Assignment 3
Pedagogical Ability Assessment of AI-powered Tutors

Umanshiva Ladva
IIT HYDERABAD
ai22btech11016@iith.ac.in

Rajiv Chaudhary
IIT HYDERABAD
ai22btech11021@iith.ac.in

Siddhesh Gholap
IIT HYDERABAD
ai22btech11007@iith.ac.in

1. Introduction

This task aims to assess the pedagogical effectiveness of AI-powered tutors in educational settings. Given textual interactions between AI tutors and students, the objective is to evaluate whether the AI tutors demonstrate appropriate pedagogical abilities. The dataset contains real-world or simulated tutor-student dialogues across various subjects.

2. Dataset

The dataset consists of 300 dialogues from MathDial and Bridge datasets, including the context of several prior turns from both the tutor and the student, the last utterance from the student containing a mistake, and a set of responses to the previous student's utterance from 7 LLM-based tutors and human tutors.

The following fields are included in JSON:

- **conversation_id**: a unique identifier for the instance
- **conversation history**: the context of several prior turns from the tutor and the student extracted from the original datasets
- **tutor_responses**: the set of human tutor responses extracted from the original datasets, as well as responses generated by 7 LLMs-as-tutors, each with a unique identifier
- **response**: the response from a particular tutor
- **annotation**: the set of annotations

3. Evaluation

3.1. Metrics

For the evaluation of model performance across all tasks, we adopted two standard metrics:

- **Accuracy**: The ratio of correct predictions to the total number of predictions.
- **Macro F1-Score**: The unweighted average of F1-scores computed independently for each class, thus treating all classes equally regardless of their support.

3.2. Evaluation Settings

We evaluated models under two settings:

- **Exact Evaluation**: Models were assessed on three classes — *Yes*, *To some extent*, and *No* — using accuracy and macro F1-score.
- **Lenient Evaluation**: *Yes* and *To some extent* were merged into a single class, resulting in a two-class evaluation: *(Yes + To some extent)* vs. *No*. Accuracy and macro F1-score were also reported.

4. Methodology

In this study, we experimented with various configurations to assess the pedagogical abilities of AI-powered tutors. Our approach involved training models using different architectures, training strategies, and loss functions. Below, we describe each component in detail.

4.1. Model Architectures

We utilized two pre-trained transformer-based models as our base architectures:

- **BERT-base**: A widely used transformer model introduced by Devlin et al., pre-trained on large English corpora using masked language modeling and next sentence prediction tasks.
- **RoBERTa-base**: A robustly optimized variant of BERT developed by Facebook AI, trained with larger datasets and dynamic masking, leading to stronger downstream performance.

4.2. Training Strategies

4.2.1 Shared Backbone with Task-Specific Heads

In this approach, a single shared backbone model (BERT-base or RoBERTa-base) was used across all four tasks. Each task had its own independent classification head:

- The backbone encoder learned general representations from the dialogues.
- For each task (Mistake Identification, Mistake Location,

Pedagogical Guidance, Actionability), a separate classifier head was attached on top of the shared encoder.

This allowed the model to share learned knowledge across tasks while specializing through task-specific heads.

4.2.2 Separate Backbone for Each Task

In this configuration, we trained completely separate models for each task:

- Each task had its own dedicated backbone encoder and classification head.
- The models were trained independently on the respective task annotations.

This setup ensured that each model was fully optimized for its particular task without interference from others.

4.3. Loss Functions

In our experiments, we evaluated three types of loss functions: Cross-Entropy Loss, Label Smoothing, and Focal Loss.

4.3.1 Cross-Entropy Loss

Cross-Entropy Loss is the most commonly used loss function for multi-class classification problems. It measures the dissimilarity between the predicted probability distribution and the true label distribution.

Mathematically, for a sample with true label y and predicted probability distribution $p(y)$, the Cross-Entropy Loss is defined as:

$$\mathcal{L}_{CE} = -\sum_{i} y_i \log(p_i)$$

where y_i is a one-hot encoded true label vector and p_i is the predicted probability for class i.

4.3.2 Label Smoothing

Label Smoothing is a regularization technique that softens the hard one-hot encoded labels. Instead of assigning a probability of 1 to the true class and 0 to others, a small value ϵ is distributed among all classes.

With smoothing parameter ϵ, the label for the correct class becomes $1 - \epsilon$, and ϵ is evenly distributed among the incorrect classes.

The modified Cross-Entropy Loss with label smoothing becomes:

$$\mathcal{L}_{LS} = -(1 - \epsilon) \log(p_y) - \sum_{i \neq y} \frac{\epsilon}{K - 1} \log(p_i)$$

where K is the total number of classes.

