**International Journal of Machine Learning (IJOML)**
E-ISSN: 3124-6362
Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

# Cold-Start Generalization in Educational Interaction Data: Comparing Student-Wise and Question-Wise Splits with Probabilistic Calibration

**Purwadi*[1], Nor Azman Bin Abu[2], Othman Bin Mohd[3]**

[1]Department of Information systems, Amikom Purwokerto University, Central Java, Indonesia
[2]Faculty of Artificial Intelligence and Cyber Security (FAIX), Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[3]Center for Advanced Computing Technology (C-ACT), Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

Email: [1]purwadi@amikompurwokerto.ac.id

Received: Jan 9, 2026; Revised: Feb 28, 2026; Accepted: Feb 28, 2026; Published: June 1, 2026

**Abstract**

Predictive models in Intelligent Tutoring Systems often face performance degradation due to sparse data and the cold-start problem, further compounded by a lack of probability calibration in standard evaluations. This study bridges this gap by systematically evaluating the trade-off between discriminative accuracy and probabilistic reliability through student-wise and question-wise splits, utilizing interaction data from the MathE platform across eight countries. By comparing identifier-based and metadata-based Logistic Regression models under a Leave-One-Country-Out protocol, we assessed generalization capabilities against distribution shifts. The results reveal a fundamental dichotomy: while identifier-based models achieve superior precision (AUC 0.687) and calibration in scenarios with historical context, they suffer from significant performance drops in student cold-start settings and exhibit negative transfer during cross-country deployment. Conversely, metadata-based models demonstrate higher robustness and invariance across varying demographics. We conclude that relying solely on accuracy metrics masks model uncertainty in new domains and recommend a "safe-start" strategy that prioritizes metadata-based features for system initialization to ensure reliable pedagogical decision-making before personalizing based on accumulated user history.

***Keywords:*** *Cold-start Problem, Domain Generalization, Educational Data Mining, Knowledge Tracing, Probability Calibration.*



## 1. INTRODUCTION

In the current era of digital transformation in education, the ability to automatically track and predict student learning progress has become a vital component of Intelligent Tutoring Systems (ITS) and online learning platforms [1], [2]. Learning outcome prediction, formally known as Knowledge Tracing (KT), aims to model students' knowledge mastery over time based on their interaction history with learning materials [2], [3]. Accurate predictive capability is essential for evaluating instructional effectiveness and providing personalized services, such as adaptive question recommendation and the construction of learning paths tailored to individual student needs, particularly in complex subjects like mathematics [4], [5].

Although deep learning-based models have demonstrated promising performance, their practical application is often constrained by the sparse and long-tailed characteristics of educational data [6], [7]. In real-world scenarios, most students interact with only a small fraction of the total items available in a question bank, while new questions are continuously added to the system [8], [9]. This sparsity of interactions hinders the model's ability to capture accurate representations of students or questions, which negatively impacts generalization to unseen data [10].

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

These challenges culminate in the cold-start problem, defined as a degradation in system performance due to the absence of historical interaction data for new entities [7]. In the educational context, this problem falls into two crucial categories: student cold-start (new students with no prior attempts) and question cold-start (new questions never attempted by any student) [7], [10]. Conventional models relying on ID embeddings often fail to provide valid predictions under these conditions because they cannot infer parameters for entities unseen during training [7], [8].

Unfortunately, evaluation protocols commonly used in the literature often employ random splits, assuming that training and testing data are independent and identically distributed (i.i.d.) [11], [12]. Such evaluations tend to yield overly optimistic performance estimates as they fail to reflect distribution shifts occurring in the real world, such as when models encounter entirely new student populations or question banks [12], [13]. Therefore, more rigorous group-based evaluation protocols—specifically student-wise and question-wise splits—are necessary to realistically measure model generalization capabilities under operational cold-start conditions [12], [13].

regarding predictive performance measurement, discriminative metrics such as the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) are widely used to assess a model's ability to rank the probability of student correctness [1], [14]. However, in high-stakes educational applications, ranking accuracy alone is insufficient; systems must possess high probabilistic reliability [15], [16]. Probability-based decisions, such as determining whether a student requires remedial intervention or has mastered a concept, require well-calibrated probability estimates where predicted probabilities reflect the actual empirical frequency of correctness [17].

A primary issue with modern predictive models, particularly deep neural networks, is the tendency toward miscalibration or overconfidence, where the model assigns high probabilities to incorrect predictions [17]. This overconfidence poses serious practical risks, such as system failure in detecting at-risk students or providing erroneous feedback because the model is confident despite being wrong [18], [19]. To mitigate this, temperature scaling (TS) has been proposed as an effective post-hoc calibration approach [17]. TS works by scaling model output logits using a single scalar parameter to align model confidence with empirical accuracy without altering the original classification accuracy [18], [20], [21].

While the urgency of cold-start and calibration issues has been acknowledged separately, a significant research gap remains: there is a lack of studies systematically comparing the impact of student cold-start versus question cold-start while simultaneously evaluating the trade-off between discriminative performance and probability calibration within a single reproducible experimental framework. Most studies focus solely on accuracy improvements, often neglecting whether the resulting probabilities are trustworthy for pedagogical decision-making. This study aims to bridge this gap by conducting a structured evaluation using group-based splits (student-wise and question-wise) accompanied by in-depth calibration analysis. The primary contribution of this research is to generate reliable methodological recommendations for educational predictive modeling, ensuring that models are not only accurate in ranking students but also trustworthy in their probability estimates under data-constrained conditions.

