ITEC at BEA 2024 Shared Task: Predicting Difficulty and Response Time of Medical Exam Questions with Statistical, Machine Learning, and Language Models

Anaïs Tack [1ad] **Siem Buseyne** [1bd, 2] **Changsheng Chen** [1bd]
Robbe D'hondt [1cd *] **Michiel De Vrindt** [1bd †] **Alireza Gharahighehi** [1cd]
Sameh Metwaly [1bd, 3] **Felipe Kenji Nakano** [1cd *] **Ann-Sophie Noreillie** [1ad]
[1] KU Leuven [a] Faculty of Arts [b] Faculty of Psychology and Educational Sciences
[c] Department of Public Health and Primary Care [d] imec research group itec
[2] Université de Lille, CIREL [3] Damanhour University

Abstract

This paper presents the results of our participation in the BEA 2024 shared task on the automated prediction of item difficulty and item response time (APIDIRT), hosted by the NBME (National Board of Medical Examiners). During this task, practice multiple-choice questions from the United States Medical Licensing Examination® (USMLE®) were shared, and research teams were tasked with devising systems capable of predicting the difficulty and average response time for new exam questions.

Our team, part of the interdisciplinary itec research group, participated in the task. We extracted linguistic features and clinical embeddings from question items and tested various modeling techniques, including statistical regression, machine learning, language models, and ensemble methods. Surprisingly, simpler models such as Lasso and random forest regression, utilizing principal component features from linguistic and clinical embeddings, outperformed more complex models. In the competition, our random forest model ranked 4th out of 43 submissions for difficulty prediction, while the Lasso model secured the 2nd position out of 34 submissions for response time prediction. Further analysis suggests that had we submitted the Lasso model for difficulty prediction, we would have achieved an even higher ranking. We also observed that predicting response time is easier than predicting difficulty, with features such as item length, type, exam step, and analytical thinking influencing response time prediction more significantly.

1 Introduction

In the medical domain, standardized tests act as crucial gatekeepers, allowing only the best healthcare professionals into the field. An example is the *United States Medical Licensing Examination*® (USMLE®), a high-stakes exam administered by the National Board of Medical Examiners (NBME) to assess a medical student's ability to provide safe and effective patient care. However, for these exams to accurately gauge the competency of medical students, organizations like the NBME meticulously design their assessments, with a specific focus on balancing the difficulty and response time of exam questions. This is essential for ensuring the fairness and validity of the exams, as test items should cover a wide range of difficulty levels, and each question should be allocated an appropriate amount of time.

Prior studies by NBME researchers have shown that predicting the difficulty and response time of medical exam questions is a challenging task (Ha et al., 2019a; Xue et al., 2020; Yaneva et al., 2020, 2021). As a result, the NBME launched an international challenge where they provided researchers with a set of retired exam questions from the USMLE®. Research teams were tasked with developing a system or model that takes as input a multiple-choice question and produces as output two estimates: (a) how challenging it is for test-takers and (b) how long it would take them to respond (see Figure 1 for an illustration). The comprehensive details and results of this shared task are outlined in the overview paper authored by Yaneva et al. (2024).

We participated in the competition with the **ITEC**[1] **team**, an interdisciplinary research group affiliated with KU Leuven and imec. Our collaborative efforts span various fields, including artificial intelligence, educational sciences, language technology, machine learning, psychometrics, and statistical modeling. Our strategy involved a fusion of statistical models, machine learning models, and language models. We integrated traditional

*supported by a fellowship from the Research Foundation Flanders (FWO)

†supported by a Baekeland mandate from the Flanders Innovation & Entrepreneurship (VLAIO)

