Learning When to Defer to Humans for Short Answer Grading

Zhaohui Li[1], Chengning Zhang[1], Yumi Jin[1], Xuesong Cang[2], Sadhana Puntambekar[2], and Rebecca J. Passonneau[1]

[1] Department of Computer Science and Engineering, Pennsylvania State University
{zjl5282,cbz5118,yfj5089,rjp49}@psu.edu
[2] Department of Educational Psychology, University of Wisconsin-Madison
xcang@wisc.edu, puntambekar@education.wisc.edu

Abstract. To assess student knowledge, educators face a tradeoff between open-ended versus fixed-response questions. Open-ended questions are easier to formulate, and provide greater insight into student learning, but are burdensome. Machine learning methods that could reduce the assessment burden also have a cost, given that large datasets of reliably assessed examples (labeled data) are required for training and testing. We address the human costs of assessment and data labeling using selective prediction, where the output of a machine learned model is used when the model makes a confident decision, but otherwise the model defers to a human decision-maker. The goal is to defer less often while maintaining human assessment quality on the total output. We refer to the deferral criteria as a *deferral policy*, and we show it is possible to learn when to defer. We first trained an autograder on a combination of historical data and a small amount of newly labeled data, achieving moderate performance. We then used the autograder output as input to a logistic regression to learn when to defer. The learned logistic regression equation constitutes a deferral policy. Tests of the selective prediction method on a held out test set showed that human-level assessment quality can be achieved with a major reduction of human effort.

Keywords: Selective prediction · Automated short answer grading

1 Introduction

Assessment is important both for student learning and for instructors' implementation of a curriculum. Asking students to show their reasoning through writing is widely believed to facilitate better understanding [9], which would argue for open-ended questions instead of fixed response ones, such as multiple choice or true/false. But the choice between the two is heavily weighted towards fixed response types because they can be graded automatically in little time, thus minimizing assessment effort and maximizing timeliness. Deep learning methods that could reduce the assessment burden also have a cost, since they depend on large datasets of reliably assessed examples (labeled data) for training and testing autograder models. To offset the costs of human assessment of student work and of providing labeled data for machine learning, we present a human-in-the-loop [15] method that also leverages transfer learning from historical data.

The dataset used in this study was from middle school students participating in a unit on physics. In this unit, students learned about the basic concepts in mechanics while designing a roller coaster using a simulation. The curriculum was designed for students to build on science knowledge, adding more concepts as students iteratively engaged in the roller coaster design process. After each design cycle, students wrote short answers to open-ended questions based on the content they just learned. The questions helped teachers and researchers assess students' current understanding, as well as how they progressively learned science ideas throughout the unit. The data was collected from a single school from students in classes taught by three teachers. The school was situated in a rural-suburban area in Midwestern United States. Students from rural as well as suburban areas attended the school. Students in this district came from diverse backgrounds representing 22 different languages. Approximately twenty two percent of students in the school district identified as economically disadvantaged

Human-in-the-loop approaches include selective prediction, which uses confident decisions from a machine learned model, but otherwise defers to a human. Recent work indicates that softmax probabilities from machine learned models can serve as a measure of relative confidence [11, 23], and that the human decision makers perform better when they are given no information about the model prediction [2], with the result that human effort can be offline. We refer to the deferral criteria as a *deferral policy*. Using an existing high-performing automated short answer grading (ASAG) model [13], we first train this model on a combination of historical data and a small amount of newly labeled data, achieving moderate accuracy. Then we compare a very simple, and therefore elegant, manually derived heuristic deferral policy with machine learned policies. On a held out test set, the learned policies generalize better than the heuristic one. Further, the learned policy has a tuning parameter to control the tradeoff between accuracy and human effort. To our knowledge, our work is the first to learn controllable selective prediction deferral policies.

