Evaluating Large Language Models through Multidimensional Item Response Theory: A Comprehensive Case Study on ENEM

Leonardo Taschetto[1] **, Renato Fileto**[1]

[1] Dept. of Computer Science, Federal Univ. of Santa Catarina, Florianópolis-SC, Brazil

Abstract. *LLM evaluations on tasks like high-stakes multidisciplinary tests still rely on raw accuracy, a metric that weights easy and difficult questions equally and ignores guessing. To help bridge this methodological gap, we repurpose the official three-parameter logistic Item Response Theory (IRT) calibration that the Brazilian education authority (INEP) uses to score humans on the Exame Nacional do Ensino Médio (ENEM), and apply it to LLM responses. We then fit a four-dimensional 3-PL model aligned with ENEM's knowledge domains. Results show that similar accuracies can mask proficiency gaps exceeding one standard deviation across domains. Mathematics remains the toughest domain for both humans and models, whereas questions on Human Sciences are systematically easier for both.*

1. Introduction

Large Language Models (LLMs) now reach state-of-the-art (SOTA) performance on high-stakes multidisciplinary tests, notably national university admission exams such as the SAT in the United States [OpenAI 2024a], the Gaokao in China [Zong and Qiu 2024], and Brazil's ENEM [Abonizio et al. 2024]. These tests have been used as benchmarks to compare LLM capabilities. However, LLM performance studies on these exams mostly report a single metric – accuracy – which offers only a coarse view of test-taker proficiency. It weighs easy and hard questions equally, ignores how well each question differentiates examinees who are stronger or weaker in particular abilities, and fails to adjust for the non-zero chance of correctly guessing answers.

Meanwhile, in the realm of human testing, where reliable ranking of candidates is critical, exam authorities apply Item Response Theory (IRT) [Baker 2001] to calibrate each question weight on the candidate score according to how sharply it discriminates high- and low-performing examinees. Thereby, it can produce scaled proficiency scores that distinguish candidates who achieve the same number of correct answers, according to the relevance of each question answered correctly and incorrectly for scoring proficiency in particular abilities. IRT also takes into account question difficulty, based on the percentage of a population that answers it correctly, and the probability of guessing.

A few recent studies applying IRT to evaluate LLMs on university-admission exams [Zhang et al. 2023, Zong and Qiu 2024] show that IRT provides finer-grained rankings than accuracy alone. However, they do not exploit IRT distinct dimensions to assess specific abilities as we propose in this work. In addition, to the best of our knowledge, no work has applied IRT to evaluate LLMs on the Portuguese-language ENEM. Motivated by this gap, we ask two linked questions. First, using uni-dimensional IRT scoring, how do SOTA LLMs of varying sizes compare with human candidates across the ENEM's four knowledge domains? Second, once baselines are established, how do multidimensional IRT (MIRT) shift each model's performance within those domains?

The major contributions of this paper are: (i) replication of ENEM's official unidimensional IRT models to evaluate LLM performance on this exam on the human scale; (ii) an extended analysis using a multidimensional IRT model aligned with the exam's official four knowledge domains. Our experimental results show that identical accuracy does not imply identical proficiency across knowledge domains, and illustrate how these differences depend on model architecture.

2. The Multidimensional Item Response Theory

The Item Response Theory (IRT) is a family of mathematical models used to measure individuals' latent abilities, i.e., unobservable characteristics or proficiencies (e.g., numerical reasoning, reading comprehension, scientific knowledge, logical thinking), indirectly from exam responses. An exam E comprises a set of N *assessment items* (e.g., questions, tasks). It can be administered to a population of size J for measuring D latent abilities from the responses of each examinee j to each assessment item i ($N, J, D \in \mathbb{N}^+$, $1 \leq i \leq N$, $1 \leq j \leq J$). The central feature of IRT is the characterization of each assessment item i using one or more parameters: *difficulty* (b_i), *discrimination* (a_i), and *guessing* (c_i). Each parameter addresses a distinct aspect of an assessment item i on measuring a latent ability k ($1 \leq k \leq D$):

Discrimination parameter (a_i): is the degree to which an assessment item i distinguishes test-takers who have high ability or proficiency levels from those who have low ones. An item i with high discrimination will provide more information about a test-taker's ability or proficiency level than an item j with low discrimination.

Difficulty parameter (b_i): is the level of ability or proficiency required to have a 50% chance of answering item i correctly. Items requiring higher proficiency are deemed difficult, whereas those answered correctly by most test-takers are considered easy.

Guessing parameter (c_i): is the probability of a test-taker answering item i correctly by chance. In multiple choice items, guessing is the probability of choosing the correct answer among the available alternatives.

