DECAF: DISENTANGLING CAPABILITY AND TRANSLATION ARTIFACTS FOR MULTILINGUAL LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While most large language models (LLMs) support a variety of languages beyond English, accurately evaluating their language-specific capabilities remains a challenge. Multilingual benchmarks are central to this evaluation; however, because they are typically translated from an English baseline using machine translation, the translation process often introduces item-level artifacts, such as translation errors and culturally sensitive content. These translation artifacts obscure the accurate estimation of a model's language-specific capabilities. To address this challenge, this paper introduces the Multilingual Differential Item Functioning Item Response Theory (MDIF-IRT) model. The MDIF-IRT framework advances multilingual LLM evaluation by disentangling language effects into distinct model capability shifts and item difficulty shifts. An application to the MMLU-ProX dataset, evaluating 25 models across 29 languages, demonstrates the utility of the proposed method. The MDIF-IRT model estimates language-specific latent capabilities while accounting for item-specific cross-lingual difficulty variation to improve estimation of capability and item parameters. Furthermore, the framework provides a tool for identifying benchmark items with significant cross-lingual translation artifacts and improves prediction of unobserved item responses compared to empirical accuracy and conventional IRT models.

1 INTRODUCTION

Despite the worldwide adoption of large language models (LLMs), their evaluation practices have primarily focused on the English language (Xuan et al., 2025; Singh et al., 2025; Xu et al., 2025). Recently, multilingual benchmarks, such as MMLU-ProX (Xuan et al., 2025), Global-MMLU (Singh et al., 2025), MGSM (Shi et al., 2022), and MGSM-Pro (Xu et al., 2026) have been developed to evaluate model capabilities across diverse languages. However, evaluating multilingual AI poses unique challenges due to artifacts introduced during benchmark item translation, such as direct translation errors and culturally sensitive item content (Peter et al., 2026; Lior et al., 2026; Singh et al., 2025; Xu et al., 2025). These item-level artifacts obscure the evaluation of the models' true language-specific capabilities (Singh et al., 2025). Consequently, a systematic approach is needed to disentangle a model's language-specific capability from the language-specific characteristics of the benchmark items.

Item response theory (IRT) has been foundational in psychometrics for measuring the latent abilities of human test-takers (Embretson & Reise, 2025). Recently, driven by the demand for systematic evaluation of LLMs, IRT has been adopted for model evaluations and benchmark refinement, providing finer-grained diagnostics (Jiang et al., 2026). For example, IRT models are used to estimate the latent capabilities of LLMs using benchmark evaluation datasets (Schilling-Wilhelmi et al., 2025; Zhou et al., 2026). Furthermore, the estimated item characteristics from IRT models can enhance the efficiency and scalability of AI evaluation by informing benchmark and item quality (Polo et al., 2024; Hofmann et al., 2025; Kipnis et al.).

To date, efforts to integrate the IRT framework into multilingual AI evaluation have been limited. One example in the multilingual AI evaluation literature is the Multilingual-IRT model (Lior et al., 2026), which adopted a multidimensional IRT framework in a multilingual setting. Nevertheless,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

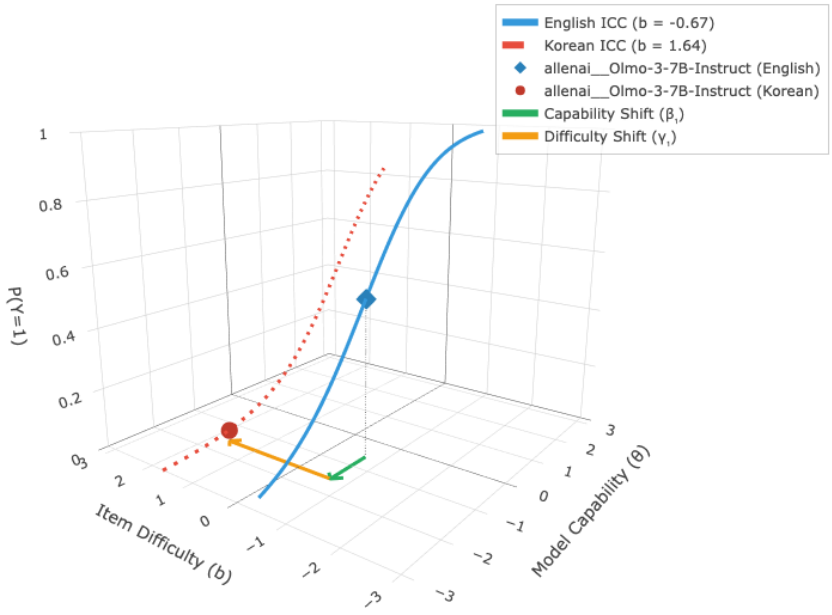


Figure 1: DECAF: Disentangling capability and translation artifacts for multilingual LLM evaluation. The proposed MDIF-IRT model disentangles the impact of language translation on benchmark performance into a model-specific capability shift (β_1) and an item-specific artificial difficulty shift (γ_1). This figure illustrates the Item Characteristic Curves (ICC) for a single benchmark item evaluated in both English (baseline) and Korean. An ICC graphically represents the probability of a correct response as a function of a model's latent capability. While the model's predicted probability of a correct response (z-axis) drops severely from 0.52 (blue dot) to 0.04 (red dot), the visualization reveals this is mostly driven by a drastic increase in item difficulty (from $b = -0.67$ to 1.64) due to translation artifacts, rather than the moderate decrease in the model's language-specific capability (from $\theta = -0.60$ to -1.56). This decomposition allows for a more rigorous evaluation of underlying language-specific capabilities and the detection of translation artifacts.

several limitations exist in this approach. First, the latent capabilities are assumed to follow a normal distribution. However, the empirical distributions of LLM capabilities may deviate from this assumption (Zhou et al., 2026; Jiang et al., 2026), potentially affecting the reliability of capability estimation. Second, although these multilingual benchmarks are initially developed in English and subsequently translated, the existing approach models item difficulty as a composite of a baseline difficulty shared across all languages and a language-specific deviation. Such a formulation prevents the language-specific deviations from being directly interpreted as artifacts of the translation process.

When items are translated into different languages, several factors can drive model performance shifts on multilingual benchmarks. Conceptually, these factors can be disentangled into two distinct sources:

- **Model Capability shifts:** True variations in a model's latent capability across languages, potentially stemming from imbalances in training data resources (Peter et al., 2026), tokenization inefficiencies (Ali et al., 2024), or differential training data contamination (e.g., MGSM translated from the English GSM8K).

- **Item Difficulty shifts:** Artificial variations in item difficulty across languages introduced by the translation process, such as direct translation errors and culturally sensitive content.

In psychometric terms, these artificial difficulty shifts manifest as Differential Item Functioning (DIF; Holland et al. (1993); Osterlind & Everson (2009)) across languages. Such cross-lingual DIF occurs when, conditional on a model's latent capability, the probability of responding correctly to an item differs across language versions. Consequently, cross-lingual DIF obscures the accurate evaluation of a model's underlying capabilities. Modeling and adjusting for these artificial difficulty shifts is therefore critical. By disentangling translation artifacts from true capability variations, researchers can obtain robust capability estimates across languages.

