PATCH! {P}sychometrics-{A}ssis{T}ed Ben{CH}marking of Large Language Models against Human Populations: A Case Study of Proficiency in 8th Grade Mathematics

Anonymous ACL submission

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 such 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. In particular, PATCH enables valid comparison between LLMs and human populations. Second, we demonstrate PATCH by measuring several LLMs' 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 current benchmarking practices. Third, we release 4 high-quality datasets to support measuring and comparing LLM proficiency in grade school mathematics and science with 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 such as *commonsense reasoning* (Zellers et al., 2019; Sakaguchi et al., 2021; Chen et al., 2021), *coding* (Chen et al., 2021; 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 like MMLU, ARC, GSM8K, DROP and MATH (Hendrycks et al., 2021; Clark et al., 2018; Cobbe et al., 2021; Dua et al., 2019; Hendrycks et al., 2021), as well as recent model technical reports' increasing focus on them (OpenAI, 2023; Google, 2023). In these benchmarks and reports, the contrast between LLM performance and human performance is often highlighted, sparking media coverage and discussions.

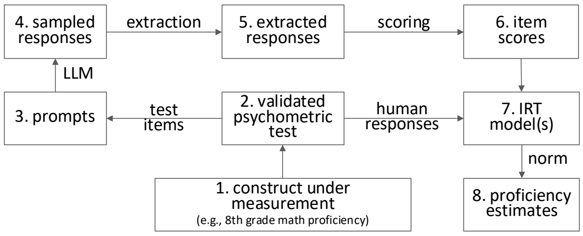


Figure 1: **PATCH: A {P}sychometrics-{A}ssis{T}ed framework for ben{CH}marking LLMs against humans.**

Despite their success in advancing LLM research and shedding light on the artificial versus human intelligence debate, existing benchmarks have notable 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, undesirable outcomes can occur, including a mismatch between a benchmark and its claimed measurement goal, missing information in a question, wrong answer keys, and low data annotation agreement (e.g., Nie et al., 2020; Wang et al., 2024; Chen, 2024).

Second, current benchmarks do not account for differences across test items, such as item discriminativeness[1] and difficulty (see Section 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 accuracy score as answering correctly to B and C, the latter (i.e., answering correctly to more difficult items) would imply higher proficiency. Furthermore, items that are too easy or too difficult (i.e., low discriminativeness) will fail to differentiate models (and humans) of different proficiency levels. Thus, without accounting for item differences, benchmarking results, especially model (versus human) rankings, can be misleading.

Third, while many benchmarks compare LLMs against humans, the human populations under comparison remain unclear (Tedeschi et al., 2023). For instance, human performance in MATH is based on the benchmark's authors; in MMLU, crowd workers; in MATH, 6 university students. Using such convenience samples (with little information about sample characteristics), the resulting human performance cannot be generalised to other human samples or populations.

To address these challenges, we propose leveraging insights from psychometrics - a field dedicated to the measurement of *latent* variables like academic proficiency - into LLM benchmarking practices. In particular, we draw on two research areas in psychometrics: *item response theory* (see Section 3.1) and *test development* (see Section 3.2 and 3.3). The former enables more accurate estimation of academic proficiency on a standardised scale by taking into account both the characteristics of the test items as well as the abilities of the LLMs and individuals being assessed, compared to common practices in LLM benchmarks (e.g., using mean scores, percentages of correct responses). It can also provide diagnostic information about the quality of each test item. The latter, test development knowledge, can help to build high quality LLM benchmarks where valid 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 (see Figure 1), which addresses the aforementioned limitations of existing benchmarks. Second, we demonstrate the implementation of PATCH by testing several LLMs' proficiency in 8th grade mathematics using the released test items and data from *Trends in International Mathematics and Science Study*[2] (TIMSS) 2011. We show empirically that a psychometrics-based approach can lead to evaluation outcomes that diverge from those obtained through conventional benchmarking practices and that are more informative, underscoring the potential of psychometrics to reshape the LLM benchmarking landscape. Third, we make our evaluation code based on the PATCH framework available[3], along with three other mathematics and science datasets based on TIMSS 2011 and 2008[4].

