UPSC2M: Benchmarking Adaptive Learning from Two Million MCQ Attempts

Kevin Shi, Karttikeya Mangalam

SigIQ.ai

Correspondence: kevin@sigiq.ai

Abstract

We present UPSC2M, a large-scale dataset comprising two million multiple-choice question attempts from over 46,000 students, spanning nearly 9,000 questions across seven subject areas. The questions are drawn from the Union Public Service Commission (UPSC) examination, one of India's most competitive and high-stakes assessments. Each attempt includes both response correctness and time taken, enabling fine-grained analysis of learner behavior and question characteristics. Over this dataset, we define two core benchmark tasks: question difficulty estimation and student performance prediction. The first task involves predicting empirical correctness rates using only question text. The second task focuses on predicting the likelihood of a correct response based on prior interactions. We evaluate simple baseline models on both tasks to demonstrate feasibility and establish reference points. Together, the dataset and benchmarks offer a strong foundation for building scalable, personalized educational systems. We release the dataset and code to support further research at the intersection of content understanding, learner modeling, and adaptive assessment: github.com/kevins-hi/upsc2m.

1 Introduction

As digital learning platforms become increasingly central to education, there is growing demand for intelligent systems that can adapt to individual learners, curate relevant content, and deliver targeted assessments. At the heart of such systems lie two fundamental modeling tasks: estimating the difficulty of educational content and predicting student performance. These capabilities underpin a wide range of applications—from personalized question selection to real-time learner diagnostics. When combined, they serve as the foundation for fully automated adaptive learning systems that dynamically tailor instruction based on both content complexity and learner proficiency.

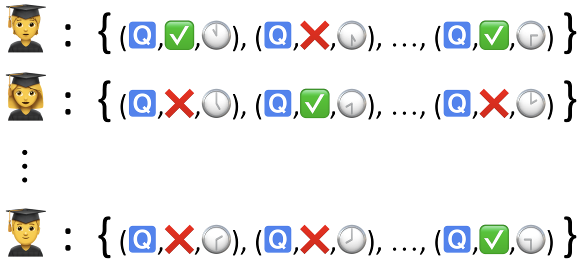

Figure 1: UPSC2M visualized as a list of students, each associated with a set of question attempts. Each attempt records the student ID, question ID, selected answer, whether it was correct, and the time taken to answer.

Statistic	Count
Unique Students	46,235
Unique Questions	8,973
Total Interactions	1,962,573

Table 1: Summary statistics for the UPSC2M dataset.

Much of the existing work in educational modeling has relied on small-scale classroom data or narrow subject domains, limiting the development of models for real-world settings. To bridge this gap, we introduce UPSC2M, a large-scale dataset of 1,962,573 question attempts from aspirants preparing for the Union Public Service Commission (UPSC) examination—one of India's most competitive standardized tests. Spanning 8,973 questions across seven subjects, UPSC2M includes correctness and timing data from 46,235 students.

We propose two core tasks supported by this dataset. The first is *Question Difficulty Estimation*, where models predict empirical difficulty from question text alone. The second is *Student Performance Prediction*, where models forecast whether a student will answer a question correctly, given their prior interactions. These tasks reflect key challenges in real-world adaptivity and serve as modular building blocks for intelligent tutoring and assessment systems.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 931–936
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

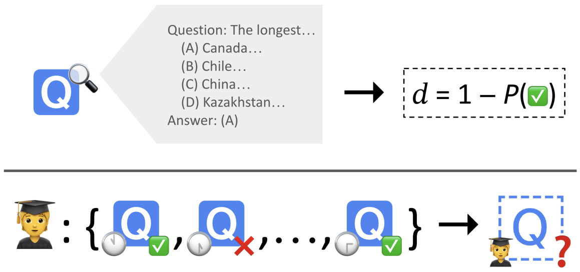

Figure 2: Illustration of the two benchmark tasks: question difficulty estimation (top) and student performance prediction (bottom). In the first task, the goal is to estimate the difficulty of a question—defined as one minus the empirical probability of a correct response—based solely on its text. In the second task, given a student's prior question attempts, predict whether the student will correctly answer a new, unseen question.