In the multidimensional IRT, $\Theta_j \in R^D$ is a D-dimensional vector containing the values of the $D > 1$ latent abilities of the examinee j in an exam E. A measure $\Theta_j[k]$ of the latent ability k for the examinee j is directly proportional to the probability $P(X)$ of this examinee correctly answering each of the N assessment items. Let X_{ij} be the binary response (1 for a correct response, 0 otherwise) of person j to item i. IRT models frequently use the standard logistic link function $\sigma(z) = \frac{1}{1+\exp(-z)}$ to represent $P(X_{ji} = 1|\Theta_j[k], z)$, i.e., the probability of the j^{th} examinee answering the i^{th} item correctly, with z being a function of parameters a_i, b_i and c_i. The three most common IRT models are:

One-parameter logistic model (1-PL): The *"Rasch" model* [Chow et al. 2024] only accounts for item difficulty.

$$P(X_{ij} = 1 \mid \Theta_j, b_i) = \frac{1}{1 + e^{-(\Theta_j - b_i)}}$$

Two-parameter logistic (2-PL): now includes *discrimination* $-\mathbf{a_i}$ and *difficulty* b_i. $z = e^{-\mathbf{a_i}(\Theta_j - b_i)}$

$$P(X_{ij} = 1 \mid \Theta_j, -\mathbf{a_i}, b_i) = \frac{1}{1 + e^{-\mathbf{a_i}(\Theta_j - b_i)}}$$

Three-parameter logistic (3-PL): A parameter c_i captures guessing:

$$P(X_{ij} = 1 \mid \Theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\,\frac{1}{1 + e^{-a_i(\Theta_j - b_i)}} \tag{1}$$

Figure 1 illustrates the isolated impact of the parameters a, b and c on the probability of a correct response from examinee j to item i: (i) shows how the difficulty b_i shifts the curve horizontally; (ii) illustrates the impact of a_i on slope; and (iii) shows the vertical shift induced by c_i.

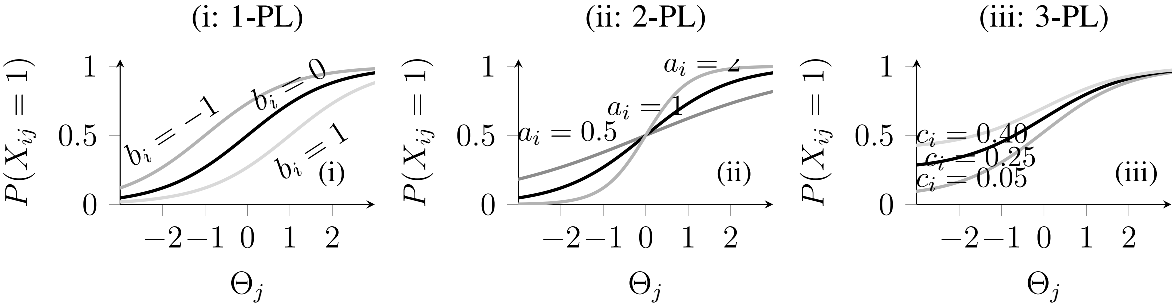

Figure 1. The impact of the parameters a, b and c.

Once parameters a_i, b_i, c_i have been calculated for every assessment item i, they are used to estimate the latent ability vector Θ_j for each examinee j, from their binary response vector X_j. The *Expected-a-Posterior (EAP)*[Bassett and Deride 2016] is a Bayesian approach to estimate proficiency across attributes by employing both observed item responses and prior information, defined as: $\widehat{\theta}_{\text{EAP},j} = \frac{\int \theta\, L_j(\theta) f(\theta)\, d\theta}{\int L_j(\theta) f(\theta)\, d\theta}, \quad where \quad L_j(\theta) = \prod_{i=1}^{N} P\big(X_{ij} = x_{ij} \mid \theta\big)$, is the joint likelihood, and $f(\theta)$ is a known prior distribution such as the standard-normal distribution. Details on the numerical implementation are given in Step (iv), Section 5 .

3. Related Works

We conducted a bibliographical search for peer-reviewed studies about the use of IRT and other metrics to assess LLMs' performance on exams like ENEM. We considered works published between Jan 2020 and Apr 2025 in the following academic databases: ACL Anthology, arXiv, IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, and ScienceDirect. The search expression employed was: *(IRT OR MIRT OR accuracy OR performance) AND (LLM or language model) AND (SAT OR Gaokao OR ENEM OR 'university-admission exam').* We also complemented this review using OpenAI's *Deep Research* tool [OpenAI].

Since we did not find in the results any study applying IRT-based LLM evaluation specifically to the ENEM exam, we broadened our inclusion criteria to address two complementary goals: (i) studies that applied IRT to evaluate LLM performance on other university-admission exams, and (ii) studies that evaluated LLM performance on the 2022 and 2023 ENEM editions, using accuracy metrics.

We found only two papers proposing IRT-based LLM evaluation on university-admission exams, both examining the Chinese Gaokao exam. [Zhang et al. 2023] mapped the performance of various LLMs onto a Rasch-calibrated (1PL) scale enabling difficulty-aware comparisons, beyond raw accuracy. Subsequently, [Liu et al. 2025] extended this approach to a different Gaokao subset, observing that even top-performing LLMs exhibited unexpected errors on easier questions, highlighting their insensitivity to item difficulty. Both studies relied exclusively on uni-dimensional 1-PL IRT models. However, as discussed in Section 2, the uni-dimensional IRT may not adequately capture the interdisciplinary nature of university-admission exams, and the 1-PL IRT model may neglect aspects such as item discrimination and guessing behavior.