2 Related Work

We are not the first to propose leveraging psychometrics for research on LLMs and other areas in NLP. 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 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 NLP 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).

The most closely related work to our paper is the use of item response theory (IRT) models in NLP for constructing more informative test datasets (Lalor et al., 2016), comparison of existing evaluation datasets and instances (e.g., difficulty, discriminativeness) (Sedoc and 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 and Yu, 2020; Lalor et al., 2019). Our work distinguishes itself from these papers in two

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

[2] http://timssandpirls.bc.edu/timss2015/encyclopedia/

[3] Anonymised url. See uploaded software code.

[4] Anonymised url. See uploaded data and Appendix C.

aspects. First, we do not apply IRT to *existing* LLM datasets/benchmarks. Instead, we introduce a framework for benchmarking LLMs by leveraging both IRT and 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 56 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 versus human comparison.

3 Preliminaries

3.1 Item Response Theory

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., academic 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 (and models) being assessed, offering advantages like item quality diagnostics and more accurate estimation of test takers' proficiency. As such, IRT models have gained widespread adoption in various fields, including education, psychology, and healthcare, where trustworthy 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 generalised 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 characterised by the latent variable θ, will respond correctly to item i:

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

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 a higher value implies a greater probability of responding correctly; a_i is the slope parameter of item i, characterising its discriminativeness (i.e., how well the item can tell test takers with higher θ from those with lower θ)[5]; b_i is the location parameter of item i, characterising 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 i^{th} item, a response x_i that is scored in the l^{th} of m_i ordered score categories:

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

where m_i is the number of response score categories for item i; x_i is the response score of item i between 0 and $m_i - 1$ (e.g., 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:

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

where $P_{i,l}(\theta)$ is of the form specific to the type of item (i.e., 3PL, 2PL or GPC); m_i equals 2 for dichotomously scored items and 3 for polytomously scored items; $u_{i,l}$ is an indicator defined as:

$$u_{i,l} = \begin{cases} 1 \text{ if response } x_i \text{ is in category } l \\ 0 \text{ otherwise} \end{cases}$$

[5]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.

This function can be viewed as a likelihood function to be maximised by the item parameters. With the estimated item parameters, θ can then be estimated via various algorithms (Reise and Revicki, 2014). In this paper, we use maximum likelihood because it gives an unbiased estimate of θ.

3.2 Test Development in Psychometrics

Test development in psychometrics concerns the process of developing and implementing a test according to psychometric principles (Irwing and Hughes, 2018). Table 1 contrasts psychometric test development (based on Irwing and Hughes (2018)) with common LLM benchmarking procedures (based on (Bowman et al., 2015; 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 normalisation standards, and testing on 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 (Programme for International Student Assessment), TIMSS and PIRLS (Progress in International Reading Literacy Study), conform 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, how 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 necessary planning: How many test items? What kind of item format (e.g., multiple choice, short answer questions)? Will the test scores be standardised? 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 kind of sampling frames and strategies should we use?

In Step 3, we develop test items, which is an iterative procedure involving five steps: (a) construct refinement, where we further clarify the definition of PGSM (e.g., What content domains should be included: number, algebra, and/or 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; (d) pilot the items with a representative sample of target 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[6]). 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. It is typical to also normalise the proficiency scores (e.g., with 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 Step 1–7 and corresponding results, to facilitate correct re-use of the response 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 the development of GSM8K (Cobbe et al., 2021) as an example. The authors of GSM8K started by specifying the need for a large, high quality mathematics test at grade school level and of moderate difficulty for LLMs (Step 1). The construct (i.e., PGSM), however, is not explicitly 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

[6]Factor structure refers to the correlational relationships between test items used to measure a construct of interest.

Psychometrics	LLM Benchmarking

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

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

analysis will be used to investigate the quality of the dataset; GPT-3 will be used to benchmark the dataset and verify dataset difficulty.

