Beyond Labels: Empowering Human with Natural Language Explanations through a Novel Active-Learning Architecture

Bingsheng Yao
Rensselaer Polytechnic Institute

Ishan Jindal
IBM Research

Lucian Popa
IBM Research

Yannis Katsis
IBM Research

Sayan Ghosh
The University of North Carolina at Chapel Hill

Lihong He
IBM Research

Yuxuan Lu
Northeastern University

Shashank Srivastava
The University of North Carolina at Chapel Hill

James Hendler
Rensselaer Polytechnic Institute

Dakuo Wang [*]
Northeastern University

Abstract

Data annotation is a costly task; thus, researchers have proposed low-scenario learning techniques like Active-Learning (AL) to support human annotators; Yet, existing AL works focus only on the **label**, but overlook the **natural language explanation** of a data point, despite that real-world humans (e.g., doctors) often need both the labels and the corresponding explanations at the same time. This work proposes a novel AL architecture to support and reduce human annotations of both labels and explanations in low-resource scenarios. Our AL architecture incorporates an explanation-generation model that can explicitly generate natural language explanations for the prediction model and for assisting humans' decision-making in real-world. For our AL framework, we design a data diversity-based AL data selection strategy that leverages the explanation annotations. The automated AL simulation evaluations demonstrate that our data selection strategy consistently outperforms traditional data diversity-based strategy; furthermore, human evaluation demonstrates that humans prefer our generated explanations to the SOTA explanation-generation system.

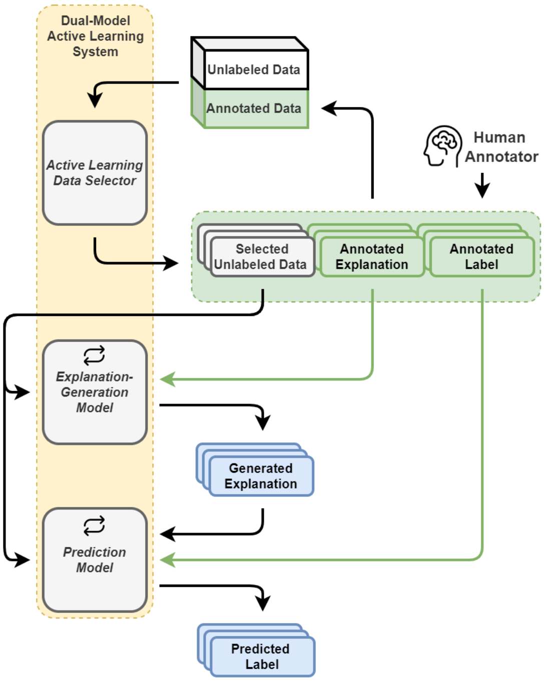


Figure 1: Our dual-model AL system architecture at every iteration: 1) the data selector selects few unlabeled examples; 2) humans provide explanations to finetune the explanation-generation model and gold labels to finetune the prediction model along with generated explanations. Green arrows denote the training target.

1 Introduction

Recent years of intense research in Natural Language Processing (NLP) has led to the emergence of State-of-the-Art (SoTA) language models (Devlin et al., 2019; Radford et al., 2019; Winata et al., 2021). These models demonstrate astonishing performance on various NLP tasks, including Question Answering (QA) and Question Generation (QG) (Rajpurkar et al., 2016; Duan et al., 2017; Kočiský et al., 2018; Yao et al., 2022), Natural Language Inference (NLI) (Bowman et al., 2015; Wang et al., 2018), et cetera. Despite the superior generative capabilities, the lack of faithful explainability within these 'black boxes' may lead to mistrust of their predictions (Lipton, 2018). Unlike end-to-end models, humans generally develop intermediate information as rationales to help decision-making, where the rationales are also faithful explanations.

[*] d.wang@northeastern.edu Corresponding Author.

The model's explainability becomes increasingly crucial because explanations are eagerly needed for humans to understand and trust model predictions. Therefore, many approaches retrospectively analyze the probability distribution within the model or ask models to generate explanations along with predictions (Ribeiro et al., 2016; Lundberg and Lee, 2017; Yu et al., 2019; Rajagopal et al., 2021; Chen et al., 2021). Meanwhile, a variety of datasets (Camburu et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021) with human-annotated explanations has been collected in the last few years.