2 Related Work

Recent machine learning for automatic short answer grading (ASAG) relies primarily on deep learning, especially transformer models such as BERT [5]. Most work trains a pipeline consisting of an encoder for the student answer and reference answer followed by a classifier layer to determine answer correctness. Two often-used benchmark datasets are SemEval [6] and ASAP [21]. SemEval has 2-way, 3-way and 5-way correctness labels in three increasingly challenging settings: unseen answers, unseen questions, and unseen domain. Using a biLSTM, Riordan et al. [17] achieved quadratic weighted kappa of 0.77 on ASAP. On the same dataset with a model that encodes rubric elements, Wang et al. [21] achieved similar results. On a large, proprietary dataset from psychology, Liu et al. [14] trained separate encoders for student and reference answers, followed by another transformer layer to merge them, followed by a multilayer perceptron (MLP) classifier. They achieved 0.89 accuracy against a simple logistic regres-

sion baseline that achieved 0.83. Also on a proprietary dataset, Sung et al. [19] achieved accuracies between 0.78 and 0.81 using BERT to separately encode the question, and the concatenation of the student and reference answer. Saha et al.[18] separately encoded the question, the reference answer and the student answer using InferSent [3], and combined the three vectors with manually engineered features as input to a simple classifier. On the nine SemEval tasks, their accuracies ranged from 0.51 to 0.79. Using only the student answer and reference answer as input, Ghavidel et al. [8] compared BERT (cased and uncased) with XLNET on the SciEntsBank subset of SemEval, achieving SOTA accuracies. On the full SemEval, Li et al. [13]'s work SFRN used BERT to separately encode the question, the reference answer, and the student answer, and learned a relation vector over the three encodings, followed by an MLP classifier. SFRN achieved accuracies on SemEval from 0.40 to 0.91, and either beat the state-of-the-art or was competitive. In particular, SFRN did far better than other approaches on the 5-way classification. Given that SFRN is superior on public datasets to the other models examined here, we use SFRN in our selective prediction system. Multiple approaches to human-machine teams exist, including human-in-the-loop, where humans intervene on items that are difficult for a given algorithm [15], and algorithm-in-the-loop, which privileges human decision-makers but incorporates algorithms for efficiency or improvements in total accuracy [10]. Factors that influence whether a human-machine team works better than either agent alone include the type of decision-making algorithm, the level of risk of the decision, the transparency of the machine decisions, and human attitudes towards AI (e.g., automation bias) or biases (e.g., racial) [4]. Evidence suggests that for the supervised learning classification task of assigning a correctness label, overall accuracy can be maximized by a selective prediction approach to human-in-the-loop AI [22]. In a selective prediction experiment on classification of images of landscapes to detect animals to benefit wildlife in the Serengeti, Bondi et al. [2] found that the total accuracy of a human-in-the-loop system was best with defer-only, meaning humans are told only that the algorithm defers, rather than showing the algorithm's decision or confidence. That it is critical to present the right information to humans in a human-machine team is consistent with initial experiments on human-in-the-loop essay grading [1]. Based on defer-only, selective prediction can be developed by acquiring human decisions prior to developing the selective prediction system, which is what we do here. Hendrycks and Gimpel [11] introduced a simple selective prediction method using the maximum softmax probability as the confidence estimator. Other work trained a calibrator model on the softmax output as a selective prediction confidence [7, 12, 20]. Our work is most similar to calibration methods, but relies on a weight hyper-parameter to control accuracy and deferral rate.

3 Description of the Data

Our dataset consists of short answer responses to open-ended questions from a middle school physics curriculum about roller coasters collected from classrooms

of seven teachers. Responses were coded on a 3-point scale as incorrect, partially correct, or correct. Here we describe the nine open-ended questions and their relative difficulty for students, the coding reliability, the size of the dataset and the labeled subset, and our historical data used in transfer learning with SFRN.

Students participated in a three to four week unit to learn about the physics of energy transfer by designing a roller coaster in simulation. There were five labs that in turn addressed the following conceptual relationships: initial drop height and energy; the effect of hills on energy; mass and energy; height and speed; mass and speed. At the end of each lab, students responded to two to five open-ended questions along with some multiple-choice questions. There were thirteen open-ended questions and six multiple choice questions across labs. The open-ended questions covered nine relations among physics concepts (e.g., greater height corresponds to more potential energy), including the law of conservation of energy. The questions thus assessed students' grasp of the interconnectedness among science concepts. Our experiments used the nine open-ended questions that assessed understanding of physics relations.

