ILDAE: Instance-Level Difficulty Analysis of Evaluation Data

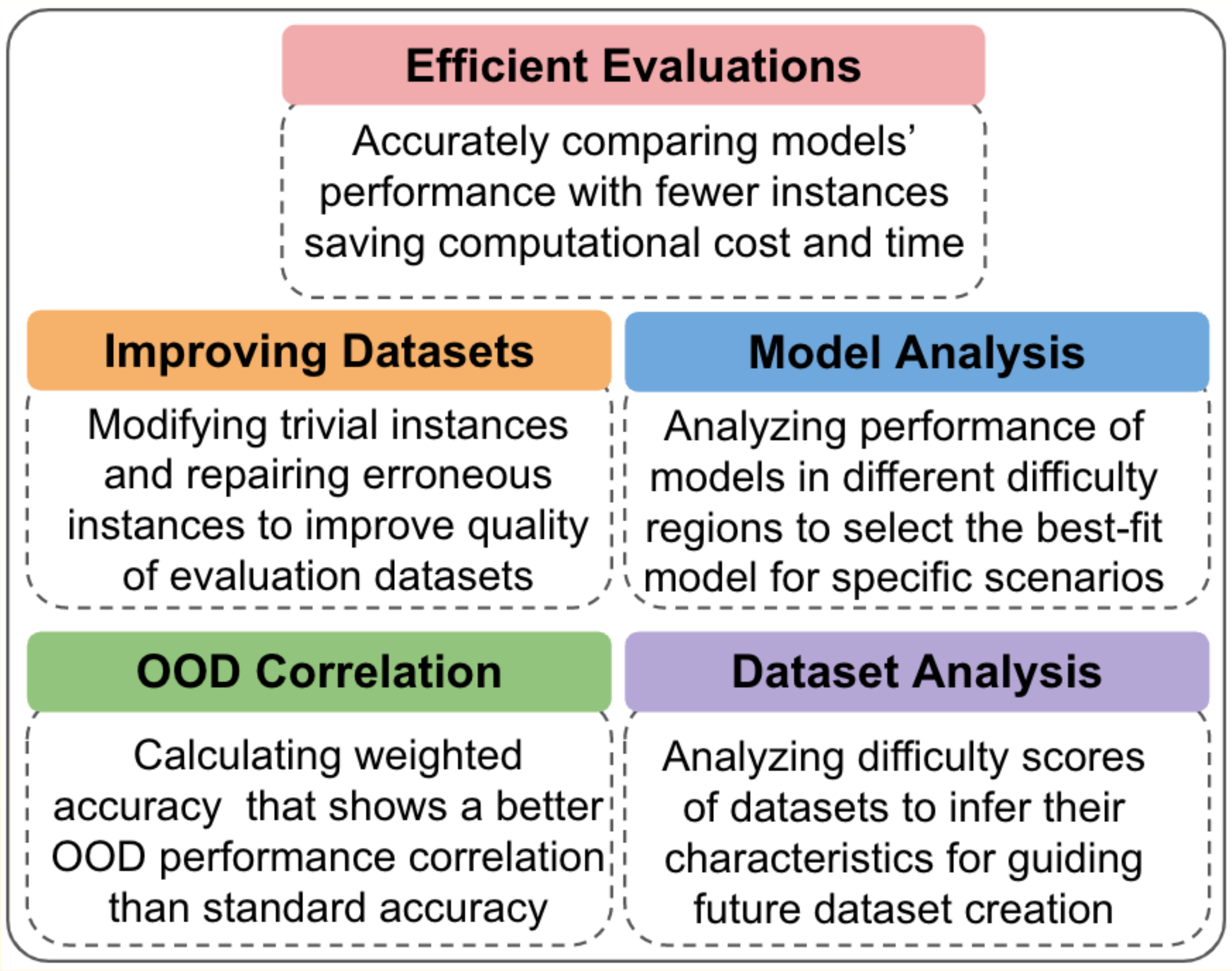
Neeraj Varshney, Swaroop Mishra, Chitta Baral
Arizona State University, USA