Since 2017, the ENEM has increasingly served as a key benchmark for evaluating AI advancements in Portuguese-language contexts [Silveira and Mauá 2017, Silveira and Mauá 2018]. Recently, substantial progress has been marked by the emergence of advanced LLMs, notably the Sabiá[1], family of models [Pires et al. 2023a, Abonizio et al. 2024], which demonstrated impressive performance on the ENEM 2022 and 2023 exams. Evaluations indicate that the largest variant, *Sabiá-3* reached an accuracy of approximately 87.7%, comparable to state-of-the-art models like GPT-4o and Claude-3.5 Sonnet. Intermediate-sized variants, *Sabiazinho-3* and *Sabiá-2 Medium*, also delivered strong results, achieving accuracies around 82.7% and 71.8%, respectively [Abonizio et al. 2024]. Additionally, smaller LLMs have been shown to significantly benefit from advanced prompting strategies, narrowing the performance gap compared to larger models [Superbi et al. 2024, Taschetto and Fileto 2024].

Our study builds directly upon the open-source[2] contributions and methodological groundwork laid by [Nunes et al. 2023] and extended by [Pires et al. 2023b]. Their curation of the question datasets detailed in Section 4.1 provided an essential foundation for our research. Additionally, their forward-looking suggestion to explore ENEM's IRT-based evaluations directly inspired our approach.

In this work, we use the official uni-dimensional parameters as a baseline. We leverage it to enable direct comparisons of human and LLM proficiency by positioning their ENEM responses into the same unified latent-ability scale. In addition, since ENEM aims to assess performance in four knowledge domains, as reflected in its structure, we also propose and implement a four-dimensional 3-PL MIRT model. This multidimensional approach explicitly models latent correlations among domains, offering a more detailed characterization of proficiency compared to the uni-dimensional baseline alone.

4. Methodology

We evaluate LLMs' performance on ENEM through uni- and multidimensional IRT using a process consisting of four sequential steps:

[1] https://www.maritaca.ai/en

[2] https://github.com/piresramon/gpt-4-enem

(i) Generate LLM responses: Collects model outputs for proficiency estimation.

(ii) Estimate LLM uni-dimensional abilities: Uses the official IRT parameters to position LLM responses onto the exam proficiency scale. This allows consistent comparison between LLMs and human examinees by estimating a single proficiency parameter (θ_j) for each LLM j.

(iii) Extend the uni-dimensional IRT to MIRT: Fits a 3-PL MIRT model (Eq. 1, Section 2) using human responses from ENEM questions. In this paper, the MIRT dimensions are the four knowledge domains of ENEM.

(iv) Estimate multidimensional abilities: Builds the latent abilities vector Θ_j within the MIRT scale calibrated for both humans and LLMs.

4.1. Datasets

The ENEM exam covers four knowledge domains – Mathematics, Natural Sciences (Physics, Chemistry, Biology), Human Sciences (History, Geography, Philosophy, Sociology), and Languages (Portuguese, Foreign Languages, Literature) – each domain comprises 45 assessment items, totaling 180. Although an additional essay impacts examinees' final score, it is not evaluated using IRT and thus beyond the scope of this study.

We rely on three datasets for each exam year analyzed (2022 and 2023): (i) ENEM's assessment items, consisting of multiple-choice questions, visual elements (e.g., tables, figures), and official answer keys; (ii) official uni-dimensional IRT parameters; and (iii) full record of examinee responses for each exam administration. Details about these datasets are in the following.

Assessment items: This dataset[3], used in the *step (ii)* of our experiments (Section 5), was assembled by [Nunes et al. 2023]. It contains assessment items from the 2022 ENEM exam. Later, [Pires et al. 2023b] added the 2023 items and official textual descriptions of visual elements to enable fair comparison between text-only and multimodal LLMs. One Mathematics item was annulled in both 2022 and 2023, reducing each year's exam to 179 valid assessment items.

IRT parameters: This dataset, used in *step (ii)* of our experiments, provides the official ENEM values for the uni-dimensional 3-PL parameters a_i (discrimination), b_i (difficulty), and c_i (guessing), described in Section 2. It covers 2022 and 2023, though the Brazilian Institute of Educational Studies and Research (INEP) calibrates the values according to the examinee responses each year, ensuring comparability of proficiency scores for all ENEM administrations [INEP 2021]. In addition to the IRT parameters, this dataset contains each assessment item's unique identifier and its position across different exam booklet versions. This information is essential to accurately map each LLM response to the corresponding assessment item parameter, enabling direct comparability with human responses.

ENEM microdata[4]: Released annually by INEP since 1998, this dataset consists of anonymized individual candidate responses to assessment items, demographic and socioeconomic information collected through a self-reported questionnaire, and administrative metadata (e.g., attendance status, exam booklet codes, foreign-language option). Our study employs the datasets from the 2022 [INEP 2022] and 2023 [INEP 2023] exam editions in Steps (iii) and (iv). Although INEP does

[3] `https://huggingface.co/datasets/maritaca-ai/enem`

not fully disclose its equating methodology for the 2009 baseline [INEP 2021], microdata enable approximate replication of official uni-dimensional scoring.

4.2. LLM selection

Given the strict page limit, we restrict our evaluation to a subset of SOTA LLMs selected according to three criteria: (i) the *model size*, defined as the number of parameters used for inference; (ii) the *licensing type* (proprietary, open-weights[5] or open-source), and (iii) the *training emphasis* (instruction-tuned vs. reasoning-oriented). Guided by these criteria, we selected ten models:

- **DeepSeek R1–0528** [DeepSeek-AI 2025] (671 B parameters, 37 B active; open-source; reasoning-optimized, released Jan 20, 2025) and its sibling **DeepSeek V3–0324** [DeepSeek-AI 2024] (identical size and license; instruction-tuned, released Dec 25, 2024) represent the open-source large-model tier.
- OpenAI's **GPT-4o** [OpenAI 2024a] (instruction-tuned), **O1** and **O3** [OpenAI 2024c] (reasoning-optimized), with undisclosed parameter counts.
- **Llama 4 Maverick** [Meta AI 2024a](400 B parameters, 17 B active; open-weights;) anchors the open-weights baseline, while its lighter sibling **Llama 4 Scout** [Meta AI 2024b](109 B parameters, 17 B active ; up to 10 M-token context; ∼40 T training tokens) provides a lightweight comparator.
- **GPT-4o Mini** [OpenAI 2024b] (instruction-tuned), **O3 Mini** [OpenAI 2025], and **O4 Mini** [OpenAI 2025] (both reasoning-optimized), each with approximately 8 B-parameters, complete the subset as compact comparators.

Prompting strategies: This study follows the pipeline proposed in [Nunes et al. 2023], which includes two prompting strategies:

Few-shot: Each prompt opens with the official ENEM question header followed by three examples, one for each domain: *Languages*, *Human Sciences*, and *Mathematics*. These examples explicitly indicate the correct alternative.

CoT: The same triad of examples is retained, but now each one includes its full step-by-step solution, implementing the CoT prompting paradigm (see [Wei 2022]). The model must: (i) produce its own reasoning, (ii) select the correct alternative, and (iii) briefly justify the exclusion of the remaining options.

4.3. Computational Implementation

INEP computes IRT parameters and estimates examinee proficiency using the proprietary software BILOG-MG, developed by *Scientific Software International (SSI)* [INEP 2021]. However, a single-user license costs US$10,920[6]. Thus, we considered open-source programming languages with dedicated MIRT libraries, such as Python (`py-irt`), Julia (`IRT.jl`), and Java (`bmirt`). However, ultimately we selected R [R Core Team 2025] due to its dedicated, peer-reviewed [Chalmers 2012] IRT modeling package, `mirt`[7]. Additionally, R provides vectorized numerical computation, built-in support for parallel computing, and an extensive package ecosystem via the Comprehensive R Archive Network (CRAN). These combined strengths make R a suitable choice for allowing easy reproduction of our IRT research.

[5] "Open-weights" licenses release the trained weights for inference and fine-tuning while withholding full training data and source code.

[6] `https://ssilive.com/bilogmg-operational` Consulted in May 2025

[7] `https://CRAN.R-project.org/package=mirt`

5. Experiments

The experiments follow the steps described in section 4. Details of each step are presented below. The code, datasets and results employed are available on GitHub[8].

Step (i): Generate LLM responses. All LLM evaluations were executed with the open-source[9] pipeline of [Nunes et al. 2023], a wrapper around the *lm-evaluation-harness framework* [EleutherAI 2024]. We instrumented the code to log every LLM complete response, because steps (ii) and (iv) require the binary response matrix. ENEM uses multiple color-coded test booklets, in which items are identical, but appear in different orders to deter copying. Thus, our logger also records (i) booklet color and (ii) item position per LLM answer. Decoding hyperparameters were set to minimize stochasticity: `temperature=0.0` (greedy) and `top_p=1.0` (no token filtering). 54 pipelines were run, yielding 19,332 item-level responses spanning all combinations of ten LLMs, 3 prompting strategies, and 358 assessment items across two exam years. Logged responses were then recoded as correct/incorrect and stacked into a 54×358 binary response matrix. This matrix, joined with the respective booklet code and item positions, was persisted for use in Steps (ii) and (iv).

Step (ii): LLM uni-dimensional θ estimation. As documented by INEP [INEP 2021], for each examinee j, θ_j is computed by the *expected-a-posteriori* (EAP) estimator using the 3-PL parameters a, b, c for each assessment item and the examinee's binary response matrix X_j. Eq. 2 defines the EAP estimator, the integral form is on the left and its Gauss–Hermite approximation on the right.

$$\widehat{\theta}_{\text{EAP},j} = \frac{\int \theta \, L_j(\theta) f(\theta) \, d\theta}{\int L_j(\theta) f(\theta) \, d\theta} \approx \frac{\sum_{k=1}^{Q_p} \theta_k L_j(\theta_k) w_k}{\sum_{k=1}^{Q_p} L_j(\theta_k) w_k}, \quad where: \tag{2}$$

$L_j(\theta) = \prod_{i=1}^{N} P\big(X_{ij} = x_{ij} \mid \theta\big)$ (see Eq. 1 in Sec. 2) is the joint likelihood, $f(\theta)$ is a defined prior distribution (see [Bassett and Deride 2016]) and Q_p is the number of quadrature points for numerical integration. INEP reported[INEP 2021] to use $Q_p = 40$ and $f(\theta) = \mathcal{N}(0, 1)$ (standard-normal distribution). The EAP score is a weighted average across all plausible proficiency levels, where each level is weighted by how likely it is, given the observed responses (likelihood) and prior assumptions about the ability distribution: $f(\theta)$. The estimation was executed independently for each LLM result from Step (i), producing 54 distinct θ_j scalars.