Four researchers working in two groups coded the answers as correct, partially correct or incorrect. One group coded labs one and three; and the other group coded the rest. Each group coded a randomly selected set of responses for each lab. Inter-rater reliability (IRR) was assessed using two-way random, consistency average-measures of intra-class correlations (ICCs). A 95% confidence interval with two-tailed tests was performed. The cutoff for qualitative ratings of agreement based on ICC values was 0.90, and for each question was 0.80. The agreement measures for each question are displayed in the ICC column of Table 1. To achieve reliability, the coders went through two to four rounds of coding. Disagreements were resolved in team discussions. The coding scheme was refined or extended in the process of making decisions for the disagreements.

The full dataset consists of 4,703 items, shown broken down by question in Table 1. Table 1 also shows the breakdown by question for two phases of coding: the first phase where reliability was developed (Coding 1), and the second phase after reliability was achieved (Coding 2). Counts per question, ICC, human accuracy and average student score are shown for Coding 1, while Coding 2 and Combined (Coding 1 + Coding 2) show the counts for each question and average student score. We split Coding 1 into a train set of 408 examples to train SFRN, and a dev set of 447 examples to evaluate SFRN and to develop deferral policies. SFRN was trained on a combination of this small subset of the spring 2022 data, and historical data that is described below. We use the Coding 2 data as a test set to evaluate the deferral policies.

As shown by the average score column in Table 1, the number of questions per lab decreased in later labs. With the exception of Lab 2, questions within a lab generally increased in difficulty, with a trend towards greater difficulty in later labs. The questions that were easiest for students (Lab1.Q1, Lab3.Q3 and Lab3.Q4) tended to have the highest ICC and human accuracy, but otherwise these rankings do not correlate well. In the results section, we show differences in selective prediction performance broken down by question.

Table 1: A breakdown by question is shown for the full data set, the first and second coding phases (Cod.1, Cod. 2), and the combination of the two codings (Comb.). For Cod. 1, the table shows ICC, average human accuracy, and average student scores. Cod. 2 and Comb. have the counts and average student scores. The last *Summary* row has column totals of counts (Full, Cod.1, Cod.2), average of averages for Hum. Acc. and Avg. Sc. columns, and the overall ICC score.

Qid	Full	Cod. 1	ICC	Hum. Acc.	Avg. Sc.	Cod. 2	Avg. Sc.	Comb.	Avg. Sc.
Lab1.Q1	539	92	0.93	0.97	1.88	56	1.68	148	1.80
Lab1.Q2	534	92	0.92	0.95	1.22	54	0.76	146	1.05
Lab1.Q4	513	92	0.92	0.92	1.12	54	0.70	146	0.97
Lab2.Q3	528	94	0.92	0.91	1.07	49	0.73	143	0.96
Lab2.Q4	514	94	0.84	0.89	1.43	50	1.00	144	1.28
Lab3.Q3	522	85	0.95	0.96	1.85	52	1.50	137	1.72
Lab3.Q4	507	85	0.94	0.94	1.72	50	1.60	135	1.67
Lab4.Q2	514	85	0.91	0.93	1.56	52	1.62	137	1.58
Lab5.Q2	532	136	0.87	0.91	1.31	50	1.46	186	1.35
Summary	4703	855	0.93	0.93	1.46	467	1.23	1322	1.37

The historical data consists of 6,956 student responses to 33 questions drawn from a decade of pre- and post-tests for assessment of middle school students' understanding of relationships among physics concepts, such as how the height of an inclined plane affects the amount of force needed to lift a load. Validity of questions was ensured through consultation with physics experts, teachers, and statisticians. As in our current data, questions were assessed on a 3-pt scale, with 10-25% of each test coded independently by two researchers who achieved at least 85% agreement, a standard metric at the time of the original studies. As the original reliability coding is no longer available, we cannot apply an agreement coefficient post-hoc. However, chance-adjusted agreement scores reduce to percent agreement as chance agreement approaches zero [16], which would hold for this data where the proportions of correct, partially correct and incorrect were only mildly skewed (respectively 25%, 42% and 33%).