As researchers look into the quality of explanation annotations (Carton et al., 2020; Yao et al., 2023), many issues were found in these datasets (Geva et al., 2019; Chmielewski and Kucker, 2020; Narang et al., 2020; Sun et al., 2022). In addition, large-scale, high-quality data annotations (Roit et al., 2020; Xu et al., 2022) are highly expensive in terms of labor, money, and time. Therefore, many works explore few-shot mechanisms like Active Learning (AL) (Settles, 2009) to minimize human annotations. AL is a human-in-the-loop framework that leverages a data selection strategy to iteratively select a small amount of examples from the candidate pool and query their annotation from humans for model training.

In this work, we propose a dual-model AL architecture for human annotations of labels and explanations inspired by human decision-making process. The dual-model system consists of:

1) an explanation-generation model guided by human-provided explanations, and

2) a prediction model that takes the data text and explanations generated by the explanation-generation model for prediction.

Our dual-model architecture incorporates AL to minimize the need for human annotations and establish human trustworthiness by actively engaging in the training process. We design a novel data diversity-based AL data selection strategy to exploit the explanation annotation. The strategy is based on data and explanation annotation similarity, which is analogous to the prevalent core-set (Sener and Savarese, 2017) strategy. At every iteration, we acquire human annotations of labels and free-form explanations for a few examples (e.g., 3 or 10) selected by the strategy.

Our system aims to support model predictions and human trustworthiness with explicitly generated natural language explanations in low-resource scenarios. As a result, the natural language explanations are explicitly generated and used as input for label prediction, signifying such explanations could faithfully explain the model's prediction.

We conduct two AL simulations with different amounts of data for each iteration. The simulations are performed on a large-scale NLI dataset with human-annotated explanations to justify the helpfulness of incorporating human-annotated explanations in AL data selection, compared with a random and traditional data diversity baseline. The AL simulation demonstrates that our AL data selector can consistently outperform both baselines at every iteration in both settings. In addition, we conduct a human study on the perceived validity, explainability, and preference of the generated explanations from our system, a SOTA explanation-generation system, and human-annotated explanations. The results show that despite ground-truth explanations being the best, our system significantly outperforms the baseline system in the generated explanation's validity and preference.

We conclude our paper by discussing limitations and future research directions.

2 Related Work

2.1 Datasets with Natural Language Explanations

Wiegreffe and Marasovic (2021) conducted a comprehensive review of 65 datasets with explanations and provided a 3-class taxonomy, namely highlights, free-text, and structured. Among the large-scale datasets with free-text explanations, **e-SNLI** (Camburu et al., 2018) is a prominent one, which extended the Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015). SNLI is a classification task to determine the inference relation between two textual contexts (premise and hypothesis): entailment, contradiction, or neutral. The e-SNLI dataset contains human-annotated free-form explanations for $549,367$ examples in train, $9,842$ in validation and $9,824$ in test split. We show some examples of e-SNLI in Appendix A.

Another popular group of datasets extended the Commonsense Question-Answering (CQA v1.0 and v1.11 versions) dataset (Talmor et al., 2019), including two variants of Cos-E: **CoS-E v1.0** and **CoS-E v1.11** (Rajani et al., 2019). However, many recent works (Narang et al., 2020; Sun et al., 2022) have found explanations in CoS-E seem to be noisy and low-quality, and thus, Aggarwal et al. (2021)

created **ECQA** by carefully designing explanation annotation protocols to create a higher quality dataset compared with CoS-E.

In this paper, we leverage the e-SNLI dataset as the benchmark dataset for our AL simulation experiment because 1) the classification task is popular and representative in NLP, 2) the massive data size ensures the diversity of data, and 3) explanations for classification task may provide more effective help compared to CQA task where training and testing data may be unrelated.

