

Fine-tuning Large Language Models for Automated Scoring: Classification, Regression, vs. Ordinal Regression

Ji Yun Zu and Akshay Badola

Educational Testing Service, Princeton, NJ

Correspondence: jzu@ets.org

Abstract

Fine-tuning large language models for automated scoring is often formulated as either a regression or a classification task. However, scores assigned by human raters are on an ordinal scale. We summarize different deep-learning ordinal regression methods and compare their performances with those from regression and classification using a real dataset.

1 Introduction

Automated scoring refers to the use of statistical and computational methods to assign scores to responses to open-ended questions (e.g., essays, math short answers). Earlier automated scoring methods include using hand-crafted features from the responses as predictors to predict human ratings. More recently, deep-learning methods that use multiple layers of neural networks have been applied to automated scoring. In particular, fine-tuning open-source large language models (LLMs) often lead to superior results. For example, the high performance of LLM approaches were a key takeaway from both the NAEP Reading Automated Scoring Challenge (Fall 2021) and the NAEP Math Automated Scoring Challenge (Spring 2023) (Whitmer et al., 2023; Whitmer and Beiting-Parrish, 2023).

A typical fine-tuning LLMs approach for automated scoring trains a pre-trained LLM (e.g., BERT family of transformer models) with either a regression or classification layer added as the last layer. They use texts of the constructed responses as input to predict human-assigned scores (see Lottridge et al., 2023, Figure 2.3 for a high-level design of neural network-based automated scoring engine). However, scores are often defined by scoring rubrics on an ordinal scale, e.g., 0 to 4 or 0 to 6. Classification treats the scores as nominal. It does not use the information that for an essay of score 3, a predicted score of 2 has smaller errors than a predicted score of 1. On the other hand, regression

treats the numeric values of the scores as interval. It ignores that the scores can only take a limited number of values, and assumes a score of 3 is equal distance from 2 as from 4.

Ordinal regression acknowledges the order of the scores, but does not assume equal distance between the adjacent scores. Researchers have applied ordinal regression models on automated scoring. For example, Haberman and Sinharay (2010) applied cumulative logistic regression models for automated essay scoring using data from four different tests. The authors concluded that the cumulative logit model performed somewhat better than the linear regression model.

More recently, ordinal regression methods for deep neural networks have also been developed. For example, Cao et al. (2020) proposed the Consistent Rank Logits (CORAL) framework, and later Shi et al. (2023) proposed Conditional Ordinal Regression for Neural Networks (CORN). In these studies, the neural network architecture are either convolutional neural networks (CNN) or multilayer perceptron (MLP), the criterion score is age rather than rubric-defined score levels, and performances of regression were not studied. In both studies, the authors applied the proposed methods to several datasets and compared their performances with classification and prior ordinal regression approaches. Results in Shi et al. (2023) suggested that CORN had the smallest or comparable average mean absolute error and root-mean-squared-error on the test sets comparing with classification, ordinal regression CNN (Niu et al., 2016), and CORAL. In these studies, the neural network architecture are either convolutional neural networks (CNN) or multilayer perceptron (MLP), the application datasets are dominated by using face images to predict a large number of age categories, and the performance of regression were not studied.

There are several differences between automated scoring and age prediction where deep-learning or-

dinal regression was studied. In automated scoring, the number of score points is usually smaller than that of the age categories; pre-trained transformer models (Vaswani et al., 2017) are often used for fine-tuning; and agreement measure between predicted and human assigned scores are often used as the metric for evaluating model performance. To our knowledge, empirical studies comparing deep-learning automated scoring models that use classification, regression, ordinal regression methods are lacking.

2 Purpose of This Study

In this study, we compare fine-tuning DeBERTa-v3-base (He et al., 2021) with a classification head, a regression head and two different ordinal regression methods (CORAL and CORN) in terms of the quality of automated scoring of a speech item using transcriptions.