Step (iii): 3-PL MIRT model. The item discrimination (a), difficulty (b) and guessing (c) parameters were estimated with a four-dimensional 3-PL MIRT model fitted to the complete binary response matrix derived from the datasets described in Section 4.1 (iii). While INEP uses the Expectation-Maximization (EM) algorithm in uni-dimensional IRT (see [INEP 2021]), the present study opts for the Quasi–Monte Carlo EM (QMCEM) algorithm (see [Chalmers 2012]), for its faster convergence. Iterations stopped when the maximum relative parameter change dropped below 10^{-4} or after 100 cycles. Parallelism was enabled via `mirtCluster` (*`mirt` v. 1.3, `R` v. 4.4.0*) to distribute threads without duplicating the response matrix.

[8] https://github.com/leonardotaschetto/enem-mirt-eval/

[9] https://github.com/piresramon/gpt-4-enem

The experiments ran for 11h on a 24-core Ryzen 9, with 64 GB RAM. Factor correlations were moderate (Languages — Human Sciences 0.48; Natural Sciences – Mathematics 0.42; cross-domain $|\rho| < 0.30$). The final set of 358 calibrated (a, b, c) items was persisted to a parquet file, keyed by the official booklet code and item position.

Step (iv): Multidimensional θ estimation. The EAP routine described in Step (ii) is generalized to estimate D-dimensional ability vectors $\hat{\mathbf{\Theta}}_j \in \mathbb{R}^D$. Given the a_i, b_i, c_i item parameters from Step (iii), the intuition is analogous to Step (ii): MIRT computes $\hat{\mathbf{\Theta}}_j$ as weighted averages over multi-dimensional proficiency profiles. It retains the same $Q_p = 40$ and $f(\Theta) = \mathcal{N}(0, I)$. Formally, $\hat{\mathbf{\Theta}}_j \approx \frac{\sum_{\mathbf{k}} \boldsymbol{\theta}_{\mathbf{k}} L_j(\mathbf{X}_j|\boldsymbol{\theta}_{\mathbf{k}}) w_{\mathbf{k}}}{\sum_{\mathbf{k}} L_j(\mathbf{X}_j|\boldsymbol{\theta}_{\mathbf{k}}) w_{\mathbf{k}}}$, where L_j is the joint likelihood for configuration j, w_k is the k-th Gauss–Hermite weight, and each vector $\boldsymbol{\theta}_k$ encodes a specific combination of proficiency levels across the four latent dimensions.

This work adopts $D = 4$ (section 4), with each MIRT dimension aligned to one of ENEM's knowledge domains: Human Sciences (HS), Natural Sciences (NS), Languages and Codes (LC) and Mathematics (MT). Running calculations for the 54 configurations generated in Step (i) yields a set of four-dimensional estimates that we stack row-wise into the matrix $\Theta \in \mathbb{R}^{54\times 4} \quad where \quad \Theta^{(j)} = \left(\Theta^{(j)}_{\text{HS}}, \Theta^{(j)}_{\text{NS}}, \Theta^{(j)}_{\text{LC}}, \Theta^{(j)}_{\text{MT}}\right)$. Row j records the latent abilities for configuration j, with the columns corresponding to the ENEM domains.

6. Results

Table 1 lists raw LLM accuracies from Step (i); the respective uni-dimensional IRT scores θ from Step (ii); and the accuracies and MIRT estimates Θ_{MT}, Θ_{NS}, Θ_{HS} and Θ_{LC} calculated in Step (iv) for the $D = 4$ ENEM domains.

Table 1. Models' accuracy, IRT and MIRT-based proficiencies.

Model	IRT		MIRT ($D = 4$)							
	(uni-dim)		Math		Natural Sc.		Human Sc.		Languages	
	acc.	θ	acc.	Θ_{MT}	acc.	Θ_{NS}	acc.	Θ_{HS}	acc.	Θ_{LC}
O3	94.44	2.74	84.44	2.68	95.56	2.64	100.00	2.79	97.78	2.32
O1	91.67	2.25	80.00	2.34	93.33	2.39	97.78	2.06	95.56	1.99
O3 Mini	90.50	2.24	84.09	2.37	93.33	2.67	97.78	2.56	86.67	1.38
O4 Mini	84.44	1.74	73.33	2.02	93.33	2.49	91.11	1.25	80.00	0.98
DeepSeek R1	75.42	1.27	68.18	1.69	66.67	1.70	82.22	1.20	84.44	1.22
GPT-4_1	74.86	1.73	29.55	-0.42	80.00	1.79	95.56	2.04	93.33	1.83
DeepSeek V3	72.07	1.70	36.36	0.41	77.78	1.81	82.22	1.07	91.11	1.47
GPT-4o Mini	63.13	1.20	20.45	-0.86	62.22	1.23	82.22	1.17	86.67	1.31
Llama 4 Maverick	58.10	0.74	31.82	-0.62	44.44	0.34	77.78	0.97	77.78	1.12
Llama 4 Scout	52.51	0.74	27.27	0.80	28.89	-0.56	80.00	0.86	73.33	0.81