2.2 Active Learning for Data Annotation

Owning to the paucity of high-quality, large-scale benchmarks for a long tail of NLP tasks, learning better methods for low-resource learning is gaining more attention in the community. Active Learning (Sharma et al., 2015; Shen et al., 2017; Ash et al., 2019; Teso and Kersting, 2019; Kasai et al., 2019; Zhang et al., 2022) is one such popular framework for low-resource learning. AL iteratively (1) selects samples from the unlabeled data pool (based on AL data selection strategy) and queries their annotation from human annotators and (2) fine-tunes the underlying model with newly annotated data.

In addition to AL, Marasovic et al. (2022) introduces a few-shot self-rationalization setting that asks a model to generate free-form explanation and the label simultaneously. Similarly, Bhat et al. (2021) proposes a multi-task self-teaching framework with only 100 train data per category. Bragg et al. (2021) provides guidance on unifying evaluation for few-shot settings.

A few AL surveys (Settles, 2009; Olsson, 2009; Fu et al., 2013; Schröder and Niekler, 2020; Ren et al., 2021) of data selection strategies provide two high-level selection concepts: data diversity and model probability. This paper focuses on the **data diversity strategies** and leverages human-annotated explanations to select data. Our data selector shares a similar concept with the established data-based clustering strategies (Xu et al., 2003; Nguyen and Smeulders, 2004) and core-set (Sener and Savarese, 2017) that aim to select the most representative data while maximizing diversity.

2.3 Natural Language Explanation Generation

Many recent works have explored different approaches to enhance the model's explainability by asking them to generate natural language explanations. Some of them (Talmor et al., 2020; Tafjord et al., 2021; Latcinnik and Berant, 2020) propose systems to generate text explanations for specific tasks. Dalvi et al. (2022) propose a 3-fold reasoning system that generates a reasoning chain and asks users for correction. Other recent works (Paranjape et al., 2021; Liu et al., 2022; Chen et al., 2022) explore different prompt-based approaches to generate additional information for the task and examine the robustness and validity. We believe our dual-model system provides explanations and uses them explicitly towards prediction while the self-rationalization setting fails short. Hase and Bansal (2022) argues that explanations are most suitable as input for predicting, and Kumar and Talukdar (2020) designed a system to generate label-wise explanations, which is aligned with our design hypothesis. Nevertheless, there exist other work (Wiegreffe et al., 2021; Marasovic et al., 2022; Zelikman et al., 2022) exploring the use of self-rationalization setting. We include the self-rationalization setting in our human evaluation on the explanation quality in section 4.5.

3 Dual-Model AL System

3.1 System Architecture

Figure 1 depicts our proposed dual-model AL framework. The proposed system consists of three primary modules: 1) **an explanation-generation model** that takes the data, fine-tunes, and generates free-form explanations; 2) **a prediction model** that takes the task content and the generated free-form explanations as input, fine-tune, and predict final label; 3) **an AL data selector** that selects a set of representative examples in terms of the semantic similarity between each unlabeled data text and labeled data's human explanations.

Since our dual-model AL framework is fine-tuned on a small amount of data, the AL data selector serves a critical role in providing the most representative and helpful set of examples at every iteration. Therefore, we design a data selector that utilizes human explanation annotations. Details of our AL data selector is in Section 3.2.

In each AL iteration, the data selector selects a representative set of unlabeled examples for human annotations, and then we fine-tune two models in order. More specifically, we first fine-tune an explanation-generation model using the data content and supervised by human-provided free-form explanations, ask this model to generate explanations on the same set of data, then fine-tune a pre-

Explanation-generation Model:	
Training Input	**explain:** what is the relationship between *[hypothesis]* and *[premise]* **choice1:** entailment **choice2:** neutral **choice3:** contradiction
Training Target	*[human annotated explanations]*
Model Generation	*[generated free-form explanation]*
Prediction Model:	
Training Input	**question:** what is the relationship between *[hypothesis]* and *[premise]* **choice1:** entailment **choice2:** neutral **choice3:** contradiction **<sep> because** *[generated free-form explanation]*
Training Target	*[human annotated label]*
Model Prediction	*[predicted category]*

Table 1: The prompt-based input templates for both models in our system, with the e-SNLI (Camburu et al., 2018) dataset as an example.

diction model with the data content and explanations generated by the previous model as input, and supervised by the human-annotated labels. The reason for fine-tuning the prediction model with generated explanations instead of human explanations is that there will be no human-annotated explanations during actual inference. The prediction model fine-tuned on model-generated explanations can better learn how to use model-generated explanations than the one fine-tuned on human explanations. After each AL iteration, we evaluate the framework on a standalone evaluation data split.