3 Methods

Let $D = \{x_i, y_i\}, i = 1, \dots, N$ be a dataset of N samples, where x_i represents the text of the constructive response of person i , and y_i is the corresponding human-assigned score that $y \in \{0, \dots, K - 1\}$ has K ordered categories. When fine-tuning an LLM such as DeBERTa for automated scoring, the general language representations learned in the pre-trained LLM is used as the starting model to connect texts of the constructed responses to representations to human assigned scores. Training samples of $\{x_i, y_i\}$ are used to update the model parameters for the specific automated scoring task. There are different ways to map the last layer of the LLM to the human assigned scores.

3.1 Classification and Regression

Classification and regression are the two most common tasks connecting the last layer of the pre-trained model to the human assigned scores. With a classification task, the human assigned scores are treated as discrete labels ignoring the relative ordering of the scores, and model parameters are learned by minimizing the multi-class cross-entropy loss. With a regression task, human assigned scores are treated as continuous, and mean-squared-error loss is used as the optimization objective during training.

3.2 Rank Consistent Ordinal Regression

The two ordinal regression methods studied in this paper, CORAL (Cao et al., 2020) and CORN (Shi et al., 2023) both formulate ordinal regression with a dependent variable of K ordered categories as $K - 1$ binary classification problems, and achieve rank consistency (i.e., the predicted probability $P(Y > k - 1) \geq P(Y > k)$). They differ in how rank consistency is achieved.

To reformulate an ordinal regression as $K - 1$ binary classification tasks, y is first recoded into $K - 1$ binary variables ($y^1, \dots, y^{K-1}$) where

$$y_i^k = \begin{cases} 0 & \text{if } y_i > k \\ 1 & \text{if } y_i \leq k \end{cases}$$

Then, $K - 1$ classifiers are simultaneously trained respectively for $P(y^k = 1|x)$, which is equivalent as a classifier for $P(y > k|x)$.

CORAL achieves rank consistency by constraining the weights for the $K - 1$ classifiers to be the same, while letting the intercepts freely estimated. Note that the same constraints are imposed in the proportional odds model. The loss function for CORAL is the sum of the $K - 1$ binary cross-entropy losses.

CORN relaxes the weight sharing constraint, instead relies on conditional training subsets and chain rule for conditional probabilities to achieve rank consistency. In other words, to train a classifier for $y > 1$, all the data is used. When training for $y > 2$, it only uses data with $y > 1$. Consequently, the predicted probabilities are conditional probabilities, e.g., $p(y > 2|y > 1)$. Given the chain rule of conditional probability, rank consistency is guaranteed. The loss function for CORN is the sum of the $K - 1$ binary cross-entropy losses weighted by the size of the conditional training subsets, and the probabilities in the loss function are conditional probabilities.

3.3 Score Prediction under Ordinal Regression

Under ordinal regression, the trained models outputs probabilities. The original papers for CAROL (Cao et al., 2020) and CORN (Shi et al., 2023) used these probabilities one way to obtain what we refer to as the median predictor. We argue there are at least two other ways of using the probabilities, leading to a mode and an expected predictor. Note that for the same model with the same loss function, these three score predictors exist.

Median predictor. In the original proposal of CAROL (Cao et al., 2020) and CORN (Shi et al., 2023), the authors use the sum of the predicted binary classifiers $\hat{y}_i^k$, which takes the value of 1 if $\hat{p}(y_i > k) > 0.5$ and 0 otherwise. This estimator is actually determined by the highest category k with $\hat{p}(y_i > k) > 0.5$, because such predicted probability for a lower category is guaranteed to have a predicted probability higher than 0.5. We refer to this estimator the median estimator due to its connection to 0.5.

Mode predictor. Given the model, we can calculate the predicted probability of each categorical score. Another predictor can be the category that has the highest predicted probability, that is $\hat{y}_i = \underset{i}{\operatorname{argmax}} \hat{p}(y_i = k)$. This is how the score from a classification is estimated.