By extending IRT theory and DIF frameworks into multilingual AI evaluation, this study introduces Disentangling Capability and Translation Artifacts for Multilingual LLM Evaluation (DECAF), an evaluation framework centered on the Multilingual Differential Item Functioning Item Response Theory (MDIF-IRT) model, to systematically evaluate the multilingual capabilities of large language models. Isolating language-specific item difficulty shifts from model capability shifts, DECAF enables more interpretable and reliable evaluations. Relaxing potentially violated normality assumptions on model capabilities, we treat them as fixed variables and implement a Gaussian Variational EM (GVEM) algorithm (Cho et al., 2021) for efficient estimation. Furthermore, the framework helps detect potentially problematic items with translation-related distortions, informing benchmark development and refinement.

The remainder of this paper is structured as follows. Section 2 briefly introduces IRT and DIF. Section 3 details the proposed MDIF-IRT model specification. In Section 4, we demonstrate the proposed model's empirical utility using an MMLU-ProX evaluation dataset. Next, Section 5 summarizes our key contributions. Finally, Section 6 discusses the limitations of the current study and outlines directions for future extensions.

2 Item Response Theory

Item response theory has been applied in LLM evaluation for estimating model capabilities and evaluating the psychometric quality of benchmarks. It models the probability of observed item response as a function of the model's capability and item characteristics such as difficulty. Let Y_{ij} be the dichotomous response of model i on a benchmark item j. One popular choice of IRT model, the Rasch model (Rasch, 1993), defines the probability of a correct response, given model latent capability θ_i and item difficulty b_j, as

$$p_{ij} = P(Y_{ij} = 1 \mid \theta_i, b_j) = \sigma(\theta_i - b_j), \tag{1}$$

where σ is the logistic function. We assume local independence, meaning that conditional on a model's latent capability, its responses to different items are independent.

In traditional psychometrics, DIF occurs when test-takers from different demographic groups who possess the exact same level of the underlying latent trait (θ_i) exhibit differing probabilities of answering a specific item correctly. In the context of evaluating LLMs across multiple languages, cross-lingual DIF manifests when an item's difficulty is not invariant across its translated versions. For example, a benchmark item containing a culturally specific idiom or a translation error may become artificially more difficult in a target language. Consequently, this artificial difficulty spike can systematically distort the model's language-specific capability estimation. Therefore, identifying cross-lingual DIF is essential for disentangling a model's true language-specific capability from translation-related artifacts. In our MDIF-IRT model described in the subsequent section, this cross-lingual DIF is handled by introducing a language-shift parameter to the item difficulty.

3 MDIF-IRT Model

Translating a benchmark to different languages can affect both the capability of the LLM and the item difficulty simultaneously. Our proposed DECAF framework models these two-way effects, allowing for 1) estimating the language-specific capability of the model and 2) quantifying the translation effect on item difficulty. The former can inform model selection and training tailored to a specific language. The latter can flag benchmark items with translation artifacts such as translation errors or culturally sensitive item content. The following details the model specification.

Let $i = 1, ..., I$ denote a model, $j = 1, ..., J$ an item, and $l = 1, ..., L$ a language. The correct response probability of model i on item j and language l follows:

$$p_{ijl} = P(Y_{ijl} = 1 \mid \theta_{il}, b_{jl}) = \sigma(\theta_{il} - b_{jl}), \tag{2}$$

where θ_{il} is language-specific model capability, b_{jl} is language-specific item difficulty.

The MDIF-IRT model accounts for shifts in both ability level and item difficulty as a function of language mode (i.e., translation effect). Let $\mathbf{X}$ be a $L \times (L-1)$ design matrix; each row is a dummy-coded vector for each language. The language-specific capability and item difficulties are functions of the language:

$$\begin{aligned} \theta_{il} &= \beta_{0i} + \mathbf{X}_l \boldsymbol{\beta}_{1i}, \\ b_{jl} &= \gamma_{0j} + \mathbf{X}_l \boldsymbol{\gamma}_{1j}, \end{aligned} \tag{3}$$

where $\mathbf{X}_l$ is the l-th row of the design matrix $\mathbf{X}$. β_{0i} is the baseline model capability, and $\boldsymbol{\beta}_{1i}$ is the $L-1$ vector of coefficients, representing language-specific ability shift. Similarly, γ_{0j} is the baseline item difficulty, and $\boldsymbol{\gamma}_{1j}$ is the $L-1$ vector of difficulty shift coefficients.

Let $\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{0i} \\ \boldsymbol{\beta}_{1i} \end{bmatrix}$ and $\boldsymbol{\gamma}_j = \begin{bmatrix} \gamma_{0j} \\ \boldsymbol{\gamma}_{1j} \end{bmatrix}$ be $L \times 1$ model parameter vectors. Let $\tilde{\mathbf{X}}_l = [1 \quad \mathbf{X}_l]$ be a $1 \times L$ design vector. As a result, equation 2 becomes:

$$p_{ijl} = \sigma(\tilde{\mathbf{X}}_l \boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l \boldsymbol{\gamma}_j). \tag{4}$$

Conventional practice in IRT is to assume the capabilities follow a normal distribution and to estimate the parameters using marginal maximum likelihood or Bayesian approaches. However, applying conventional IRT estimation directly to AI benchmark data introduces two primary challenges. First, AI evaluations typically feature a data structure where the number of items vastly exceeds the number of evaluated models ($J \gg I$). This high-dimensional setting renders conventional estimation methods highly unstable and often infeasible (Jiang et al., 2026). Furthermore, a small I frequently yields uniform response vectors (all 0s or all 1s) within specific item-language combinations, rendering maximum likelihood estimates undefined. Overcoming this data sparsity necessitates borrowing information across languages to stabilize parameter estimation. Second, traditional methods assume that latent capabilities θ_i are normally distributed. Because empirical evaluations of LLMs often reveal severe skewness or multimodality in performance, imposing this normality assumption introduces systematic bias into item parameters and capability estimates. While alternative approaches might employ joint maximum likelihood estimation (treating all parameters as fixed), these methods often struggle with computational scalability in high-dimensional settings. Addressing these limitations, our study implements a Gaussian Variational EM (GVEM) algorithm that treats the item parameters as random variables, enabling scalable estimation while relaxing the restrictive normality assumption on the latent capabilities. A multivariate normal prior is set for the item parameters:

$$\boldsymbol{\gamma}_j \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_0), \tag{5}$$

where

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \sigma^2_{\gamma_0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_\gamma \end{bmatrix}. \tag{6}$$

To ensure model identifiability and resolve location indeterminacy, the MDIF-IRT model uses the zero-centered prior on the item parameters ($\boldsymbol{\gamma}_{1j} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\gamma)$). This constraint fixes the expected average difficulty shift within each language to zero. As a result, the framework disentangles translation impacts by loading systemic language proficiency onto the model capability shifts ($\boldsymbol{\beta}_{1i}$), while $\boldsymbol{\gamma}_{1j}$ isolates item-relative artificial shifts.