[1] https://itec.kuleuven-kulak.be

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 512–521

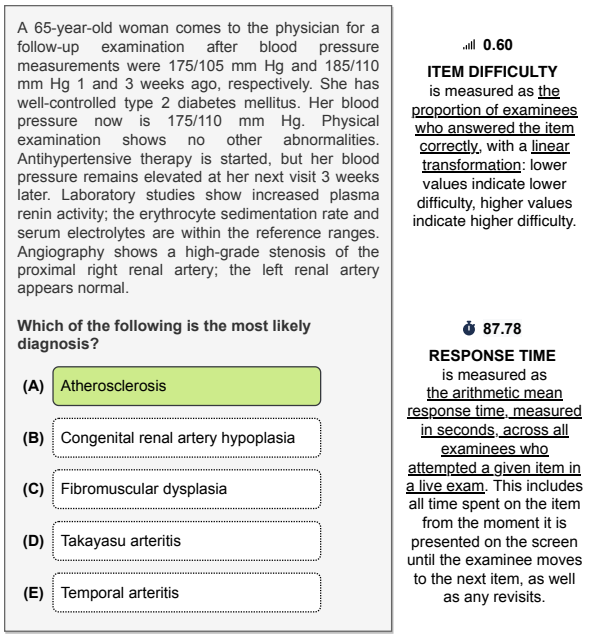

Figure 1: Example of a multiple-choice question from the USMLE® Step 1 provided by the NBME during the shared task's training phase. Each question had an item stem and up to ten possible answers and was labeled with item difficulty and average response time.

feature engineering with contemporary fine-tuning and transfer learning approaches. The following sections will delve into our method and results.

2 Method

The shared task unfolded into two phases. During the training phase, spanning from January 15 to February 9, 2024, we received 466 multiple-choice questions along with additional metadata, such as the item type and exam step. Our goal in this phase was to develop models that could predict two key targets: item difficulty and response time. Transitioning to the evaluation phase, which took place from February 10 to February 16, 2024, we received an additional set of 201 multiple-choice question items, accompanied by the same supplementary metadata, excluding the two targets. Utilizing our top-performing models from the previous phase, our focus was on predicting the unknown targets of difficulty and response time. Each target allowed up to three final predictions to be submitted, and the submissions were then ranked based on the Root Mean Squared Error (RMSE). In this section, we provide more detailed information about our methodology.

2.1 Feature Extraction

As an initial step, we began by extracting features from the multiple-choice questions. Drawing from prior studies (e.g., Ha et al., 2019b), we employed various methods to transform the test items and answer choices into meaningful representations.

First, we used features that we could extract and compute directly from the data provided by the organizers. We defined a set of raw features including the **answer key** (A, B, C, D, E, F, G, H, I, or J), **item type** (Text or PIX), **exam** (Step 1, 2, or 3), the **number of answer options** (4, 5, 6, 7, 8, 9, or 10), the **ordinal position of the correct key within the sequence of answers options**, normalized between 0 and 1 (0.0, 0.11, 0.17, 0.2, 0.25, 0.29, 0.33, 0.4, 0.43, 0.5, 0.57, 0.6, 0.67, 0.75, 0.8, or 1.0).

Apart from the initial set of basic features, we generated more sophisticated features using various natural language processing tools. These tools encompass **Linguistic Inquiry and Word Count 2022** (LIWC-22; Pennebaker et al., 2022), evaluations of lexical sophistication relying on **TAALES 2.2** (Kyle and Crossley, 2015), and the extraction of text embeddings with the **Bio_ClinicalBERT** model (Alsentzer et al., 2019).

2.1.1 Linguistic Inquiry and Word Count

LIWC-22, created by Pennebaker et al. (2022), is a text analysis tool that facilitates the exploration of diverse linguistic dimensions within textual data. Its utility extends across various fields, including psychology and communication.

LIWC-22 offers variables including word count (total words in a text), words per sentence (average number of words per sentence), big words (percentage of words with seven letters or more), and dictionary words. The 2022 version employed in this study also evaluates newer summary variables such as analytical thinking (Pennebaker et al., 2014), clout, authenticity, and emotional tone. These metrics, derived from previous research, are calculated using standardized scores from extensive comparison corpora (Boyd et al., 2022).

In addition to the summary variables, LIWC provides valuable insights into linguistic dimensions by examining the relative frequencies of different word categories such as personal pronouns and negations, represented as percentages.

For this study, LIWC features were independently extracted for (1) the item stem of the multiple-choice question and (2) the aggregated answer options.

2.1.2 Lexical Sophistication

TAALES 2.2, developed by Kyle and Crossley (2015), is a tool designed for the automated analysis of lexical sophistication, calculating over 400 measures in this domain. Its indices have found applications in various fields such as educational psychology, cognitive science, and artificial intelligence. The tool addresses challenges associated with both second language (L2) and first language (L1) writing proficiency, L2 speaking proficiency, as well as spoken and written lexical proficiency.