In Step 3, where dataset development[7] 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 in GSM8K, create own items from scratch (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; conducting a (pilot) study to collect the items and/or annotations; check annotator agreement; and assessing errors associated with the items or annotations. This step is iterated until a sufficient number of items and datasets are reached while meeting desired quality standards (e.g., high annotator agreement, low error rate). In total, GSM8K includes 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, typically with training, evaluation and testing partitions. In Step 5, selected LLMs are evaluated on the dataset. 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, current 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. Again, 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 curricula and domain experts to develop test items.

Second, despite 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 the test's target user populations.

Third, besides 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 not yet the norm, there have been promising attempts (see Section 2).

Lastly, the released benchmark often does not contain sufficient details about the process of benchmark creation. For instance, the GSM8K paper does not report instructions for item generation and annotation, results of the pilot study, agreement statistics, or annotator characteristics, all of which are important for external researchers to independently verify the quality of the benchmark.

4 PATCH: Psychometrics-AssisTed benCHmarking of LLMs

Figure 1 illustrates PATCH, our conceptualisation of a Psychometrics-AssisTed framework for

[7] Note that we use the term "dataset development" here, contrasting "item development" in psychometrics, because of LLM benchmarks' typical emphasis on large and multiple datasets rather than concrete test items.

benCHmarking LLMs against human populations.[8] Under PATCH, the first step is to define the construct of interest (e.g., proficiency in 8th grade mathematics). The second step is to find an existing validated psychometric test measuring this property; alternatively, a test can be developed from scratch, following the procedures described in Section 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 fulfils psychometric quality requirements (e.g., sufficiently many discriminative items 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)).

Next (Step 3→4), we use the items from 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).

For Step 2→7, the responses of human test takers (and of LLMs, if a sufficient number of LLMs are involved) 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 used when different types of test items are used in a test. These latent proficiency scores are typically standardised z-scores (i.e., mean of 0 and standard deviation of 1), which can optionally go through further normalisation (e.g., re-scaling to a mean of 500 and a standard deviation of 100) (Step 6→7). These final proficiency scores enable comparison with other models and human populations.

At the heart of PATCH lies a validated psychometric test, which not only provides the basis for accurate measurement of the 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; utilising existing tests can be a shortcut, which, however, should satisfy three conditions: clear human population reference; test items available; human responses and/or item parameter estimates available. The second and third 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 alleviate this problem.

To the best of our knowledge, among 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 geographical regions with representative student samples, enabling population-level comparisons. In the following section, we demonstrate PATCH by measuring several LLMs' proficiency in 8th grade mathematics, using the latest available data from TIMSS 2011.

5 Demonstration: Measuring LLM Proficiency in 8th Grade Mathematics

5.1 Data: TIMSS 2011 8th Grade Mathematics

56 geographical regions participated in TIMSS 2011, with typically a random sample of about 150 schools in each 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 region'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[9], 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 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.

[8] PATCH is partly inspired by the Hexagon Framework of scientific measurements proposed by Mari et al. (2023).

[9] https://timssandpirls.bc.edu/timss2011/international-database.html

These test items are only available in a PDF file that can be downloaded from the NCES website, which includes also scoring instructions.[10] To extract 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 have released this LLM-compatible version of test items, as well as an eighth grade science test dataset from TIMSS 2011, an advanced secondary school mathematics test dataset from TIMSS 2008, and an advanced secondary school physics test dataset from TIMSS 2008. See Appendix C for details.

IRT and Item Parameters The TIMSS 2011 database also specifies the IRT model used for each test item and contains the item parameter estimates (e.g., discriminativeness, difficulty), which we use to reconstruct the final IRT model for proficiency estimation and verification.

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

5.2 LLMs: GPT-4, Gemini-Pro and Qwen with Vision Capability

Considering that more than 1/3 of the test items contain visual elements, we selected four competitive vision language models: GPT-4 with Vision (GPT-4V), Gemini-Pro-Vision, as well as the open-source Qwen-VL-Plus and Qwen-VL-Max (Bai et al., 2023). There are more LLMs with vision capability. However, our goal is to showcase PATCH, not to benchmark as many LLMs as possible.