Let $f(\boldsymbol{\gamma}_j)$ denote the probability density function of the prior distribution. The marginal log-likelihood is:

$$\ell = \sum_{j=1}^{J} \log \left(\int \prod_{i=1}^{I} \prod_{l=1}^{L} \left[p_{ijl}^{Y_{ijl}} (1 - p_{ijl})^{1 - Y_{ijl}} \right] f(\boldsymbol{\gamma}_j)\, d\boldsymbol{\gamma}_j \right). \tag{7}$$

Appendix A presents the details of the GVEM algorithm implemented in this study. A parameter recovery study, presented in Appendix B, demonstrates that the GVEM algorithm reliably recovers the true model parameters under simulation conditions that closely mirror the empirical data structure analyzed in the subsequent section.

4 EXPERIMENT

4.1 DATASET

To demonstrate the empirical utility of the DECAF framework, we evaluate it on the MMLU-ProX (Xuan et al., 2025) dataset, a large-scale multilingual multiple-choice benchmark. Originally developed in English as the MMLU-Pro, the benchmark was extended into 28 additional languages using a semi-automatic translation pipeline. The benchmark consists of $J = 11{,}759$ items spanning 14 subjects, including biology, business, computer science, economics, and psychology. Each item is evaluated across $L = 29$ languages, encompassing the English reference language and a diverse array of high- and low-resource languages across Europe, Asia, and Africa. The observed response data for the $I = 25$ models evaluated in Lior et al. (2026) are publicly available via Hugging Face[1]. The comprehensive list of evaluated models and the specific languages included in the benchmark are detailed in Appendix C.

4.2 ANALYSIS

The proposed MDIF-IRT model was fitted to the $I \times J \times L$ response tensor using the GVEM algorithm described in Appendix A. The computation time for the full dataset ($I = 25$, $J = 11{,}759$, $L = 29$) was 0.11 hours on a single MacBook Pro (Apple M1 processor, 16GB RAM).

For comparison, two variants of the Rasch model were also estimated. First, an independent Rasch model was fitted to the $I \times J$ response matrix for each language to estimate language-specific capabilities separately. Second, a pooled Rasch model was fitted to the flattened $I \times (J \times L)$ response matrix, assuming item difficulties are invariant across languages. Both baseline Rasch models were estimated via joint maximum likelihood using the `sirt` R package (Robitzsch, 2017). The code used to run the analysis in this paper is available at `https://anonymous.4open.science/r/MDIF-IRT-7164`.

4.3 MODEL PARAMETER ESTIMATES

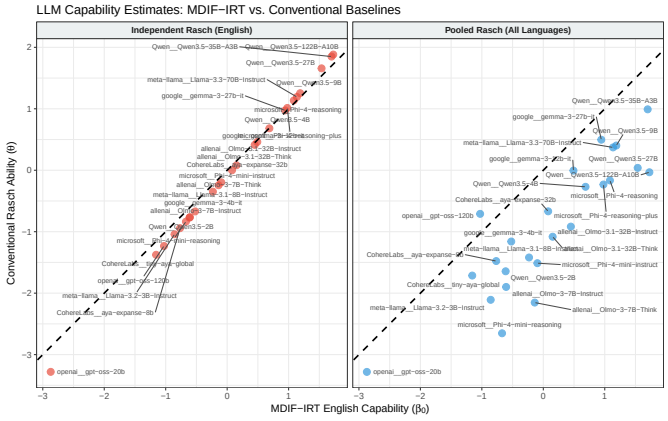


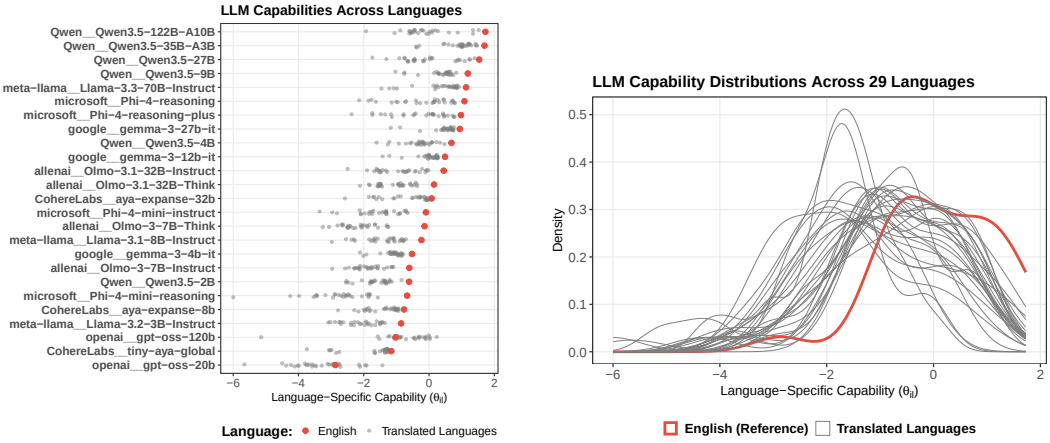
Figure 2: LLM capability estimates: MDIF-IRT vs. Rasch models.

Figure 2 compares the baseline capability estimates for the reference language (English) obtained from the MDIF-IRT model against those from conventional Rasch models. The left panel demonstrates that the English capabilities align closely between the MDIF-IRT model and the Independent Rasch model. In contrast, the right panel illustrates that the Pooled Rasch model generally underestimates the capabilities of the models compared to the MDIF-IRT model by ignoring language-specific effects on item difficulty.

As illustrated in Figure 3a, the reference English capabilities generally represent an upper bound of performance for most models. However, English does not represent the peak capability for a few

[1] `https://huggingface.co/datasets/gililior/mmlu-prox-eval-predictions`

lower-performing models, such as the OpenAI `gpt-oss` variants (20B and 120B). Furthermore, Figure 3b shows that the empirical distributions of these language-specific capabilities (θ_{il}) are non-normal with severe skewness, multimodality, and varying kurtosis. This suggests that imposing a standard continuous normal prior, as is conventional in standard IRT, would distort the capability estimates of these models. Notably, the estimated latent capabilities preserve the global model rankings derived from raw empirical accuracy (Spearman's $\rho = 1.0$).



(a) Capability distributions by models

(b) Capability densities by languages

Figure 3: Distribution of LLM capabilities. Panel (a) illustrates the individual language-specific capabilities (θ_{il}) for each model, with the English baseline highlighted in red. Panel (b) displays the estimated density curves for each language, demonstrating that the distributions are highly non-normal and asymmetric.