The five areas of lexical sophistication covered by TAALES 2.2 include lexical frequency, range (indicating how widely a word or word family is used), n-gram frequency (measuring the frequency of combinations of n number of words), academic vocabulary, and psycholinguistic word properties (e.g., age of acquisition, concreteness, familiarity).

The tool takes a single text as input and produces a list of features for that text. In our study, we utilized the tool to extract the same set of features for five distinct input types: (1) for the item stem text, (2) for the item stem text combined with the correct answer, (3) for all the answer options combined, (4) for the correct answer, and (5) for the combined distractors.

2.1.3 Clinical Embeddings

In addition to the interpretable linguistic features outlined in Sections 2.1.1 and 2.1.2, we also considered the feature dimensions of clinical embeddings extracted from the publicly available pre-trained **Bio_ClinicalBERT** model (Alsentzer et al., 2019). These embeddings consist of 768-dimensional vectors for each token within an input text. We extracted identical features for four distinct input types:

1. For the item stem text.
2. For the scenario extracted from the item stem text (i.e., the clinical case description, excluding the final sentence; e.g., *A 65-year-old woman comes to the physician for a follow-up examination (...) the left renal artery appears normal.* in Figure 1).
3. For the question extracted from the item stem text (i.e., retaining only the final sentence in the item stem text; e.g., *Which of the following is the most likely diagnosis?* in Figure 1).
4. For each of the at most ten different answer options separately.

For each of these input types, we used Hugging Face's feature extractor pipeline to extract token embeddings and compute the average vector over all token embeddings in the input.

2.1.4 Features Summary

Utilizing the features outlined in Sections 2.1.1 to 2.1.3, we obtained a total of 4,479 features for each of the 466 multiple-choice questions in the training set. These features encompass:

1. 5 raw features
2. 235 LIWC-22 features (118 for the item stem text, 117 for the answers)
3. 1,166 TAALES 2.2 features (202 for the item stem text, 241 for the item stem text combined with the correct answer, 241 for all answer options combined, 241 for the correct answer, and 241 for the combined distractors)
4. 3,072 clinical features extracted from the BERT embeddings (768 for the scenario, 768 for the question, 768 for the correct answer, 768 for the aggregated distractors)

2.2 Model Development

Following the extraction of a comprehensive set of features, as outlined in the preceding section, our next step involved the development of various models. We conducted experiments with both statistical (see Section 2.2.1) and machine learning (see Section 2.2.2) models utilizing the set of extracted features (refer to Section 2.1). Additionally, we explored the fine-tuning of biomedical and clinical language models (see Section 2.2.3). Furthermore, we constructed an ensemble model (detailed in Section 2.2.4) by leveraging the strengths of these diverse models. Finally, we ran some feature importance analysis (Section 2.4).

2.2.1 Statistical Models

In statistical models, adhering to Occam's Razor principle (Ortner and Leitgeb, 2011), the goal was to find a simple yet effective model through two steps: filtering features and building models using the stepwise regression procedures (Venables and Ripley, 2002). The two steps were conducted on pre-processed data, where all features were normalized to maintain consistency in their scale. Additionally, features with missing values were excluded from the analyses. Ultimately, 3,952 features were utilized for the subsequent analyses

conducted under 10-fold cross-validation (see Section 2.3).

Specifically, within each cross-validation fold, we initially conducted feature selection based on Pearson's correlation coefficients between the target and all features, setting a minimum threshold of 0.12 (Lovakov and Agadullina, 2021). This step aimed to identify the most relevant features, considering that stepwise regression procedures for building models could become unstable with an excessively high number of features. The number of selected features ranged from 60 to 90 across folds. Next, the selected features underwent stepwise regression analysis, which involved developing a series of simple linear regression models by iteratively removing or adding features to the baseline model. These models were then compared based on the information criteria, Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). The final selected model in each fold was determined based on the lowest value of either AIC or BIC. Since AIC or BIC could recommend different models for each fold, we calculated the average RMSE across the 10 folds to compare them. Interestingly, we found that the models recommended by BIC yielded a lower RMSE compared to those recommended by AIC. Finally, with the correlation filtering and BIC setting applied, 10 simple regression models were recommended for item difficulty and response time tasks respectively. The final chosen model was the one with the lowest RMSE across all folds.