Computing Instance-Level Difficulty Scores

"Not All Instances are Equally Difficult"

- **Desiderata for Difficulty Scores:**
  - **Interpretation**: Human perception of difficulty may not always correlate well with machine's interpretation. Thus, difficulty scores must be computed via a model-in-the-loop technique so that they directly reflect machine's interpretation.
  - **Relationship with Predictive Correctness**: Difficulty scores must be negatively correlated with predictive correctness since a difficult instance is less likely to be predicted correctly than a relatively easier instance.
- We consider model's **prediction confidence** in the ground truth answer (indicated by softmax probability assigned to that answer) as the measure of its predictive correctness.
- We compile an ensemble of models trained with varying configurations and use their mean predictive correctness to compute difficulty scores.
- **Configurations**: Data Size, Data Corruption, and Training Steps.

Applications of ILDAE



Efficient Evaluations

Accurately comparing models' performance with fewer instances saving computational cost and time

Improving Datasets

Modifying trivial instances and repairing erroneous instances to improve quality of evaluation datasets

Model Analysis

Analyzing performance of models in different difficulty regions to select the best-fit model for specific scenarios

OOD Correlation

Calculating weighted accuracy that shows a better OOD performance correlation than standard accuracy

Dataset Analysis

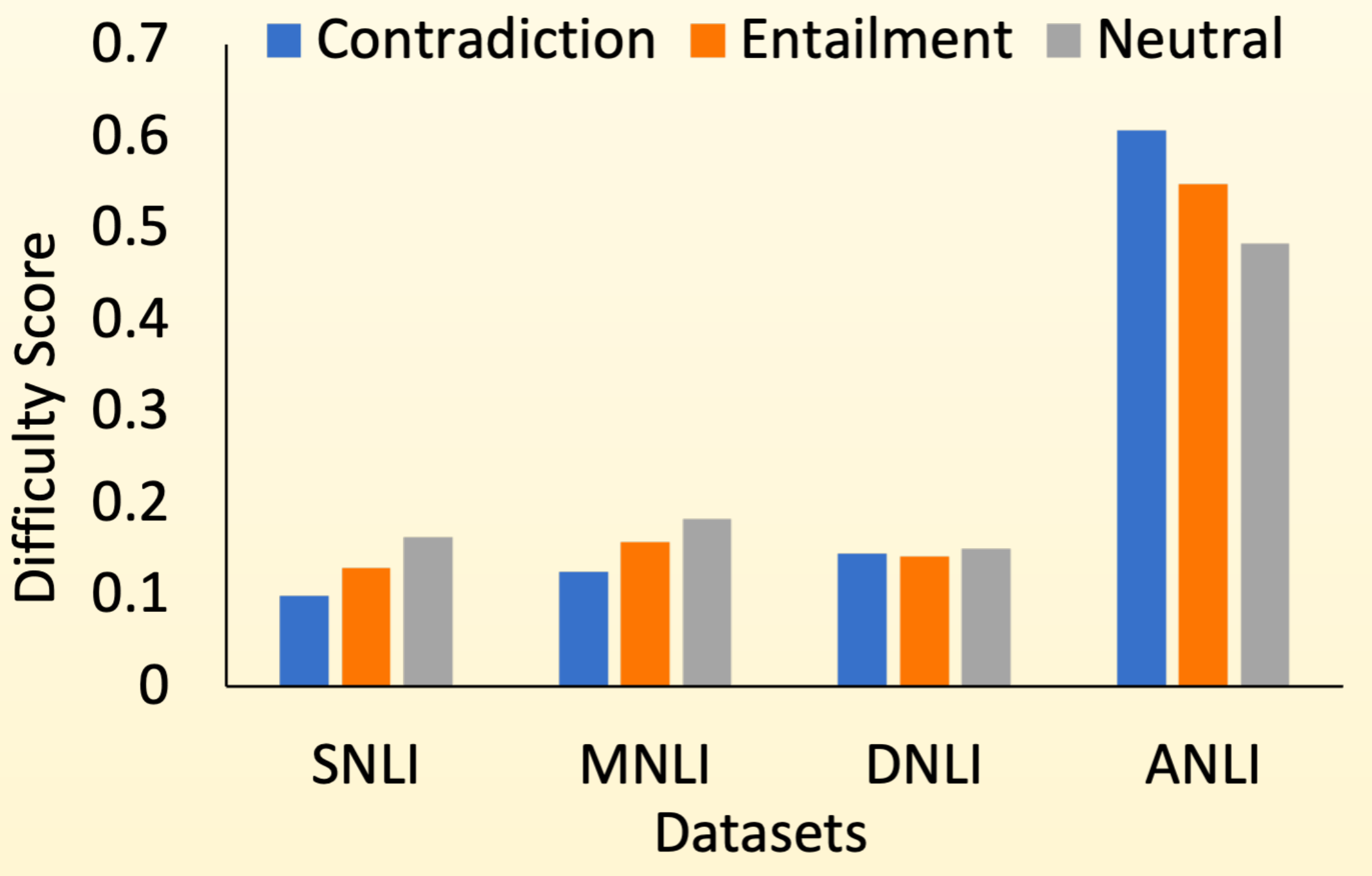
Analyzing difficulty scores of datasets to infer their characteristics for guiding future dataset creation