Figure 4 compares the item difficulty estimates across the $L = 29$ languages obtained from the MDIF-IRT model against those from the Independent Rasch model. The horizontal banding observed on the y-axis demonstrates a limitation of independent estimation. With only $I = 25$ models evaluated, a substantial number of items exhibit zero variance within a specific language (i.e., all models answer correctly or incorrectly). Conventional maximum likelihood estimation for the Independent Rasch model cannot produce difficulty estimates for such response patterns. In contrast, MDIF-IRT models a hierarchical prior on the language shift and infers difficulty estimates for items with homogeneous response patterns by borrowing information from the cross-lingual covariance structure $\boldsymbol{\Sigma}_\gamma$.

Figure 5 visualizes the estimated prior covariance matrix $\boldsymbol{\Sigma}_\gamma$ transformed into a correlation matrix, along with the average variational posterior mean of the language-specific difficulty shifts ($\boldsymbol{\mu}_j$). The block-diagonal structure of this matrix, alongside the mean shift vector, reveals distinct language clusters. For instance, the top-left cluster comprises low-resource African languages (Swahili, Wolof, Yoruba, and Zulu), which also exhibit the highest overall difficulty shifts, as indicated by the dark green in the last column. Conversely, the cluster exhibiting the lowest mean shifts consists of Romance languages (Italian, French, Spanish, and Portuguese). As high-resource languages, they align closely with the English baseline, resulting in negligible difficulty shifts.

4.4 Cross-Lingual DIF Detection

To detect items exhibiting significant cross-lingual DIF, the language shift parameter vector $\boldsymbol{\gamma}_{1j}$ for each item j is evaluated. Specifically, we compute a standardized DIF statistic, D_j, using the variational posterior mean $\boldsymbol{\mu}_j$ and covariance matrix $\mathbf{S}_j$ of these shift parameters:

$$D_j = \boldsymbol{\mu}_j^\top \mathbf{S}_j^{-1} \boldsymbol{\mu}_j. \tag{8}$$

In this study, we evaluate D_j against a χ^2 distribution with $L - 1$ degrees of freedom and apply the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) to correct for multiple comparisons

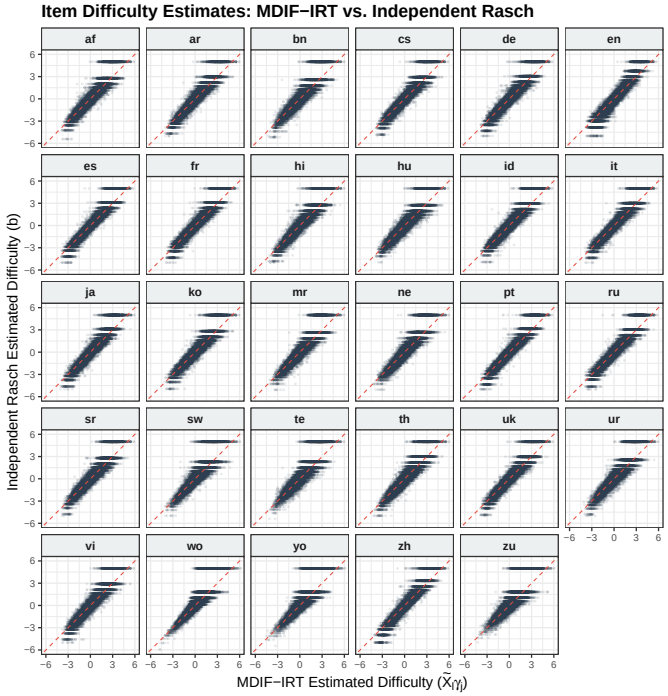


Figure 4: Item difficulty parameter estimates across languages.

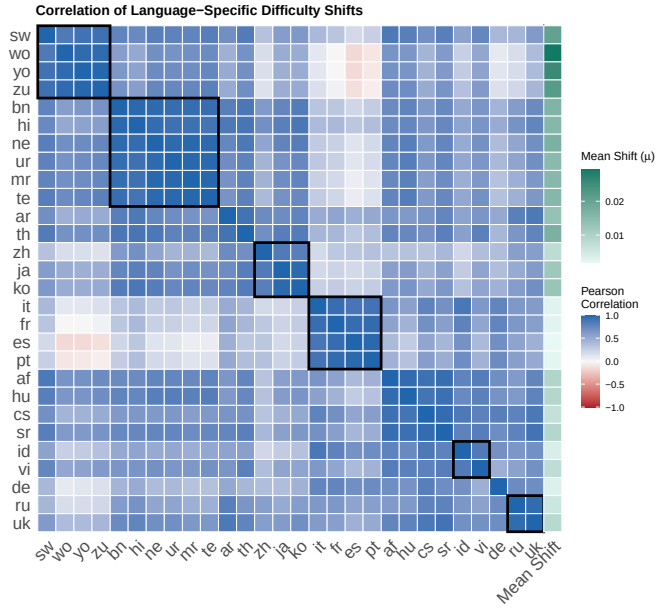


Figure 5: Language-specific difficulty shifts and correlations.

across items. An item j is flagged if its adjusted p-value is below 0.05.[2] In other words, this method flags an item when the posterior mean of its difficulty shift from the English baseline is significantly different from zero in at least one language, suggesting the presence of potential translation-related artifacts. Applied to the $11{,}759$ benchmark items, 748 items (6.4%) were flagged. This cross-lingual DIF detection method provides a fast screening tool for isolating potentially problematic items. In subsequent benchmark development and refinement, these items can either be removed from the dataset or subjected to qualitative review and revision.

To illustrate the qualitative investigation of detected cross-lingual DIF, we conducted an automated review of the flagged items. For each item, we identified the specific language exhibiting the largest difficulty shift in absolute value, $\max|\boldsymbol{\gamma}_{1j}|$ (i.e., the most severe artifact). We then extracted the original English item content alongside its translated counterpart in that target language. An AI assistant (GPT-5.5) was prompted to compare the two versions and classify the primary source of the difficulty shift into one of three categories: 1) Translation Error (e.g., semantic loss, formatting breakdown, or grammatical failure), 2) Culturally Specific Content (e.g., localized idioms, U.S.-centric history, or pop culture references), or 3) Other. Additionally, we instructed the AI to provide a brief explanation detailing its diagnostic reasoning for each classification.

Of the 748 items flagged for significant cross-lingual DIF, the automated qualitative review classified the vast majority ($n = 537$, 71.8%) as containing direct Translation Errors, while an additional 52 items (7.0%) were flagged for Culturally Specific Content. The remaining 159 items (21.3%) revealed a pervasive issue with Lexical Gaps and Specialized Terminology, where target languages lacked standardized scientific or medical vocabulary. Appendix D presents the prompt used to run this qualitative review, alongside the 10 items with the highest standardized DIF statistics and their corresponding diagnostics.

4.5 Predictive Accuracy Comparison

Table 1 demonstrates the predictive superiority of the MDIF-IRT model evaluated via a Monte Carlo cross-validation (5 iterations). In each iteration, the observed response tensor was randomly partitioned into an 80% training set and a 20% holdout set ($\sim$1.6 million points per iteration). Parameters estimated from the training set generated predictions for the holdout set, with an Empirical Accuracy baseline computed from training-set means. Performance was quantified via Classification Accuracy, Brier Score, and Binary Cross-Entropy (BCE), reporting means and standard errors. Substantial improvements over the baseline confirm the necessity of modeling item-level difficulty. Furthermore, the Independent Rasch model outperforms the Pooled Rasch model (Accuracy: 0.8032 vs. 0.7960), indicating that item difficulty is not invariant across languages but shifts due to translation effects. Finally, MDIF-IRT achieves the strongest overall predictive performance.

