Task Difficulty Prediction for USMLE Dataset (Amboss)

Rahul Saxena
UMass Amherst
rahulsaxena@umass.edu

Anirudh Hariharan
UMass Amherst
anirudhharih@umass.edu

Abstract

This study explores the task of difficulty prediction for medical MCQs in the USMLE dataset (Amboss). The project is divided into three parts: training the Electra model for classification, leveraging GPT-4o-mini with DSPy for prompt optimization, and developing a custom training pipeline for the LLama model. Challenges related to dataset imbalance and task complexity were identified, and various approaches were tested to improve prediction accuracy.

1 Introduction

Accurately predicting the difficulty of medical MCQs is critical for optimizing educational resources and ensuring effective learning. This research focuses on building and evaluating machine learning models to classify difficulty levels in the Amboss dataset. The dataset comprises 5000 samples, with difficulty levels ranging from 1 to 5. The project investigates three methodologies:

1. Training a BERT-based model (Electra-Large).
2. Utilizing GPT-4o-mini with DSPy for automatic prompt generation.
3. Developing a fine-tuned LLama model with a curated dataset.

2 Dataset

The Amboss dataset consists of 5000 medical MCQ samples labeled with difficulty levels ranging from 1 (easiest) to 5 (hardest). A notable class imbalance was observed, with difficulty levels 2 and 3 dominating the distribution. This imbalance posed significant challenges for model training and evaluation.

3 Training Electra-Large Model for Difficulty Prediction

3.1 Objective

To train a classification model using the Google Electra-Large architecture to predict the difficulty levels of medical MCQs.

3.2 Methodology

Model Configuration: Electra-Large model was fine-tuned for the classification task.

Hyperparameter Tuning: Various configurations were tested, including:

- Learning rates: $10^{-5}, 10^{-4}, 10^{-3}$
- Batch sizes: 1, 4, 8, 16
- Epochs: 1, 2, 3

Training Details: The final configuration used a batch size of 16 and 2 epochs.

3.3 Results

Despite hyperparameter tuning, the model achieved an accuracy of approximately 33%. The performance was skewed towards majority classes (difficulty levels 2 and 3), indicating poor generalization for minority classes (difficulty levels 1, 4, and 5).

3.4 Challenges

1. **Class Imbalance:** Predicted results favored majority classes, reducing accuracy for minority classes.
2. **Task Complexity:** The task required understanding nuanced medical terminology and context, which the model struggled to capture.

3.5 Conclusion

The Electra-Large model was insufficient for this task due to class imbalance and inherent complexity. Future work includes exploring alternative architectures and resampling techniques.

4 Leveraging GPT-4o-mini with DSPy

4.1 Objective

To utilize GPT-4o-mini for difficulty prediction by generating optimized prompts using DSPy.

4.2 Dataset Modification

The dataset was transformed to explore multi-class and binary classification:

- **Five-class classification:** Retained original labels (1 to 5).
- **Three-class classification:** Combined classes as follows:
  - Levels 1 and 2 $\rightarrow$ Class 1
  - Level 3 $\rightarrow$ Class 2
  - Levels 4 and 5 $\rightarrow$ Class 3
- **Two-class classification:** Simplified into:
  - Levels 1, 2, 3 $\rightarrow$ Class 1
  - Levels 4, 5 $\rightarrow$ Class 2

4.3 Methodology

Prompt Optimization: DSPy was used to automatically generate optimal prompts for each classification task.

Evaluation: GPT-4o-mini was tested on the modified datasets.

4.4 Results

- **Five-class classification:** Accuracy – 20%
- **Three-class classification:** Accuracy – 35%
- **Two-class classification:** Accuracy – 54%

4.5 Conclusion

The GPT-4o-mini model performed poorly, with results comparable to random guessing. Further investigation is needed into the limitations of prompt-based models for this task.

5 Training LLama Model for Difficulty Prediction

5.1 Dataset Preparation

Initial Predictions:

- GPT-4o-mini was used to predict difficulty levels (two-class) for 1000 samples, with reasoning.
- Correct predictions (542 samples) were retained.

Ground Truth Reasoning:

- For the remaining 458 samples, GPT-4o-mini was provided ground truth labels to generate reasoning.

Cross-Evaluation:

- Samples were reevaluated using reasoning. Predictions matching ground truth were retained, while the rest were discarded.

5.2 Model Training

We took the 542 samples of correctly predicted questions with options and the correct answer along with the predicted label and gpt-reasoning. This file was the dataset.csv that was used to fine-tune the llama model.

A 4-bit Llama 3-8B model was used for fine-tuning on the difficulty prediction downstream task. LoRA(Low Rank Adaptation) was applied for parameter efficient fine-tuning.

Training Configuration

SFT-Trainer was configured with settings optimized for memory and computational efficiency.
Learning Rate: 3e-4
Batch Size: 2
Warmup Steps: 50
Training Epochs: 5
Optimizer: AdamW with 8-bit precision.
Weight Decay: 0.01

Model Output: Model output was a fine-tuned model and tokenizer saved for downstream task, and inference.

5.3 Evaluation

For the evaluation metrics and inference, A inference.csv was created with 100 rows (50 easy and 50 difficult) of questions with options and correct answer, along with the ground truth difficulty level and ground truth reasoning.

An evaluation pipeline was developed to assess the following metrics:

- **Difficulty Prediction:**
  - **Accuracy:** 0.5521
  - **F1 Score:** 0.5221
  - **Precision:** 0.5621
  - **Recall:** 0.5521
  - **Confusion Matrix:**

$$\begin{bmatrix} 39 & 10 \\ 33 & 14 \end{bmatrix}$$

- **Reasoning Quality:**
  - **BERT Score:**
    - **Precision:** 0.9000
    - **Recall:** 0.8631
    - **F1 Score:** 0.8811
  - **METEOR Score:** 0.2766
- **LLM Judge Scores:**
  - **Mean Score:** 2.3958

The model demonstrates moderate performance in predicting difficulty, with an accuracy of 0.5521, F1 score of 0.5221, and balanced precision and recall around 0.55. The confusion matrix suggests that the model struggles with identifying positive instances, resulting in relatively low recall. However, its reasoning capabilities are strong, as evidenced by high BERT scores (precision: 0.9000, recall: 0.8631, F1: 0.8811), indicating effective handling of reasoning tasks.

Despite these strengths, the METEOR score of 0.2766 and the average human evaluator score of 2.4 reflect that the model's generated outputs still have room for improvement in fluency and coherence.

Overall, while the model excels in reasoning quality, it needs refinement in difficulty prediction and human evaluation to achieve more consistent and reliable results.

6 Discussion

The findings highlight the challenges of difficulty prediction for medical MCQs, particularly with imbalanced datasets and nuanced task requirements. While Electra-Large and GPT-4o-mini provided baseline insights, the LLama model's custom training pipeline shows promise for addressing these challenges.

7 Conclusion

This research underscores the need for advanced techniques in handling imbalanced datasets and complex linguistic tasks. Future work will focus on:

1. Fine-tuning the LLama model with additional curated data.
2. Exploring ensemble models and regression-based approaches.
3. Enhancing evaluation metrics to better capture reasoning quality.

References