Expected predictor. We can also use the expected value of the predicted scores as the expected predictor. That is $\hat{y}_i = \sum_{k=1}^{K-1} k \times \hat{p}(y_i = k)$.

4 Empirical Study

We conducted an empirical study to compare the quality of automated scoring models trained by fine-tuning DeBERTa-v3-base (He et al., 2021) with a classification head, a regression head and two different ordinal regression methods (CORAL and CORN) using data of a speaking item from an English proficiency test.

4.1 Item Type

We used data from a speaking item type in a standardized English proficiency test. For each item, test takers are presented with a prompt and asked to express their opinion on the topic within a 45-second time limit. Spoken responses are recorded via computer as audio files. Each response is scored on a scale from 0 to 4. The scoring rubrics concern the delivery, language use, and topic development aspects of the spoken response.

4.2 Datasets

We obtained 99,506 audio recordings with corresponding human-assigned raw scores of this item type. These are test takers’ responses to 195 different items.

In terms of the score distribution of human ratings, 3% of the responses received a score of 1, 32% received a 2, 55% received a 3, 10% received a 4, and there is only 0.1% of 0’s.

We first transcribed the audio files into text using NVIDIA Parakeet (Sekoyan et al., 2025), an open-source high performing automatic speech recognition model. In the transcription, the number of words per response ranged from 0 to 299, with the median of 94 words. Note that fillers e.g., uh, em and repetitions are kept in the transcription.

We randomly split the dataset into a training (70%, 69,654), an evaluation (15%, 14,926), and a test (15%, 14,926) set.

4.3 Model Training

We fine-tuned DeBERTa-v3-base (He et al., 2021) with four tasks: regression, classification, CORAL and CORN. In each task, transcriptions were used as input to predict human scores. Batch size was fixed at 16. Models with three learning rate ($5e-6$, $1e-5$, $5e-5$) were trained on the training set to up to 4 epochs. For CORAL and CORN, median, mode, and expected predicted scores were obtained. For each task, hyperparameters that led to the best performance on the evaluation set were selected to train the final model using the combined training and evaluation dataset.

Python libraries transformers (Wolf et al., 2020), PyTorch (Paszke et al., 2019) and coral-pytorch (Raschka et al., 2020) were used.

4.4 Model Evaluation

For model evaluation, we used quadratic weighted kappa (QWK), which captures the level of agreement above what would be obtained by chance and weighted by the extent of disagreement, as the evaluation metrics to describe the agreement between predicted scores and human ratings. We also reported accuracy (the proportion of exact agreement), bias, root-mean-squared error (RMSE) and correlation between the predicted and criterion scores. Predicted scores from regression, CORAL expected, and CORN expected are continuous. They were rounded and truncated to between 0 and 4 to calculate accuracy and QWK, and kept as is for calculating bias, RMSE, and correlation.

5 Results

5.1 Results on the Evaluation Set

The QWKs of the eight estimators (classification, regression, CORAL expected, CORAL median, CORAL mode, CORN expected, CORN median, and CORN mode) on the evaluation set were plotted against training learning rates and number of

epochs and shown as Figure 1. These plots clearly showed that a) CORAL median and CORAL mode predicted scores had low QWKs; b) Most training runs reached stable QWK within two epochs.

Hyperparameters that led to the best performing classification, regression, CORAL, and CORN estimators and the corresponding evaluation statistics on the evaluation set are reported in Table 1. Although results on the evaluation set suggested superiority of the CORN method, because data in the evaluation set were used for selecting hyperparameters, they should not be used to make conclusions on model performance. These results are reported as a reference for comparison to results on the test set.

5.2 Results on the Test Set

Performances of the final models using regression, classification, CORAL and CORN methods on the test set and their corresponding hyperparameters are reported in Table 2. These results suggest that a) CORN mode and Classification led to the highest and comparable QWK, followed by CORAL, and regression had the smallest QWK; b) CORN mode, CORAL expected, and Classification had comparable accuracy; c) Regression showed a positive bias; d) comparing with the QWK on the evaluation set, there was a 0.01 to 0.02 drop on the test set.