A major concern in using these LLMs is data contamination, which is difficulty to check due to inaccessible (information about) 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 source PDF. Third, to the best of our knowledge, these test items do not exist in current LLM mathematics benchmarks. Fourth, we prompted the selected LLMs to explain or provide solutions to the test items' IDs (available in the source PDF). All failed to recognise 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, utilising 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)).[11] 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 (Sclar et al., 2024; 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.[12]

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 Scoring and Proficiency Estimation

We manually scored the sampled responses from the LLMs following the official scoring rubrics 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 corresponding to 11 system message variants. We then use inverse variance weighting (Marín-Martínez and Sánchez-Meca, 2010) to combine these estimates. Inverse variance weighting gives more weight to estimates that are more precise (i.e., having lower variance) and less weight to those that are less pre-

[10] https://nces.ed.gov/timss/pdf/TIMSS2011_G8_Math.pdf

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

[12] We are aware of other prompt engineering techniques like few-shot prompting and self-consistency. We did not experiment with them, as our focus is on demonstrating PATCH.

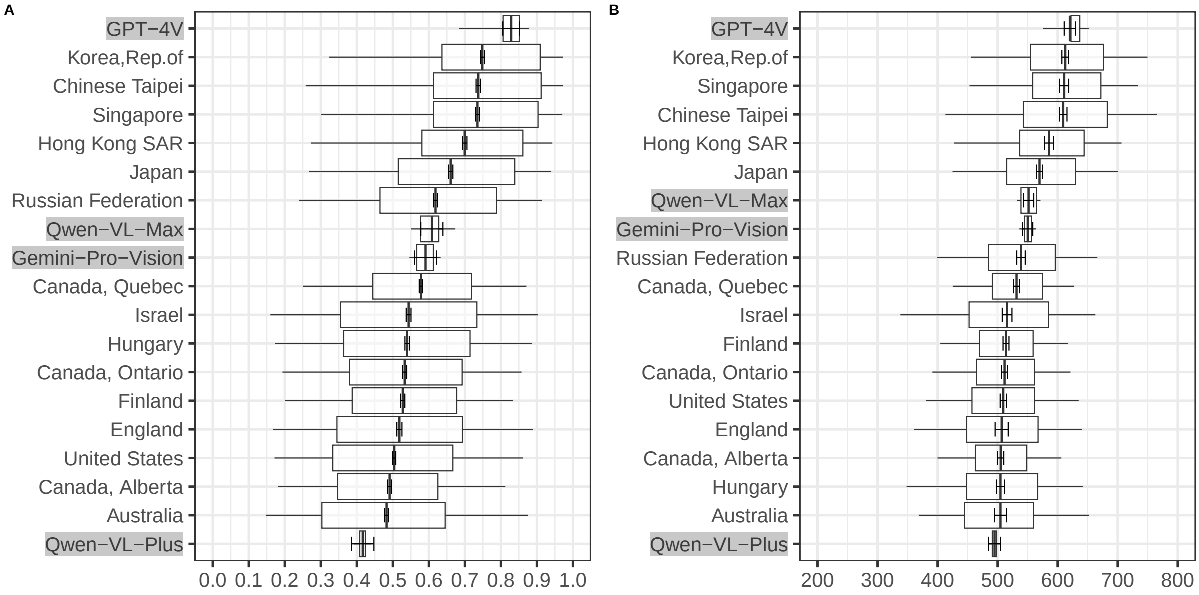


Figure 2: **Distribution of proficiency estimates for GPT-4V, Gemini-Vision-Pro, Qwen-VL-Plus, Qwen-VL-Max and selected participating regions of the TIMSS 2011 8th grade mathematics test.** Left figure (A) shows the proficiency estimates based on the percentages of correct responses. Right figure (B) 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. *Note that we adhere to the official naming conventions of TIMSS 2011 when reporting the names of participating regions, with no intent to offend anyone*.

cise (i.e., having higher variance). This way, we obtain a more accurate *overall* θ estimate and its 95% confidence interval (CI) for each model.