Both the explanation-generation model and the prediction model could be any SoTA sequence-to-sequence models, such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). We leverage T5 as the backbone for both models in this work and design a prompt-based input template for both models, as shown in Table 1. To explain how each prompt addresses a different part of data content:

1) '*explain:*' and '*predict:*' are the leading prompts in the explanation-generation model and the prediction model, indicating different tasks for both models and are followed by the original task content. For the e-SNLI dataset, the task content becomes "what is the relationship between" the hypothesis and premise sentences;

2) '*choiceN*' is followed by candidate answers, where $N \in [1, 3]$ for the e-SNLI dataset corresponds to entailment, neutral, and contradiction. We pass choices to the explanation-generation model expecting it will learn to generate free-text explanations that may reflect potential relationships between the data content and the task;

3) for the prediction model, an additional prompt '*because:*' is followed by the explanations generated by the explanation-generation model, and we use a special token to separate the original task content and the explanation.

3.2 AL Data Selector

Algorithm 1 AL Data Selector

```
Variables:
  $D_{train} \Rightarrow$ unlabeled data in train split
  $D_{prev} \Rightarrow$ previously-annotated data
  $d_p^{data} \Rightarrow$ data content as a string of $d_p$ (for e-SNLI, it is the premise and hypothesis
  $d_p^{exp} \Rightarrow$ previously-annotated free-form explanation of $d_p$
  $x \Rightarrow$ number of data to be selected each iteration
  $n_{train} = len(D_{train});\ n_{prev} = len(D_{prev})$
  for $D_i \in D_{train}$ do
    if $iteration == 0$ then
      $score_{d_i} = \frac{1}{n_{train}} \cdot \sum_{d_p \in D_{train}} similarity(d_i^{data}, d_p^{data})$
    else
      $score_{d_i} = \frac{1}{n_{prev}} \cdot \sum_{d_p \in D_{prev}} similarity(d_i^{a}ta, d_p^{exp})$
    end if
  end for
  $D_{train}^{'} = rank\ D_{train}\ by\ score$
  $D_{selected} =\ select\ x\ data\ from\ D_{train}^{'} with\ equal\ intervals$
  Human annotation on $D_{selected}$
  $D_{train} - = D_{selected}; D_{prev} + = D_{selected}$
```

According to a few AL surveys (Settles, 2009; Olsson, 2009; Fu et al., 2013; Schröder and Niekler, 2020; Ren et al., 2021), there are two primary concepts of AL data selection: data diversity and model probability. Data diversity approaches take advantage of various data features, such as data distribution and similarity, so as to select a representative set of examples from the candidate pool while maximizing diversity, such as core-set.

Model probability-based approaches, on the other hand, aim to select the examples that the models are least confident about. The model probability approaches require conducting inference on the unlabeled data at every iteration for data selection, thus taking more time and computing resources. In addition, the helpfulness of model probability-based approaches is generally model-specific, while the data diversity-based approaches are model agnostic.

This paper proposeS a **data diversity-based** AL selection strategy that shares similar concept with traditional data-based clustering strategy Nguyen and Smeulders (2004) and core-set strategy. Our strategy differs from traditional strategies because ours leverages human-annotated explanations for selection. More specifically, our data selector strategy aims to select examples that are representative of the unlabeled data pool in terms of average similarity to human-annotated explanations of all previously-labeled data while maximizing the diversity of newly-selected data.

We assume **the human-annotated explanations have an important contribution to the model's prediction and convey more information than the original data content alone**, such that the explanation could be relations between the data content and the choices. For example, in e-SNLI dataset, the data content is the concatenation of the hypothesis and premise sentences. Later, we construct the baseline selector in AL simulation experiment (Sec. 4.2) with the same setup, except that it only compares the similarity between data content, and we demonstrate that using human-annotated explanations for data selection leads to better prediction performance than using data content alone.