4 Methods

To produce a selective prediction system, we first trained a classifier to predict correctness of each student response. Here we describe how we used historical data combined with a small amount of labeled data to train the classifier (N=408; from Coding 1 in Table 1), testing it on our dev data (N=447; also drawn from Coding 1). Using this classifier, the SFRN output on the dev data was then used to develop deferral policies, which we evaluated on the test data (N=467; Coding 2). Here we describe the classifier, a heuristic deferral policy, and our use of linear machine learning to learn deferral policies.

For the classifier, we chose the SFRN algorithm for its superior performance, as reviewed above in section 2. In the SemEval dataset that SFRN was developed on, there are often multiple reference answers per question prompt. For

each student response, three text strings are first encoded using BERT [5]: the question prompt, a reference answer, and the student answer. This triple of embeddings is the input to a multilayer linear perceptron (MLP) that learns a single relation vector for each reference answer for a given student answer. A fusion relation is then learned that merges all the relation vectors for a given student answer, which is then classified for correctness by a final learned MLP. As well as creating the question prompts for the historical and spring 2022 data, we also created reference answers for each correctness class. We first trained SFRN on the historical data alone (6,956 student answers, 33 questions), then in combination with the training subset of Coding 1. When trained only on the historical data, SFRN had 45.86% accuracy on dev. By adding in the training subset of the spring 2022 data, SFRN performance improved to 62.41%. This is a substantial increase that shows the power of transfer learning, but it is still well below average human accuracy of 93%.

Given that the softmax probability from a neural model's output layer performs well as a confidence estimator for selective prediction [11, 23], we develop a heuristic deferral policy D based on a probability threshold θ:

$$D(\mathcal{P}(x);\theta) = \begin{cases} 1, & \text{if } \max \mathcal{P}(x) \leq \theta \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $P(x)$ is the probability distribution over the three classes, 1 represents the decision to defer, and 0 triggers use of the model decision. We test deferral policies by retrieving the correct label when the policy says to defer.

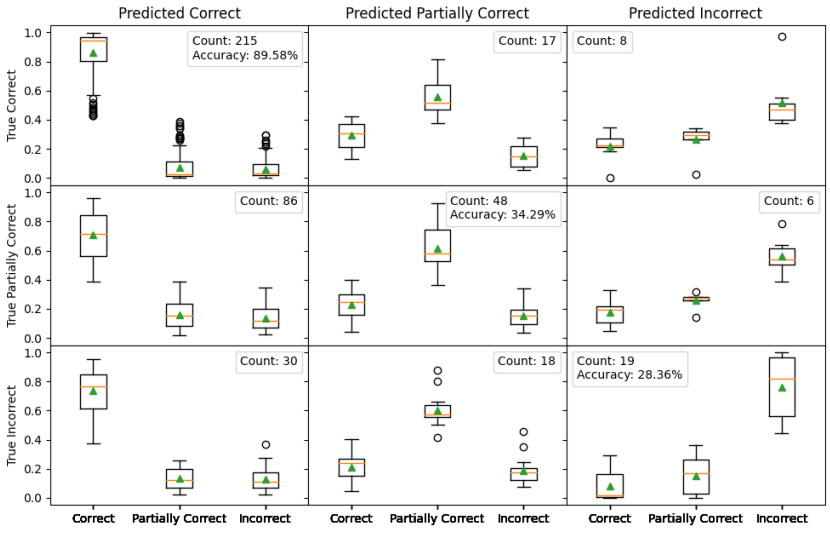


Fig. 1: Confusion matrix of SFRN on dev (N=447). Each cell contains the count and three whisker plots for the probability distribution over the three classes.

Fig. 1 is a confusion matrix of the SFRN predictions on dev, where each cell has the count and whisker plots for the probabilities of the three classes. The cells on the diagonal show that when SFRN predicts correct, it is both accurate (89.58%) and quite certain (mean probability of correct above 0.90), but has lower accuracy and confidence on its other predictions. The first column shows that a single threshold of around 0.85 or so when SFRN predicts correct would include most of the true correct, and exclude the predictions of correct that are best to defer on. However, a single θ is not obvious from this figure.