2.2.2 Machine Learning Models

The machine learning pipeline consisted of a dimensionality reduction step followed by a model fitting step. As dimensionality reduction, we used a separate principal component analysis (PCA) for each extracted feature set (i.e., LIWC, TAALES, and BERT). The number of principal components retained for each feature set equaled the number of components required to explain at least 60% of the variance in the original features. On these preprocessed features, we trained 4 different machine learning models: Lasso (regularized linear regression), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN).

The Lasso model was used with regularization $\alpha = 0.1$. RF was used with default hyperparameters. The hyperparameters for SVM and KNN were tuned using a grid search with nested 5-fold cross-validation. For SVM we considered an RBF kernel with regularization parameter $C \in \{0.1, 1, 10, 100\}$ and kernel width $\gamma \in \{1, 0.1, 0.01, 0.001\}$. For KNN we considered the number of neighbors $K \in \{1, 5, 10, 15, 20, 100\}$.

2.2.3 Language Models

In addition to traditional statistical and machine-learning models, we also experimented with fine-tuning a transformer model to predict response time and item difficulty as a multi-target prediction task. Previously, Xue et al. (2020) utilized a pre-trained model for a similar purpose, demonstrating the benefits of transfer learning in enhancing predictions. However, our methodology diverged in two important ways. On the one hand, we framed this as a multi-target regression task, contrary to treating response time and item difficulty as separate regression tasks, thus capturing their interdependencies. This approach is particularly meaningful as the relationship between the two variables is not strictly linear (Yaneva et al., 2021, p. 223).

On the other hand, we deliberately selected domain-specific pre-trained models tailored for biomedical or clinical texts, known to outperform nonspecific models (Alsentzer et al., 2019). The domain-specific pre-trained language models under consideration were trained on datasets from clinical sources such as MIMIC-III, as well as biomedical corpora like PubMed and PMC full-text articles and abstracts. These models encompassed **BERT-ClinicalQA** (exafluence, 2021), **Bio_ClinicalBERT** (Alsentzer et al., 2019), **Bio_ClinicalBERT_emrqa** (aaditya, 2022), **Clinical-BigBird** (Li et al., 2022), **Clinical-Longformer** (Li et al., 2023), and **ClinicalBERT** (Wang et al., 2023).

For model training, we initiated the pre-trained models sourced from Hugging Face (version 4.36), employing a PyTorch backend (version 2.1). The models were fine-tuned on an NVIDIA GeForce RTX 3090 (CUDA 12.2). We employed a BERTForSequenceClassification architecture equipped with two regression outputs, tailored for predicting both item difficulty and response time. We utilized the RMSE loss function to minimize the predicted item difficulty and response time over three epochs, assigning equal weight to both targets within the loss function. The optimization process employed the AdamW optimizer with a learning rate of 5^{e-5}, alongside a linear scheduler and weight decay. To accommodate the LongFormer and Big-Bird mod-

els, a batch size of one was used.

It should be noted that, given the substantial difference in scale between the two targets, we rescaled response time from seconds to minutes before training, thereby aiding smoother model convergence. Subsequently, during the inference stage, response time was transformed back from minutes to seconds for accurate interpretation.

It is also important to note that we chose to utilize a fixed initialization seed (15012024) for conducting post-hoc predictions after the winner announcement, aiming to ensure the reproducibility of the final reported predictions. However, it is important to acknowledge that the absence of a more comprehensive hyperparameter search on model initialization represents a limitation we intend to address in future work.

We conducted experiments using two different input formats: (1) solely focusing on the item stem and (2) concatenating the item stem with the list of answer options. Initial results suggested that including the answers led to slightly improved predictions across all models.

Moreover, we investigated whether integrating the classification of the exam step as an auxiliary task could improve the accuracy of predicting item difficulty and response time. To facilitate this classification, we introduced three extra output dimensions, indicating the probability of belonging to each exam step. This model, denoted as Bio_ClinicalBERT_FTMT, was initialized with Bio_ClinicalBERT (Alsentzer et al., 2019) and was optimized over ten epochs.