Table 1: Predictive accuracy of MDIF-IRT model compared to baselines empirical accuracy and Rasch variants (mean and standard error across 5 iterations).

MODEL	ACCURACY	BRIER SCORE	BCE
Empirical Accuracy	0.6991 (0.0001)	0.1939 (0.0001)	0.5688 (0.0002)
Pooled Rasch	0.7960 (0.0002)	0.1446 (0.0000)	0.4602 (0.0001)
Independent Rasch	0.8032 (0.0002)	0.1397 (0.0001)	0.4473 (0.0001)
MDIF-IRT	**0.8131 (0.0001)**	**0.1337 (0.0001)**	**0.4224 (0.0001)**

To further evaluate whether the cross-lingual covariance structure of MDIF-IRT facilitates information borrowing across translations, we conducted an additional block-wise holdout experiment. In this procedure, for 20% of the benchmark items, all model responses in one randomly selected translated language were masked. The models were then tasked with predicting these unobserved item-language blocks relying solely on the items' performance in the remaining available languages. Results from this evaluation confirm that MDIF-IRT maintains the highest predictive accuracy (81.5%) and lowest Brier score (0.133), demonstrating that its cross-lingual covariance structure provides

[2] A bootstrap-based covariance matrix could alternatively be used in place of $\mathbf{S}_j$ to adjust for the potential underestimation of variational posterior variance.

a distinct advantage over traditional methods when inferring entirely unobserved translations (Appendix E).

5 CONCLUSION

This study proposes the DECAF framework for evaluating the language-specific capabilities of LLMs using multilingual benchmarks. The proposed MDIF-IRT model enhances multilingual AI evaluations by disentangling language-specific model capability shifts from language-specific item difficulty shifts. Further investigating these item-level difficulty shift parameters serves as a diagnostic tool to detect translation artifacts within benchmark items that inadvertently inflate task difficulty.

Empirical evaluations validate the predictive and diagnostic advantages of this approach. By explicitly modeling cross-lingual covariance, MDIF-IRT borrows information across languages to enhance parameter estimation. This hierarchical structure allows the model to reliably infer item difficulties for homogeneous response patterns and to accurately predict item responses to unobserved language translations. Consequently, MDIF-IRT demonstrated superior predictive accuracy and lower Brier scores compared to empirical baselines and conventional IRT models. Furthermore, leveraging these cross-lingual structures enabled the identification of distinct linguistic clusters, revealing severe artificial difficulty spikes in low-resource African languages versus minimal shifts in high-resource Romance languages.

The implementation of a GVEM algorithm ensures highly scalable estimation across large $I \times J \times L$ response tensors. This approach also provides the flexibility to relax the often-violated normality assumptions inherent in traditional IRT, yielding more robust estimates for LLM capabilities.

Cross-lingual DIF detection provides diagnostics for benchmark developers. Manually verifying the translation quality of high-dimensional benchmarks across dozens of languages is cost-prohibitive and practically infeasible. On the other hand, our proposed method serves as a fast, data-driven screening tool. Demonstrating this utility, the framework identified 6.4% of the benchmark items as exhibiting significant DIF indicative of translation artifacts; subsequent qualitative review confirmed that nearly 72% of these flagged items suffered from direct translation errors, while the remainder exposed cultural biases and lexical gaps. By automatically flagging items that are empirically prone to severe difficulty shifts, researchers can construct highly reliable multilingual benchmarks more cost-effectively. Potentially, it can guide efficient sampling decisions when subsetting an extensive baseline benchmark for resource-intensive human translation or post-editing.

6 LIMITATIONS

While the DECAF framework provides a robust method for evaluating multilingual LLMs, several limitations exist in the present study. First, the empirical application in this study is constrained to the MMLU-ProX evaluation dataset. Although this benchmark covers a diverse set of subjects and languages, future work should validate the framework across other prominent multilingual benchmarks, such as Global MMLU, MMMLU, and MGSM.

Second, our comparative analysis focused on independent and pooled Rasch baselines to demonstrate the necessity of modeling language-specific difficulty shifts. Future research should expand this to include comparisons with more complex existing approaches, such as the multilingual IRT model (Lior et al., 2026).

Third, the proposed MDIF-IRT model assumes uniform item discrimination across all models and languages, similar to a Rasch framework. An extension would be the inclusion of an item discrimination parameter, analogous to a 2PL IRT model. This would enable the detection of non-uniform DIF, capturing instances where translation artifacts not only shift the difficulty of an item but also alter its ability to effectively differentiate between high- and low-capability models.

Finally, while this study introduces a standardized DIF statistic for detecting cross-lingual DIF, a comprehensive benchmark refinement procedure remains future work. The flagged items can undergo a systematic pipeline for qualitative review by AI or human linguists to establishing protocol for revising, re-translating, or discarding problematic items.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, assisting with translation, supporting qualitative and thematic data analysis, and designing or providing feedback on research methodology or experiments. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses, or interpreting results, and cleaning and reformatting datasets are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We declare no ethical concerns regarding the methodology, applications, or findings of this work.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our results, the complete source code used to process and analyze the data in this paper is publicly available through an anonymous repository at `https://anonymous.4open.science/r/MDIF-IRT-7164/README.md`.

REFERENCES

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, et al. Tokenizer choice for llm training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3907–3924, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

April E Cho, Chun Wang, Xue Zhang, and Gongjun Xu. Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 74:52–85, 2021.

Susan E Embretson and Steven P Reise. *Item response theory: Foundations for psychologists and social scientists*. Taylor & Francis, 2025.

Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A Smith. Fluid language model benchmarking. *arXiv preprint arXiv:2509.11106*, 2025.

Paul W Holland, Howard Wainer, et al. *Differential item functioning*, volume 538. Lawrence Erlbaum Hillsdale, NJ, 1993.

Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Han Jiang, Sunbeom Kwon, Jinwen Luo, Ziang Xiao, and Susu Zhang. Can we trust item response theory for ai evaluation? *arXiv preprint arXiv:2607.15190*, 2026.

Alex Kipnis, Konstantinos Voudouris, Luca M Schulze Buschoff, and Eric Schulz. metabench–a sparse benchmark of reasoning and knowledge in large language models, 2025. *URL https://arxiv.org/abs/2407.12844*.

Gili Lior, Tzviel Frostig, Gabriel Stanovsky, and Matan Eyal. Extending item response theory for efficient and meaningful multilingual evaluation. *arXiv preprint arXiv:2606.15643*, 2026.