5.5 Results

Figure 2 shows the proficiency score distribution and ranking of the top 15 performing participating regions, as well as GPT-4V, Gemini-Pro-Vision, Qwen-VL-Plus and Qwen-VL-Max. The complete figures can be found in Appendix E. The proficiency scores (x-axis) on the left panel are percentages of correct responses, corresponding to the default approach in current LLM benchmarking; the proficiency estimates on the right panel are based on IRT. We make 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 region. Second, the method of proficiency estimation affects the ranking results. For instance, while Chinese Taipei is ranked 3rd on the left, it is ranked 4th on the right; Gemini-Pro-Vision is ranked 8th on the left, but ranked 7th on the right. Similarly, while Hungary is ranked 11th on the left, it drops to the 16th place 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, Korea, Rep.of, 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, especially in contrast with human performances.

6 Conclusion

In this paper, we propose PATCH, a psychometrics-inspired framework to address current limitations of LLM benchmarks, including questionable measurement quality, lack of quality assessment on the item level and unwarranted comparison between humans and LLMs. We demonstrate PATCH with an 8th grade mathematics proficiency test and show evaluation outcomes that diverge from those based on existing benchmarking practices, especially when comparison with human test takers is made. This underscores the potential of PATCH to reshape the LLM benchmarking landscape. Furthermore, we release 4 datasets that meet the requirements of PATCH, supporting the measurement of LLM proficiency in grade school math and science and its comparison with human performance.

Limitations

Our paper has the following limitations, among others. 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. Third, our experiment only includes four LLMs and one proficiency test. We consider this sufficient for demonstrating PATCH, but not enough if the goal is to benchmark as many LLMs as possible across different tests.

References

AERA, APA, and NCME. 2014. *The Standards for Educational and Psychological Testing*. American Educational Research Association.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966.

Lukas Birkenmaier, Clemens Lechner, and Claudia Wagner. 2023. ValiTex - A uniform validation framework for computational text-based measures of social science constructs. *arXiv*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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

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

Edwin Chen. 2024. Hellaswag or hellabad? 36% of this popular llm benchmark contains errors. Accessed: 2025-02-12.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *arXiv*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv*.

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. 2021. Training verifiers to solve math word problems. *arXiv*.

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

Yupei Du, Qixiang Fang, and Dong Nguyen. 2021. Assessing the reliability of word embedding gender bias measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10012–10034, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Qixiang Fang, Dong Nguyen, and Daniel L. Oberski. 2022. Evaluating the construct validity of text embeddings with application to survey questions. *EPJ Data Science*, 11(1):1–31.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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

Paul Irwing and David J. Hughes. 2018. Test development. In *The Wiley Handbook of Psychometric Testing*, pages 1–47. John Wiley & Sons, Ltd.

Anne Kreuter, Kai Sassenberg, and Roman Klinger. 2022. 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*, pages 315–323. Association for Computational Linguistics.

John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. 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*, pages 4711–4716. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

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.

John P. Lalor and Hong Yu. 2020. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 545–555, Online. Association for Computational Linguistics.

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

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

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

Eiji Muraki. 1992. A generalised partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv*.

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

Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. 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)*, pages 4486–4503.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. 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*.

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

Indira Sen, Fabian Flöck, and Claudia Wagner. 2020. 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)*, pages 1413–1426, Online. Association for Computational Linguistics.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

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

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel Bowman. 2021. 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)*, pages 1141–1158.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Prompts

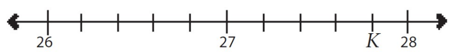
A.1 Example Test Items (User Messages)

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

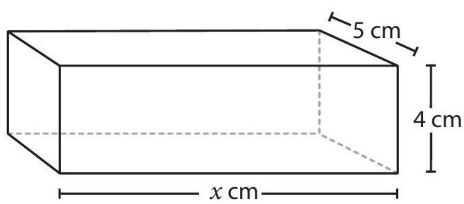
Example 2



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 mathematics 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 mathematics 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 mathematics 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 mathematics 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 order of) 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 labelled 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, Qwen-VL-Max, correct)

