middle vertical line in each box plot represents the weighted mean proficiency score, with the error bars indicating its 95% confidence interval. The borders of each box indicate the range of the middle 50% of all values, with the two whiskers indicating the 5th and 95th percentiles.

three observations. First, regardless of the method of proficiency estimation, GPT-4V has the overall best performance relative to Gemini-Pro-Vision and the average proficiency of 8th grade students of each participating country/region. Second, the method of proficiency estimation affects the ranking results. For instance, while Chinese Taipei is ranked third on the left, it is ranked 4th on the right; Gemini-Pro-Vision is ranked the 8th on the left, but ranked the 7th on the right. Third, the method of proficiency estimation affects the estimated 95% CIs, which are usually wider when IRT is used (as it accounts for both item and test taker variances). Notably, while on the left panel the CI of GPT-4V does not overlap with the second best, South Korea, indicating a statistically significant difference, they overlap on the right panel, suggesting otherwise. This finding shows that the adoption of PATCH is likely going to make a difference to LLM benchmark results.

6 Conclusions

In this paper, we propose PATCH, a psychometrics-inspired framework to address current limitations of LLM benchmarks, especially for the purpose of model and human comparison. We demonstrate PATCH with a 8th grade mathematics proficiency test, where PATCH yields evaluation outcomes that diverge from those based on existing benchmarking practices. This underscores the potential of PATCH to reshape the LLM benchmarking landscape. Nevertheless, our paper has several limitations. First, PATCH requires validated tests, which can be resource-intensive if tests need to be developed from scratch. However, this also opens up opportunities for collaboration between LLM researchers, psychometricians and test institutes. Second, the validity, reliability, and fairness of using tests validated solely on humans for LLM benchmarking are debatable due to possibly differing notions of proficiency and cognitive processes between LLMs and humans. Nonetheless, such tests are still better than non-validated benchmarks, particularly for comparison of model and human performance. Advancing LLM benchmarking further requires tests validated on LLMs (and humans for model-human comparisons), necessitating theoretical work on LLM-specific constructs and the development of LLM-specific IRT models and testing procedures.

References

American Educational Research Association AERA, American Psychological Association APA, and National Council on Measurement in Education NCME. *The Standards for Educational and Psychological Testing*. American Educational Research Association, 2014.

Jacopo Amidei, Paul Piwek, and Alistair Willis. Identifying annotator bias: A new irt-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4787–4797, 2020.

Lukas Birkenmaier, Clemens Lechner, and Claudia Wagner. Valitex–a uniform validation framework for computational text-based measures of social science constructs. *arXiv preprint arXiv:2307.02863*, 2023.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, 2021.

Gregory Camilli. Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. *Journal of Educational Statistics*, 19(3):293–295, 1994.

R Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29, 2012.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID:3922816.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Yupei Du, Qixiang Fang, and Dong Nguyen. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10012–10034, 2021.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:67855846.

Qixiang Fang, Dong Nguyen, and Daniel L Oberski. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, 11(1):39, 2022.

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel Oberski. On text-based personality computing: Challenges and future directions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10861–10879, 2023a.

Qixiang Fang, Zhihan Zhou, Francesco Barbieri, Yozen Liu, Leonardo Neves, Dong Nguyen, Daniel L Oberski, Maarten W Bos, and Ron Dotsch. Designing and evaluating general-purpose user representations based on behavioral logs from a measurement process perspective: A case study with snapchat. *arXiv preprint arXiv:2312.12111*, 2023b.

Gemini Team Google. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jentse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=H3UayAQWoE.

Paul Irwing and David J. Hughes. Test Development. In *The Wiley Handbook of Psychometric Testing*, pp. 1–47. John Wiley & Sons, Ltd, 2018. ISBN 978-1-118-48977-2. doi: 10.1002/9781118489772.ch1. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118489772.ch1.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Anne Kreuter, Kai Sassenberg, and Roman Klinger. Items from psychometric tests as training data for personality profiling models of twitter users. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pp. 315–323, 2022.

John P Lalor and Hong Yu. Dynamic data selection for curriculum learning via ability estimation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, pp. 545. NIH Public Access, 2020.

John P Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pp. 648. NIH Public Access, 2016.

John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, pp. 4711. NIH Public Access, 2018.

John P Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, pp. 4240. NIH Public Access, 2019.

Manikanta Loya, Divya Sinha, and Richard Futrell. Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3711–3716, 2023.

Luca Mari, Mark Wilson, and Andrew Maul. *Measurement across the sciences: Developing a shared concept system for measurement*. Springer Nature, 2023.

Fulgencio Marín-Martínez and Julio Sánchez-Meca. Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, 70(1):56–73, 2010.

Eiji Muraki. A generalized partial credit model: Application of an em algorithm. *Applied psychological measurement*, 16(2):159–176, 1992.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9131–9143, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.734. URL https://aclanthology.org/2020.emnlp-main.734.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, pp. 17456916231214460, 2023.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. Ai and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Steven P Reise and Dennis A Revicki. *Handbook of item response theory modeling*. Taylor & Francis New York, NY, 2014.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4486–4503, 2021.