Steven J Osterlind and Howard T Everson. *Differential item functioning*. Sage Publications, 2009.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 191–204, 2026.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Alexander Robitzsch. Package 'sirt'. *Computer software. https://www. maths. bris. ac. uk/R/web/packages/sirt/sirt. pdf*, 2017.

Mara Schilling-Wilhelmi, Nawaf Alampara, and Kevin Maik Jablonka. Lifting the benchmark iceberg with item-response theory. In *AI for Accelerated Materials Design-ICLR 2025*, 2025.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Tianyi Xu, Kosei Uemura, Alfred Malengo Kondoro, Tadesse Destaw Belay, Catherine Nana Nyaah Essuman, Ifeoma Okoh, Ganiyat Afolabi, Ayodele Awokoya, and David Ifeoluwa Adelani. Mgsm-pro: A simple strategy for robust multilingual mathematical reasoning evaluation. *arXiv preprint arXiv:2601.21225*, 2026.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual large language models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362, 2025.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1513–1532, 2025.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35085–35093, 2026.

A GAUSSIAN VARIATIONAL EXPECTATION-MAXIMIZATION ALGORITHM FOR THE MDIF-IRT MODEL

The marginal log-likelihood of the MDIF-IRT model is:

$$\ell = \sum_{j=1}^{J} \log \left(\int \prod_{i=1}^{I} \prod_{l=1}^{L} \left[p_{ijl}^{Y_{ijl}} (1 - p_{ijl})^{1 - Y_{ijl}} \right] f(\boldsymbol{\gamma}_j) \, d\boldsymbol{\gamma}_j \right). \tag{9}$$

The evidence lower bound (ELBO) is:

$$\mathcal{L}(q) = \sum_{j=1}^{J} \mathbb{E}_{q_j} \left[\sum_{i=1}^{I} \sum_{l=1}^{L} (Y_{ijl} \log p_{ijl} + (1 - Y_{ijl}) \log(1 - p_{ijl})) + \log f(\boldsymbol{\gamma}_j) - \log q_j(\boldsymbol{\gamma}_j) \right], \tag{10}$$

where $q_j(\boldsymbol{\gamma}_j)$ is the variational distribution approximating the posterior for item j.

Note that the term $-\mathbb{E}_{q_j}[\log q_j(\boldsymbol{\gamma}_j)]$ does not depend on the model parameters. Therefore, maximizing the ELBO with respect to the model parameters is equivalent to maximizing the objective function

$$Q = \sum_{j=1}^{J} \mathbb{E}_{q_j} \left[\sum_{i=1}^{I} \sum_{l=1}^{L} (Y_{ijl} \log p_{ijl} + (1 - Y_{ijl}) \log(1 - p_{ijl})) + \log f(\boldsymbol{\gamma}_j) \right]. \tag{11}$$

To derive a closed-form solution for the logistic likelihood, auxiliary parameters $\xi_{ijl} > 0$ are introduced to construct a quadratic lower bound on each logistic log-likelihood term (Jaakkola & Jordan, 2000):

$$Y_{ijl} \log p_{ijl} + (1 - Y_{ijl}) \log(1 - p_{ijl}) \geq \log \sigma(\xi_{ijl}) + \left(Y_{ijl} - \frac{1}{2} \right) (\tilde{\mathbf{X}}_l \boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l \boldsymbol{\gamma}_j) - \frac{1}{2} \xi_{ijl} - \eta(\xi_{ijl}) \left[(\tilde{\mathbf{X}}_l \boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l \boldsymbol{\gamma}_j)^2 - \xi_{ijl}^2 \right], \tag{12}$$

where $\eta(\xi) = \frac{1}{2\xi} [\sigma(\xi) - \frac{1}{2}]$.

As a result, an approximation of the objective function Q, denoted by Q', is derived as:

$$\begin{aligned} Q' = \sum_{j=1}^{J} \sum_{i=1}^{I} \sum_{l=1}^{L} \mathbb{E}_{q_j} \Bigg[& \log \sigma(\xi_{ijl}) + \left(Y_{ijl} - \frac{1}{2} \right) (\tilde{\mathbf{X}}_l \boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l \boldsymbol{\gamma}_j) - \frac{1}{2} \xi_{ijl} - \eta(\xi_{ijl}) \left((\tilde{\mathbf{X}}_l \boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l \boldsymbol{\gamma}_j)^2 - \xi_{ijl}^2 \right) \Bigg] \\ & - \sum_{j=1}^{J} \mathbb{E}_{q_j} \left[\frac{1}{2} \boldsymbol{\gamma}_j^\top \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\gamma}_j \right] - \frac{J}{2} \log(2\pi \sigma_{\gamma_0}^2) - \frac{J}{2} \log |2\pi \boldsymbol{\Sigma}_\gamma|. \end{aligned} \tag{13}$$

A.1 E-STEP: VARIATIONAL POSTERIOR UPDATE

The goal of the E-step is to update the variational posterior $q_j(\boldsymbol{\gamma}_j)$ for each j. Under the mean-field approximation, the joint variational posterior factorizes across the J items:

$$q(\boldsymbol{\gamma}) = \prod_{j=1}^{J} q_j(\boldsymbol{\gamma}_j). \tag{14}$$

For fixed auxiliary parameters ξ, maximizing the Jaakkola-Jordan variational lower bound with respect to $q_j(\boldsymbol{\gamma}_j)$ yields a Gaussian coordinate-optimal variational factor $q_j(\boldsymbol{\gamma}_j) = \mathcal{MVN}(\boldsymbol{\mu}_j, \mathbf{S}_j)$ with

$$\begin{aligned}\mathbf{S}_j &= \left[\boldsymbol{\Sigma}_0^{-1} + 2\sum_{i=1}^{I}\sum_{l=1}^{L}\eta(\xi_{ijl})\tilde{\mathbf{X}}_l^\top\tilde{\mathbf{X}}_l\right]^{-1}, \\ \boldsymbol{\mu}_j &= \mathbf{S}_j\left[\sum_{i=1}^{I}\sum_{l=1}^{L}\left(2\eta(\xi_{ijl})\tilde{\mathbf{X}}_l\boldsymbol{\beta}_i - \left(Y_{ijl}-\frac{1}{2}\right)\right)\tilde{\mathbf{X}}_l^\top\right].\end{aligned} \tag{15}$$

The auxiliary variable ξ_{ijl} is updated by:

$$\xi_{ijl} = \sqrt{\mathbb{E}_{q_j}\left[(\tilde{\mathbf{X}}_l\boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l\boldsymbol{\gamma}_j)^2\right]} = \sqrt{(\tilde{\mathbf{X}}_l\boldsymbol{\beta}_i - \tilde{\mathbf{X}}_l\boldsymbol{\mu}_j)^2 + \tilde{\mathbf{X}}_l\mathbf{S}_j\tilde{\mathbf{X}}_l^\top}. \tag{16}$$

A.2 M-STEP: MODEL CAPABILITY PARAMETER AND HYPERPARAMETER UPDATE

The M-step updates the model capability parameter and the hyperparameter values by maximizing Q'.