To better understand these results, the confusion matrices of the predicted scores and the criterion scores are plotted and shown in Figures 2 and 3. These plots clearly reveal that a) all four scoring models have the tendency to predict the most frequent score point 3; b) regression actually provided higher accuracy for predicting score 3 and 4. However, it overestimated more for score points 1 and 2; c) as expected, none of the models predict 0 well, due to the lack of training data; d) the wrongly predicted scores are often an adjacent score.

6 Discussion

In this study, we compared automated scoring models trained by fine-tuning DeBERTa with a regression, a classification, and two ordinal regression methods using the same real dataset. For the ordinal regression methods, we also proposed to estimate the scores as the score point that has the highest predicted probability (what we referred to as the mode estimator), and the expected value given model predicted probabilities (expected estimator). Results show that for this dataset a) the

CORAL median score estimator as used in the original CORAL paper had low QWK, while CORAL expected had QWKs closer to the other methods; b) CORN mode and Classification led to the highest and comparable QWK; c) CORN mode, CORAL expected, and Classification had comparable accuracies.

We hypothesized ordinal regression would performed the best for this dataset. However, our results did not suggest ordinal regression as a clear winner. Instead, classification provided comparable results. We understood regression may not be the best choice, because the constructed response item used in this study has a relative small number of categories. Research on Likert scale ratings usually recommended not to treat the responses as continuous when the number of score points is small (Rhemtulla et al., 2012). The comparable results between ordinal regression and classification suggests to us for this dataset predicted scores from classification are rarely to be wrong by more than one score point. Thus, treating them as nominal doesn't lose much information. Whether fine-tuning an LLM can yield predicted scores within one score point depending on the item type, how the rubrics are written, and the observed score distribution.

The QWKs on the test set are lower than those on the evaluation set, which is as expected. However, the drop for the two ordinal regression models are slightly larger than those for regression and classification. This could suggest more concerns for overfitting for ordinal regression due to their complexity.

One lesson we learned from this study was that it is still the best practice to try out regression, classification, and ordinal regression on the training and evaluation set before making a decision.

For future directions, we would like to explore weighted loss function that takes into account the relative frequencies of different scores and to evaluate more datasets.

7 Limitations

This study only analyzed one real dataset. We plan to study automated scoring of other item types to try to improve automated scoring performances and explain the differences among predicted scores from different tasks.

The data analyzed in this study were collected from a speaking item type. However, the automated

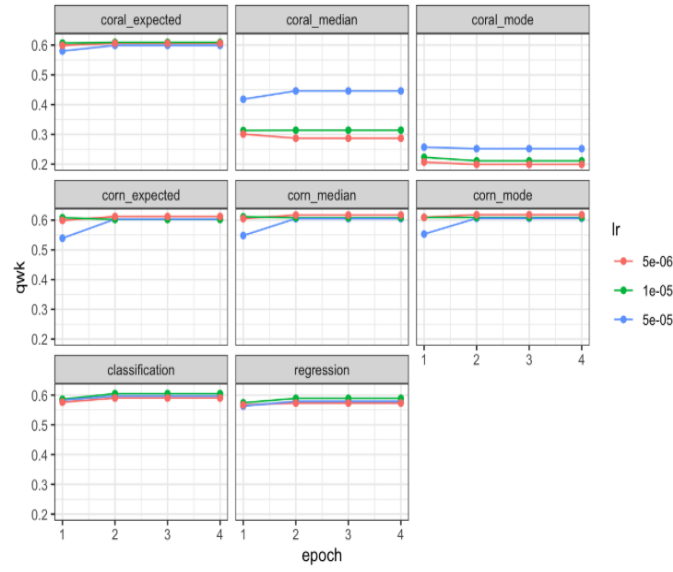


Figure 1: QWKs of the eight estimators on the evaluation set at different learning rate and number of epochs.