Here we explain our data-based AL data selector in detail (shown in Algorithm 1). For each data in the unlabeled data pool, we use sentence-transformers (Reimers and Gurevych, 2019) to calculate the semantic similarity between its data content and every previously-annotated explanation. Then, we take the averaged similarity scores for each unlabeled example and rank all the unlabeled data in terms of the averaged similarity score. Finally, to select the most representative data in the candidate pool while maximizing diversity, we select examples from ranked data list with equal intervals. Worth pointing out that there is no previously-annotated explanation at the first iteration, so we compare the similarity between the data content with each other instead.

4 Evaluation

We leverage the e-SNLI (Camburu et al., 2018) dataset as the benchmark dataset for our AL simulation experiment. The primary goal of this experiment is to justify **incorporating human annotated explanations in AL data selection can select more representative and helpful data** among a reasonably large-scale dataset with our proposed dual-model framework.

However, because e-SNLI has 549, 367 examples in the train split, it is unrealistic to use all training data in our simulation because it will take a huge amount of time and computing resources for the data selection at every iteration.

We would like to find a reasonable number of unlabeled candidate data for the AL simulation with a preliminary experiment on our dual-model framework. The size of the unlabeled data pool should be big enough to avoid biases by having certain overwhelming data features in the data, and we expect to reduce data without significantly degrading the performance compared to the model fine-tuned on the full dataset. For all the experiments, we leverage the pre-trained T5-base (Raffel et al., 2020) as the backbone model. All the experiment settings and hyperparameters are shown in Appendix C.

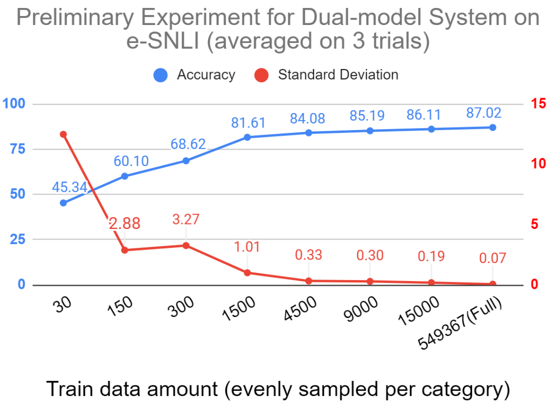


Figure 2: Preliminary experiment result of our dual-model system on e-SNLI (Camburu et al., 2018) dataset.

4.1 Preliminary Experiment

As aforementioned, the goal of the preliminary experiment is to 1) find out the upper bound of and how the performance of our dual-model system drops as we gradually reduce the amount of training data, and 2) to find a reasonable amount of training data pool for the AL simulation.

In order to minimize the potential bias caused by the uneven number of data for each category in the training data, we randomly sample the same amount of data for each category from the full e-SNLI training split and choose eight different settings ranging from [10, 50, 100, 500, 1500, 3000, 5000] and the complete data per category. Since the e-SNLI dataset consists of three categories: entailment, neutral, and contradiction, the training data size in each setting becomes [30, 150, 300, 1500, 4500, 9000, 15000, and 549, 367 (full train split size)] correspondingly, as denoted in Figure 2.

In the preliminary experiment, we only fine-tune the explanation generation model and the prediction model once with the training data; thus, we conduct a hyper-parameter search for each setting and experiment three trials with randomly sampled data on each setting to get an averaged result. The evaluation is conducted on the full testing split in e-SNLI, which contains 9, 824 examples.