Again using SFRN output on dev, Fig. 2 shows a separate curve for each correctness class with values of θ on the x-axis and class accuracy on the y-axis. The incorrect and partially correct curves intersect at θ=0.68 with accuracy above 90%, but with lower class accuracy for correct. A heuristic policy with θ=0.68 applied to dev yields a total accuracy of 80%, and a low deferral rate of 30%. We report performance of this policy on the test set in the next section.

Based on the small size of the dev data for policy training, we use linear classifiers to learn deferral policies from SFRN's output on the dev data. We selected logistic regression and random forest for their interpretability and good performance. We tested a variety of feature representations, drawing from a set of 11: SFRN probabilities for each class (P0, P1, P2), SFRN prediction (Max), class with the next highest SFRN probability (Mid), SFRN accuracy on the given question (SFRN-Qdiff), average student score per question (stud-Qdiff), ICC score per question (coder-Qdiff). We also tested a feature for noise injection, sampling numbers from a Gaussian distribution with mean and standard deviation of 0.5. Max is the inverse of the proportion of each class in SFRN output. This captures the class influence, and puts the values on the same scale as the other features (0 to 1) for better interpretability of the learned weights.

As a proxy for a ground truth label of whether to defer, we use the SFRN correctness to label each example as +defer if the model was incorrect and -defer otherwise. We use a weighted cross entropy loss, where a weight hyperparameter λ on the positive class allows us to increase the weight on the decision to defer, resulting in a higher deferral rate but also higher accuracy. In the next section we show how accuracy and deferral rate on the test set varies with values of λ.

5 Results

Here we report performance of the deferral policies on the test data (N=467). Because logistic regression (LR) models performed somewhat better than random forest, we only discuss the LR policies. Below we present the LR training parameters, followed by a discussion of the features that performed best. For the best performing set of features, we then discuss how the λ weight in our loss function affects the tradeoff between accuracy and deferral rate. Then we compare the performance of SFRN alone, the heuristic policy, and several of the LR policies, showing that the heuristic policy does not generalize well. An LR policy that has the same accuracy as humans has less reduction of human effort,

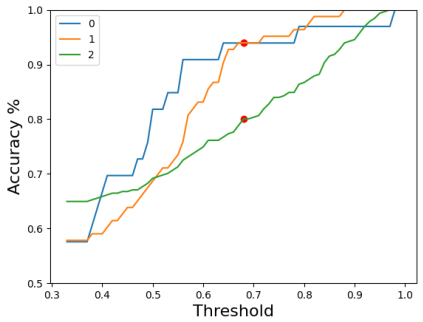


Fig. 2: Deferral threshold (θ) by accuracy for correct (2), partially correct (1), and incorrect (0) (dev data).

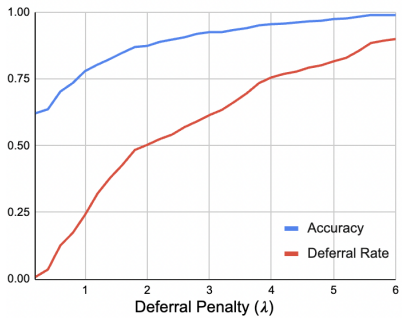


Fig. 3: Sensitivity plot for the loss weight (λ) by accuracy (blue) and deferral rate (red) (test data).

Table 2: Feature sets (N=3 to 5), weights and performance for three LR models.

Learned Weights					Intercept	Performance	
SFRN-QDiff	Max	P0	P1	P2	Intercept	Accuracy	Deferral rate
NA	*NA*	-0.0963	0.8442	-0.8959	5.4066	71.30	17.98
NA	-13.2667	2.6016	2.4180	-1.9343	5.3704	75.18	21.61
-3.4324	-13.6363	3.4137	3.5853	-1.5924	5.9448	**77.97**	**24.17**

however reasonable accuracy is achieved with a significant reduction of effort. Finally, we report differences in performance across the nine questions prompts.