Update $\boldsymbol{\beta_i}$:

$$\boldsymbol{\beta}_i = \mathbf{A}_i^{-1}\mathbf{c}_i, \tag{17}$$

where

$$\begin{aligned}\mathbf{A}_i &= 2\sum_{j=1}^{J}\sum_{l=1}^{L}\eta(\xi_{ijl})\tilde{\mathbf{X}}_l^\top\tilde{\mathbf{X}}_l, \\ \mathbf{c}_i &= \sum_{j=1}^{J}\sum_{l=1}^{L}\left(2\eta(\xi_{ijl})\tilde{\mathbf{X}}_l\boldsymbol{\mu}_j + Y_{ijl} - \frac{1}{2}\right)\tilde{\mathbf{X}}_l^\top.\end{aligned} \tag{18}$$

Update $\sigma_{\gamma_0}^2$:

$$\sigma_{\gamma_0}^2 = \frac{1}{J}\sum_{j=1}^{J}\left(\mu_{0j}^2 + S_{j[1,1]}\right), \tag{19}$$

where μ_{0j} is the first element of $\boldsymbol{\mu}_j$ and $S_{j[1,1]}$ is the top-left element of $\mathbf{S}_j$.

Update $\boldsymbol{\Sigma}_\gamma$:

$$\boldsymbol{\Sigma}_\gamma = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{\mu}_{1j}\boldsymbol{\mu}_{1j}^\top + \mathbf{S}_{1j}\right), \tag{20}$$

where $\boldsymbol{\mu}_{1j}$ is the $(L-1)$ vector containing the remaining elements of $\boldsymbol{\mu}$, and $\mathbf{S}_{1j}$ is the $(L-1)\times(L-1)$ lower-right block of the covariance matrix $\mathbf{S}_j$.

B PARAMETER RECOVERY CHECK

This section evaluates the parameter recovery of the MDIF-IRT model using the GVEM algorithm. In each condition, 100 observed response data sets were randomly generated for $I \in \{25, 50\}$, $J \in \{500, 1000\}$, and $L \in \{5, 10\}$. True model parameters were generated to simulate realistic variance and cross-lingual covariance. Specifically, the English capabilities (β_{i0}) and item difficulties (b_{0j}) were drawn from $N(0, 1.2^2)$ and $N(0, 1.0)$, respectively. The language-specific difficulty shifts (γ_{lj}) were sampled from a multivariate normal distribution with a variance of 0.5 and an off-diagonal covariance of 0.2, while model capability shifts (β_{il}) were drawn with a variance of 0.3.

The results in Table A1 indicate that the GVEM algorithm reliably recovers both the item and capability parameters of the MDIF-IRT model, yielding negligible bias across all conditions. Further more, increasing the number of items (J) decreases the RMSE of the model capability parameters, and increasing the number of models (I) improves the recovery of the item difficulty parameters. Conversely, the number of language groups (L) showed little effect on the accuracy of parameter estimates. In addition, the MDIF-IRT model consistently yields lower RMSE of model capability and item difficulty estimates, compared to the Rasch baselines (Table A2).

Table A1: Parameter recovery results for MDIF-IRT

CONDITIONS			BIAS				RMSE				CORRELATION			
I	J	L	β_0	$\boldsymbol{\beta}_1$	γ_0	$\boldsymbol{\gamma}_1$	β_0	$\boldsymbol{\beta}_1$	γ_0	$\boldsymbol{\gamma}_1$	β_0	$\boldsymbol{\beta}_1$	γ_0	$\boldsymbol{\gamma}_1$
25	500	5	.002	-.002	.002	-.003	.120	.166	.356	.497	.996	.955	.935	.722
25	500	10	-.001	.003	.000	.001	.118	.164	.340	.499	.996	.957	.941	.715
25	1000	5	.002	.003	.003	.000	.086	.119	.354	.495	.998	.977	.935	.724
25	1000	10	.000	-.001	-.001	.001	.085	.117	.339	.498	.998	.978	.941	.719
50	500	5	.003	.000	.002	.000	.123	.166	.283	.402	.995	.958	.960	.825
50	500	10	.010	-.005	.006	-.002	.116	.164	.277	.405	.996	.958	.962	.821
50	1000	5	.001	.003	.001	.004	.085	.116	.283	.402	.998	.978	.960	.825
50	1000	10	.002	.003	.003	.002	.085	.117	.275	.402	.998	.978	.962	.824

Table A2: Comparison of model capability ($\boldsymbol{\theta}$) and item difficulty ($\mathbf{b}$) recovery across MDIF-IRT, Independent Rasch, and Pooled Rasch models

CONDITIONS			BIAS						RMSE					
			$\boldsymbol{\theta}$			$\mathbf{b}$			$\boldsymbol{\theta}$			$\mathbf{b}$		
I	J	L	MDIF	Ind.	Pool	MDIF	Ind.	Pool	MDIF	Ind.	Pool	MDIF	Ind.	Pool
25	500	5	-.002	.000	-.002	-.001	-.002	-.002	.129	.153	.438	.403	.566	.553
25	500	10	.004	-.001	.002	.001	.003	.003	.128	.155	.493	.396	.568	.552
25	1000	5	.004	.000	.003	.003	.005	.005	.095	.118	.443	.400	.564	.551
25	1000	10	.001	.000	.001	.000	-.002	-.002	.094	.118	.495	.392	.565	.550
50	500	5	.002	.003	.003	.002	.003	.003	.128	.138	.448	.313	.382	.374
50	500	10	.004	.000	.001	.004	.002	.002	.126	.135	.497	.309	.384	.374
50	1000	5	.003	.004	.003	.004	.004	.004	.090	.098	.438	.310	.381	.373
50	1000	10	.005	.004	.004	.005	.004	.004	.091	.097	.495	.308	.383	.373

C EXPERIMENT SETTINGS

Table A3: List of 29 languages evaluated in the MMLU-ProX dataset

CODE	LANGUAGE	CODE	LANGUAGE	CODE	LANGUAGE
af	Afrikaans	id	Indonesian	te	Telugu
ar	Arabic	it	Italian	th	Thai
bn	Bengali	ja	Japanese	uk	Ukrainian
cs	Czech	ko	Korean	ur	Urdu
de	German	mr	Marathi	vi	Vietnamese
en	English	ne	Nepali	wo	Wolof
es	Spanish	pt	Portuguese	yo	Yoruba
fr	French	ru	Russian	zh	Chinese
hi	Hindi	sr	Serbian	zu	Zulu
hu	Hungarian	sw	Swahili		