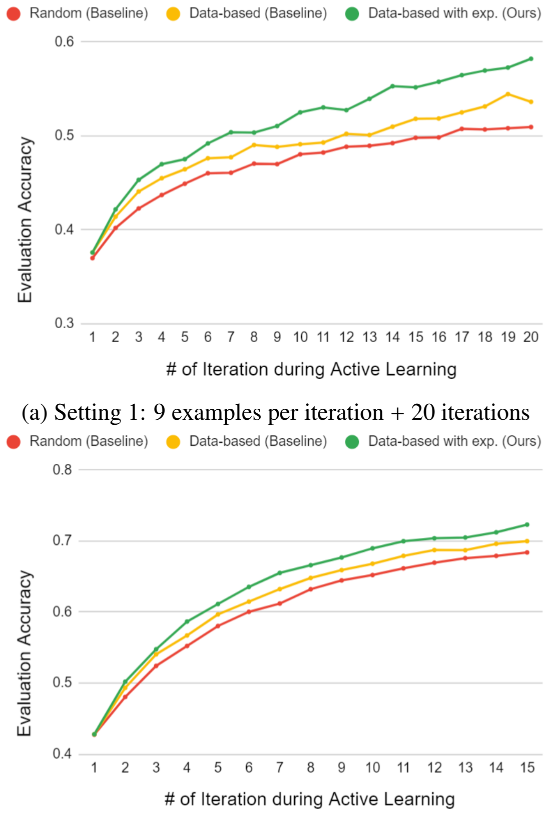
The results are shown in Figure 2, where the blue dots denote the averaged prediction accuracy at each setting, and the red dots indicate the standard deviation of accuracy among three trials. Both units are percentages. We observe that with more than $1,500$ data per category, the performance drop compared to using full train split is inconspicuous (84.08% to 87.02%), while the standard deviation is below 0.5%. This observation indicates that using 1% of the original training data size only leads to a performance drop of merely 3%. In addition, we observe that with only 10 data per category (30 data in total), our system can still achieve an average of 45% accuracy, though the deviation is reasonably big; Furthermore, when we extend the training data size from 100 to 500 data per category, which is a reasonable amount of data that could be applicable in real-world scenarios, the accuracy can reach over 80% accuracy, which is very promising considering the amount of training data is much smaller than the amount of data for evaluation.

4.2 Simulation Experiment: Evaluation Setup

Based on the findings from the preliminary experiment, we decide to use $3,000$ examples per category ($9,000$ in total) as the candidate unlabeled train data pool for the Active Learning simulation.

Inspired by the guidance on unifying evaluation for few-shot settings provided by Bragg et al. (2021), we conduct **80** trials for each AL setting, then calculate the averaged performance for ours and the baseline data selectors at every iteration. During each trial of the experiment, we first randomly sample $3,000$ examples per category from the complete train dataset, then use the same data to conduct AL simulation on different data selectors with our dual-model system. This way, we can ensure the performance differences during each trial are not due to different unlabeled data pools but to differences in the performance of AL data selectors. For the evaluation, we randomly shuffle 300 examples per category (900 in total) from the test split of e-SNLI every trial and evaluate with the same test data after each iteration.

The AL simulation consists of two settings where we simulate annotating a total of 180 and 450 data correspondingly. These two amounts of data annotation requirements can mimic real-world scenarios reasonably that users have a limited budget, annotator, and data for annotation. More specifically, we experiment with two following settings:



(a) Setting 1: 9 examples per iteration + 20 iterations

(b) Setting 2: 30 examples per iteration + 15 iterations

Figure 3: Results of AL Simulation experiment on our Dual-model system with different data selectors.

1) For every iteration, select **3** examples per category (9 in total) with **20** iterations, which results in **180** examples altogether;

2) For every iteration, select **10** examples per category (30 in total) with **15** iterations, which results in **450** examples altogether.

We fix the same set of hyper-parameters across each setting (Appendix C). As mentioned before, all the experiments leverage a pre-trained T5-base model as the backbone and the prompts-based input structure we designed in Section 3.1.

Our AL simulation experiment compares the performance of our data selector with two baselines. Our data selector is a data diversity-based algorithm that leverages human-annotated explanations. The details has been explained in Section 3.2. In comparison, we have a random data selector as the basic benchmark and another traditional data diversity-based algorithm that shares the same procedures except that it only compares the similarity between each unlabeled data's content and the previously-labeled data's content, not using the human-annotated explanations.

Worth mentioning that our data selector does not use task content in previously-labeled examples;

instead, we only use the human-annotated explanations to justify that the explanations are more helpful than task content. At the first iteration, ours and the data diversity baseline perform the same way because no previously-annotated data exists.

4.3 Simulation Experiment: Result

The evaluation results are shown in Figure 2. To explain the diagrams in detail, each dot in the diagram is the average accuracy on 80 trials at every iteration for each data selector. The green/yellow/red line denotes our data selector/data diversity-based baseline/random selector.

