Pedagogical Ability Assessment of AI-powered Tutors

Assignment-3

Sai Kiran - AI24MTECH02003
Dhana Lakshmi - AI22BTECH11012
Jayadeep - CS24MTECH14009

April 27, 2025

Course: Natural Language Processing
IIT HYDERABAD

Outline

Introduction

Problem Statement

Dataset

Methodology

Architecture & Implementation

Results & Discussions

Comparitive Study

References

Introduction

Introduction

- **Task goal**: Assess the *pedagogical effectiveness* of AI-powered tutors by analyzing their dialogue with students.
- **Context**: LLM-based tutors (e.g. GPT-4, LLaMA) offer scalable, personalized support but standard NLG metrics miss true teaching impact.
- **Approach**: Examine 300 real/simulated math dialogues—each with a student error, one human response, and seven LLM replies—using four pedagogical tasks (mistake identification, location, guidance, actionability).

Problem Statement

Problem Statement

- **Core challenge**: There is no standardized way to verify whether AI tutors can accurately detect and correct student mistakes, provide meaningful guidance, and suggest clear next steps.
- **Evaluation tasks**: We focus on four essential pedagogical abilities:
  - *Mistake Identification*
  - *Mistake Location*
  - *Pedagogical Guidance*
  - *Actionability*

Dataset

Dataset Overview

- **Size & source**: 300 math dialogues drawn from the MathDial and Bridge collections.
- **Instance structure** (JSON):
  - `conversation_id`, full `history` of prior turns
  - `student_utterance`: final turn containing an error
  - `tutor_responses`: one human "gold" reply + 7 LLM-generated replies
  - `annotations`: human labels for mistake identification, location, guidance, actionability
- **Use**: Provides a standardized testbed for comparing how well different tutors diagnose and remediate student mistakes.

Methodology

Methodology Overview

- **Data**: 300 MathDial & Bridge dialogues, 80/20 split
- **Tasks**: Mistake ID, Location, Guidance, Actionability
- **Evaluation**:
  - Exact (3-class) vs. Lenient (2-class) metrics
  - Accuracy, macro-F1

Architecture & Implementation

Model & Training

- **Backbone**: Pre-trained RoBERTa-base
- **Heads**: Four binary/3-class linear heads (one per task)
- **Regularization**: Dropout (0.2) on pooled classes
- **Loss**: Sum of cross-entropy (or focal) over tasks
- **Optimizer**: AdamW, LR=2e-5, batch=8
- **epochs**:50
- **Framework**: PyTorch

Model & Training

RoBERTa: Robustly Optimized BERT Pretraining Approach - RoBERTa is a transformer-based NLP model developed by Meta AI. It builds on BERT by using more training data, improved training strategies, and removing components that were found to be unnecessary. RoBERTa keeps the same model architecture as BERT but enhances performance significantly through better optimization.

Here better optimization refers to :

- It was trained on 10x more data than BERT i.e, RoBERTa trained on 160GB of data.
- It re-masks text at each epoch
- It removes NSP(Next Sentence Prediction), leading to more focused training on token-level understanding

How Inputs Are Given to the Model

Step-by-Step Input Preparation:

1. Take the **entire conversation history** up to the current tutor response.
2. Append the **current tutor response** separately.
3. Combine them as a single input text:

   `[Conversation History] + ‘‘Tutor Response: ’’ + [Current Response]`

Resulting Input to Model:

- Full conversation history **and** the new tutor response concatenated.
- Tokenized, truncated/padded to `max_length=512`.

Results & Discussions

Exact vs. Lenient Performance

Tasks	Accuracy		Macro F1	
	3 Class	**2 Class**	**3 Class**	**2 Class**
Mistake Identification	89.11	**93.9**	69.9	**96.5**
Mistake Location	75.6	**81.2**	54.5	**87.3**
Providing Guidance	67.14	**82.9**	58.45	**89.2**
Actionability	73.6	**84.9**	64.48	**88.5**

Table 1: Performance of different tasks across 3-class and 2-class classification settings.

Discussions

Overall Task Performance

From Table 1

- 2-Class classification consistently outperforms 3-Class in both Accuracy and Macro F1.
- Mistake Identification and Mistake Location are easier tasks, with higher scores.
- Providing Guidance and Actionability remain more challenging, especially in 3-Class settings.

Accuracy Comparison of Models

Table 2: Comparison of model performance across four NLP tasks using 3-class and 2-class accuracy metrics.