Our contributions are threefold: (1) We release UPSC2M, a large-scale dataset capturing both question content and behavioral interaction data in a high-stakes, multi-subject testing context. (2) We define two core prediction tasks that capture key challenges in adaptive education. (3) We establish baselines and outline directions for future work. Together, UPSC2M and its benchmark tasks provide a robust foundation for research in scalable personalized education. By supporting more accurate models of question difficulty and student performance, this work lays the groundwork for educational platforms that adapt to individual needs at scale, expanding access to high-quality, personalized learning for students regardless of background.

2 Related Work

Large-scale Interaction Datasets A number of publicly available datasets have driven progress in student modeling and adaptive learning. The PSLC DataShop repository provides tens of thousands of student–problem interactions across diverse domains (Stamper et al., 2011), and the ASSISTments dataset offers fine-grained logs of middle-school mathematics practice. More recently, EdNet—a hierarchical dataset of over 130 million interactions from an online tutoring platform—has enabled deep sequence models at unprecedented scale (Choi et al., 2020). Our dataset, UPSC2M, complements these by focusing on a highly competitive, multi-subject exam context, capturing both correctness and response-time signals for UPSC aspirants.

Question Difficulty Estimation Classical item response theory (IRT) models difficulty as a latent parameter estimated from response patterns (Lord, 1980), but they rely solely on interaction counts. Recent work has explored textual and semantic features to predict question difficulty directly from content (Blum and Corter, 2014). By pairing a large, annotated UPSC question bank with empirical accuracy rates, UPSC2M supports both purely content-based difficulty regression and hybrid approaches that integrate behavioral priors.

Student Performance Prediction Predicting learner outcomes has a long history in educational data mining. Bayesian Knowledge Tracing (BKT) (Corbett and Anderson, 1994) and Performance Factor Analysis (PFA) (Pavlik Jr et al., 2009) established early probabilistic frameworks for tracking mastery. The advent of neural methods—e.g. Deep Knowledge Tracing (DKT) (Piech et al., 2015) has further improved sequence-based prediction. The UPSC2M dataset, with its detailed question content, student attempt outcomes, and rich temporal metadata, offers a new testbed for benchmarking such models on high-stakes exam data.

Applications for Adaptive Testing Adaptive testing algorithms—such as computerized adaptive testing (CAT) (Weiss, 2011)—depend critically on calibrated item difficulties and real-time performance estimates. Datasets that combine content features with large-scale attempt logs enable more responsive and personalized CAT systems. We anticipate that UPSC2M will spur advances in adaptive exam design, question selection strategies, and real-time learner diagnostics.

3 Proposed Dataset

3.1 Motivation and Collection

The UPSC exam is among the most competitive and high-stakes assessments in India, attracting over one million aspirants annually. The exam begins with Paper 1, a 2-hour, 100-question multiple-choice test that spans a broad spectrum of subjects, including history, polity, economy, science, geography, environment, and current affairs. Questions are carefully crafted to assess not only factual recall, but also higher-order reasoning, elimination strategies, and nuanced interpretive understanding under strict time constraints.