Efficient Evaluations

- Success of BERT has fostered development of several other pre-trained language models such as RoBERTa, DistilBERT, XLNET, ALBERT.
- Though, it has resulted in the availability of numerous model options for a task, comparing the performance of such a large number of models has become computationally expensive and time-consuming.
- *How can we make the evaluations **efficient**?*

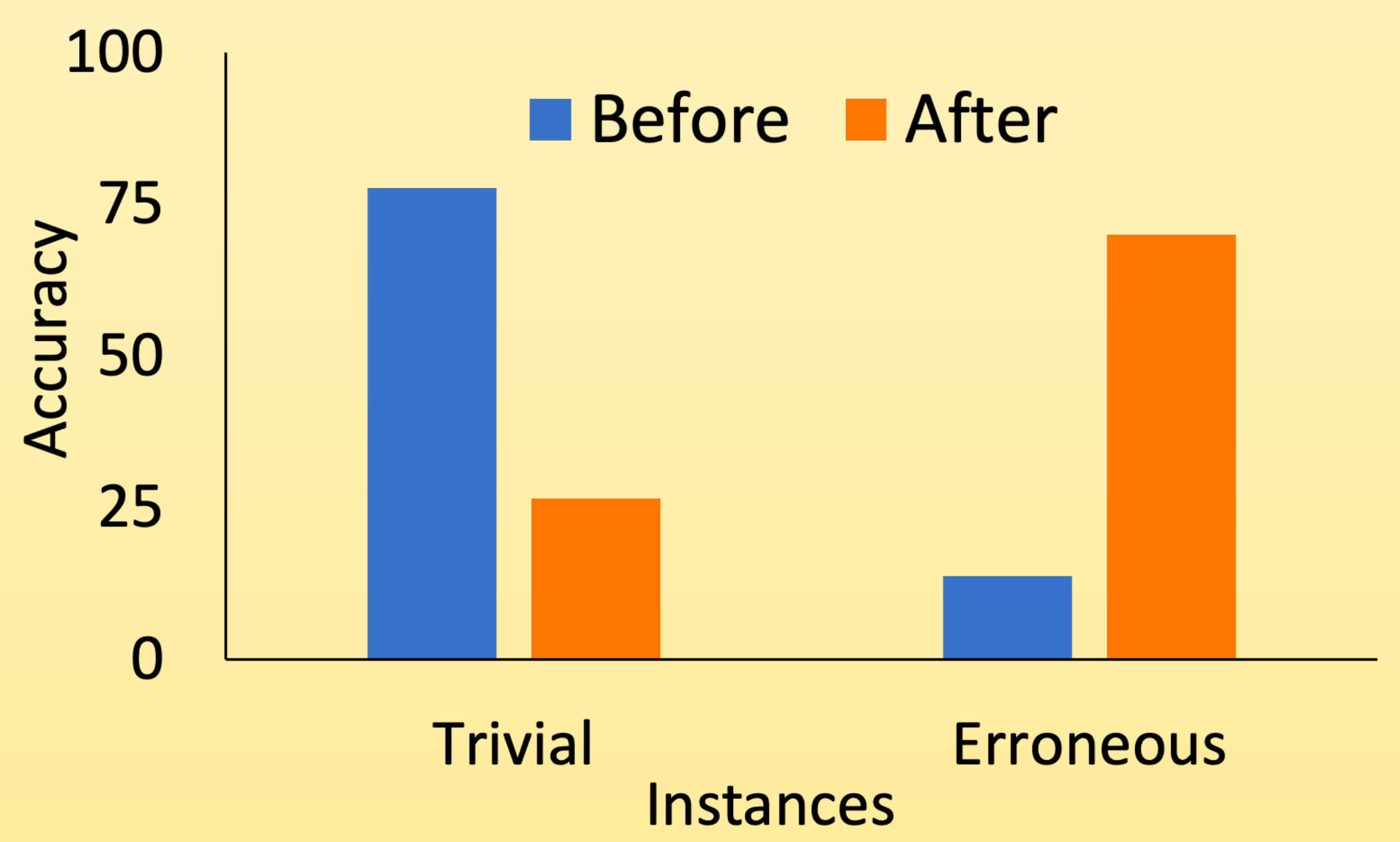
- We propose an instance selection technique that makes the selection based on the difficulty scores.
- We argue that instances with extreme difficulty scores (very low and very high scores) would not be effective in distinguishing between the candidate models.
- This is because the former instances are trivial and would be answered correctly by many/all candidate models, while the latter ones are hard and would be answered correctly by only a few/none models.
- Therefore, **we select a majority of instances for evaluation with moderate difficulty scores.**
- Our approach uses as little as 5% instances to achieve up to 0.93 Kendall correlation with evaluations conducted using the complete dataset.
- **Thus, without considerably impacting the effectiveness of evaluations, our approach saves computational cost and time.**

Dataset Analysis:



- For SNLI and MNLI datasets, **contradiction examples receive lower average difficulty score.**
- Therefore, while enhancing these datasets, more effort should be invested on contradiction examples as they are relatively easier.

Improving Evaluation Datasets



- We show that **trivial** and **erroneous** instances can be identified using our difficulty scores and present a model-and-human-in-the-loop technique to modify/repair such instances resulting in improved quality of the datasets.
- In case of SNLI dataset, **on modifying the trivial instances, accuracy drops** from 77.58% to 26.49%, and **on repairing the erroneous instances, it increases** from 13.65% to 69.9%. Thus, improving the dataset quality.