2.2.4 Ensemble Model

Motivated by the 'no free lunch' theorem (Wolpert and Macready, 1997), we aimed to leverage the predictive power of the diverse models introduced in the previous sections, including statistical, machine learning, and language models. The goal was to create an ensemble where individual models, each with its specific errors, could compensate for one another. Following the stacking concept, we used predictions from all individual models on training instances as features in the ensemble model.

Consistent with our approach in machine learning models, we applied dimensionality reduction to the extracted features (i.e., LIWC, TAALES, and BERT) using PCA. These reduced features were then incorporated into the ensemble model as part of its input. As for the choice of ensemble model, we experimented with Lasso, RF, Extra Trees, multi-layer perception, and gradient boosting regressor.

2.3 Model Selection

During the training phase, we ran a 10-fold cross-validation experiment on the training data, utilizing the *scikit-learn* library. The data was divided into ten folds, with these identical folds utilized for both training and evaluating each of the models outlined in Section 2.2. To maintain consistency, we utilized a fixed random seed (15012024) for shuffling the data before the splitting process. Subsequently, we calculated the average RMSE to assess and compare the performance of our various models.

2.4 Feature Importances

To better understand how the models used the input features, we performed some post-hoc interpretation techniques. One of the model-agnostic tools that we used is a permutation feature importance analysis. Such an analysis first randomly shuffles (i.e., permutes) the values for one of the features in the dataset. Then, using the models trained before on the non-shuffled data, cross-validated predictions can be regenerated with the shuffled feature and performance can be recalculated. In this way, we can see the impact on the performance of the model when one of the input features is 'randomized', and thus get a univariate feature importance metric. To counter variability, the whole procedure is repeated 5 times per feature with a different random permutation each time, and the average impact on performance is then reported.

3 Results

3.1 Phase 1: Cross-Validation

Regarding the difficulty, as reported in Figure 2, the Ensemble and RF methods managed to provide slightly superior results. However, the difference in performance was rather subtle, as most of the models reached RMSE values of approximately 0.30 in most of the cases. Slightly differently from that, Bio_ClinicalBERT_emrqa and Clinical-Longformer performed marginally worse by achieving an RMSE of 0.31, followed by the statistical model and Bio_ClinicalBERT_FTMT, which yielded roughly 0.32 and 0.33 of RMSE, respectively.

As for the response time, as can be seen in Figure 3, a more noticeable difference in performance was observed where Lasso and BERT-ClinicalQA

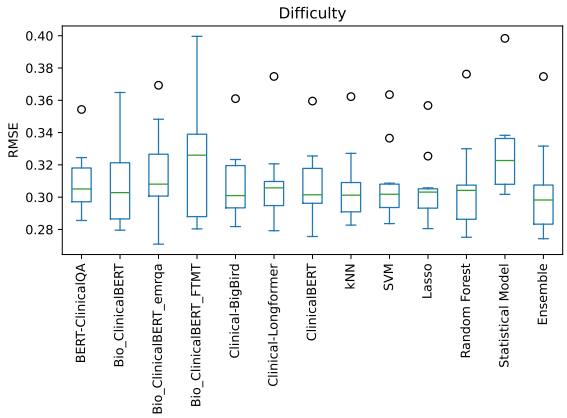

Figure 2: Performance of models in predicting difficulty on the training set, evaluated with 10-fold cross-validation.

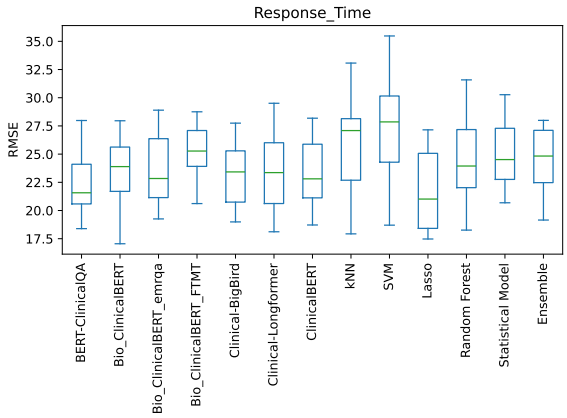

Figure 3: Performance of models in predicting response time on the training set, evaluated with 10-fold cross-validation.

had the upper hand since both of them achieved an RMSE score of approximately 21. As opposed to that, Bio_ClinicalBERT_FTM, KNN, and SVM performed relatively poorly, reaching RMSE values above 27.5. All the other compared methods provided rather overlapping results.