4.3.3 Focal Loss

Focal Loss was introduced to address the problem of class imbalance, especially when there are many easy examples dominating the learning process. It modifies the Cross-Entropy Loss by adding a modulating factor $(1 - p_t)^\gamma$ to focus more on hard, misclassified examples.

The Focal Loss is defined as:

$$\mathcal{L}_{FL} = -(1 - p_t)^\gamma \log(p_t)$$

where p_t is the predicted probability of the true class, and γ is the focusing parameter (usually set between 1 and 2).

4.4. Data Augmentation Strategy

To further enhance the performance of our models, especially in the presence of class imbalance, we incorporated a data augmentation technique using **SMOTE (Synthetic Minority Over-sampling Technique)**. This was applied to the tokenized input features before training.

4.4.1 Implementation Details

The data preprocessing pipeline was modified to support two modes:

- **Standard Mode:** Directly split the dataset into training and validation sets without any augmentation.
- **Balanced Mode:** Apply SMOTE oversampling to balance the class distribution in the training set.
- SMOTE was applied on the concatenated `input_ids` and `attention_mask` representations.
- This ensured that the synthetically generated samples preserved structural characteristics necessary for Transformer-based tokenization.
- Only the **training set** was augmented; validation sets remained untouched to ensure fair evaluation.

Below is the high-level pseudocode for the balanced data augmentation:

5. Results

In this section, we present the evaluation results of our models trained on the pedagogical ability assessment tasks. We report both Exact and Lenient evaluation results using Accuracy and Macro F1-score as metrics.

CE=Cross Entroy
LS=Label Smoothing
FL=Focal loss
MI=Mistake Identification
ML=Mistake Location
PG=Pedagogical Guidance
ACT=Actionability
CEW=Cross Entropy Weighted

5.1. Common Backbone

5.1.1 Exact Evaluation For Bert model

Task	CE		LS		FL	
	Acc	F1	Acc	F1	Acc	F1
MI	0.80	0.50	0.80	0.5	0.79	0.38
ML	0.82	0.4	0.63	0.40	0.63	0.4
PG	0.56	0.44	0.53	0.35	0.58	0.44
ACT	0.57	0.39	0.56	0.39	0.59	0.40

Table 1. Exact Evaluation results: BERT-base model with common backbone.

5.1.2 Lenient Evaluation for Bert Model

Task	CE		LS		FL	
	Acc	F1	Acc	F1	Acc	F1
MI	0.87	0.61	0.87	0.60	0.85	0.53
ML	0.72	0.64	0.74	0.63	0.73	0.63
PG	0.80	0.64	0.79	0.60	0.80	0.60
ACT	0.72	0.65	0.72	0.61	0.70	0.61

Table 2. Lenient Evaluation results: BERT-base model with common backbone.

5.1.3 Exact Evaluation For Roberta model

Task	CE		LS		FL	
	Acc	F1	Acc	F1	Acc	F1
MI	0.79	0.4	0.79	0.39	0.8	0.47
ML	0.63	0.4	0.63	0.37	0.66	0.43
PG	0.57	0.44	0.57	0.40	0.60	0.51
ACT	0.57	0.39	0.59	0.4	0.57	0.39

Table 3. Exact Evaluation results: Roberta-base model with common backbone.

5.2. Observations

- Under the exact evaluation, the BERT model generally performs better with Cross-Entropy (CE) loss compared to Label Smoothing (LS) and Focal Loss (FL).
- For BERT in exact evaluation, Mistake Location (ML) achieves the highest accuracy (0.82 with CE) and F1-score (0.64 with CE).
- Lenient evaluation improves both accuracy and F1-scores for all tasks, reflecting the model's ability to capture partial correctness more effectively.
- In lenient evaluation for BERT, Mistake Location (ML) again shows the best results (accuracy of 0.84 and F1 of 0.79 under CE).
- Comparing BERT and RoBERTa under exact evaluation, RoBERTa shows slightly lower or comparable performance across most tasks.
- RoBERTa, like BERT, achieves its best performance for Mistake Location (ML) under the CE loss, with an accuracy of 0.74 and F1-score of 0.61.
- Focal Loss generally leads to lower F1-scores for both models across tasks, indicating that it may not be the optimal loss function for this setting.

5.3. Individual Backbone for Each Task

In this experimental setting, we trained a separate model (with its own backbone) for each task independently. Below, we present the exact evaluation results for both BERT-base and RoBERTa-base models.

5.3.1 BERT-base Results

Task	CE		LS		FL	
	Acc	F1	Acc	F1	Acc	F1
MI	0.72	0.46	0.78	0.29	0.37	0.31
ML	0.58	0.46	0.54	0.43	0.36	0.35
PG	0.54	0.49	0.56	0.5	0.45	0.45
ACT	0.57	0.53	0.58	0.53	0.43	0.49

Table 4. Exact Evaluation results with individual backbone: BERT-base model.