Pedro Rodriguez, Phu Mon Htut, John P Lalor, and João Sedoc. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pp. 100–112, 2022.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande. *Communications of the ACM*, 64:99 – 106, 2019. URL https://api.semanticscholar.org/CorpusID:198893658.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2023.

João Sedoc and Lyle Ungar. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 21–33, 2020.

Indira Sen, Fabian Flöck, and Claudia Wagner. On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1413–1426, 2020.

Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. Undesirable biases in nlp: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40, 2024.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1141–1158, 2021.

Huy Vu, Suhaib Abdurahman, Sudeep Bhatia, and Lyle Ungar. Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1512–1524, 2020.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1131–1142, 2021.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:159041722.

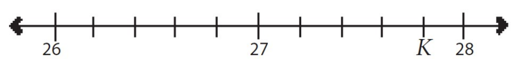
A Prompts

A.1 Example Test Items (User Messages)

Example 1

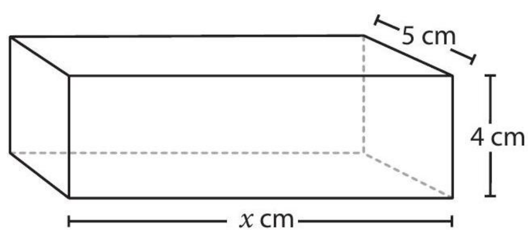
> The fractions $\frac{4}{14}$ and $\frac{\square}{21}$ are equivalent. What is the value of $\square$?
> [A] 6 [B] 7 [C] 11 [D] 14

Example 2



> Which number does K represent on this number line?
> [A] 27.4 [B] 27.8 [C] 27.9 [D] 28.2

Example 3



> The volume of the rectangular box is 200 cm^3. What is the value of x ?

A.2 Example System Messages

Base prompt:

> You are given a math question written in LaTeX format.
> Instructions:
> 1. Type of question: Is it multiple choice, free text response, or drawing?
> 2. Think step by step, and describe your thought process and reasoning.
> 3. Answer:
> - For multiple choice: [selected answer key].
> - For free-text response: [provide your short answer].
> - For drawing: [describe clearly the steps to complete the drawing].
> - If uncertain, make an educated guess.

Variant 1 (item type reordered):

> You are given a math question written in LaTeX format.
> Instructions:
> 1. Type of question: Is it drawing, free text response, or multiple choice?

2. Think step by step, and describe your thought process and reasoning.
3. Answer:
- For drawing: [describe clearly the steps to complete the drawing].
- For free-text response: [provide your short answer].
- For multiple choice: [selected answer key].
- If uncertain, make an educated guess.

Variant 2 (new lines removed):

> You are given a math question written in LaTeX format. Instructions: 1. Type of question: Is it multiple choice, free text response, or drawing? 2. Think step by step, and describe your thought process and reasoning. 3. Answer: - For multiple choice: [selected answer key]. - For free-text response: [provide your short answer]. - For drawing: [describe clearly the steps to complete the drawing]. - If uncertain, make an educated guess.

Variant 3 (upper case heading):

You are given a math question written in LaTeX format.
INSTRUCTIONS:
1. Type of question: Is it multiple choice, free text response, or drawing?
2. Think step by step, and describe your thought process and reasoning.
3. Answer:
- For multiple choice: [selected answer key].
- For free-text response: [provide your short answer].
- For drawing: [describe clearly the steps to complete the drawing].
- If uncertain, make an educated guess.

B Example Responses

Note that the example responses below correspond to the example test items in Appendix A.1.

Example 1 (base prompt, GPT-4-V, correct)

> Type of question: Multiple choice
>
> To find the value of the square that makes the fractions equivalent, we need to find a number that, when divided by 21, gives the same value as 4 divided by 14.
>
> To simplify the fraction $\frac{4}{14}$, we divide both the numerator and the denominator by their greatest common divisor, which is 2. This gives us $\frac{2}{7}$.
>
> Now, we need to find a number that, when placed in the square and divided by 21, gives us the fraction $\frac{2}{7}$.
>
> Since $21 = 7 \times 3$, we need to find a number that, when multiplied by $\frac{1}{3}$, gives us 2. That number is $2 \times 3 = 6$.
>
> Therefore, the value of the square that makes the fractions equivalent is 6.
>
> Answer: [A] 6
>
> Answer: - For multiple-choice: [A] 6

Example 2 (base prompt, Gemini-Vision-Pro, incorrect)

> 1. This is a multiple-choice question.
>
> 2. First, I notice that the number line is labeled from 26 to 28. This means that the value of K must be between 26 and 28.
>
> 3. Next, I look at the answer choices. [A] 27.4 is the closest to 27, so it is a possibility. [B] 27.8 is also close, but it is further away from 27 than [A] 27.4. [C] 27.9 is even further away from 27 than [B] 27.8, so it is unlikely to be the correct answer. [D] 28.2 is the furthest away from 27, so it is the least likely to be the correct answer.
>
> 4. Based on this reasoning, I would select [A] 27.4 as the best answer.