Table 1: Performance on the Evaluation Set

Estimator	Learning Rate	Number of Epoch	Accuracy	Bias	RMSE	Correlation	QWK
Classification	1e-5	2	0.678	0.084	0.572	0.615	0.605
Regression	1e-5	2	0.655	0.188	0.596	0.618	0.589
CORAL expected	1e-5	2	0.683	0.075	0.567	0.619	0.609
CORN mode	5e-5	2	0.681	0.053	0.569	0.623	0.618

Table 2: Performance on the Test Set

Estimator	Learning Rate	Number of Epoch	Accuracy	Bias	RMSE	Correlation	QWK
Classification	1e-5	2	0.674	0.070	0.580	0.603	0.593
Regression	1e-5	2	0.651	0.190	0.602	0.598	0.561
CORAL expected	1e-5	2	0.676	0.081	0.577	0.594	0.576
CORN mode	5e-5	2	0.675	0.050	0.578	0.603	0.594

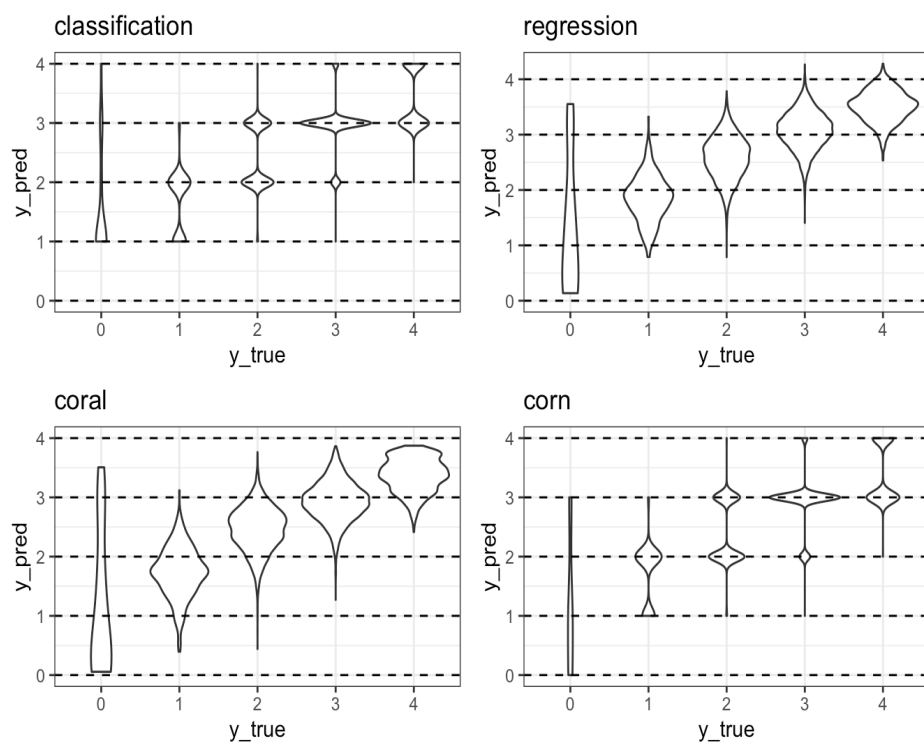


Figure 2: Violin plots of predicted and criterion scores.