This examination offers a rich environment for studying educational modeling tasks. In particular, Paper 1 presents a uniquely challenging setting: questions span multiple knowledge domains, often

		Students per Question			Questions per Student		
Subject	**Question Count**	Mean	Median	Max	Mean	Median	Max
Current Affairs	1793	127.79	112	3576	20.13	5	1502
Polity	1487	348.00	79	3284	19.31	5	1425
History	1449	259.72	77	2559	20.94	5	1227
Economy	1111	183.86	72	1728	20.17	5	1069
Science	1094	139.81	19	2869	11.48	5	1008
Environment	1022	181.63	104	2801	11.70	4	913
Geography	1017	291.82	145	3055	19.93	5	956
Overall	**8973**	**218.72**	**91**	**3576**	**42.45**	**8**	**6553**

Table 2: Per-subject statistics in the UPSC2M dataset, including the number of questions and summary statistics for student and question engagement—measured as students per question and questions per student.

require implicit reasoning, and are attempted by a large student body interacting with a shared question bank. These characteristics make it an ideal testbed for developing, benchmarking, and evaluating adaptive educational technologies at scale.

To support research on adaptive learning algorithms, we deployed a custom learning platform targeting UPSC aspirants. Students engaged with a curated bank of 8,973 multiple-choice questions. Over a 2-year period, we collected interaction data from 46,235 students, totaling 1,962,573 question attempts. The resulting dataset has been rigorously cleaned and anonymized to ensure student privacy while retaining the signals necessary for downstream modeling tasks.

3.2 Dataset Schema

UPSC2M is a large-scale dataset comprising two components: an *attempts dataset* and a *questions dataset*. Each row in the attempts dataset represents a single interaction between a student and a question, capturing key fields including `user_id`, `question_id`, `user_answer`, `user_correct`, and `time_taken`. The accompanying questions dataset provides metadata for each question, including its `id`, `subject`, question `stem`, multiple-choice `options`, and the correct `answer`. While no student metadata is included, the dataset enables rich behavioral analysis: the `user_answer` field supports investigations into distractor effectiveness and common misconceptions, while the `time_taken` field—measured in seconds—offers a proxy for question engagement and fluency under time pressure. Each question is constrained to a 60-second limit, mirroring the real-world pacing of the UPSC exam.

3.3 Dataset Statistics

UPSC2M exhibits substantial scale and diversity in learner behavior across content categories. As shown in Table 2, each question is attempted by an average of 219 students, with some questions receiving over 3,000 attempts. This breadth of coverage stems from both the temporal dynamics of question exposure—where older or more prominently featured questions accumulate more interactions—and varying levels of learner interest across subject areas. Such variation necessitates models capable of generalizing across both high-frequency and low-frequency questions.

The average student attempted 42 questions, with the most active student answering over 6,500. This long-tailed distribution, typical of open educational platforms, supports modeling across a wide range of engagement levels. However, the low median number of questions per student indicates that many students engage only briefly, emphasizing the need for models that are robust to cold-start scenarios and sparse interaction histories.

4 Question Difficulty Estimation

4.1 Problem Formulation

We propose a task to estimate the empirical difficulty of a multiple-choice question using only its textual content. Each question is represented as a tuple `(id, subject, stem, options, answer)`, where `stem` denotes the question prompt, `options` is a list of four candidate choices, and `answer` specifies the index of the correct option.

The empirical difficulty of a question is defined as $1 - p_{\text{correct}}$, rounded to two decimal places, where p_{correct} denotes the proportion of students in UPSC2M who answered the question correctly among those who attempted it. This definition re-

Method	RMSE	MAE	R^2
Training Mean	0.2057	0.1699	-0.0001
Text Embedding	**0.1910**	**0.1543**	**0.1375**

Table 3: Test set performance of regression models for question difficulty estimation. The *Training Mean* baseline predicts the mean difficulty for all training samples.

flects the intuition that more difficult questions are associated with lower observed accuracy.

Setup To support reproducible evaluation, the questions dataset includes a predefined `split` field designating train, validation, and test partitions in a 70/15/15 split. Each question is also annotated with a precomputed `difficulty` score based on the formulation above.

4.2 Text Embedding Regression

As a baseline for question difficulty estimation, we adopt a simple regression approach. Specifically, we encode the question using a frozen pretrained text encoder and train a small MLP to predict the associated difficulty.