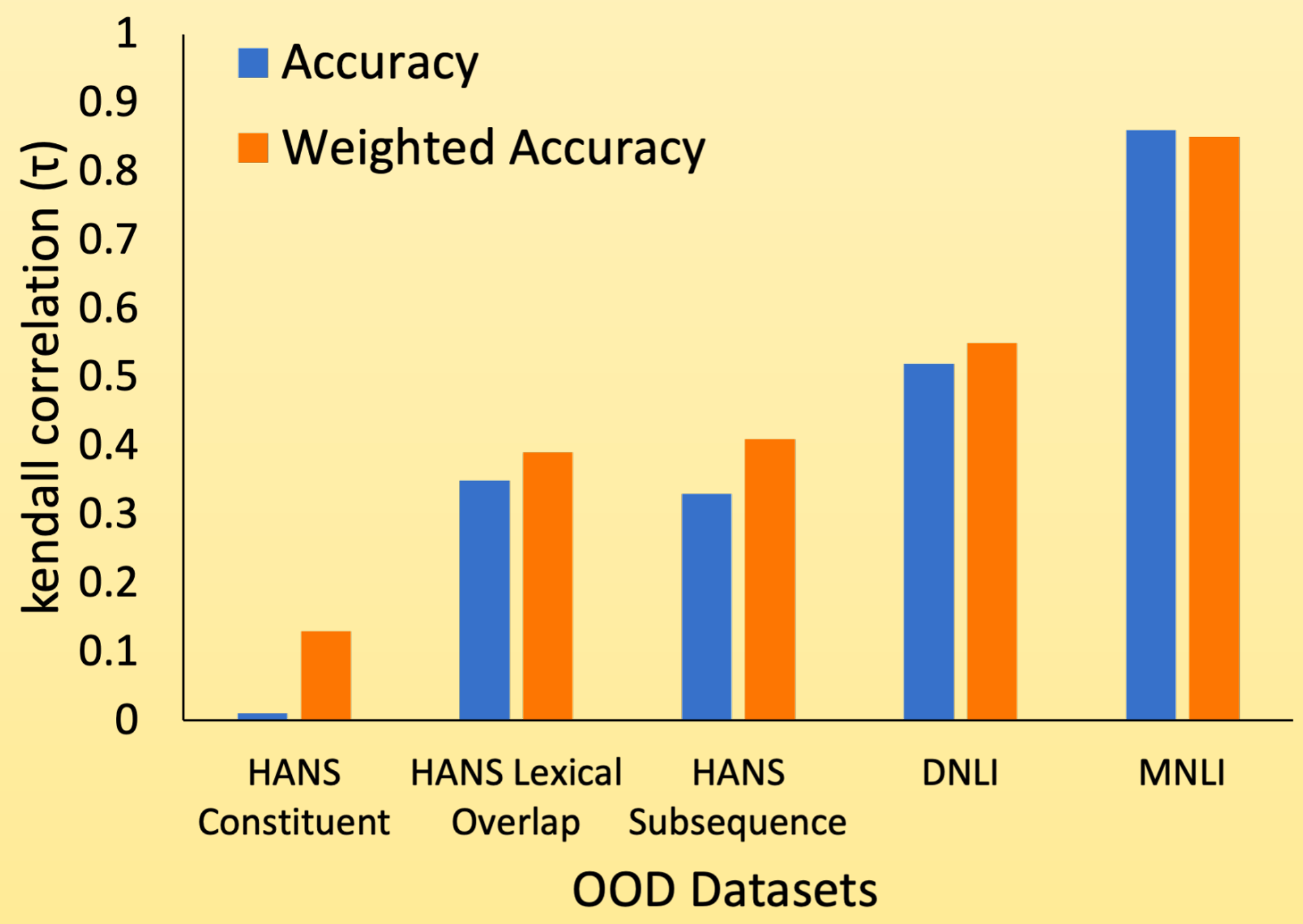
Model Analysis

- We divide instances into different regions based on difficulty scores and analyze models' performance in each region.
- **A single model does not achieve the highest accuracy in all difficulty regions.**
- The model that achieves the highest performance on easy instances may not necessarily achieve the highest performance on difficult instances.
- Such analyses could benefit in **model selection**. For instance, in scenarios where a system is expected to encounter hard instances, the model that performs well in high difficulty regions could be selected and for scenarios containing easy instances, the model that has the highest accuracy in easy regions can be selected.

Code and Resources

https://github.com/nrjvarshney/ILDAE

OOD Correlation



- We compute **weighted accuracy** leveraging the difficulty scores and show that it leads to 5.2% **higher Kendall correlation** with Out-of-Domain performance than the standard unweighted accuracy.
- Thus, ILDAE helps in getting a more reliable estimation of OOD performance.