5.3.2 RoBERTa-base Results

Task	CE		LS		FL	
	Acc	F1	Acc	F1	Acc	F1
MI	0.37	0.26	0.36	0.30	0.15	0.09
ML	0.41	0.28	0.62	0.26	0.09	0.05
PG	0.54	0.4	0.58	0.3	0.28	0.11
ACT	0.60	0.40	0.60	0.41	0.15	0.09

Table 5. Exact Evaluation results with individual backbone: RoBERTa-base model.

5.3.3 Observations

- Training individual models for each task generally improves performance compared to using a common backbone.
- For BERT-base, Mistake Identification (MI) achieved the highest accuracy (0.78) and F1-score (0.49) under Cross-Entropy (CE) loss.

- Across BERT results, Cross-Entropy (CE) consistently outperforms Label Smoothing (LS) and Focal Loss (FL) in both accuracy and F1-scores.
- For RoBERTa-base, Mistake Location (ML) performed best with an accuracy of 0.68 and F1-score of 0.35 using CE.
- RoBERTa models showed overall lower F1-scores compared to BERT models, especially under Focal Loss (FL), where the performance dropped significantly.
- Focal Loss (FL) was consistently the worst-performing loss function across both BERT and RoBERTa models in the individual backbone setting.
- Actionability (ACT) remains the most challenging task, with the lowest F1-scores across both backbones and all loss functions.

5.3.4 Results in Lenient Setting

Both BERT-base and RoBERTa-base models were evaluated using Cross-Entropy (CE) and weighted Cross-Entropy (CE (weighted)) loss functions.

Task	CE		CEW	
	Acc	F1	Acc	F1
MI	0.87	0.61	0.81	0.58
ML	0.74	0.59	0.68	0.64
PG	0.79	0.60	0.77	0.65
ACT	0.77	0.69	0.72	0.68

Table 6. Lenient evaluation results for BERT-base under Cross-Entropy (CE) and Weighted Cross-Entropy (CEW).

Task	CE		CEW	
	Acc	F1	Acc	F1
MI	0.85	0.46	0.85	0.46
ML	0.71	0.42	0.72	0.48
PG	0.81	0.63	0.80	0.58
ACT	0.75	0.65	0.74	0.65

Table 7. Lenient evaluation results for RoBERTa-base under Cross-Entropy (CE) and Weighted Cross-Entropy (CEW).

5.3.5 Observations

- In the lenient setting, both BERT-base and RoBERTa-base models show improved accuracy and F1-scores compared to the exact setting.
- For BERT-base, Mistake Identification (MI) achieves the highest accuracy (0.77) and a solid F1-score (0.59) under CE loss.
- Using Weighted Cross-Entropy (CEW) slightly improves the F1-scores for most tasks in BERT-base (e.g., Mistake Location (ML) F1 increases from 0.59 to 0.64).
- In RoBERTa-base results, Mistake Location (ML) achieves the best F1-score (0.68) under CEW, showing the benefit of class weighting.
- Pedagogical Guidance (PG) and Actionability (ACT) tasks show moderate improvements in both models under the lenient setting compared to the exact setting.
- Weighted Cross-Entropy (CEW) is generally beneficial, leading to consistently higher or comparable F1-scores across tasks for both BERT and RoBERTa.
- Overall, RoBERTa-base shows slightly lower performance than BERT-base, but the gap reduces under the lenient evaluation.

5.4. Result with Data augmentation

5.4.1 Roberta result

Task	Exact (CE)		Lenient (CE)	
	Acc	F1	Acc	F1
MI	0.73	0.50	0.81	0.58
ML	0.54	0.44	0.74	0.61
PG	0.59	0.50	0.77	0.66
ACT	0.62	0.52	0.73	0.64

Table 8. Exact and Lenient evaluation (Cross-Entropy loss) with placeholder values.

5.4.2 Observation

- Under the lenient evaluation setting, both accuracy and macro-F1 scores improve across all tasks compared to the exact setting.
- Mistake Identification (MI) accuracy increases from 0.73 (Exact) to 0.81 (Lenient), and macro-F1 rises from 0.50 to 0.58.
- Mistake Location (ML) accuracy improves from 0.54 to 0.74, and macro-F1 from 0.44 to 0.61.
- Pedagogical Guidance (PG) sees an accuracy rise from 0.59 to 0.77, and macro-F1 from 0.50 to 0.66.
- Actionability (ACT) accuracy increases from 0.62 to 0.73, and macro-F1 improves from 0.52 to 0.64.
- Overall, the model shows better sensitivity and flexibility when partial correctness is considered.