## 2. METHOD

This study employs a structured methodological framework to evaluate the generalization capabilities of predictive models under cold-start scenarios as well as the reliability of their probability estimates. The methodology encompasses data processing, the establishment of group-based evaluation protocols, model training, probability calibration, and the aggregation of results in a consistent and reproducible manner. The overall workflow of the research system is illustrated in Figure 1.

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

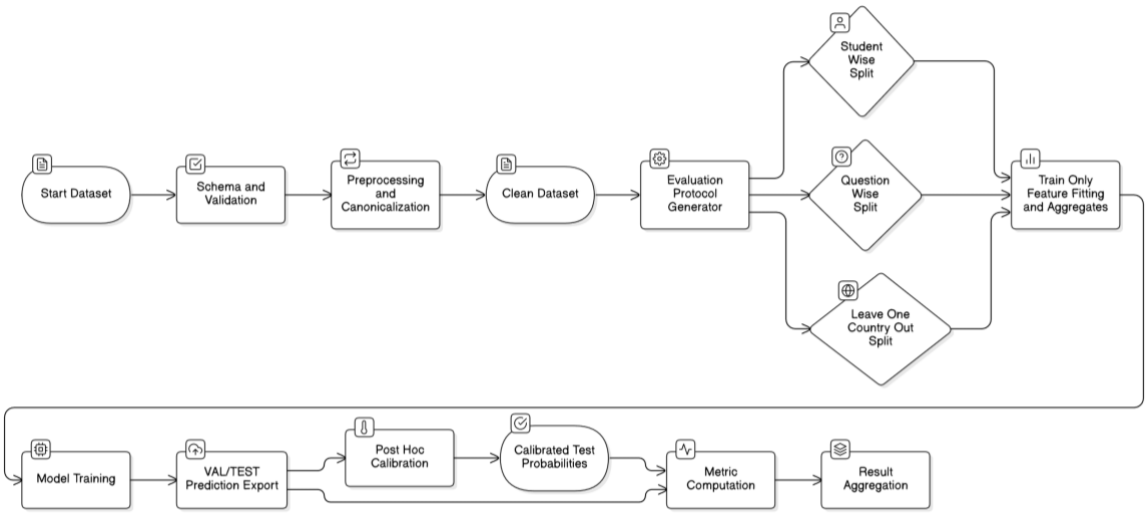


Figure 1. Overview of The Research Workflow

### 2.1. Dataset and Task Definition

This study utilizes the mathematics learning interaction dataset from the collaborative MathE platform [22], which comprises 9,546 activity logs from 372 higher education students across eight countries between January 2019 and December 2023. This dataset was selected due to its heterogeneous data representation and hierarchical content structure, consisting of 14 main topics and 24 subtopics. Each data entry records granular attributes, including student identity, question metadata, and the difficulty level of the items, categorized as either *basic* or *advanced* by the instructional team.

The prediction task is formulated as a binary classification problem, where the target variable is determined by the correctness of the student's response (correct or incorrect) during each interaction. To accommodate the variability in material complexity, the experimental framework divides the analysis into two independent data subsets based on the *basic* and *advanced* difficulty labels.

### 2.2. Preprocessing and Experimental Protocols

To ensure reproducibility and valid evaluation under dynamic operational conditions, we applied consistent data standardization and generalization testing protocols targeting cold-start scenarios and domain shift. The dataset was cleaned and canonicalized by normalizing schema definitions, resolving inconsistent labels and missing values, and harmonizing categorical attributes into a global coding scheme so that observed performance differences primarily reflect modeling choices rather than data artifacts [1], [12]. To reduce noise, extremely short interaction sequences and duplicate or invalid sessions were removed [23], [24].

Cold-start evaluation employed group-based splits to prevent data leakage, as random splits often yield overly optimistic estimates of generalization [25]. Two complementary settings were considered. First, a student-wise split enforced disjoint sets of learners between the training and test partitions to assess generalization to new users without historical interactions [7], [26]. Second, a question-wise split fully isolated a subset of items into the test partition to simulate the introduction of new items or curriculum content and to evaluate whether the model can generalize without item-specific historical support [5], [10]. Train, validation, and test partitions were constructed consistently according to the group definition in each scenario.

To assess robustness to distribution shifts across populations, we additionally adopted a Leave-One-Country-Out (LOCO) protocol, in which the model was trained on the pooled data from all countries except one and evaluated on the held-out country, following the Domain Generalization

DOI: https://doi.org/10.66472/ijoml.v1i1.4

framework [13], [27]. This protocol tests whether the model learns representations that are relatively invariant across domains or instead relies on domain-specific, non-transferable patterns that may reflect spurious correlations [26], [28].

### 2.3. Feature Representation

Features were constructed under two complementary schemes to examine the trade-off between personalization and cross-domain generalization under cold-start conditions [6], [7]. The identifier-based scheme employed sparse one-hot representations for student and item entities, augmented with metadata and keyword features, to capture user–item interaction patterns that are effective when interaction histories are sufficiently rich [23]. In contrast, the semantic-only scheme removed all unique identifiers and relied exclusively on content metadata and keywords to encourage more stable inductive inference when generalizing to previously unseen domains [10], [29].