The LLMs with the second and the third highest general accuracy – O3 Mini (90.5 %) and O1 (91.7 %) – achieve almost identical uni-dimensional θ (2.24 vs 2.25). However, MIRT reveals that O3 Mini is slightly stronger than O1 in Mathematics ($\Theta_{MT} = 2.37$) and much more in Natural Sciences ($\Theta_{NS} = 2.67$), while O1 has its edge in NS ($\Theta_{NS} = 2.39$). Conversely, GPT-4o and DeepSeek R1 differ by less than one percentage point in accuracy (74.9% vs 75.4%), but GPT-4o's $\Theta_{MT} = -0.42$ contrasts with DeepSeek R1's solid $+1.69$, reversing their ranking for any math-heavy use case.

Figure 2 enables comparison of human and LLM performance on the four domains, in the official ENEM score scale (300 – 1000). The 5th–95th percentile of human candidates is in translucent cyan; darker close to the median. For each model, the *solid* line links its latent abilities, while the *dashed* line links its accuracies on the four domains, projected into the same scale. For most models and domains, the ability score (Θ) is lower than the respective accuracy projected in the same scale. One exception is O3 in Math, highlighting its superior abilities.

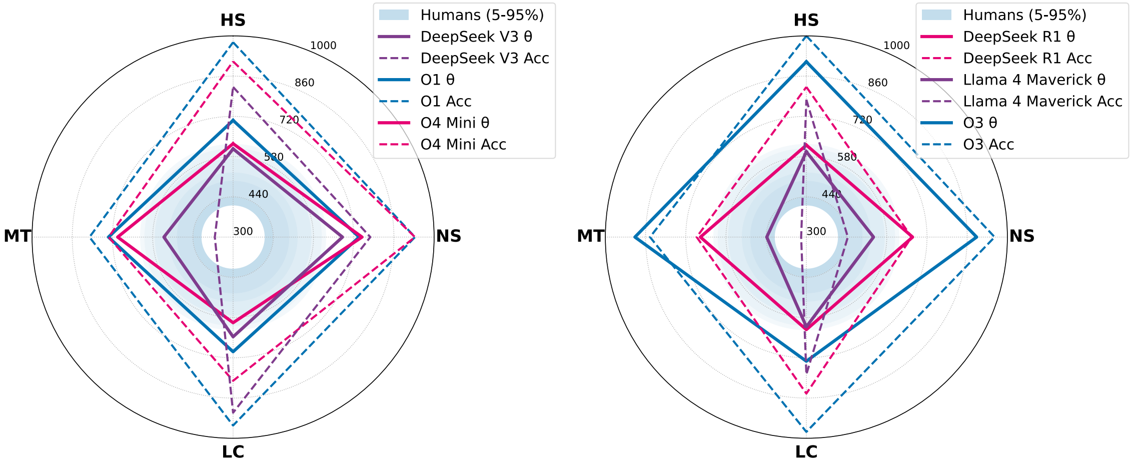

Figure 2. Four-domain abilities versus accuracies of ENEM candidates and LLMs

Figure 3 compares MIRT results from Step (iv) with a baseline of 665 421 candidates who were tested in the same booklet codes used for model evaluation, yielding a mean accuracy of 39.8% and $P_{75} = 47.8\%$. In each panel the vertical axis shows raw item-level accuracy (0–1), while the horizontal axis shows the multidimensional-IRT score mapped to the official ENEM scale (300–1000). Once raw accuracies are projected onto the ENEM proficiency scale, top-tier models (e.g.,O3) cluster within the top one human percentile, yet the grey student cloud reveals that many weaker candidates can share the same score band if they happen to guess well on high-information items.

Domain-wise, five LLMs exceed 95% accuracy where students average 28% and 49%, respectively. Mathematics remains the most challenging section; while leading engines break 80%, *GPT-4o Mini* and the two Llama variants fall below the human mean of 34%, highlighting a persistent gap in quantitative reasoning.

Table 2 quantifies how much the joint $D = 4$ MIRT model re-weights proficiency with respect to the conventional, domain-isolated approach adopted by INEP. The domain-specific columns of the four-factor MIRT output were aligned and subtracted from their uni-dimensional counterparts, yielding the difference matrix. Positive values ($\Delta > 0$) indicate that the joint model *up-weights* the ability estimate after borrowing information from inter-domain covariance; negative values mean that the uni-dimensional model was overly optimistic once cross-domain evidence is considered.