Submission Strategy At the end of the training phase, we devised a submission strategy. This plan entailed submitting our top two models for each track, with the final submission reserved for a model distinct from the leading two. This approach was particularly crucial for the difficulty prediction, given its inherent complexity and the difficulty in discerning superior models during the training. With all models demonstrating performance close to an average baseline (which we calculated ourselves), uncertainty arose regarding which models would outperform on the test set. Therefore, maximizing the diversity in our model selection became paramount.

For the difficulty prediction, our Ensemble and RF models achieved the lowest RMSE values. Consequently, these two models were chosen for submission during the test phase. To introduce diversity into our approach, we included BERT-ClinicalQA in the final run submission because its predictions were different from the previous runs and it was one of the top models for predicting response time. In terms of the response time prediction, our Lasso and BERT-ClinicalQA models achieved the highest scores, exhibiting the lowest RMSE values, and were thus submitted during the test phase. Additionally, to further diversify our strategy, we utilized the final run to submit a completely different model, i.e., Bio_ClinicalBERT_emrqa.

3.2 Phase 2: Leaderboard

Tables 1 and 2 present the final evaluation results for all teams based on the test datasets released by the leaderboard. For the difficulty prediction (See Table 1), the baseline model achieved an RMSE of 0.311. Our team's RF model reached 0.305, which was better than the baseline. Compared to the RMSE results of other teams, the RF model was in the 4th place out of 43 teams, and its RMSE was slightly higher than the best (0.299). Apart from our best model, our team's ensemble model also had a good performance, with an RMSE of 0.308 (slightly better than the baseline), ranking 12th out of 43 teams. For the response time prediction (See Table 2), most teams achieved better results in terms of RMSE compared to the baseline model (31.68). Our team's Lasso model performed impressively better than other models, coming the 2nd out of 34 teams with an RMSE of 24.116, significantly better than the baseline and close to the best RMSE.

Tables 3 and 4 show the performance results of all our models on the test set. It is evident from these findings that the Lasso model exhibited superior predictive capability for both difficulty and response time. Despite our knowing of the Lasso model's effectiveness during the cross-validation experiment (Figure 3) and subsequent winner announcement (Table 2), its unexpected success in predicting difficulty on the test set was surprising. Throughout the training phase, we encountered challenges in distinguishing between models for predicting difficulty, as all models performed simi-

#	Team	Run	RMSE
1	EduTec	electra	0.299
2	UPN-ICC	run1	0.303
3	EduTec	roberta	0.304
4	ITEC	RandomForest	0.305
5	BC	ENSEMBLE	0.305
12	ITEC	Ensemble	0.308
16	Baseline	DummyRegressor	0.311
43	ITEC	BERT-ClinicalQA	0.393

Table 1: Our three submissions to the leaderboard on difficulty prediction. The top 5 submissions are given as well as the shared task baseline.

#	Team	Run	RMSE
1	UNED	run2	23.927
2	ITEC	Lasso	24.116
3	UNED	run1	24.777
4	UNED	run3	25.365
5	EduTec	roberta	25.64
25	Baseline	DummyRegressor	31.68
32	ITEC	BERT-ClinicalQA	53.844
33	ITEC	Bio_ClinicalBERT_emrqa	54.719

Table 2: Our three submissions to the leaderboard on the response time prediction. The top 5 submissions are given as well as the shared task baseline.

larly close to baseline levels. This initial difficulty hindered our recognition of the Lasso model as the optimal choice, despite its strong performance in predicting response time. Had we submitted the Lasso model to the difficulty leaderboard, we would have outperformed the second-best model, securing the second position on difficulty as well. As the Lasso model demonstrated superior performance in predicting both difficulty and response time, we will delve deeper into examining the feature importance of this model in the subsequent section.

3.3 Feature Importances

The best models based on cross-validated training RMSE turned out to be the RF for item difficulty and the Lasso for response time. Therefore, we conducted a permutation feature importance analysis for these two models. For the item difficulty, the top features for the RF were the word count from LIWC (with an average increase in RMSE of 0.034

#	Model	RMSE
FEATURE-BASED MODELS		
4	Lasso $*$	0.301
7	Random Forest $\bullet$	0.305
9	kNN	0.307
11	SVM	0.310
12	Statistical Model	0.343
FINE-TUNED LANGUAGE MODELS		
1	Clinical-Longformer $\circ*$	0.294
2	ClinicalBERT $\circ*$	0.299
3	Bio_ClinicalBERT $\circ*$	0.300
5	Bio_ClinicalBERT_emrqa $\circ*$	0.302
6	Clinical-BigBird $\circ*$	0.303
8	BERT-ClinicalQA $\bullet\circ$	0.306
13	Bio_ClinicalBERT_FTMT	0.350
ENSEMBLE MODEL		
10	Ensemble $\bullet$	0.308

Table 3: Performance and ranking of our models in predicting difficulty on the test set. Models denoted by $\bullet$ were submitted to the leaderboard. Models marked with $\circ$ are reported with post-hoc predictions. Models labeled with $*$ surpassed our best leaderboard model.

when this feature is randomly shuffled), one of the BERT answer embeddings (0.022), one of the BERT distractor embeddings (0.017), and the analytical thinking measure from LIWC (0.016). All other features lead to an RMSE increase of at most 0.010. For the response time, the top features of the Lasso model were the word count from LIWC (with an average increase in RMSE of 6.8), the exam step (1.0), the item type (0.69), the number of answers (0.60), the analytical thinking measure from LIWC (0.30), and the position of the correct answer (0.20). All other features lead to an RMSE increase of at most 0.03.

Both these models also have built-in feature importance metrics: the RF through the heuristic values observed during training and the Lasso model through the magnitude of its coefficients. These metrics revealed that most of the total feature importance weight for the RF and Lasso models was given to the principal components (PCs) coming from the BERT embeddings (78% and 76% respectively). However, these PCs also represent 43 out of the 58 features (74%) remaining after PCA. For the RF model, each feature had a similar importance of on average 2.0% $\pm$ 1.0% (mean $\pm$ standard deviation). On the other hand, for the Lasso model, there

#	Model	RMSE
FEATURE-BASED MODELS		
1	Lasso •	24.116
8	Random Forest	26.527
11	Statistical Model	27.020
12	kNN	28.919
13	SVM	31.101
FINE-TUNED LANGUAGE MODELS		
2	Clinical-Longformer ○	24.829
4	ClinicalBERT ○	25.643
5	BERT-ClinicalQA •○	26.014
6	Bio_ClinicalBERT ○	26.310
7	Bio_ClinicalBERT_FTMT	26.504
9	Clinical-BigBird ○	26.555
10	Bio_ClinicalBERT_emrqa •○	26.771
ENSEMBLE MODEL		
3	Ensemble ○	25.298

Table 4: Performance and ranking of our models in predicting response time on the test set. Models denoted by • were submitted to the leaderboard. Models marked with ○ are reported with post-hoc predictions.

was a clear ranking of feature sets, with the raw features first (4.7% $\pm$ 4.3%) followed by the BERT PCs (1.8% $\pm$ 1.4%) and the LIWC PCs (0.10% $\pm$ 0.069%). Interestingly, while the LIWC PCs seem to have a low importance to the Lasso model based on their coefficients, they had a big impact on predictive performance based on the permutation feature importance test.

4 Discussion

As previous research by the shared task organizers has shown (Ha et al., 2019a; Xue et al., 2020; Yaneva et al., 2020, 2021), predicting response time and difficulty of multiple-choice questions for medical licensing exams is a challenging task. In this study, our team tried to solve this challenge by adopting a multidisciplinary perspective, combining insights from statistical modeling, machine learning, and natural language processing.

While previous studies have primarily concentrated on examining the influence of exam and item metadata, along with certain linguistic complexity features (e.g., Ha et al., 2019a; Yaneva et al., 2021), we explored the integration of several novel, unexplored features. While our results validate the importance of specific raw metadata features (such as the number of answer options), they also highlight the significance of features derived from LIWC and TAALES, as well as embeddings from biomedical language models. Notably, the LIWC feature indicating the degree of "analytical thinking" for answers emerged as particularly noteworthy for predicting response time.

Regarding the models, it is noteworthy that the more sophisticated ones did not surpass the less intricate models. Simple models proved more accurate in predicting the response time of multiple-choice questions. This resonates with Occam's Razor principle, which favors simpler models as long as their performance matches or exceeds that of more complex alternatives (e.g., Ortner and Leitgeb, 2011). In our study, models utilizing Lasso or RF with principal component features outperformed the fine-tuned language model with embeddings. This suggests that, for this specific task, traditional machine learning methods incorporating dimensionality reduction were more effective and robust compared to complex statistical models.

5 Conclusion

Our team's contribution to the shared task of predicting the difficulty and response time of medical exam questions demonstrates that simpler models like Lasso (l_1-regularized) or RF regression, which utilize principal component features derived from linguistic features and clinical embeddings, outperform more complex, fine-tuned NLP models. In the winner announcement, the RF model secured the 4th position out of 43 submissions for difficulty, while the Lasso model attained the 2nd position out of 34 for response time. Post-hoc analyses revealed that if we had submitted the predictions of the Lasso model of difficulty to the leaderboard, we would have surpassed the second position in predicting difficulty as well.

Moreover, predicting the response time for medical multiple-choice questions has proven to be a more straightforward task compared to predicting the difficulty of such questions. Response time primarily hinges on item length (i.e., word count and number of answers), item type, exam step, and the level of analytical thinking required for the answers, as illustrated by permutation feature importance analyses. Conversely, predicting item difficulty poses greater challenges, with all models approaching an average baseline performance. Nevertheless, post-hoc analyses suggest that more extensive experimentation with fine-tuned language models

could potentially aid in discerning the difficulty of multiple-choice questions. While response time can be more accurately predicted from linguistic features like word count, predicting difficulty may require more intricate modeling of deep clinical text representations.

6 Limitations

In the future study, we could deepen our understanding of our findings, potentially shedding light on the circumstances in which simpler models might be advantageous.

One initial limitation we would have liked to tackle is the utilization of student responses instead of percentage- and mean-aggregated targets. This limitation stems from the fact that we only received aggregated or summarized data for difficulty and response time per item, rather than the individual-level data. Access to the individual-level data would have allowed us to explore more advanced psychometric models that consider interactions between items and students.

Another limitation we aim to address is conducting a more comprehensive study on fine-tuning language models. Specifically, we plan to delve into a more exhaustive grid search, which could potentially illuminate the most optimal model initialization and hyperparameters.

Finally, another constraint of our study is the possibility of overlooked features in the data. This limitation arises from our focus on a predetermined set of features, including LIWC, TAALES, and the BERT clinical model, for feature selection. In future research, additional methods for feature extraction could be explored.

References

aaditya. 2022. Bio_clinicalbert_emrqa. https://huggingface.co/aaditya/Bio_ClinicalBERT_emrqa. Accessed: 03/14/2024.

H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ryan L. Boyd, Ashwini Ashokkumar, Shiva Seraj, and James W. Pennebaker. 2022. The development and psychometric properties of liwc-22.

exafluence. 2021. Bert-clinicalqa. https://huggingface.co/exafluence/BERT-ClinicalQA. Accessed: 03/14/2024.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019a. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019b. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

Kristopher Kyle and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Andrey Lovakov and Elena R. Agadullina. 2021. Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, 51:485–504.

Ronald Ortner and Hannes Leitgeb. 2011. Mechanizing induction. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 719–772. North-Holland.

James W. Pennebaker, Ryan L. Boyd, Roger J. Booth, Ashwini Ashokkumar, and Martha E. Francis. 2022. Linguistic inquiry and word count: Liwc-22. *Pennebaker Conglomerates*.

James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9(12):e115844.

Gideon Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Springer New York.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642.

David H Wolpert and William G Macready. 1997. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197, Seattle, WA, USA → Online. Association for Computational Linguistics.

Victoria Yaneva, Le An Ha, Peter Baldwin, and Janet Mee. 2020. Predicting Item Survival for Multiple Choice Questions in a High-Stakes Medical Exam. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6812–6818, Marseille, France. European Language Resources Association.

Victoria Yaneva, Daniel Jurich, Le An Ha, and Peter Baldwin. 2021. Using Linguistic Features to Predict the Response Process Complexity Associated with Answering Clinical MCQs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–232, Online. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, An Ha Le, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.