To prevent data leakage [28], all transformations followed strict train-only fitting: encoders, scalers, and vectorizers were fitted solely on the training set, and the validation and test sets were transformed using training-derived parameters. Historical and aggregate features were also computed causally from the training history, denoted as $\mathcal{H}_{s,k}^{\text{train}}$, without accessing validation or test labels [30], [31]. Student–topic ability estimates were obtained using a smoothed mean defined in Equation (1):

$$\hat{\theta}_{s,k} = \frac{\sum_{(i,y)\in\mathcal{H}_{s,k}^{\text{train}}}(y+\alpha)}{|\mathcal{H}_{s,k}^{\text{train}}|+2\alpha}, y \in \{0,1\} \qquad (1)$$

### 2.4. Models and Training Procedure

To establish valid lower-bound references and support a comprehensive evaluation, we assessed deterministic and statistical baselines before the primary models. The baselines included a Majority Predictor that always outputs the global dominant class, and a Topic-level Prior that estimates prior probabilities from training-set statistics within each topic, thereby characterizing the underlying label distribution without feature learning [23], [32]. Our main approach employed logistic regression, which is interpretable and robust for sparse educational data, using one-hot features derived from metadata. This choice is computationally efficient for modeling item–response patterns and is conceptually aligned with multidimensional IRT formulations, while reducing the risk of overfitting relative to neural networks in data-scarce regimes [15], [33].

Training was conducted consistently for every combination of evaluation protocol, fold, and random seed to ensure fair comparison and reproducibility. We followed domain generalization practices inspired by DomainBed, emphasizing a clean validation split for model selection without test-set bias, alongside tight control of non-deterministic sources of variation [12], [13], [25].

### 2.5. Post-hoc Probability Calibration

To improve probabilistic reliability and reduce overconfidence, we applied post-hoc temperature scaling to the model outputs. Temperature scaling was selected because it has been shown empirically to align predictive confidence with accuracy without degrading discriminative performance such as AUC [25], [34], [35]. Formally, the logit $z$is calibrated by a positive scalar $T$, yielding calibrated probabilities $\hat{q} = \sigma(z/T)$, or a softmax transformation in the multi-class setting [21], [36]. The temperature parameter $T$was estimated exclusively on a held-out validation set by minimizing the Negative Log-Likelihood, which is typically stable under limited-data regimes [17], [37]. For degenerate cases, such as near single-class validation labels or severe validation bias, the optimization was constrained and $T$was reverted to the identity setting $T = 1$ to avoid worsening miscalibration due to an

DOI: https://doi.org/10.66472/ijoml.v1i1.4

unrepresentative validation set [19]. The selected $T$ was then applied globally to the test-set logits to produce the final calibrated probabilities [18].

### 2.6. Evaluation

Model performance was assessed using a comprehensive framework that jointly considers discrimination, uncertainty quality, and probability calibration. As the primary discrimination metric, we used the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) due to its robustness to class imbalance and its widespread adoption in the Knowledge Tracing literature [26]. To evaluate the intrinsic quality of probabilistic estimates, we employed proper scoring rules, namely Negative Log-Likelihood (NLL) and the Brier Score. NLL quantifies the divergence between predicted probabilities $p_i$and labels $y_i$by penalizing overconfident yet incorrect predictions, as defined in Equation (2):

$$\mathrm{NLL} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(p_i) + (1-y_i)\log(1-p_i)] \quad (2)$$

The Brier Score was used to simultaneously capture prediction reliability and resolution. Calibration was quantified using the Expected Calibration Error (ECE) under a fixed-binning scheme, which measures the deviation between model confidence and empirical accuracy across $M$bins, as defined in Equation (3):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \; |\, \mathrm{acc}(B_m) - \mathrm{conf}(B_m) \,| \quad (3)$$

These metrics were complemented with reliability diagrams to visualize systematic miscalibration. For more granular diagnostics, we stratified performance by item difficulty, reporting $\mathrm{AUC}_{basic}$and $\mathrm{AUC}_{advanced}$to detect variability across data subpopulations. All experimental results are reported as mean and standard deviation $\mathrm{mean} \pm \mathrm{std}$over multiple random-seed initializations to support reproducibility. For the Leave-One-Country-Out (LOCO) setting, we additionally adopted robustness-oriented summaries, including macro-averaged performance, worst-country performance, and the generalization gap. These metrics were selected to explicitly assess stability under distribution shift and to discourage over-optimization toward majority domains, consistent with fairness and robustness principles [12].

## 3. RESULT

The empirical evaluation of student performance prediction models was conducted through a set of controlled experiments designed to probe model limits under three fundamental scenarios: prediction on previously unseen items given known student histories, referred to as Cold Question; generalization to entirely new student populations, referred to as Student Cold-Start; and cross-country robustness under a Leave-One-Country-Out protocol. The quantitative results reveal distinct performance patterns, highlighting a critical trade-off between leveraging individual-specific information and achieving robust generalization.