Each question is serialized as a single string combining the stem and options, which is then passed through OpenAI's `text-embedding-3-large` model—a general-purpose text embedding model. The resulting fixed-dimensional embedding serves as input to an MLP trained to minimize mean squared error against ground-truth difficulty scores. This approach offers a lightweight text-to-score mapping that sets a lower bound for models leveraging richer representations.

4.3 Results and Discussion

Our baseline achieves modest gains over a dummy regressor, reducing RMSE by 7.1% and MAE by 9.2%. While this demonstrates that semantic features carry some signal, the limited improvement underscores the difficulty of estimating question difficulty from text alone. These results motivate the incorporation of richer features—such as behavioral priors and structural cues.

Beyond benchmarking, automatic estimation of question difficulty has broad value in educational applications, enabling adaptive learning systems to personalize content to learner proficiency and maintain engagement. It also aids large-scale content management by facilitating question bank auditing, difficulty calibration, and the efficient construction of balanced assessments with minimal manual effort. In generative settings, difficulty estimation models can act as verifiers to ensure that newly created questions meet predefined pedagogical goals. As educational platforms scale across diverse curricula and learner populations, automated question difficulty estimation will become a cornerstone of personalized adaptive learning infrastructure.

5 Student Performance Prediction

5.1 Problem Formulation

We propose a task to predict whether a student will answer a given multiple-choice question correctly, based on their prior interaction history. Each row in the attempts dataset represents a single interaction and is formatted as a tuple (`user_id`, `question_id`, `user_answer`, `user_correct`, `time_taken`), where `user_correct` is a binary label indicating whether the response was correct.

For evaluation, the fields `user_answer`, `user_correct`, and `time_taken` are treated as target variables—models may access them during training but must not use them as input features at inference time. At test time, each example is defined solely by the pair (`user_id`, `question_id`), and the model must predict whether the student answers the question correctly.

Formally, this task involves estimating the conditional probability that a student answers a question correctly, given their historical behavior. This formulation mirrors real-world scenarios in adaptive learning systems, where predicting a learner's performance is essential.

Setup To facilitate reproducible evaluation, the attempts dataset includes a predefined `split` field that assigns each interaction to the training, validation, or test set, following an 80/10/10 ratio. The split is randomized at the interaction level, with post-processing to ensure that all students and questions in the validation and test sets also appear in the training set. This constraint ensures that models are evaluated on their ability to generalize to new interactions, rather than on cold-start cases with unseen students or questions.

5.2 Baselines

To contextualize the performance of more sophisticated models, we evaluate several simple baselines for this task.

Random and Zero Predictors As naive reference points, we consider two trivial classifiers. The *Random* baseline predicts correctness by sampling from the empirical label distribution in the training set, which shows a slight class imbalance (59.81% incorrect). The *Zero Predictor* always predicts the majority class (0 for incorrect), thereby serving as a worst-case lower bound on accuracy and calibration. While uninformative, these baselines are useful for verifying that more complex models exploit meaningful structure in the data.

Difficulty-Based Heuristic As a simple yet informative baseline, we ignore the student's interaction history and estimate the probability of a correct response based solely on the difficulty of the target question. Specifically, we compute the predicted probability as $1 - d$, where d denotes the difficulty score of the question, defined in Section 4.1. This formulation assumes that all students have an equal chance of answering a question correctly, modulated only by how empirically difficult the question is for the population.

Despite its simplicity, this baseline captures coarse priors over questions and highlights the influence of item difficulty on student performance. Comparing it to history-aware models underscores the value of incorporating personalized signals.

5.3 Collaborative Filtering

To assess the utility of standard recommender system techniques for modeling student performance, we evaluate several collaborative filtering (CF) (Su and Khoshgoftaar, 2009) methods that treat the task as a matrix completion problem. The student-question interaction matrix is constructed from observed correctness labels, and models are trained to predict whether a student will answer a given question correctly.