Table A4: List of the 25 large language models evaluated

DEVELOPER	MODELS
Alibaba	Qwen3.5-2B, Qwen3.5-4B, Qwen3.5-9B, Qwen3.5-27B, Qwen3.5-35B-A3B MoE, Qwen3.5-122B-A10B MoE
AllenAI	Olmo-3-7B-Instruct, Olmo-3-7B-Think, Olmo-3.1-32B-Instruct, Olmo-3.1-32B-Think
Cohere	aya-expanse-8b, aya-expanse-32b, tiny-aya-global
Google	gemma-3-4b-it, gemma-3-12b-it, gemma-3-27b-it
Meta	Llama-3.1-8B, Llama-3.2-3B, Llama-3.3-70B-Instruct
Microsoft	Phi-4-mini-instruct, Phi-4-mini-reasoning, Phi-4-reasoning, Phi-4-reasoning-plus
OpenAI	gpt-oss-20b, gpt-oss-120b

D CLASSIFICATIONS AND DIAGNOSTICS OF CROSS-LINGUAL DIF ITEMS

To automate the qualitative classification of the flagged items, we queried the GPT-5.5 model using the following prompt:

The following AI evaluation benchmark item is flagged as having a significant cross-lingual DIF. Compare the two versions and classify the primary source of the difficulty shift into one of three categories:

1. Translation Error (e.g., semantic loss, formatting breakdown, or grammatical failure)
2. Culturally Specific Content (e.g., localized idioms, U.S.-centric history, or pop culture references)
3. Other

You must respond ONLY with a valid JSON object using the exact keys below:

```
{
  "category": "Select exactly one: 'Translation Error',
  'Culturally Specific Content', or 'Other'",
  "specific_label": "If category is 'Other', provide a 2-4
  word label for the source of difficulty.
  Otherwise, leave empty.",
  "explanation": "A brief 1-2 sentence explanation of why it
  falls into this category."
}
```

Below is the English item content and the translated version:

Table A5: Top 10 items with the largest standardized DIF statistics for cross-lingual DIF, including AI-generated qualitative classifications.

Item	DIF Stat.	Lang.	English Question	Category	Label	Explanation
2322	229.79	wo	What is an experimental control? Why is it important to have experimental control?	Other	Technical terminology	The translation is largely faithful and not culturally specific, but it relies on a literal/loanword rendering of the scientific term "experimental control." The difficulty shift likely comes from specialized terminology being less familiar or less standardized in Wolof rather than a clear semantic translation error.
5238	201.21	hi	By what name is Bob Keeshan better known?	Culturally Specific Content	—	The Hindi translation accurately preserves the meaning of the English question. The difficulty shift likely comes from the U.S.-specific pop culture reference to Bob Keeshan, better known as Captain Kangaroo.
4668	176.22	wo	This question refers to the following information. (...) The perspective of the passage above would most directly support which of the following political goals?	Translation Error	—	The final question mistranslates "political goals" as "political leaders" (*njiit politig*), changing what the respondent is being asked to identify. There are also smaller lexical awkwardnesses, but this stem error directly shifts the task difficulty.
4798	173.25	hi	During the Pleistocene era, the islands of Java, Sumatra, Bali, and Borneo formed _____. The landmass connecting Australia, New Guinea, and Tasmania was called _____.	Other	Specialized terminology	The Hindi translation preserves the meaning and structure of the English item without a clear semantic or grammatical error. Any difficulty shift is more likely due to specialized paleogeographic terms and background knowledge.
7008	171.99	wo	What is fiscal policy?	Translation Error	—	The Wolof translation renders "fiscal policy" as roughly "government money policy," which loses the specific meaning of taxation and government spending and may blur it with monetary policy.
5428	171.87	wo	The second stage of the advertising campaign is to establish _____, or to become known as trustworthy and believable.	Translation Error	—	The Wolof translation introduces awkward and semantically confusing wording for "advertising campaign," making the sentence less clear than the English. This likely shifts difficulty due to translation quality rather than cultural content.
10761	169.91	wo	Nussbaum claims that in cross-cultural communication, inhabitants of different conceptual schemes tend to view their interaction:	Translation Error	—	The translation weakens and shifts key philosophical wording: 'inhabitants of different conceptual schemes' becomes roughly 'people of countries with different thoughts,' and 'tend to' is omitted. This semantic loss can change how the item is understood.
4553	168.11	wo	Opportunistic foraging involves:	Translation Error	—	The Wolof translation is awkward and does not accurately convey the ecological concept of "opportunistic foraging." It appears to render the phrase as something like searching/seizing opportunities or gatherings, causing semantic loss.

Continued on next page

Table A.5

Item	DIF Stat.	Lang.	English Question	Category	Label	Explanation
2684	167.62	th	Distinguish between the terms “in situ,” “in vitro,” and “in vivo.”	Other	Terminology transliteration	The translation preserves the basic task, but renders the Latin scientific terms phonetically in Thai rather than keeping the standard Latin forms. This may change difficulty because learners may recognize the original Latin/English spellings more readily than the transliterations.
5265	164.78	yo	The latitude and longitude coordinates of a place are an example of	Translation Error	—	The Yoruba translation changes or obscures key terms: “place” is rendered as “aye,” which more often means “world/life,” and “coordinates” is reduced to “measure.” This semantic loss can make the item harder or confusing.

E PREDICTING UNOBSERVED ITEM-LANGUAGE BLOCKS

E.1 EXPERIMENTAL DESIGN

To test whether the MDIF-IRT framework effectively borrows structural information across languages, we implemented a block-wise holdout experiment. In this setup, we randomly selected 20% of the items from the benchmark. For each selected item, one translated language (excluding the reference language, English) was chosen at random. We then masked all I model responses for that specific item-language combination in the training set. This creates a scenario where an item is entirely unobserved in one language, forcing the algorithms to infer the missing item parameters based solely on that item's behavior in the remaining observed languages.

E.2 PREDICTION GENERATION

Predictions for the completely masked item-language blocks were generated as follows:

- **MDIF-IRT**: The MDIF-IRT framework naturally handles this block-wise missingness through its hierarchical prior. Because the variational posterior mean ($\boldsymbol{\mu}_j$) for the item parameters incorporates the cross-lingual covariance matrix ($\boldsymbol{\Sigma}_0$), the model infers the missing language-specific difficulty shift by borrowing information from the item's observed responses in the unmasked languages. Predictive probabilities were computed using the standard logistic function on the recovered model capabilities and item parameters.
- **Independent and Pooled Rasch**: In the traditional Rasch frameworks, completely masking an item-language block results in undefined item difficulty estimates. To allow these models to generate predictions on the test set, the missing difficulty parameter for a masked block was imputed using the average estimated difficulty of that specific item calculated across all available, unmasked languages. While this imputation strategy provides an item-specific anchor, the traditional Rasch frameworks inherently lack an ability to model cross-lingual covariance.
- **Empirical Accuracy**: As a baseline, predictions for the missing blocks were generated using the overall average accuracy for each specific model-language pair, derived from the remaining training data.

Table A6: Predictive accuracy of MDIF-IRT model compared to baseline empirical accuracy and Rasch variants on unobserved item-language blocks.

MODEL	ACCURACY	BRIER SCORE	BCE
Empirical Accuracy	0.701	0.193	0.566
Pooled Rasch	0.792	0.146	0.455
Independent Rasch	0.811	0.135	0.430
MDIF-IRT	**0.815**	**0.133**	**0.421**