### 3.1. Cold Question Scenario

In the Cold Question setting, where the model has access to students' interaction histories but must predict outcomes for items not observed during training, the model that incorporates student-identity features consistently outperformed alternative approaches. As shown in Table 1, the ID-based logistic regression achieved the highest mean ROC-AUC of 0.687 and exhibited strong stability, as indicated by a low standard deviation of 0.009. In contrast, the metadata-only logistic regression lagged

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

substantially, with a mean ROC-AUC of 0.584 and higher variability of 0.014. This performance gap widened markedly on the advanced-item subset: the ID-based model maintained a high ROC-AUC of 0.712, whereas the metadata-only model degraded to 0.539, approaching near-random discrimination.

Table 1. Comparative model performance under unseen-item evaluation, Cold Question scenario

| Model specification | ROC-AUC | Log Loss | ECE | Advanced-item ROC-AUC |
|---|---|---|---|---|
| Logistic Regression (Student-ID Features) | 0.687 ± 0.009 | 0.636 ± 0.006 | 0.034 ± 0.005 | 0.712 ± 0.015 |
| Logistic Regression (Metadata Features) | 0.584 ± 0.014 | 0.683 ± 0.003 | 0.030 ± 0.013 | 0.539 ± 0.029 |
| Topic-level Prior (Student-ID Features) | 0.552 ± 0.018 | 0.689 ± 0.006 | 0.021 ± 0.009 | 0.548 ± 0.012 |
| Topic-level Prior (Metadata Features) | 0.552 ± 0.018 | 0.689 ± 0.006 | 0.021 ± 0.009 | 0.548 ± 0.012 |
| Global Majority Baseline (Student-ID Features) | 0.500 ± 0.000 | 0.692 ± 0.002 | 0.014 ± 0.014 | 0.500 ± 0.000 |
| Global Majority Baseline (Metadata Features) | 0.500 ± 0.000 | 0.692 ± 0.002 | 0.014 ± 0.014 | 0.500 ± 0.000 |

Beyond discrimination, probability quality assessment via Expected Calibration Error indicated that the ID-based model remained competitively calibrated. Consistent with the reliability diagram in Figure 2, the calibration curve of the ID-based model closely tracked the ideal diagonal, yielding an average calibration error of 0.034. This suggests that the predicted probabilities are meaningfully aligned with empirical correctness rates. In contrast, the topic-level prior exhibited systematic deviations in certain probability ranges, reflecting under-confidence and over-confidence depending on the operating region.

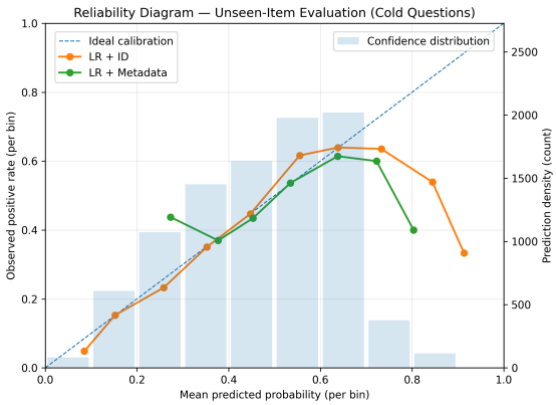


Figure 2. Reliability diagram for unseen-item evaluation, Cold Question scenario

### 3.2. Student Cold-Start Scenario

Transitioning to the Student Cold-Start setting, which simulates a cold-start condition for new learners without historical interaction records, led to a global performance degradation across all model variants. As reported in Table 2, the performance of the ID-based and metadata-based models converged sharply. The ID-based logistic regression dropped to a mean ROC-AUC of 0.556, only marginally higher than the metadata-based logistic regression with a mean ROC-AUC of 0.548. This degradation was accompanied by increased performance uncertainty, most notably reflected in the metadata-based

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

model, whose ROC-AUC standard deviation rose to 0.042, substantially higher than its variability under the Cold Question setting.

Table 2. Model performance under unseen-learner evaluation, Student Cold-Start scenario

| Model specification | ROC-AUC | Log Loss | ECE |
|---|---|---|---|
| Logistic Regression (Student-ID Features) | 0.556 ± 0.010 | 0.692 ± 0.002 | 0.039 ± 0.024 |
| Logistic Regression (Metadata Features) | 0.548 ± 0.042 | 0.692 ± 0.008 | 0.053 ± 0.017 |
| Topic-level Prior (Student-ID Features) | 0.520 ± 0.029 | 0.782 ± 0.199 | 0.039 ± 0.031 |
| Topic-level Prior (Metadata Features) | 0.520 ± 0.029 | 0.782 ± 0.199 | 0.039 ± 0.031 |
| Global Majority Baseline (Student-ID Features) | 0.500 ± 0.000 | 0.695 ± 0.006 | 0.043 ± 0.026 |
| Global Majority Baseline (Metadata Features) | 0.500 ± 0.000 | 0.695 ± 0.006 | 0.043 ± 0.026 |

Moreover, on the advanced-item subset, the advantage of the ID-based model diminished considerably, reaching a ROC-AUC of 0.585 compared with 0.534 for the metadata-based model. The calibration analysis in Figure 3 further underscores the difficulty of generalization in this setting: calibration curves for all models are more dispersed and deviate more from the ideal diagonal than in the Cold Question scenario. The higher mean ECE values of 0.039 for the ID-based model and 0.053 for the metadata-based model confirm that the absence of personal learner history substantially reduces the accuracy of probabilistic estimates.

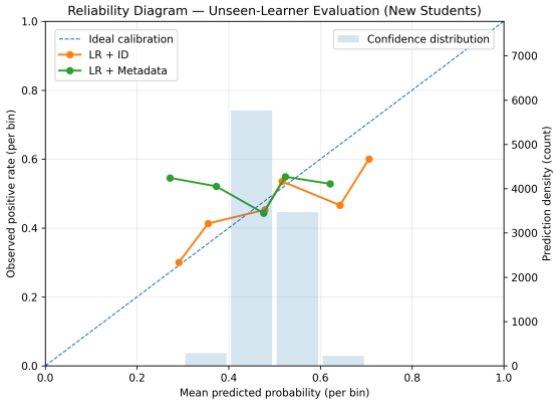


Figure 3. Reliability diagram for unseen-learner evaluation, Student Cold-Start scenario

### 3.3. Cross-Country Evaluation Under Leave-One-Country-Out

Robustness testing under the Leave-One-Country-Out protocol revealed a counterintuitive performance pattern consistent with negative transfer. As summarized in Table 3, the metadata-based model unexpectedly outperformed the ID-based model in macro-averaged ROC-AUC, achieving 0.576 versus 0.568. This advantage is reinforced by the worst-country ROC-AUC: the metadata-based model maintained a minimum performance level of 0.510, whereas the ID-based model dropped below the random baseline to 0.483 in the most adverse case.

The country-wise breakdown in Table 4 provides a granular view of this variability. In countries with larger populations and potentially more homogeneous data characteristics, such as Spain, both models achieved peak performance with ROC-AUC exceeding 0.73. In contrast, in countries with more distinct data characteristics, such as Romania, the ID-based model exhibited severe generalization failure, with ROC-AUC dropping to 0.483, far below the metadata-based model, which remained at 0.552. Supporting analyses of cross-country variability further indicate that the metadata-based model generally exhibited smaller performance fluctuations across most target countries.

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

Table 3. Cross-country generalization summary under Leave-One-Country-Out evaluation

| Model specification | Macro ROC-AUC | Worst-country ROC-AUC | Generalization gap | ECE |
|---|---|---|---|---|
| Logistic Regression (Metadata Features) | 0.576 | 0.51 | 0.262 | 0.128 |
| Logistic Regression (Student-ID Features) | 0.568 | 0.483 | 0.253 | 0.116 |
| Topic-level Prior (Student-ID Features) | 0.535 | 0.497 | 0.19 | 0.104 |
| Topic-level Prior (Metadata Features) | 0.535 | 0.497 | 0.19 | 0.104 |
| Global Majority Baseline (Student-ID Features) | 0.5 | 0.5 | 0 | 0.083 |
| Global Majority Baseline (Metadata Features) | 0.5 | 0.5 | 0 | 0.083 |

Table 4. Country-wise ROC-AUC matrix for Leave-One-Country-Out evaluation

| Target country | LR + ID | LR + Metadata | Topic Prior + ID | Topic Prior + Metadata | Majority + ID | Majority + Metadata |
|---|---|---|---|---|---|---|
| Ireland | 0.538 | 0.547 | 0.521 | 0.521 | 0.5 | 0.5 |
| Italy | 0.54 | 0.549 | 0.507 | 0.507 | 0.5 | 0.5 |
| Lithuania | 0.564 | 0.535 | 0.524 | 0.524 | 0.5 | 0.5 |
| Portugal | 0.549 | 0.542 | 0.499 | 0.499 | 0.5 | 0.5 |
| Romania | 0.483 | 0.552 | 0.506 | 0.506 | 0.5 | 0.5 |
| Russia | 0.625 | 0.6 | 0.54 | 0.54 | 0.5 | 0.5 |
| Slovenia | 0.511 | 0.51 | 0.497 | 0.497 | 0.5 | 0.5 |
| Spain | 0.736 | 0.772 | 0.688 | 0.688 | 0.5 | 0.5 |

Figure 4 visualizes these cross-country disparities, showing that the metadata-based model yields more stable performance across countries, whereas the ID-based model exhibits larger fluctuations. This trade-off is further summarized by the Pareto frontier in Figure 5, which places the metadata-based model closer to the optimal region for systems that prioritize worst-case protection under a maximin strategy. In contrast, the ID-based model appears comparatively high-risk for cross-population deployment without additional adaptation.