7. Conclusions and Future Work

This study demonstrates two key advantages of IRT scoring over raw accuracy when evaluating LLMs on the ENEM:

(i) Uni-dimensional IRT: weighting each response according to item difficulty, discrimination, and guessing, produces global proficiency scores that are substantially

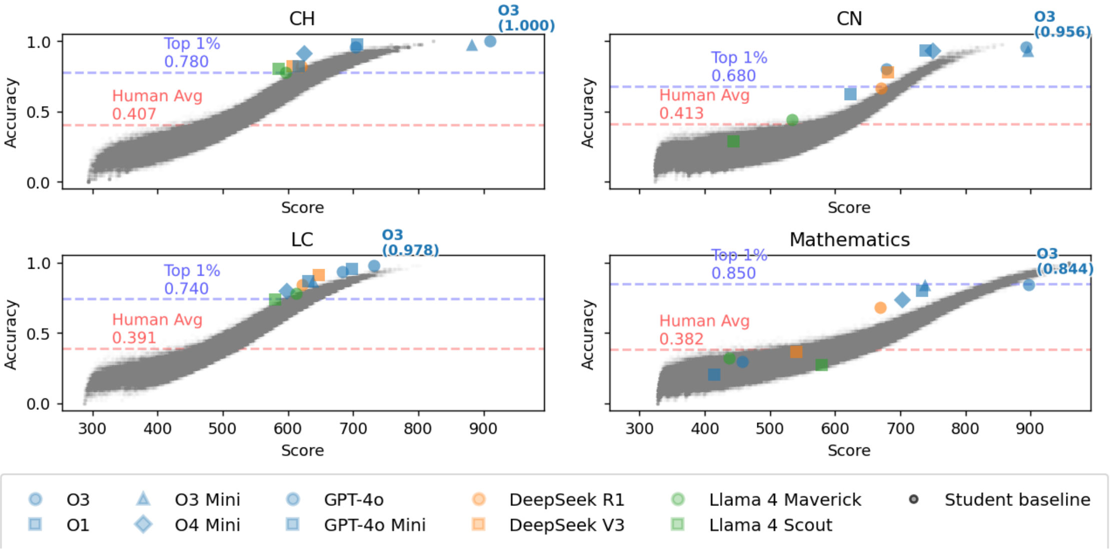

Figure 3. Accuracy x MIRT score of ENEM students and language models.

Table 2. Domain-wise comparison between independent uni-dimensional IRT (θ^{Uni}) and joint 4-factor MIRT (Θ). $\Delta = \Theta - \theta^{\text{Uni}}$.

Model	Math (MT)			Natural Sc. (NS)			Human Sc. (HS)			Languages (LC)		
	θ^{Uni}	Θ	Δ	θ^{Uni}	Θ	Δ	θ^{Uni}	Θ	Δ	θ^{Uni}	Θ	Δ
O3	2.63	2.68	+0.05	2.54	2.64	+0.10	2.69	2.79	+0.10	2.22	2.32	+0.10
O1	2.29	2.34	+0.05	2.31	2.39	+0.08	1.96	2.06	+0.10	1.89	1.99	+0.10
O3 Mini	2.32	2.37	+0.05	2.59	2.67	+0.08	2.46	2.56	+0.10	1.33	1.38	+0.05
O4 Mini	2.00	2.02	+0.02	2.41	2.49	+0.08	1.17	1.25	+0.08	0.93	0.98	+0.05
DeepSeek R1	1.71	1.69	–0.02	1.72	1.70	–0.02	1.15	1.20	+0.05	1.17	1.22	+0.05
GPT-4_1	–0.30	–0.42	–0.12	1.74	1.79	+0.05	1.94	2.04	+0.10	1.75	1.83	+0.08
DeepSeek V3	0.51	0.41	–0.10	1.79	1.81	+0.02	1.02	1.07	+0.05	1.39	1.47	+0.08
GPT-4o Mini	–0.74	–0.86	–0.12	1.25	1.23	–0.02	1.12	1.17	+0.05	1.26	1.31	+0.05
Llama 4 Maverick	–0.52	–0.62	–0.10	0.42	0.34	–0.08	0.95	0.97	+0.02	1.10	1.12	+0.02
Llama 4 Scout	0.92	0.80	–0.12	–0.44	–0.56	–0.12	0.81	0.86	+0.05	0.79	0.81	+0.02

more reliable and comparable to human results than simple percent-correct.

(ii) Four-factor MIRT: retains that reliability while decomposing proficiency into four domains: MT (Mathematics), NS (Natural Sciences), HS (Human Sciences), and LC (Languages and Codes). This profile highlights domain-specific strengths and weaknesses that a single-number IRT score may conceal.

Future work includes a deeper statistical comparison of the MIRT 3-PL model with the official uni-dimensional (e.g., using likelihood-ratio tests, information criteria, and item/ability fit plots) model providing visual and numerical diagnostics that clarify when the richer multidimensional view materially changes conclusions about model or human proficiency; it might also extend the calibration to other ENEM years and examine how specific prompting strategies interact with latent abilities across domains.

Acknowledgements: Thanks to the authors of [Pires et al. 2023b, Nunes et al. 2023] for their commitment to open-source, which significantly facilitated this work. This work was supported by a 2022 CNPq Universal grant, FAPESC grant 2021TR1510, and indirectly by the Céos project, financed by the Public Ministry of Santa Catarina State (MPSC).

References

[Abonizio et al. 2024] Abonizio, H. Q., Almeida, T. S., Laitz, T. S., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2024). Sabiá-3 technical report. *CoRR*, abs/2410.12049.

[Baker 2001] Baker, F. B. (2001). *The Basics of Item Response Theory ISBN 1-886047-03-0*. Heinemann, second edition.