We include matrix factorization methods such as Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF), which learn low-dimensional embeddings for students and questions based on historical responses. We also evaluate a bias-only model that estimates correctness using additive student and item biases, as well as a K-nearest neighbors (KNN) approach that aggregates correctness labels from similar students. Together, these methods span a spectrum of personalization strategies, from global baselines to fine-grained models that exploit relational structure in the data.

Method	Accuracy	AUC	Brier
Random	0.5204	0.5000	0.2400
Zero Predictor	0.6002	0.5000	0.3998
Heuristic	0.6698	0.7118	0.2080
KNN CF	0.6429	0.6461	0.2330
SVD CF	0.6755	0.7133	0.2076
NMF CF	0.6757	0.7157	0.2100
Bias Only CF	**0.6788**	**0.7210**	**0.2051**

Table 4: Test set performance of baseline methods on the student performance prediction task. *CF* denotes collaborative filtering.

These models serve as a classical baseline for student performance prediction, illustrating how much signal can be captured from past interactions alone, without access to question content.

5.4 Results and Discussion

Table 4 reports the performance of all baseline models on the student performance prediction task. The *Heuristic* model substantially outperforms trivial baselines, demonstrating that question difficulty alone provides a strong prior for estimating student success. This suggests that well-estimated item-level difficulty can serve as a meaningful signal, even without any personalization.

Among collaborative filtering methods, *Bias Only* yields the highest overall performance, while more expressive models such as *SVD*, *NMF*, and *KNN* fail to produce significant gains in accuracy. The high sparsity of the student-question matrix (99.62%) likely inhibits the ability of these models to learn effective representations or student neighborhoods, constraining their ability to capture student-specific patterns beyond simple item and student-level tendencies.

Predicting student performance is vital to adaptive educational systems, enabling personalized question selection, targeted review, and adaptive pacing to support diverse learners. When paired with difficulty estimation, it lays the groundwork for fully automated instruction by combining item-level insights with behavioral modeling. As educational platforms scale, these predictive capabilities are key to delivering truly individualized learning—ensuring each student receives the right content at the right time. Together, these tasks form the backbone of scalable, data-driven education.

Limitations

Our question difficulty estimation labels are based solely on correctness rates and ignore temporal or student-specific variation; future work may redefine difficulty through joint modeling of student and item characteristics, potentially incorporating response times. Our collaborative filtering models are likely hindered by the high prevalence of low-activity learners—the median questions attempted per student is just 8—which may limit generalization and overall performance. None of our current models incorporate response time features, which could offer valuable signals related to fluency or hesitation. Finally, while UPSC2M is large and diverse, its focus on one high-stakes exam context may limit direct transferability to other educational domains. Despite these limitations, we view our dataset and task formulations as a strong foundation for building more expressive, interpretable, and personalized models of learner behavior.

References

Au Blum and James E. Corter. 2014. Estimating question difficulty and user ability in a collaborative question answering community. In *Workshop on Personalized and Adaptive Learning in EDM*.

Youngduck Choi, Youngnam Lee, Dongmin Shin, Junghyun Cho, Seoyon Park, Seewoo Lee, Jineon Baek, Chan Bae, Byungsoo Kim, and Jaewe Heo. 2020. Ednet: A large-scale hierarchical dataset in education. In *Proceedings of the 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, volume 12164 of *Lecture Notes in Computer Science*, pages 69–73. Springer.

Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum.

Philip I. Pavlik Jr, Hui Cen, and Kenneth R. Koedinger. 2009. Performance factors analysis: A new alternative to knowledge tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pages 505–513.

John C. Stamper, Kenneth R. Koedinger, Ryan S. J. d. Baker, Alida Skogsholm, Brett Leber, Sandy Demi, Shawnwen Yu, and Duncan Spencer. 2011. Datashop: A data repository and analysis service for the learning science community. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED)*, page 628, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:1–19.

David J. Weiss. 2011. *Adaptive Testing*. Oxford University Press.