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

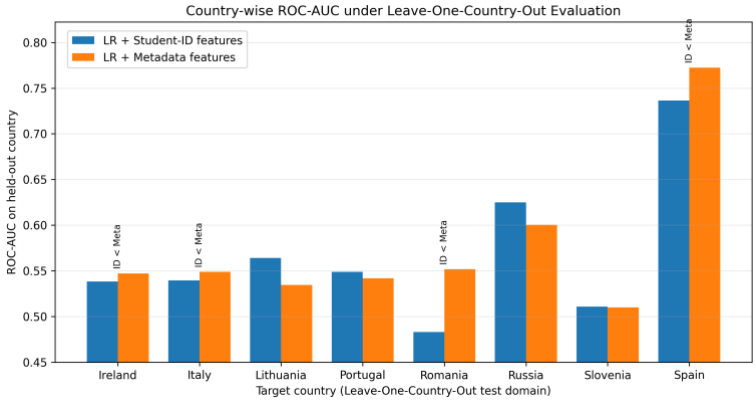


Figure 4. Country-wise ROC-AUC comparison under LOCO evaluation

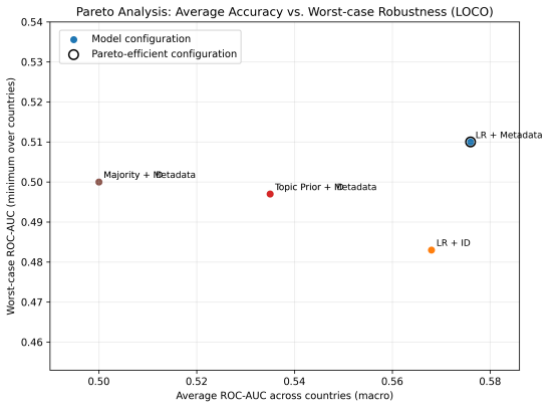


Figure 5. Pareto analysis of average accuracy versus worst-case robustness (LOCO)

## 4. DISCUSSIONS

The central findings of this study reveal a fundamental tension between representational specificity and generalization capacity in student performance modeling. The dominance of identity-feature models in the Cold Question setting reinforces the theoretical premise that, in relatively closed educational environments where interaction histories are available, estimates of students' latent proficiency provide substantially stronger predictive signal than item characteristics alone. This is consistent with core principles of Item Response Theory and its modern extensions such as Deep Knowledge Tracing, where a large portion of response variability is explained by individual ability parameters learned from historical data [38], [39]. The combination of high discrimination and precise calibration further indicates that when the test distribution aligns with the training distribution, the model can capture fine-grained learning patterns that cannot be represented by generic metadata.

In contrast, the sharp performance degradation under the Student Cold-Start setting highlights the classic vulnerability of cold-start, which remains a key challenge for adaptive learning systems. When access to students' historical embeddings is removed, the discriminative advantage of ID-based models largely disappears, effectively forcing the model to rely on weak population-level averages. This observation aligns with prior findings that excessive reliance on ID features without effective knowledge transfer mechanisms leads to poor adaptation for unseen entities [6], [7]. The convergence between ID-based models and simpler metadata-based models in this setting suggests that even sophisticated architectures cannot compensate for missing information. This strengthens the case for hybrid approaches or richer semantic signals, such as problem text or concept graphs, to bridge the information gap during early interactions [5], [10].

Vol. 1, No. 1, June 2026, Page. 38-50
https://ijoml.com/
DOI: https://doi.org/10.66472/ijoml.v1i1.4

The most consequential paradox emerges in the Leave-One-Country-Out analysis, where metadata-based models exhibit stronger robustness than ID-based models. This provides empirical evidence of negative transfer for ID-based representations when deployed across heterogeneous demographic populations. ID-based models appear to overfit to country-specific proficiency distributions and curricular characteristics observed during training, thereby capturing spurious correlations that do not generalize across countries. In contrast, item metadata such as topic and difficulty are comparatively more invariant to domain shift, enabling the model to retain a stable baseline level of performance when transferred to a new national context. This is consistent with the domain generalization perspective, which emphasizes that models overly dependent on domain-specific features are prone to catastrophic failure under fundamental distributional changes [13], [28].

From a reliability standpoint, the calibration analysis offers a critical insight: higher accuracy does not necessarily imply more trustworthy predictive confidence. Although ID-based models achieve superior accuracy under local data conditions, their pronounced calibration instability in cross-country settings indicates that they can become systematically overconfident in unseen domains. In high-stakes applications such as education, where model outputs can shape students' learning trajectories, calibration metrics such as ECE should be treated as co-primary objectives alongside accuracy [17], [40]. The more stable calibration behavior of metadata-based models supports their role as a safer, more conservative choice for initial deployment in new regions, particularly before sufficient local interaction data is available to enable effective model adaptation.

Practically, these results suggest a paradigm shift for globally scaled Intelligent Tutoring Systems. Developers should no longer assume that a single monolithic model can serve all populations equally well. Instead, we recommend a safe-start strategy: initialize deployment in a new country or population using a metadata-driven model that is comparatively invariant to domain shift, thereby reducing worst-case risk, and progressively transition toward personalized ID-based modeling as local interaction data accumulates. Theoretically, this study motivates future work on approaches such as causal representation learning or invariant risk minimization in educational prediction [27], with the goal of learning student representations that are not only accurate in-distribution but also robust to demographic and curricular variation at global scale.

## 5. CONCLUSION

This study concludes that student performance modeling entails a fundamental trade-off between local predictive precision and global generalization robustness. The empirical evidence shows that while identity-based models achieve superior accuracy and well-calibrated probabilities in closed settings with rich interaction histories, these advantages are fragile and degrade substantially under cold-start conditions and cross-country domain shifts. In contrast, metadata-based models, despite a lower performance ceiling, provide markedly stronger robustness and reduced exposure to negative transfer, serving as a critical stability anchor in heterogeneous learning environments. A key contribution of this work is the empirical demonstration that conventional accuracy-focused metrics can obscure model uncertainty on newly encountered populations, motivating an evaluation shift toward balancing discrimination and calibration.

For future work, we recommend exploring hybrid architectures that can transition dynamically from metadata-driven prediction to personalized modeling as student-specific data accumulates, thereby narrowing the initial performance gap. In addition, integrating causal representation learning or invariant risk minimization is strongly encouraged to disentangle instructionally relevant causal signals from demographically biased correlations, enabling models that are genuinely insensitive to geographic variation. Finally, enriching semantic features through natural language processing of problem content

presents a strategic opportunity to reduce dependence on static metadata, paving the way for more adaptive and globally scalable Intelligent Tutoring Systems.

## CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

## ACKNOWLEDGEMENT

The authors sincerely appreciate the contributions, assistance, and encouragement from all parties who supported the research process and the development of this manuscript, which played an important role in the completion of this study.

## REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge Tracing: A Survey," *ACM Comput Surv*, vol. 55, no. 11, pp. 1–37, Nov. 2023, doi: 10.1145/3569576.

[2] Y. Bai, J. Zhao, T. Wei, Q. Cai, and L. He, "A survey of explainable knowledge tracing," *Applied Intelligence*, vol. 54, no. 8, pp. 6483–6514, Apr. 2024, doi: 10.1007/s10489-024-05509-8.

[3] S. Yao, Y. Song, Y. Liu, J. Chen, D. Zhao, and X. Wang, "ANT-KT: Adaptive NAS Transformers for Knowledge Tracing," *Electronics (Basel)*, vol. 14, no. 21, p. 4148, Oct. 2025, doi: 10.3390/electronics14214148.

[4] B. Xu *et al.*, "Learning Behavior-oriented Knowledge Tracing," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2023, pp. 2789–2800. doi: 10.1145/3580305.3599407.

[5] G. Liu, H. Zhan, and J. Kim, "Question Difficulty Consistent Knowledge Tracing," in *Proceedings of the ACM Web Conference 2024*, New York, NY, USA: ACM, May 2024, pp. 4239–4248. doi: 10.1145/3589334.3645582.

[6] Y. Guo *et al.*, "Mitigating Cold-Start Problems in Knowledge Tracing with Large Language Models: An Attribute-aware Approach," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Oct. 2024, pp. 727–736. doi: 10.1145/3627673.3679664.

[7] H. Jung, J. Yoo, Y. Yoon, and Y. Jang, "CLST: Cold-Start Mitigation in Knowledge Tracing by Aligning a Generative Language Model as a Students' Knowledge Tracer," 2024. [Online]. Available: https://arxiv.org/abs/2406.10296

[8] L. Fu *et al.*, "SINKT: A Structure-Aware Inductive Knowledge Tracing Model with Large Language Model," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, New York, NY, USA: ACM, Oct. 2024, pp. 632–642. doi: 10.1145/3627673.3679760.

[9] H. Ma, C. Wang, H. Zhu, S. Yang, X. Zhang, and X. Zhang, "Enhancing Cognitive Diagnosis Using Un-interacted Exercises: A Collaboration-Aware Mixed Sampling Approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, pp. 8877–8885, Mar. 2024, doi: 10.1609/aaai.v38i8.28735.

[10] L. Ni *et al.*, "Enhancing Student Performance Prediction on Learnersourced Questions with SGNN-LLM Synergy," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, pp. 23232–23240, Mar. 2024, doi: 10.1609/aaai.v38i21.30370.

[11] J. Cha *et al.*, "SWAD: Domain Generalization by Seeking Flat Minima," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 22405–22418. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/bcb41ccdc4363c6848a1d760f26c28a0-Paper.pdf

[12] P. W. Koh *et al.*, "WILDS: A Benchmark of in-the-Wild Distribution Shifts," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in

Proceedings of Machine Learning Research, vol. 139. PMLR, Dec. 2021, pp. 5637–5664. [Online]. Available: https://proceedings.mlr.press/v139/koh21a.html

[13] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," 2020. [Online]. Available: https://arxiv.org/abs/2007.01434

[14] T. Long, Y. Liu, J. Shen, W. Zhang, and Y. Yu, "Tracing Knowledge State with Individual Cognition and Acquisition Estimation," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2021, pp. 173–182. doi: 10.1145/3404835.3462827.

[15] T. S. Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Mach Learn*, vol. 112, no. 9, pp. 3211–3260, Sep. 2023, doi: 10.1007/s10994-023-06336-7.

[16] Z. Jiang, A. Liu, and B. Van Durme, "Addressing the Binning Problem in Calibration Assessment through Scalar Annotations," *Trans Assoc Comput Linguist*, vol. 12, pp. 120–136, Feb. 2024, doi: 10.1162/tacl_a_00636.

[17] S. A. Balanya, J. Maroñas, and D. Ramos, "Adaptive temperature scaling for Robust calibration of deep neural networks," *Neural Comput Appl*, vol. 36, no. 14, pp. 8073–8095, May 2024, doi: 10.1007/s00521-024-09505-4.