We can easily observe that our data selector can maintain a consistent advantage in prediction performance over the traditional feature-based selector baseline, while the traditional one can consistently beat the random baseline by a significant margin. To summarize the observation, our data selector outperforms both baselines in every iteration for both AL settings, indicating that **using human-annotated explanations in data selector with our dual-model AL framework is more beneficial than using data content alone.** Even with only 180 and 450 data to be annotated in each setting, our system is able to achieve 55% and 72% accuracy on average, correspondingly. We anticipate that our experiment will reach a similar performance around 85% as shown in Figure 2 but converge much faster than the random selector if we continue the active learning process.

4.4 Ablation Study: Transfer to Multi-NLI

In addition to the AL simulation experiment on e-SNLI dataset, we further conduct an exploratory transfer learning AL simulation, as ablation study on another dataset for the same NLI task, Multi-NLI (Williams et al., 2018).

The ablation study consists of the following steps: 1) fine-tune an explanation-generation model using AL with our framework on e-SNLI dataset; 2) use the explanation-generation model in AL simulation on Multi-NLI dataset by freezing the model to generate free-text explanations on selected Multi-NLI examples; 3) fine-tune the prediction model for Multi-NLI task at every iteration. Unlike e-SNLI AL experiment, our AL data selection algorithm will use model-generated explanations to select examples at every iteration in the transfer learning AL simulation. We fine-tune the explanation-generation models on e-SNLI with both settings in the previous experiment, average the result on 15 trials of experiments, and keep consistent with every other experiment hyper-parameters.

The ablation results are shown in Figure 4 of Appendix B. The blue/red lines denote the explanation-generation model is fine-tuned on e-SNLI with setting 1/2 in Section 4.2 correspondingly. We observe the explanation-generation model can indeed provide helpful explanations to consistently improve the system's prediction performance and reach up to more than 65% accuracy. In comparison, we trained a single prediction model with the full Multi-NLI dataset ($392,702$ examples), achieving 86% accuracy. In addition, the explanation-generation model fine-tuned on more data can indeed perform better, hypothesizing it learned to generate more helpful explanations.

4.5 Human Evaluation Setup and Results

To qualitatively evaluate the explainability of the generated explanations from our system against a SoTA few-shot explanation-generation system, the self-rationalization baseline (Marasovic et al., 2022), and the human ground-truth, we recruited three human participants to conduct a human evaluation. The self-rationalization baseline is a T5-base model, which uses the same input template of our explanation-generation model shown in Table 1 but asks the model to generate both the label and explanation simultaneously.

We leverage AL setting 1 described in Section 4.2 to fine-tune our system with a total of 180 examples over 20 epochs and use the same 180 examples to fine-tune the self-rationalization baseline with the same set of hyper-parameters as for our prediction model. Both systems are used to infer the complete test split of e-SNLI dataset after fine-tuning; then, we randomly sample 80 examples for the human study.

For each data instance, the rater is presented with the textual content of the *premise* and *hypothesis* of the original data paired with three sets of *labels* and *explanations* from our system, baseline system, and the human-annotated ground-truth from the e-SNLI dataset. Participants who are not aware of the source of each label-explanation pair are asked to answer four questions with [Yes/No]:

1) Is the Prediction correct?

2) Is the Explanation itself a correct statement?

3) Regardless of whether the AI Prediction and Explanation is correct or not, can the Explanation help you to understand why AI has such Prediction?

Yes / No Count	Label	Exp.	Exp. → Label	Trustworthy AI
Ground-truth	83 / 7	86 / 4	87 / 3	78 / 12
Dual-model (ours)	64 / 26	68 / 22	48 / 42	35 / 55
Self-rationalization	42 / 48	67 / 23	51 / 39	21 / 69

Table 2: Human evaluation results.

4) Will you trust & use this AI in real-world decision-making?

To ensure inter-coder consistency, we first conduct a 30-min tutorial session to educate all three participants with 10 examples to build a consensus among them. In the actual experiment, each of the three participants is then asked to rate 30 data instances (20 unique ones and 10 shared ones), which make up a total of 70 data instances, and 360 ratings (3 rater*30 instances*4 questions). We first calculated the Inter-Rater Reliability score (IRR) among them for each of the four questions. With the IRR score of (Q1: 1, Q2: 0.89, Q3: 0.98, Q4: 0.87), we are confident that the three coders have the same criteria for further result analysis.