> Is it drawing, free text response, or multiple choice? It's a multiple-choice question.
>
> To find the volume of a rectangular prism (box), we use the formula:
>
> $$\text{Volume} = \text{Length} \times \text{Width} \times \text{Height}$$
>
> In this case, the length is x, the width is 5 cm, and the height is 4 cm. So,
>
> $$200 \text{ cm}^3 = x \times 5 \text{ cm} \times 4 \text{ cm}$$
>
> Now, let's solve for x:
>
> $$x = \frac{200 \text{ cm}^3}{5 \text{ cm} \times 4 \text{ cm}}$$
>
> $$x = \frac{200 \text{ cm}^3}{20 \text{ cm}^2}$$
>
> $$x = 10 \text{ cm}$$
>
> So, the value of x is 10 cm. Answer Key: A) 10 cm

C TIMSS Datasets

TIMSS 2011 Mathematics Eighth Grade This dataset was used in this study to demonstrate the PATCH framework. See Section 5.1 for details.

> Source: User Guide, Items and International Database for TIMSS 2011: Science - Eighth Grade. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Our study contributes three additional datasets. Similar to the dataset above, they are also based on officially released items by TIMSS but differ in the test subject, school grade level and/or test year. We constructed each dataset by using a mix of manual labour and OCR tools to extract item details from the official PDFs of the released items. The resulting dataset consists of a LaTeX file ("main.tex") and a folder of item-related images. The test items are formatted in LLM-friendly format. With these three additional datasets, we hope to facilitate interested researchers to benchmark LLMs using these datasets with our PATCH framework. See below for more detail.

TIMSS 2011 Mathematics Fourth Grade This dataset is similar to the one we used to demonstrate PATCH but focuses on a different fourth grade mathematics with 73 items covering three domains: number, geometric shape and measures, and data display. It can be used to benchmark LLMs against representative samples of fourth-grade students from 57 regions.

> Source: User Guide, Items and International Database for TIMSS 2011: Mathematics - Fourth Grade. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

TIMSS 2008 Advanced Mathematics This dataset focuses on assessing proficiency in advanced mathematics at the end of secondary high school. It can be used to benchmark LLMs against representative samples of final-year students in secondary school from 10 countries who have taken an advanced mathematics course. There are 40 items in total, covering algebra, calculus and geometry.

> Source: TIMSS Advanced 2008 User Guide and Items for the International Database: Advanced Mathematics. Copyright © 2009 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

TIMSS 2008 Advanced Physics This dataset focuses on assessing proficiency in advanced physics at the end of secondary high school. It can be used to benchmark LLMs against representative samples of final-year students in secondary school from 10 countries who have taken an advanced physics course. There are 39 items in total, covering mechanics, atomic and nuclear physics, electricity and magnetism, as well as heat and temperature.

> Source: TIMSS Advanced 2008 User Guide and Items for the International Database: Advanced Physics. Copyright © 2009 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Licences According to the website of TIMSS 2011[13] and 2008[14]:

> TIMSS and PIRLS are registered trademarks of IEA. Use of these trademarks without permission of IEA by others may constitute trademark infringement. Furthermore, the website and its contents, together with all online and/or printed publications and released items by TIMSS, PIRLS, and IEA are and will remain the copyright of IEA.
>
> All publications and released items by TIMSS, PIRLS, and IEA, as well as translations thereof, are for non-commercial, educational, and research purposes only. Prior notice is required when using IEA data sources or datasets for assessments or learning materials. IEA reserves the right to refuse copy deemed inappropriate or not properly sourced.

Therefore, our use of TIMSS data in this research is in accordance with the intended use.

D Use of AI Assistants

We used ChatGPT to improve the writing of limited parts of the paper. We also used Mathpix to perform OCR on the PDFs containing the TIMSS released items before further processing into appropriate format. No AI was used for coding or analyses.

E Detailed Result Figure

See next page.