[Bassett and Deride 2016] Bassett, R. and Deride, J. (2016). Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, 174.

[Chalmers 2012] Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6):1–29.

[Chow et al. 2024] Chow, J. C., Cheng, T. Y., Chien, T.-W., and Chou, W. (2024). Assessing chatgpt's capability for multiple choice questions using raschonline: Observational study. *JMIR Form Res*, 8:e46800.

[DeepSeek-AI 2024] DeepSeek-AI (2024). Deepseek-v3 technical report. arXiv:2412.19437. https://arxiv.org/abs/2412.19437.

[DeepSeek-AI 2025] DeepSeek-AI (2025). Deepseek-r1: Incentivizing reasoning capability in llms via sparse mixture-of-experts. arXiv:2501.12948. https://arxiv.org/abs/2501.12948.

[EleutherAI 2024] EleutherAI (2024). The language model evaluation harness.

[INEP 2021] INEP (2021). Accessed 8 May 2025 https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_procedimentos_de_analise.

[INEP 2022] INEP (2022). Accessed 8 May 2025 "https://download.inep.gov.br/microdados/microdados_enem_2022.zip.

[INEP 2023] INEP (2023). Accessed 8 May 2025 https://download.inep.gov.br/microdados/microdados_enem_2023.zip.

[Liu et al. 2025] Liu, Y., Bhandari, S., and Pardos, Z. A. (2025). Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, (Early View):1–25.

[Meta AI 2024a] Meta AI (2024a). The Llama4 herd: The beginning of a new era of natively multimodal AI innovation. Accessed 16Jun2025.

[Meta AI 2024b] Meta AI (2024b). The Llama4 herd: The beginning of a new era of natively multimodal AI innovation. Announcement of the Llama4 family—including *Llama4Scout*. 17B activated / 109B total parameters; 10M-token context; knowledge-cutoff August 2024. Accessed 16Jun2025.

[Nunes et al. 2023] Nunes, D., Primi, R., Pires, R., Lotufo, R. A., and Nogueira, R. (2023). Evaluating GPT-3.5 and GPT-4 models on brazilian university admission exams. *CoRR*, abs/2303.17003.

[OpenAI] OpenAI. Deep research. Acessed Jun-16,2025. openai.com/index/introducing-deep-research/.

[OpenAI 2024a] OpenAI (2024a). GPT-4o System Card. Accessed 13Jun2025. https://openai.com/index/gpt-4o-system-card/.

[OpenAI 2024b] OpenAI (2024b). GPT-4o mini: Advancing Cost-Efficient Intelligence. Released 18Jul2024 — accessed 13Jun2025. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[OpenAI 2024c] OpenAI (2024c). OpenAI o1 System Card. Updated 5Dec2024 — accessed 13Jun2025 https://openai.com/index/openai-o1-system-card/.

[OpenAI 2025] OpenAI (2025). OpenAI o3 and o4-mini System Card. https://openai.com/index/o3-o4-mini-system-card/. Published 16Apr2025 — accessed 13Jun2025.

[Pires et al. 2023a] Pires, R., Abonizio, H., Almeida, T., and Nogueira, R. (2023a). Sabiá: Portuguese large language models. In *Anais da XII Brazilian Conference on Intelligent Systems*, pages 226–240, Porto Alegre, RS, Brasil. SBC.

[Pires et al. 2023b] Pires, R., Almeida, T. S., Abonizio, H. Q., and Nogueira, R. (2023b). Evaluating gpt-4's vision capabilities on brazilian university admission exams. *CoRR*, abs/2311.14169.

[R Core Team 2025] R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Silveira and Mauá 2017] Silveira, I. C. and Mauá, D. D. (2017). University entrance exam as a guiding test for artificial intelligence. In *2017 Brazilian Conference on Intelligent Systems, BRACIS 2017, Uberlândia, Brazil, October 2-5, 2017*, pages 426–431. IEEE Computer Society.

[Silveira and Mauá 2018] Silveira, I. C. and Mauá, D. D. (2018). Advances in automatically solving the ENEM. In *7th Brazilian Conference on Intelligent Systems, BRACIS 2018, São Paulo, Brazil, October 22-25, 2018*, pages 43–48. IEEE Computer Society.

[Superbi et al. 2024] Superbi, J., Pinto, H., Santos, E., Lattari, L., and Castro, B. (2024). Enhancing large language model performance on enem math questions using retrieval-augmented generation. In *Anais do XVIII Brazilian e-Science Workshop*, pages 56–63, Porto Alegre, RS, Brasil. SBC.

[Taschetto and Fileto 2024] Taschetto, L. and Fileto, R. (2024). Using retrieval-augmented generation to improve performance of large language models on the brazilian university admission exam. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 799–805, Porto Alegre, RS, Brasil. SBC.

[Wei 2022] Wei, J. e. a. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

[Zhang et al. 2023] Zhang, X., yan Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X. (2023). Evaluating the performance of large language models on gaokao benchmark. *ArXiv*, abs/2305.12474.

[Zong and Qiu 2024] Zong, Y. and Qiu, X. (2024). GAOKAO-MM: A Chinese human-level benchmark for multimodal models evaluation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8817–8825, Bangkok, Thailand. Association for Computational Linguistics.