Using the logistic regression from the scikit learn python library, we found good performance with the default hyperparameters, apart from assigning no regularization penalty (versus the default L2 norm), and a high inverse regularization strength C=20 versus the default of 1. Of the 11 features we experimented with, we found that six did not contribute to model performance. We compared multiple feature sets from the remaining five.

The five features that had an impact on performance were the SFRN prediction (Max), the SFRN training accuracy on the question (SFRN-QDiff), and the softmax probabilities (P0, P1, P2). A logistic regression deferral policy $\mathcal{D}_{\mathcal{LR}}$ is the learned linear equation for the log odds of the positive class (+defer):

$$\mathcal{D}_{\mathcal{LR}} = \alpha + (\sum_i W_i \times F_i) + \epsilon \tag{2}$$

where α is the intercept, F_i represents the value of feature i for a given input, W_i represents the learned weight for that feature, and ϵ is the residual error. The sign of W_i indicates whether F_i increases or decreases the log odds of +defer, and higher magnitudes of W_i indicate F_i has greater influence.

Table 2 shows the learned weights and the performance of the LR policies for the softmax probability features alone or in combination with other features; NA indicates the feature was not used. The 5-feature model has the highest accuracy,

Table 3: Accuracies and deferral rates of the heuristic and LR deferral policies on the Dev and Test sets, compared to SFRN alone (None). For the learned policies, we show 95% confidence intervals (CIs) on the test set, and also the ICC score with the ground truth labels to compare with human ICC.

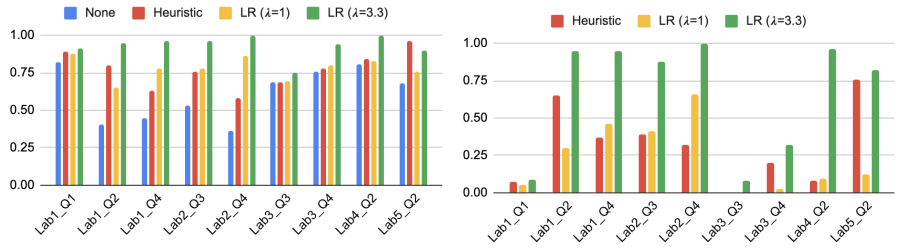
	Dev		Test		
Deferral Policy	Accuracy	Def. Rate	Accuracy (95% CI)	Def. Rate (95% CI)	ICC
None	63.08	*N*A	61.75 (60.71, 62.85)	*N*A	0.64
Heuristic	80.08	30.42	76.33 (75.47, 77.38)	27.09 (26.18, 28,09)	0.80
LR ($\lambda = 1$)	77.79	25.75	77.97 (76.66, 79.28)	24.17 (22.85, 25.24)	0.80
LR ($\lambda = 0.9$)	75.32	20.38	75.82 (74.52, 77.14)	**20.48** (19.04, 21.66)	0.78
LR ($\lambda = 1.1$)	79.36	28.88	79.49 (78.33, 80.71)	27.19 (25.70,28.33)	0.82
LR ($\lambda = 3.3$)	95.29	66.72	**92.94** (92.38, 93.57)	65.03 (63.56, 66.43)	**0.94**

but also the highest (worst) deferral rate. Because we prioritize accuracy first, our remaining results pertain to this LR. Two of the three negative features (Max, SQRN-Qdiff) have high weights, meaning high values of these features reduce the log odds more: the policy is increasingly less likely to defer for each next class ordered as incorrect, partially correct, correct, and the policy is less likely to defer when it is more accurate on the given question. It has a somewhat lower negative weight on the probability of the correct class. The high positive weights on P0 and P1 mean that the higher the probabilities of incorrect or partially correct, the more often the policy will defer, which is consistent with what we saw in Fig. 1 of lower SFRN accuracies on incorrect and partially correct.

Fig. 3 shows the sensitivity of accuracy and deferral rate to values of λ for the 5-feature LR policy. For $\lambda = 1$, accuracy is 77.97% with a low deferral rate of 24.17%. As λ increases, the deferral rate increases faster than accuracy, meaning that in our setting reducing human effort without reducing accuracy is very challenging. Reducing λ below 1 does not increase the accuracy.

Table 3 compares SFRN with no selective prediction, the heuristic policy, and the 5-feature LR policy with 4 different values of λ. To assess variance, we computed confidence intervals (CIs) using 200 iterations of bootstrapped samples (with replacement) using 90% of the test data. All policies are clearly better than the classifier on its own, showing it is possible to greatly increase the accuracy through selective prediction at less than 100% human effort. The heuristic policy, which had higher dev accuracy than LR $\lambda = 1$, does not generalize as well to the test set, and while the LR $\lambda = 1$ policy does not have confidently better accuracy and has equal ICC, it has a confidently lower deferral rate. Compared to LR $\lambda = 1$ policy, LR $\lambda = 1.1$ is about as accurate (overlapping CIs), but has a confidently higher deferral rate, further evidence that our setting is challenging. For $\lambda = 3.3$, the policy is as accurate as expert humans with somewhat higher ICC (0.94 versus 0.93), at two thirds the human effort.

Because human reliability varied by question, we examined the performance of the deferral policies broken down by question. Fig. 4a) shows that SFRN accuracy varies a lot by question. On questions that are harder for SFRN, there



(a) Accuracy. (b) Deferral rate.

Fig. 4: Performance broken down by question.

are also greater accuracy differences across policies, especially for Lab2_Q4. The heuristic policy, with an overall deferral rate of 27.09%, defers most for questions Lab1_Q2 and Lab5_Q2, and otherwise much less often. In contrast, LR $\lambda = 3.3$, which has a high overall deferral rate of 65.03%, has low deferral rates only on questions Lab1_Q1, Lab3_Q3 and Lab3_Q4.

6 Discussion and Conclusion

In this paper, we have shown it is possible to learn a selective prediction policy that can achieve expert human accuracy and ICC, while reducing human effort. Our immediate goal is to produce a reliable coding of our dataset of 4,703 items. By labeling 1,322 items, we have learned a selective prediction policy LR $\lambda = 3.3$ that can be expected to achieve expert human performance on the remaining unlabeled data (N=3,381), where the expert humans would need to label 65.03%, or 2,199 more items. In sum, experts would have labeled 3,521 items, for a 25% reduction in human effort and expert human accuracy. In practice, we plan to learn new LR policies where we split our combined set of 1,322 labels into 50% for training and 50% for development. Instead of a reserved test set, we could verify the quality by re-coding random samples after we use the new LR policy to code the entire dataset using online hu. By retraining SFRN using the historical data combined with 611 labeled examples instead of 408, SFRN should improve. An LR policy learned on the remaining 611 examples should perform better than our current LR $\lambda = 3.3$ policy, given improved performance of SFRN and a larger dataset for policy training of 611 instead of 447 (an increase of 28%).

The main limitation of our work is that it would be preferable to find a learning method for selective prediction that can jointly optimize accuracy and the deferral rate. Another limitation is that both classifier performance and policy performance might have been higher if we had been able to code relatively more of the questions that turned out to be difficult for the classifier. Future work could investigate these two issues, as well as whether selective prediction methods could work in the classroom, which could potentially increase timeliness and accuracy of grading, and avoid the subjectivity of human graders.

References

1. Baikadi, A., Becker, L., Budden, J., Foltz, P.W., Gorman, A., Hellman, S., Murray, W., Rosenstein, M.: An apprenticeship model for human and AI collaborative essay grading. In: Trattner, C., Parra, D., Riche, N. (eds.) Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019). vol. 2327 (2019), http://ceur-ws.org/Vol-2327/IUI19WS-UIBK-2.pdf
2. Bondi, E., Koster, R., Sheahan, H., Chadwick, M., Bachrach, Y., Cemgil, T., Paquet, U., Dvijotham, K.: Role of Human-AI Interaction in Selective Prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36 (5), pp. 5286–5294 (Jun 2022). https://doi.org/10.1609/aaai.v36i5.20465
3. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 670–680. ACL, Copenhagen (Sep 2017). https://doi.org/10.18653/v1/D17-1070
4. De-Arteaga, M., Fogliato, R., Chouldechova, A.: A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12. ACM, Honolulu HI USA (Apr 2020). https://doi.org/10.1145/3313831.3376638
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL. pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423
6. Dzikovska, M., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H.T.: SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 263–274. ACL (2013), https://aclanthology.org/S13-2045
7. Garg, S., Moschitti, A.: Will this question be answered? question filtering via answer model distillation for efficient question answering. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7329–7346. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.583, https://aclanthology.org/2021.emnlp-main.583
8. Ghavidel, H., Zouaq, A., Desmarais, M.: Using BERT and XLNET for the Automatic Short Answer Grading Task:. In: Proceedings of the 12th International Conference on Computer Supported Education. pp. 58–67. SCITEPRESS - Science and Technology Publications, Prague, Czech Republic (2020). https://doi.org/10.5220/0009422400580067
9. Graham, S., Kiuhara, S.A., MacKay, M.: The Effects of Writing on Learning in Science, Social Studies, and Mathematics: A Meta-Analysis. Review of Educational Research **90**(2), 179–226 (Apr 2020). https://doi.org/10.3102/0034654320914744
10. Green, B., Chen, Y.: Algorithm-in-the-loop decision making. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34 (9), pp. 13663–13664 (Apr 2020). https://doi.org/10.1609/aaai.v34i09.7115
11. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks, http://arxiv.org/abs/1610.02136, number: arXiv:1610.02136

12. Kamath, A., Jia, R., Liang, P.: Selective question answering under domain shift. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5684–5696. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.503, https://aclanthology.org/2020.acl-main.503
13. Li, Z., Tomar, Y., Passonneau, R.J.: A Semantic Feature-Wise Transformation Relation Network for Automatic Short Answer Grading. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6030–6040. ACL (2021). https://doi.org/10.18653/v1/2021.emnlp-main.487
14. Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G.Y., Liu, Z.: Automatic short answer grading via multiway attention networks. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds.) Artificial Intelligence in Education. pp. 169–173. Springer International Publishing, Cham (2019)
15. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., Fernández-Leal, A.: Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review (Aug 2022). https://doi.org/10.1007/s10462-022-10246-w
16. Passonneau, R.J., Carpenter, B.: The benefits of a model of annotation. Transactions of the ACL **2**, 311–326 (2014). https://doi.org/10.1162/tacl_a_00185, https://doi.org/10.1162/tacl_a_00185
17. Riordan, B., Horbach, A., Cahill, A., Zesch, T., Lee, C.M.: Investigating neural architectures for short answer scoring. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 159–168. ACL, Copenhagen (Sep 2017). https://doi.org/10.18653/v1/W17-5017
18. Saha, S., Dhamecha, T.I., Marvaniya, S., Sindhgatta, R., Sengupta, B.: Sentence level or token level features for automatic short answer grading?: Use both. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (eds.) Artificial Intelligence in Education. pp. 503–517. Springer International Publishing, Cham (2018)
19. Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., Arora, R.: Pre-training BERT on domain resources for short answer grading. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6071–6075. ACL, Hong Kong (2019). https://doi.org/10.18653/v1/D19-1628
20. Varshney, N., Mishra, S., Baral, C.: Investigating Selective Prediction Approaches Across Several Tasks in IID, OOD, and Adversarial Settings. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 1995–2002. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.findings-acl.158
21. Wang, T., Inoue, N., Ouchi, H., Mizumoto, T., Inui, K.: Inject rubrics into short answer grading system. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 175–182. ACL, Hong Kong (Nov 2019). https://doi.org/10.18653/v1/D19-6119
22. Wiener, Y.: Theoretical foundations of selective prediction. Ph.D. thesis, Technion - Israel Institute of Technology, Israel (2013)
23. Xin, J., Tang, R., Yu, Y., Lin, J.: The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing (IJCNLP). pp. 1040–1051. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.84