[13] https://timssandpirls.bc.edu/timss2011/international-database.html

[14] https://timssandpirls.bc.edu/timss_advanced/idb.html

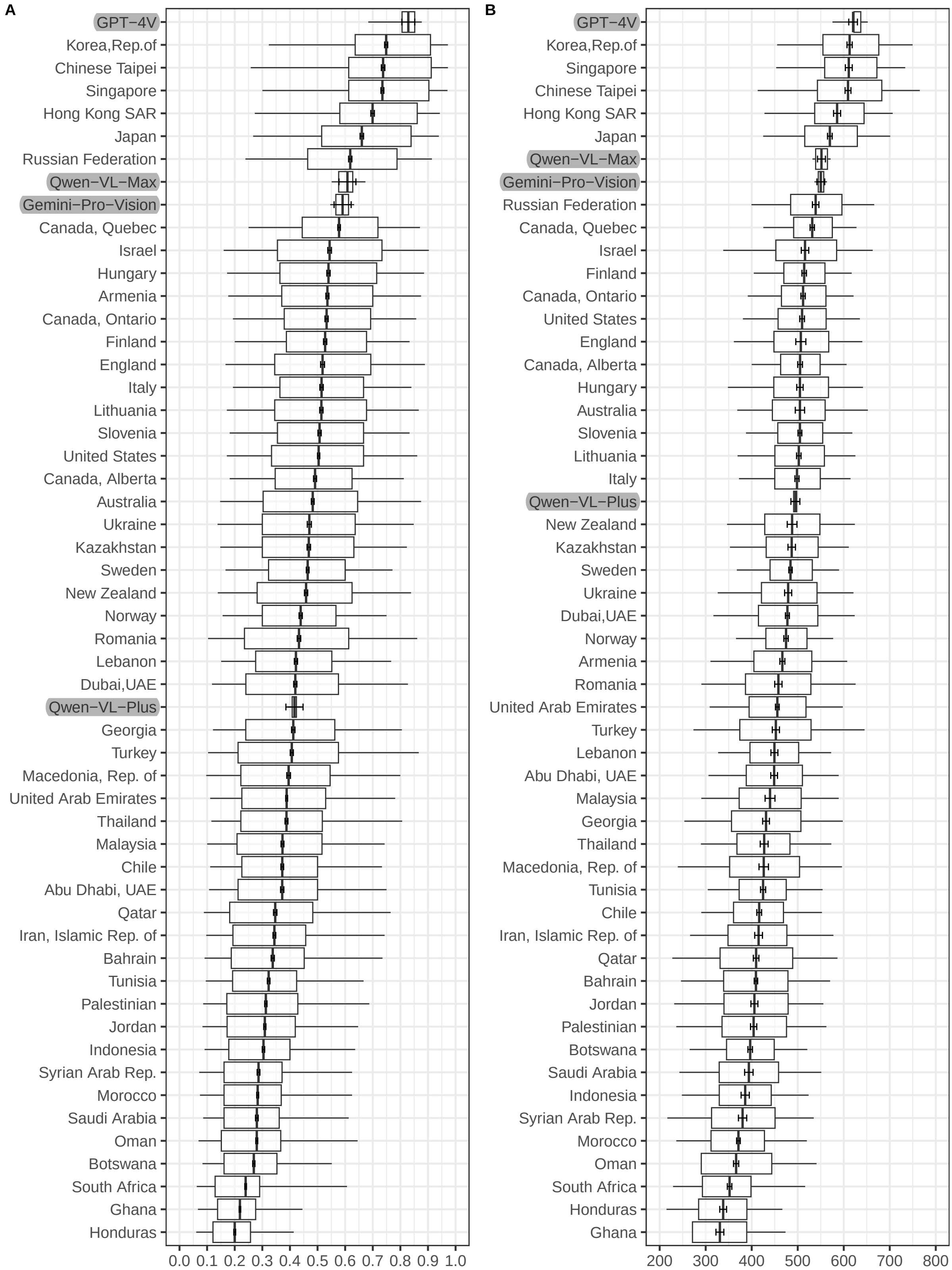


Figure 3: **Distribution of proficiency estimates for GPT-4V, Gemini-Vision-Pro, Qwen-VL-Plus, Qwen-VL-Max and all participating regions of TIMSS 2011 8th grade mathematics test.** Left figure (A) shows the proficiency estimates based on the percentages of correct responses. Right figure (B) shows the 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. *Note that we adhere to the official naming conventions of TIMSS 2011 when reporting the names of participating regions, with no intent to offend anyone*.