Example 3 (prompt variant 1, GPT4-V, correct)

> Type of Question: [Free Text Response] This is a free-text response question.
>
> To find the value of x, we will use the formula for the volume of a rectangular box, which is:
>
> $$\text{Volume} = \text{length} \times \text{width} \times \text{height}$$
>
> The volume of the box is given as $200\,\text{cm}^3$, the width is $5\,\text{cm}$, and the height is $4\,\text{cm}$. We can set up the equation as follows:
>
> $$200 = x \times 5 \times 4$$
>
> To solve for x, we divide both sides of the equation by 5×4:
>
> $$x = \frac{200}{5 \times 4}$$
> $$x = \frac{200}{20}$$
> $$x = 10\,\text{cm}$$
>
> Answer: $x = 10\,\text{cm}$

C Detailed Result Figures

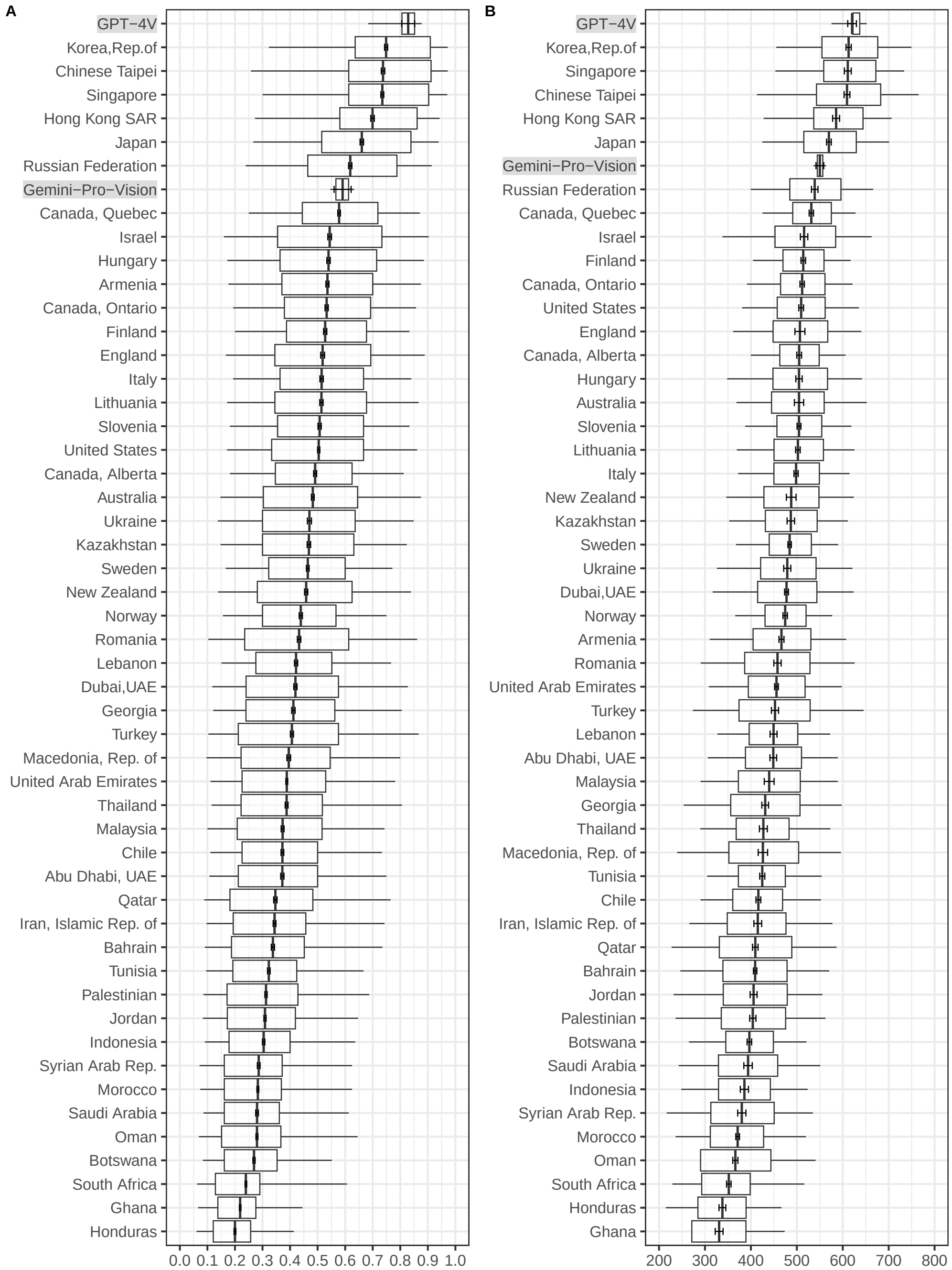


Figure 3: Distribution of proficiency estimates for GPT-4V, Gemini-Pro-Vision and all participants of TIMSS 2011 8th grade mathematics test. Left figure (A) shows the proficiency estimates based on the percentages of correct responses. Right figure (B) shows the IRT-based proficiency estimates. The middle vertical line in each box plot represents the weighted mean proficiency score, with the error bars indicating its 95% confidence interval. The borders of each box indicate the range of the middle 50% of all values, with the two whiskers indicating the 5th and 95th percentiles.