Model Name	No. of Samples	Mistake Identification		Mistake Location		Providing Guidance		Actionability	
		3 Class	2 Class	3 Class	2 Class	3 Class	2 Class	3 Class	2 Class
Novice	16	62.5	75	75	75	**87.5**	**87.5**	**87.5**	87.5
Phi3	60	95	95	**95**	96.6	78.3	85	83.3	88.3
Gemini	60	88.3	88.3	50	68.3	46.6	78.3	43.3	73.3
Expert	60	66.6	95	55	70	66.6	83.3	71.6	**95**
Mistral	60	91.6	98.3	65	76.6	46.6	68.3	55	73.3
GPT-4	60	95	95	80	86.6	61.6	80	60	66.6
Llama31405B	60	**98.6**	**98.6**	85	**93.3**	73.3	**95**	75	85
Llama318B	60	81.6	90	53.3	58.3	40	68.3	50	56.6
Sonnet	60	85	97.2	66.6	80	48.3	78.3	53.3	88.3

Macro F1 Comparison of Models

Table 3: Macro F1 scores of models across NLP tasks under 3-class and 2-class

Model Name	No. of Samples	Mistake Identification		Mistake Location		Providing Guidance		Actionability	
		3 Class	2 Class	3 Class	2 Class	3 Class	2 Class	3 Class	2 Class
Novice	16	43.3	42.8	**66.6**	66.6	**79.4**	**79.4**	**57.4**	**79.4**
Phi3	60	94.3	94.3	63.1	**96.1**	67.6	**84.2**	45.3	78.1
Gemini	60	46.9	46.9	31.8	54.5	31.8	43.9	33	58.3
Expert	60	26.6	48.7	32.4	53.1	50	59.5	27.8	48.7
Mistral	60	31.8	49.5	31.1	49.5	30.7	48.9	31.9	47.7
GPT-4	60	48.7	48.7	36.5	56.63	41.9	56.7	41.6	66.7
Llama31405B	60	**98.6**	**98.6**	50.6	48.2	28.2	48.7	44.1	69.1
Llama318B	60	29.9	47.3	35.1	52.5	36	48.9	**48.3**	55.8
Sonnet	60	45.9	97.2	37.9	60.7	34.6	50.4	34.5	47

Note: Macro F1 scores were multiplied by 100 for better readability and comparison.

Discussions

- **2-Class** tasks consistently outperform **3-Class** in both Accuracy and Macro F1.
- **Llama31405B** is the best overall model, achieving top scores across all tasks.
- **Phi3** is a strong runner-up, especially in Mistake Location.
- **GPT-4** and **Expert** perform well in Accuracy but show lower F1, indicating class imbalance.
- **Providing Guidance** and **Actionability** are the hardest tasks, particularly in 3-Class settings.
- Simpler tasks like **Mistake Identification** yield higher, more stable performance across models.

Comparitive Study

Comparitive Study

We conducted a comparative study to evaluate the performance of three transformer-based language models:

- **BERT**
- **RoBERTa**
- **DistilBERT**

Each model was tested on multiple tasks such as Mistake Identification, Mistake Location, Providing Guidance, and Actionability under both 2-Class and 3-Class settings.

Among the three, **RoBERTa** consistently achieved the highest performance across tasks, making it the most effective model for our use case.

DistilBERT Model & Training

- **Backbone**: Pre-trained distilbert-base-uncased
- **Heads**: Four binary/3-class linear heads (one per task)
- **Regularization**: Dropout (0.2) on pooled `[CLS]`
- **Loss**: Sum of cross-entropy (or focal) over tasks
- **Optimizer**: AdamW, LR=4e-5, batch=32, epochs=25
- **Framework**: PyTorch + HuggingFace Transformers

Exact vs. Lenient Performance

	Exact (3-class)		Lenient (2-class)	
Task	Acc	Macro-F1	Acc	Macro-F1
Mistake ID	0.87	0.67	0.93	0.96
Mistake Loc	0.74	0.54	0.80	0.86
Guidance	0.67	0.58	0.82	0.88
Actionability	0.72	0.63	0.84	0.88

Table :Performance of different tasks across 3-class and 2-class classification.

Performance Comparison: RoBERTa vs. DistilBERT

Table:Model performance across tasks (2-Class Setting)

Tasks	RoBERTa		DistilBERT	
	Accuracy	Macro F1	Accuracy	Macro F1
Mistake Identification	**94.7**	**96.9**	92.7	95.2
Mistake Location	**83.1**	**88.6**	80	86
Providing Guidance	**84.3**	**89.9**	82	88
Actionability	**85.5**	**89.1**	84	88

RoBERTa slightly outperforms DistilBERT in all tasks and metrics, especially in Macro F1.

Conclusion and Future Scope

Comparative Study: We evaluated BERT, RoBERTa, and DistilBERT across four pedagogical tasks. **RoBERTa** consistently outperformed the others in both Accuracy and Macro F1, making it the most effective model for our use case.

Key Challenges:

- Low performance in 3-class settings due to increased label complexity.
- Providing Guidance and Actionability remain difficult tasks across models.
- Class imbalance affects macro F1, especially in underrepresented categories.

Future Scope:

- Plan to model tutor responses sequentially from last to first, capturing backward conversational dependencies to enhance understanding and prediction accuracy.
- Incorporate multi-turn dialogue modeling for richer pedagogical context.
- Explore instruction-tuned and task-adaptive LLMs (e.g., GPT-4 Turbo, Claude).

References

References

- Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019.
- Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
- Maurya, K.K., Srivatsa, K.V., Petukhova, K. and Kochmar, E., 2024. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach.