Predicting Item Difficulty and Item Response Time with Scalar-mixed Transformer Encoder Models and Rational Network Regression Heads

Sebastian Gombert[1], **Lukas Menzel**[2], **Daniele Di Mitri**[1], **and Hendrik Drachsler**[1,2,3,4]
[1]DIPF: Leibniz Institute for Research and Information in Education, Frankfurt, Germany
[2]studiumdigitale & [3]Computer Science Department, Goethe University Frankfurt, Germany
[4]Department of Online Learning and Instruction, Open University, Heerlen, Netherlands
{s.gombert,d.dimitri,h.drachsler}@dipf.de
menzel@sd.uni-frankfurt.de

Abstract

This paper describes a contribution to the *BEA 2024 Shared Task on Automated Prediction of Item Difficulty and Response Time*. The participants in this shared task are to develop models for predicting the difficulty and response time of multiple-choice items in the medical field. These items were taken from the United States Medical Licensing Examination® (USMLE®), a high-stakes medical exam. For this purpose, we evaluated multiple BERT-like pre-trained transformer encoder models, which we combined with Scalar Mixing and two custom 2-layer classification heads using learnable Rational Activations as an activation function, each for predicting one of the two variables of interest in a multi-task setup. Our best models placed first out of 43 for predicting item difficulty and fifth out of 34 for predicting Item Response Time.

1 Introduction

According to Madaus and Airasian (1970), assessments are arguably among the core components of education. They help diagnose and monitor learners' skill levels and, thus, function as a basis for downstream educational decisions. Depending on their concrete function, they can be further categorized. Placement assessments are needed to recommend courses for learners at an appropriate level. Formative assessments are required to monitor learning progress. Summative assessments are needed to measure learners' outcomes.

Each assessment comprises multiple items, i.e., individual tasks test-takers must complete. For standardized assessments, items must be evaluated to guarantee fair and comparable outcomes. In this context, multiple factors must be assessed as listed in the *Standards for educational and psychological testing* (Association et al., 1985).

Among these factors are *Item Difficulty*, i.e., a numerical variable describing the overall difficulty of solving a given item, and *Item Response Time*, which encodes the overall time needed to solve an item measured in seconds. Traditionally, *Item Difficulty* has been assessed using methods such as *Rasch Analysis* (Rasch, 1960) or *Item Response Theory* (An and Yung, 2014). Both of them rely on collection data from pre-evaluations with cohorts of test takers. As administering respective pre-evaluation steps is still a labour-intensive and costly process (Settles et al., 2020), there has been ongoing research on automating these procedures using machine learning methods with a higher potential for generalization.

One of these instances is the *First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions* (Yaneva et al., 2024). In this paper, we describe a submission to this shared task, which placed first for predicting *Item Difficulty* and fifth for predicting *Item Response Time*.

2 Related Work

Both the prediction of *Item Difficulty* and *Item Response Time* for multiple choice questions utilizing natural language processing are comparably novel tasks. Earlier research on predicting *Item Difficulty* tackled mostly other item formats such as C-tests, a form of fill-in-the-blank test aimed at testing language proficiency, (Beinborn et al., 2015) or constructed response items (Padó, 2017). In the context of language learning, Settles et al. (2020) developed a method to assess the difficulty of various types of items for language learning in terms of the *CEFR* framework.

Early research on predicting *Item Difficulty* for multiple choice questions was conducted by Ha et al. (2019), who fit various feature-based models using heterogeneous sets of features incorporating embeddings, as well as lexical, syntactic, semantic, cohesion-based, and psycholinguistic features to predict *Item Difficulty* for a large-scale dataset comprised of *United States Medical Licensing Ex-*

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 483–492

amination® (USMLE®) items. The authors also use features derived from information retrieval systems. They reason that retrieving an answer for a given question through Information Retrieval might predict the difficulty of cognitively retrieving an answer. Subsequent work by Yaneva et al. (2020) and Yaneva et al. (2021) used similar approaches to predict *Item Survival* and *Item Response Complexity*.

For predicting *Item Response Time*, Baldwin et al. (2021) used feature-based models using primarily the same features and algorithms which Ha et al. (2019) applied for predicting *Item Difficulty*. They found that embeddings and linguistic features were robust in predicting *Item Response Time*, with IR-based features being less predictive while still holding some degree of predictive power. Yaneva et al. (2023) combined linguistic features with static embeddings produced by word2vec and contextual word embeddings produced by non-fine-tuned BERT models to predict a range of item characteristics, including *Item Response Time*.

What becomes apparent when reviewing the past literature on the topic is that transformer-encoder language models such as BERT (Devlin et al., 2019) have not been fine-tuned for the prediction of *Item Difficulty* and *Item Response Time* as of now. This can be regarded as a clear research gap, given that transformer encoders could push the state of the art for a wide range of tasks in natural language processing and outperformed more traditional feature-based approaches for these (Rogers et al., 2020).

3 Method

To close this gap, we aim to evaluate the overall predictiveness of pre-trained transformer encoder language models for *Item Difficulty* and *Item Response Time* in our submission for the *First Shared Task on Automated Prediction of Difficulty and Response Time for Multiple Choice Questions* (Yaneva et al., 2024).

3.1 Dataset

The dataset used for this task was provided by Yaneva et al. (2024) and consists of multiple choice items that were previously used for the *United States Medical Licensing Examination® (USMLE®)*. It is divided into a training and a test set, with the training set comprising 466 and the test set comprising 201 items. Each item consists of a prompt with up to 10 different response options, of which a single one is correct. Moreover, for each item, it is remarked whether the response options come in the form of texts or images (the images are not provided with the dataset; instead, there are descriptions of what is depicted) and if the items belong to the first, second or third step of the *USMLE®*.

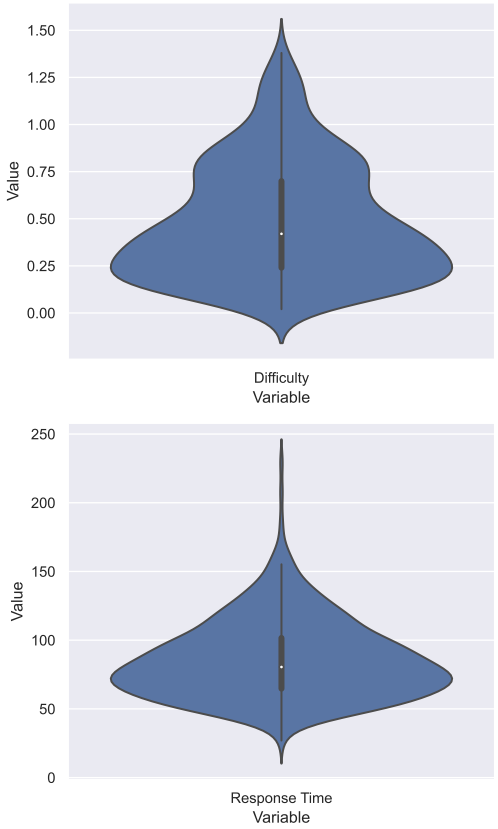

Figure 1: Violinplots depicting the general distributional properties of *Item Difficulty* and *Item Response Time*.

Each item is given a single rating for *Item Difficulty* and one for *Item Response Time*. Figure 1 shows the distribution of values for both of them. Going by *Shapiro-Wilk*, neither *Item Difficulty* ($W = 0.93$, $p < 0.000$) nor *Item Response Time* ($W = 0.94$, $p < 0.000$) follow a normal distribution. However, as Figure 2 reveals, both are significantly correlated, which is also confirmed by *Pearson's* ($r = 0.49$, $p < 0.000$) and *Spearman's* ($r = 0.52$, $p < 0.000$) correlation coefficients.

As the difficulty of an item very likely influences the time needed to think about the correct answer, it can be speculated that there is, to a certain degree, a causal relationship between both variables. However, given that the r values are not higher, it can also be concluded that this is not the only factor

influencing the exact outcome of both variables for each item.

3.2 System Description

The architecture we implemented for this shared task is derived from the modified transformer-based model implemented by Gombert et al. (2022) for automated short answer scoring, where it outperformed regular transformer-based models for this task. Our architecture can be flexibly applied to various regression and classification tasks. It is a deep neural network architecture based upon regular *BERT*-like transformer-encoder language models (Devlin et al., 2019). The typical BERT regression setup uses a single output neuron. This neuron is fed with the last layer's classification token output. Our setup, however, is modified.

The first difference to the standard BERT implementation is the usage of scalar mixing. Scalar mixing calculates a weighted mean of all hidden layers of a transformer language model. The weights from which this mean is calculated are fit during training. This technique was mainly applied to investigate the influence of different pre-trained layers on a given prediction (Tenney et al., 2019; Kuznetsov and Gurevych, 2020). Still, it can also be used as a regular neural network building block.

Different layers of BERT-like models learn representations for different linguistic phenomena (Tenney et al., 2019). Using scalar mixing lets us exploit all these representations, instead of only the output of the last layer, while simultaneously learning their importance for the final output. Scalar mixing can be depicted using the following equation with tensors $t_1, ..., t_n$ being the hidden layer outputs, and γ and $w_1, ..., w_n$ being the learnable parameters:

$$S(t_1, ..., t_n) = \gamma \sum_{j=0}^{n} softmax(w_j) t_j \quad (1)$$

The second adjustment to the classification heads is to use a two-layer setup. The output of the intermediate layer runs through a *Rational Activation* (Molina et al., 2020), a form of learnable activation function whose shape is optimized during training; thus, a "Rational Network". This activation function outperformed non-learnable activation functions for multiple architectures and benchmarks. Rational Activations are based upon Padé approximants (Brezinski et al., 1995), which can generally be optimized to approximate various functions, including typical activation functions. Given a hypothetical optimal activation function $f(x)$ for a problem at hand, one can approximate this function by learning a Padé approximant $F(x)$ of the pre-defined orders n and m using the following equation where coefficients a_j and a_k are learned during training:

$$F(x) = \frac{\sum_{j=0}^{m} a_j x^j}{1 + |(\sum_{k=1}^{m}) a_k x^k|} \quad (2)$$

Another important aspect of our model is the use of multi-task learning. As Peng et al. (2020) put it, "[m]ulti-task learning (MTL) is a field of machine learning where multiple tasks are learned in parallel while using a shared representation", with "representation" referring to the internal embeddings put out by the different model layers. Although the shared task rules prevented using one of the two variables to predict the other directly, they did not prevent implementing a system simultaneously predicting both. As shown in section 3, *Item Difficulty* and *Item Response Time* are significantly correlated in the training set. While this does not necessarily prove a causal relationship, it implies that the internal representations used to predict one of the two variables can likely benefit the prediction of the other. Therefore, using shared representations will likely lead to improved predictions for both variables.

Multi-task learning is usually conducted by attaching multiple prediction heads to the base model for transformer-encoder models. Our setup involves the usage of a complete distinct regression head per variable, each with separate units for *Scalar Mixing* and *Rational Activations*, and distinct linear layers. We reason, while the transformer encoder learns shared representations during fine-tuning, both variables might require a stronger or weaker emphasis on different model layers during *Scalar Mixing*. Moreover, an optimal learned activation function $F(x)$ might look different for both.

Given an item k, the model receives the following corresponding input $I(k)$, with $\oplus$ referring to the separation token of a given model, $s_k \in \{1, 2, 3\}$ to the exam step, $t_k \in \{TEXT, PIX\}$ to the item type, p_k to the item prompt, $r_{k1}, ..., r_{kn}$ to the possible answers, and $c_k \in \{r_{k1}, ..., r_{kn}\}$ to the correct answer:

$$I(k) = s_k \oplus t_k \oplus p_k \oplus r_{k1} \oplus ... \oplus r_{kn} \oplus c_k \quad (3)$$

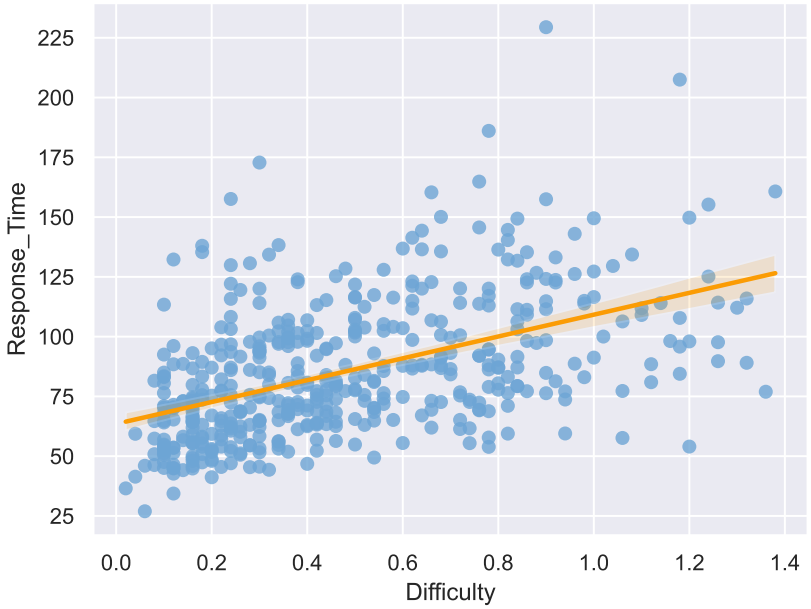

Figure 2: This scatterplot depicts the relationship between *Item Difficulty* and *Item Response Time*.

During training, the mean loss for both variables is calculated to acquire the gradients for backpropagation. Since *Item Response Time* and *Item Difficulty* are on different scales, a naïve approach would strongly bias the model towards *Item Response Time*. For this reason, we divide *Item Response Time* by 100 to get similar scales for both variables. Consequently, the model outputs for *Item Response Time* must be multiplied by 100 again to acquire the actual item response time. With a model $M(x)$ receiving an input as defined by $I(k)$, v_k being the *Item Difficulty*, and w_k being the *Item Response Time* of k, the following equation illustrates this:

$$M(I(k)) = (v_k, \frac{w_k}{100}) \quad (4)$$

Figure 3 illustrates the overall system setup.

3.3 Evaluation

3.3.1 Pre-Evaluation (Model Selection)

In a pre-evaluation step, we aimed to select the most appropriate transformer language model to use as the basis for our shared task submission. Therefore, we evaluated the architecture described in the System Description section with different pre-trained transformer-encoder language models. All models were implemented using the *Huggingface Transformers* framework (Wolf et al., 2020). However, we implemented our own training and evaluation procedures. These are the following models:

- BERT-large[1]: this model is the original BERT model as described in Devlin et al. (2019).
- RoBERTa-large[2]: This model is an established BERT variant that was pre-trained on a larger data set without the usage of next sentence prediction and outperforms regular BERT on established benchmarks such as *SuperGLUE* (Wang et al., 2019).
- ELCTRA-large[3]: this model was published by Clark et al. (2020). Unlike BERT and RoBERTa, it is pre-trained in an adversarial setup using two models that implement a variation of masked language modelling. One model, the generator, predicts masked tokens. The other model, the discriminator, then must classify random input tokens concerning whether they were generated or ground truth.
- DeBERTa-v3-large[4]: this model was published by He et al. (2023). It uses *disentangled attention* to separately encode the content and

[1]https://huggingface.co/google-bert/bert-large-uncased
[2]https://huggingface.co/FacebookAI/roberta-large
[3]https://huggingface.co/google/electra-large-discriminator
[4]https://huggingface.co/microsoft/deberta-v3-large

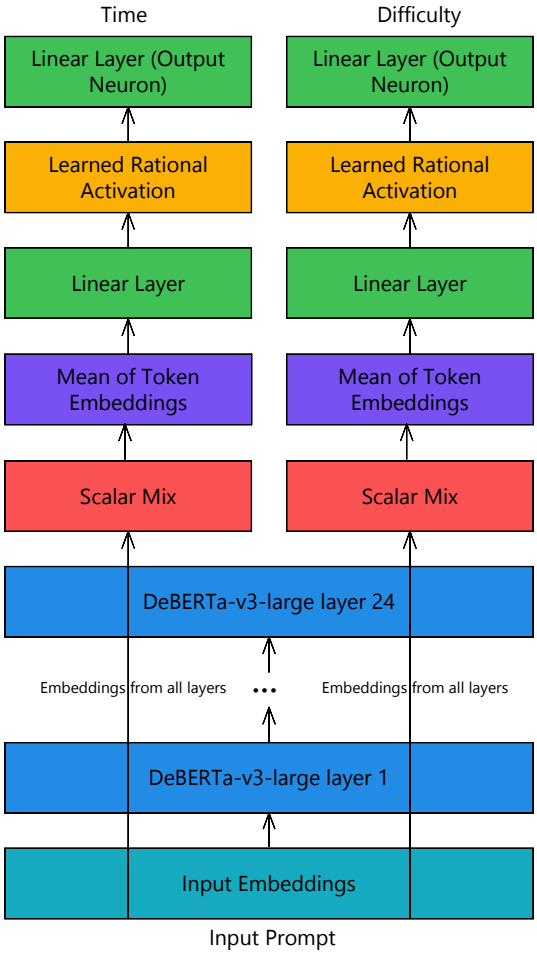

Figure 3: This diagram depicts the general architecture of our models. A given input is encoded into static embeddings. These are then propagated through all layers of a given pre-trained transformer encoder language model. The static embeddings and the outputs of all layers are propagated into the respective scalar mixing units, where a weighted mean is calculated from the individual tensors per variable. These are then propagated into the individual regression heads.

position of a token within an input text. Moreover, it is pre-trained using a specialized adversarial setup similar to ELECTRA. We chose this model since it is the best-performing open BERT-like model on the *SuperGLUE* (Wang et al., 2019) leaderboard[5].

- BiomedBERT-large[6]: This model is a BERT variant which was published by Tinn et al. (2023). It is trained identically to BERT but uses biomedical data exclusively (abstracts crawled from PubMed). We evaluated this model for the shared task since its dataset also stems from the biomedical domain.

- BiomedELECTRA-large[7]: This model is an ELECTRA variant which was published by Tinn et al. (2023). It is trained identically to ELECTRA but uses biomedical data exclusively (abstracts crawled from PubMed). We evaluated this model for the shared task since its dataset also stems from the biomedical domain.

We also added two simpler baseline models. We used *Linear Regression* and *Random Forests* as algorithms, which both are given the following features:

- *Tf-ifd*-encoded character trigrams for the item prompt and each answer option, motivated by the fact that character *n*-gram frequencies can provide valuable signals in terms of predicting readability (Imperial and Kochmar, 2023), which should be correlated with *Item Difficulty* and *Item Response Time*, given the results from Ha et al. (2019) and Baldwin et al. (2021).

- The overall number of tokens of the item prompt, motivated by the general observation of text length being correlated to text complexity as reported by DuBay (2007).

Additionally, we added dummy regressors that consistently predict the respective mean.

The evaluation was conducted solely on the training set using 5x5 cross-validation implemented via the *RepeatedKFold* class from *Scikit-learn* (Pedregosa et al., 2011). We trained for four epochs and reported the best results achieved during one of these epochs. All runs used the same random seed, namely *1*, to keep the results perfectly comparable. For each model, we measured *RMSE* (the primary evaluation metric of the shared task), *MAE* and r. To this, we added r_s to measure to which degree the models can correctly rank the items by the predicted variables without explicitly considering the exact predictions. Table 1 shows the respective results, ranked by *RMSE*.

It is visible that the correct prediction of the *Item Difficulty* is nearly impossible using our proposed method with the given data. None of the models

[5]https://super.gluebenchmark.com/leaderboard

[6]https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-large-uncased-abstract

[7]https://huggingface.co/microsoft/BiomedNLP-BiomedELECTRA-large-uncased-abstract

Item Difficulty				
Model	RMSE ↓	MAE	r	r_s
ELECTRA	**0.31**	**0.25**	**0.19**	**0.16**
RoBERTa	**0.31**	**0.25**	0.17	**0.16**
DeBERTa-v3	**0.31**	0.26	0.17	0.15
Dummy (Mean)	**0.31**	*0.26*	-	-
Random Forests	**0.31**	*0.26*	*0.09*	*0.07*
BERT	0.32	0.27	0.16	0.14
BiomedBERT	0.32	0.26	0.11	0.11
Linear Regression	*0.32*	*0.26*	*0.11*	*0.07*
BiomedELECTRA	0.33	0.27	0.12	0.10
Item Response Time				
Model	RMSE ↓	MAE	r	r_s
DeBERTa-v3	**23.05**	**17.48**	**0.63**	**0.65**
BERT	23.52	17.76	0.60	0.64
RoBERTa	23.76	17.79	0.61	0.64
BiomedELECTRA	23.88	17.87	0.61	0.63
BiomedBERT	23.97	18.02	0.59	0.62
ELECTRA	24.68	18.57	0.60	0.64
Dummy (Mean)	*46.87*	*37.77*	-	-
Random Forests	*47.13*	*38.56*	*0.19*	*0.22*
Linear Regression	*47.60*	*38.87*	*0.17*	*0.17*

Table 1: The results of our pre-evaluation experiments to determine the strongest models ranked by RMSE. All results were calculated during 5x5 cross-validation runs.

Item Difficulty					
Model	RMSE ↓	MAE	r	r_s	Rank
ELECTRA	**0.29**	**0.24**	**0.27**	**0.25**	**1/43**
RoBERTa	0.30	0.24	0.24	0.20	3/43
Dummy	*0.31*	-	-	-	16/43
DeBERTa-v3	0.31	0.25	0.21	0.19	17/43
Item Response Time					
Model	RMSE ↓	MAE	r	r_s	Rank
UNED run2	**23.92**	-	-	-	**1/34**
RoBERTa	25.64	17.94	0.60	0.67	5/34
ELECTRA	25.87	19.14	0.57	0.65	6/34
DeBERTa-v3	27.30	21.48	0.56	0.63	14/34
Dummy	*31.68*	-	-	-	25/34

Table 2: The final shared task evaluation results. For *Item Difficulty*, we report the results of our models and the baseline dummy model of the shared task organizers. For *Item Response Time*, we also report the results of the overall winning system from a competing team called *UNED run2*.

we tested achieved a better *RMSE* score than the dummy regressor, meaning the models hold almost no predictive power. The model based on *BioMED-BERT-large* and the *Linear Regression* baseline are outperformed by this dummy regressor in terms of *RMSE*. Nonetheless, the r and r_s results show that all transformer-based models are at least more successful in modelling the *Item Difficulty* than the baselines. However, this success is still minimal.

Our pre-evaluations yielded better results for *Item Response Time*. Here, all transformer-based models significantly outperformed the baseline models. This means it is possible – to a certain degree – to model *Item Response Time* with our proposed method and the given data. While models based on *BioMED-BERT-large* and *DeBERTa-v3-large* achieve a similar *RMSE*, the model based on DeBERTa-v3-large outperforms all other models in terms of r and r_s, meaning it is the overall best model.

3.3.2 Shared Task Evaluation

The shared task organizers allowed the submission of up to three predictions per variable. We submitted results predicted with models based upon *ELECTRA*, *RoBERTa* and *DeBERTa-v3*. *BERT*, *Biomed-BERT* and *Biomed-ELECTRA* were not used since they performed worse for the prediction of the *Item Difficulty* while achieving very similar results to the other models for the *Item Response Time*. For this purpose, all three models were re-trained on the whole training set for four epochs. While the models based upon *ELECTRA* and *RoBERTa* achieved very high placements on the shared task leaderboard for both variables, the model based on *DeBERTa-v3* performed worse, which is a surprising outcome.

The overall trends observed during our pre-evaluation steps continued into the final shared task evaluations. While for the *Item Difficulty*, barely any system could show a performance superior to a dummy regressor baseline, the *Item Response Time* was easier to predict. Interestingly, the model based on *DeBERTa-v3* ranks the worst out of our models for both variables despite being the best-performing approach for predicting the *Item Response Time* during the pre-evaluations. However, except for this, the results line up.

Going by r and r_s, it is visible that predictions and ground truth values are positively correlated for both variables. However, a trend that is observable for all models and both variables is revealed in Figure 4. On average, the predicted values are lower than the ground truth. This pattern is more drastic for the *Item Difficulty* but also visible for the *Item Response Time*.

4 Discussion

The research at hand has multiple implications. First, we proved that using established pre-trained transformer-encoder language models for predicting the *Item Difficulty* and the *Item Response Time* can be a viable choice overall. Moreover, we could also show that our adjustments to the typical BERT

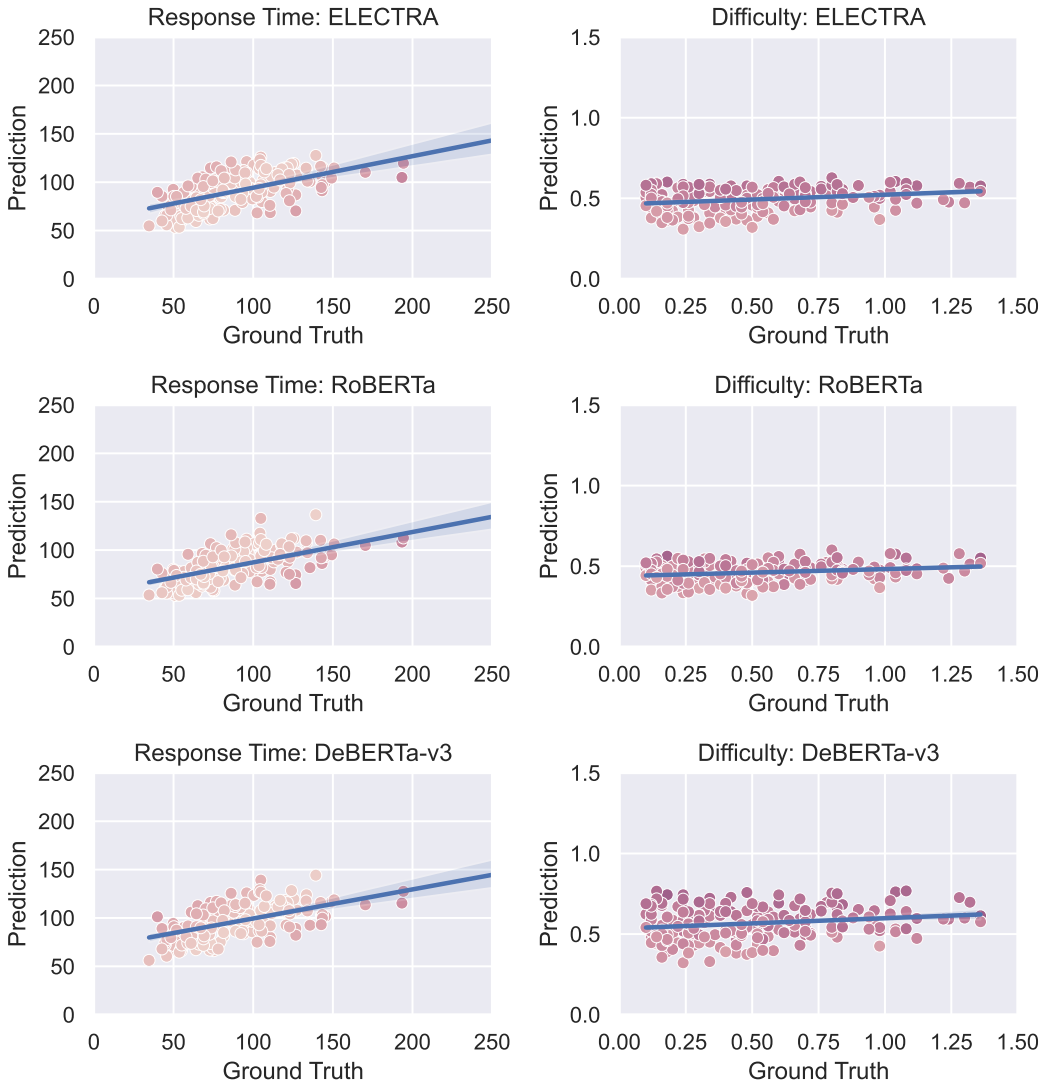

Figure 4: These regression plots illustrate the relationship between ground truth values and model predictions. The x-axis refers to the ground truth value of a given data point, while the y-axis refers to the respective predicted value. The individual data points' colour coding indicates the differences between ground truth and prediction, with a darker colour indicating a larger difference.

architecture proved fruitful. These adjustments let our models achieve very competitive performance in the shared task, with our best model even winning one of the two tracks (*Item Difficulty*).

In theory, our approach can be easily integrated with the past feature-based models published by Ha et al. (2019), Yaneva et al. (2020), Yaneva et al. (2021) and Baldwin et al. (2021). For this purpose, one needs to fine-tune a respective model. One can then use the output of all intermediate layers as embeddings. Using an algorithm such as Random Forests or Gradient Boosting, selecting appropriate features from these internal representations should be possible. Works as those by Minixhofer et al. (2021), Gombert and Bartsch (2021), Rotaru (2021), Smolenska et al. (2021) or Gombert (2021) show that the integration of task-specific transformer-based contextualized embeddings with more traditional feature-based algorithms can yield fruitful outcomes. Considering systems such as the ones published by Ha et al. (2019), one could easily replace the generalized embeddings they use with task-specific ones. Future work could thus involve testing whether such embeddings can add to a more traditional feature set to improve the overall predictive power of a given model.

It is also visible that the prediction of the *Item Difficulty* remains a challenging task since even the best participating models barely outperformed a dummy baseline model. On average, the models

underestimate the difficulty of input items. In the case of this shared task, this effect might result from the data set being comparably small and from a highly specialized domain, namely the biomedical one with its comparably complicated and specialized language.

However, since all past work on predicting the difficulty and required response time of multiple choice questions using machine learning models was aimed at assessments from this domain, it is hard to make generalized judgements on the overall difficulty of this problem. What is required here is the publication of additional datasets from different domains and the evaluation of models using these. In this context, cross-domain evaluations especially would be of high use.

Predicting the *Item Response Time* was a more fruitful endeavour, with models outperforming the dummy baseline by a larger margin. However, with an RMSE rate of 23.92 for the best-performing model, one still needs to consider that the predicted *Item Response Time* is far from accurate. The same issue for predicting the *Item Difficulty* holds true for the *Item Response Time*: the dataset at hand is from a highly specialized domain, and data from other domains is not generally available.

5 Conclusion

This paper explains our submissions for the BEA 2024 shared task on predicting the *Item Difficulty* and the *Item Response Time*, of which the best placed first for predicting the *Item Difficulty* and fifth for predicting the *Item Response Time*. Our architecture combines pre-trained transformer encoder models with multi-task learning and custom regression heads, expanding upon an architecture published by Gombert et al. (2022) by combining them with *Scalar Mixing* and *Rational Activations*.

The results suggest predicting *Item Response Time* and especially *Item Difficulty* are comparably difficult tasks. However, the dataset used for this paper stems from the biomedical domain. This domain uses a very specialized language. For this reason, the tasks need to be evaluated with data from more domains to make a general claim. This could be the objective of future work.

6 Limitations

The limitations of our systems have already been discussed in the Discussion section. First, the dataset used is from a narrow domain. For this reason, results might not translate to datasets from other domains. So far, datasets from domains other than the medical one are unavailable. This is a clear research gap that must be addressed in future work. Second, even though our systems won one of the two shared tracks and generally achieved high ranks, the results suggest that the problems of predicting *Item Difficulty* and *Item Response Time* are far from solved.

References

Xinming An and Yiu-Fai Yung. 2014. Item response theory: What it is and how you can use the irt procedure to apply it. *SAS Institute Inc. SAS364-2014*, 10(4):1–14.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, et al. 1985. Standards for educational and psychological testing. *APA*.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Claude Brezinski, Ufr Ieea, and Jim Van Iseghem. 1995. A taste of padé approximation. *Acta Numerica*, 4:53–103.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William H DuBay. 2007. *Smart language*. Booksurge Publishing.

Sebastian Gombert. 2021. Twin BERT contextualized sentence embedding space learning and gradient-boosted decision tree ensembles for scene segmentation in german literature. In *Proceedings of the*

Shared Task on Scene Segmentation co-located with the 17th Conference on Natural Language Processing (KONVENS 2021), Düsseldorf, Germany, September 6th, 2021, volume 3001 of *CEUR Workshop Proceedings*, pages 42–48. CEUR-WS.org.

Sebastian Gombert and Sabine Bartsch. 2021. TUDA-CCL at SemEval-2021 task 1: Using gradient-boosted regression tree ensembles trained on a heterogeneous feature set for predicting lexical complexity. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 130–137, Online. Association for Computational Linguistics.

Sebastian Gombert, Daniele Di Mitri, Onur Karademir, Marcus Kubsch, Hannah Kolbe, Simon Tautz, Adrian Grimm, Isabell Bohm, Knut Neumann, and Hendrik Drachsler. 2022. Coding energy knowledge in constructed responses with explainable nlp models. *Journal of Computer Assisted Learning*, 39(3):767–786.

Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

George F Madaus and Peter W Airasian. 1970. Placement, formative, diagnostic, and summative evaluation of classroom learning. In *Proceedings of the AERA Annual Meeting, 1970*. ERIC.

Benjamin Minixhofer, Milan Gritta, and Ignacio Iacobacci. 2021. Enhancing transformers with gradient boosted decision trees for NLI fine-tuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 303–313, Online. Association for Computational Linguistics.

Alejandro Molina, Patrick Schramowski, and Kristian Kersting. 2020. Padé activation units: End-to-end learning of flexible activation functions in deep networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ulrike Padó. 2017. Question difficulty – how to estimate without norming, how to use for automated grading. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online. Association for Computational Linguistics.

Georg Rasch. 1960. *Probabilistic models for some intelligence and attainment tests*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Armand Rotaru. 2021. ANDI at SemEval-2021 task 1: Predicting complexity in context using distributional models, behavioural norms, and lexical resources. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 655–660, Online. Association for Computational Linguistics.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Greta Smolenska, Peter Kolb, Sinan Tang, Mironas Bitinis, Héctor Hernández, and Elin Asklöv. 2021. CLULEX at SemEval-2021 task 1: A simple system goes a long way. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 632–639, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Victoria Yaneva, Peter Baldwin, Christopher Runyon, et al. 2023. Extracting linguistic signal from item text and its application to modeling item characteristics. In *Advancing Natural Language Processing in Educational Assessment*, pages 167–182. Routledge.

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.