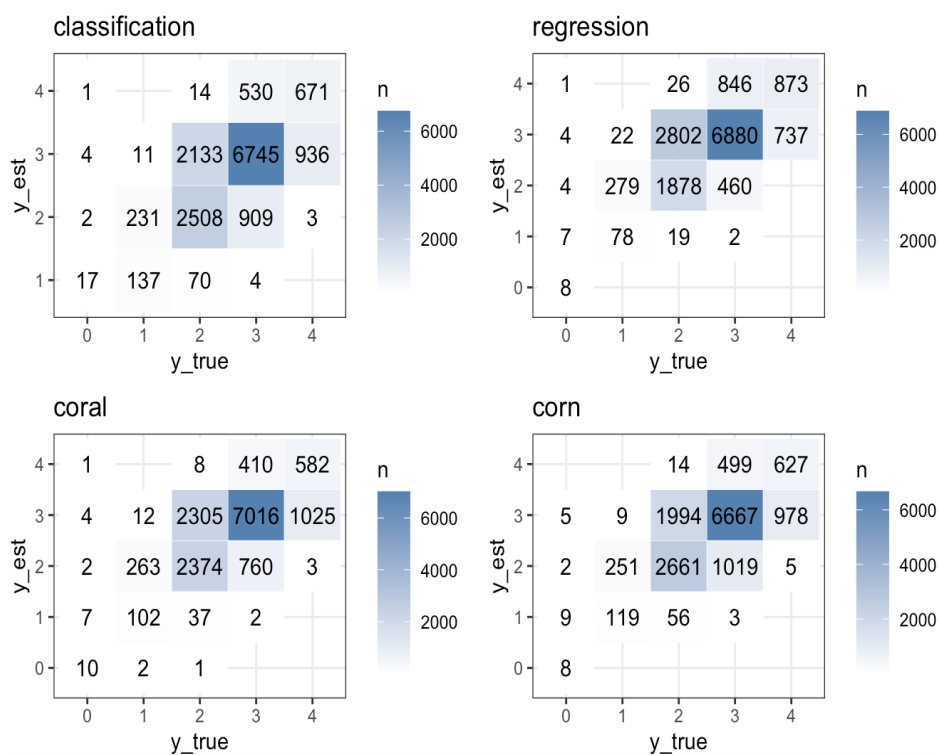


Figure 3: Heat plots of confusion matrix between predicted and criterion scores.

scoring models examined used only transcripts of spoken responses as input. Scores of speaking items usually come from three main dimensions, delivery, language use, and topic development. Using transcripts solely may underrepresent aspects of the delivery construct, such as pronunciation and prosody (Chen et al., 2018). Future research could address this limitation by incorporating temporal features alongside transcripts, or leveraging pre-trained speech representation models for scoring directly from audio recordings.

References

- Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.
- Lei Chen, Klaus Zechner, Su-Youn Yoon, Keelan Evanini, Xinhao Wang, Anastassia Loukina, Jidong Tao, Lawrence Davis, Chong Min Lee, Min Ma, Robert Mundkowsky, Chi Lu, Chee Wee Leong, and Binod Gyawali. 2018. Automated scoring of non-native speech using the speechratersm v. 5.0 engine. *ETS Research Report Series*, 2018(1):1–31.
- Shelby J. Haberman and Sandip Sinharay. 2010. The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics*, 35(5):586–602.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.
- Susan Lottridge, Chris Ormerod, and Amir Jafari. 2023. Psychometric considerations when using deep learning for automated scoring. In Victoria Yaneva and Matthias von Davier, editors, *Advancing Natural Language Processing in Educational Assessment*, pages 15–30. Routledge, New York.
- Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.
- Sebastian Raschka, Wenzhi Cao, and Vahid Mirjalili. 2020. coral-pytorch: Coral and corn implementations for ordinal regression in pytorch. <https://github.com/Raschka-research-group/coral-pytorch>. GitHub repository.
- Mijke Rhemtulla, Pénélope É. Brosseau-Liard, and Victoria Savalei. 2012. When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under characteristics of non-normality. *Psychological Methods*, 17(3):354–373.
- Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 & parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*.
- Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2023. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *arXiv preprint arXiv:2111.08851*. Version 5, June 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.
- John Whitmer and Maggie Beiting-Parrish. 2023. Results of NAEP math item automated scoring data challenge & comparison between reading & math challenges.
- John Whitmer, Evelyn Deng, Charles Blankenship, Maggie Beiting-Parrish, Ting Zhang, and Paul Bailey. 2023. Results of NAEP reading item automated scoring data challenge.
- 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, and 3 others. 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.