[18] L. Dabah and T. Tirer, "On Temperature Scaling and Conformal Prediction of Deep Classifiers," 2025. [Online]. Available: https://arxiv.org/abs/2402.05806

[19] R. Roelofs, N. Cain, J. Shlens, and M. C. Mozer, " Mitigating Bias in Calibration Error Estimation ," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Eds., in Proceedings of Machine Learning Research, vol. 151. PMLR, Dec. 2022, pp. 4036–4054. [Online]. Available: https://proceedings.mlr.press/v151/roelofs22a.html

[20] A. Karandikar *et al.*, "Soft calibration objectives for neural networks," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[21] Y. Yu, S. Bates, Y. Ma, and M. I. Jordan, "Robust calibration with multi-domain temperature scaling," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, in NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[22] B. F. Azevedo, A. I. Pereira, F. P. Fernandes, and M. F. Pacheco, "Mathematics learning and assessment using MathE platform: A case study," *Educ Inf Technol (Dordr)*, vol. 27, no. 2, pp. 1747–1769, Mar. 2022, doi: 10.1007/s10639-021-10669-y.

[23] S. Minn, J.-J. Vie, K. Takeuchi, H. Kashima, and F. Zhu, "Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 12810–12818, Jun. 2022, doi: 10.1609/aaai.v36i11.21560.

[24] F. Liu, C. Bu, H. Zhang, L. Wu, K. Yu, and X. Hu, "FDKT: Towards an Interpretable Deep Knowledge Tracing via Fuzzy Reasoning," *ACM Trans Inf Syst*, vol. 42, no. 5, pp. 1–26, Sep. 2024, doi: 10.1145/3656167.

[25] F. Wang *et al.*, "NeuralCD: A General Framework for Cognitive Diagnosis," *IEEE Trans Knowl Data Eng*, vol. 35, no. 8, pp. 8312–8327, Aug. 2023, doi: 10.1109/TKDE.2022.3201037.

[26] L. Hu *et al.*, "PTADisc: a cross-course dataset supporting personalized learning in cold-start scenarios," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[27] R. Dai, Y. Zhang, Z. Fang, B. Han, and X. Tian, "Moderately distributional exploration for domain generalization," in *Proceedings of the 40th International Conference on Machine Learning*, in ICML'23. JMLR.org, 2023.

[28] A. Robey, G. J. Pappas, and H. Hassani, "Model-based domain generalization," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[29] Y. Lu, L. Tong, and Y. Cheng, "Advanced Knowledge Tracing: Incorporating Process Data and Curricula Information via an Attention-Based Framework for Accuracy and Interpretability.," *Journal of Educational Data Mining*, vol. 16, no. 2, pp. 58–84, 2024.

[30] M. Chen, Q. Guan, Y. He, Z. He, L. Fang, and W. Luo, "Knowledge Tracing Model with Learning and Forgetting Behavior," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, New York, NY, USA: ACM, Oct. 2022, pp. 3863–3867. doi: 10.1145/3511808.3557622.

[31] Y. Yin *et al.*, "Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer," in *Proceedings of the ACM Web Conference 2023*, New York, NY, USA: ACM, Apr. 2023, pp. 855–864. doi: 10.1145/3543507.3583255.

[32] S. P. Neshaei, R. L. Davis, A. Hazimeh, B. Lazarevski, P. Dillenbourg, and T. Käser, "Towards Modeling Learner Performance with Large Language Models ," in *Proceedings of the 17th International Conference on Educational Data Mining* , International Educational Data Mining Society, Jul. 2024, pp. 759–768. doi: 10.5281/zenodo.12729942.

[33] S. Frick, A. Krivosija, and A. Munteanu, "Scalable Learning of Item Response Theory Models," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, S. Dasgupta, S. Mandt, and Y. Li, Eds., in Proceedings of Machine Learning Research, vol. 238. PMLR, Dec. 2024, pp. 1234–1242. [Online]. Available: https://proceedings.mlr.press/v238/frick24a.html

[34] M. Minderer *et al.*, "Revisiting the calibration of modern neural networks," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.

[35] O. Bohdal, Y. Yang, and T. Hospedales, "Meta-Calibration: Learning of Model Calibration Using Differentiable Expected Calibration Error," 2023. [Online]. Available: https://arxiv.org/abs/2106.09613

[36] T. Joy, F. Pinto, S.-N. Lim, P. H. S. Torr, and P. K. Dokania, "Sample-Dependent Adaptive Temperature Scaling for Improved Calibration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14919–14926, Jun. 2023, doi: 10.1609/aaai.v37i12.26742.

[37] L. Clarté, B. Loureiro, F. Krzakala, and L. Zdeborová, "Expectation consistency for calibration of neural networks," in *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, in UAI '23. JMLR.org, 2023.

[38] S. Pu, Y. Yan, and B. Zhang, "Predicting Students' Future Success: Harnessing Clickstream Data with Wide & Deep Item Response Theory.," *Journal of Educational Data Mining*, vol. 16, no. 2, pp. 1–31, 2024.

[39] J. Chen, Z. Liu, S. Huang, Q. Liu, and W. Luo, "Improving Interpretability of Deep Sequential Knowledge Tracing Models with Question-centric Cognitive Representations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14196–14204, Jun. 2023, doi: 10.1609/aaai.v37i12.26661.

[40] I. Arrieta-Ibarra, P. Gujral, J. Tannen, M. Tygert, and C. Xu, "Metrics of calibration for probabilistic predictions," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.