Our questions all have binary responses, and we rely on Chi-square analysis (Elliott and Woodward, 2007) to examine the statistical significance of the rating groups' differences. As shown in Table 2, human groundtruth explanations are rated highest across all four dimensions by the participants. Between our system and the few-shot self-rationalization system (baseline), participants believe our systems' predicted labels are more likely to be correct, 64 'valid' out of 90 ratings of our system versus 42 out of 90 ratings of baseline system. Chi-square test indicates such a difference is statistically significant ($\chi^2(1) = 21.61, p < 0.01$).

When asked whether they would trust the AI if they were faced with such AI system to support their real-world decision-making, only 35 out of 90 answered 'Yes' to our system, but it is still significantly better than the baseline system (21 'Yes' out of 90) ($\chi^2(1) = 12.17, p < 0.01$). In comparison, 78 out of 90 times people voted they would trust the explanations of human ground-truth quality.

As for Question 2 ("the validity of the generated explanation") and Question 3 ("whether the generated explanation is supporting its prediction"), the human evaluation fails to suggest statistical meaning results between our system and the baseline system ($\chi^2(1) = 0.06, p = 0.89$ for explanation validity, and $\chi^2(1) = 0.41, p = 0.52$ for explanation supporting prediction). In summary, human participants believe our system can outperform the baseline system on the label prediction's quality and the trustworthiness of AI dimensions, but there is a large space to improve as human evaluators believes the ground-truth label and explanation quality much better than either AI systems.

5 Limitations

In this paper, we demonstrate the effectiveness of our framework on a representative large-scale classification dataset (e-SNLI), but there are many other NLP tasks, such as question answering and commonsense reasoning, the generalizability of our system on other NLP tasks remains unexplored. Another limitation is that this work proposed a data diversity-based AL selector design and we only benchmarked it with a traditional data diversity-based selector to demonstrate the usefulness of explanations. Prior literature have proposed other model probability-based designs for AL data selectors, which have not been evaluated in this paper.

6 Conclusion and Future Work

In summary, this paper proposes a novel dual-model AL system to support the common real-world use case that domain experts need to both annotate classification labels and provide explanations. Our system consists of a specifically designed data diversity-based AL example selector and two sequence-to-sequence language models, one for explanation generation and the other for label prediction. With an AL simulation evaluation and a human evaluation of the e-SNLI dataset, the results suggest that our dual-model system outperforms both baseline data selectors.

Our work is the first step towards a human-centered interactive AI solution (it can be easily implemented as an interactive system as shown in Fig 5 in Appendix D) that helps domain experts to annotate their text data with labels and explanations. No matter how well the SOTA AI models perform, many real-world tasks still require domain experts to review and annotate each single data instance and leave their signature for accountability purposes (e.g., a lawyer reviews and signs off a legal document). We call for researchers to join us to move forward with this line of research to support this ubiquitous real-world task.

Acknowledgements

This work was supported by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), which is part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*.

Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee. 2021. Self-training with few-shot rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10702–10712, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. Flex: Unifying evaluation for few-shot nlp. *Advances in Neural Information Processing Systems*, 34:15787–15800.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9294–9307, Online. Association for Computational Linguistics.

Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. 2021. Generate natural language explanations for recommendation. *arXiv preprint arXiv:2101.03392*.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Michael Chmielewski and Sarah C Kucker. 2020. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems. *arXiv preprint arXiv:2204.13074*.

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.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Alan C Elliott and Wayne A Woodward. 2007. *Statistical analysis quick reference guidebook: With SPSS examples*. Sage.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems*, 35:249–283.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39, Dublin, Ireland. Association for Computational Linguistics.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 8730–8742, Online. Association for Computational Linguistics.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79.

Fredrik Olsson. 2009. A literature survey of active machine learning in the context of natural language processing.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. SELFEXPLAIN: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. *arXiv preprint arXiv:2008.07267*.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Burr Settles. 2009. Active learning literature survey.

Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active learning with rationales for text classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 441–451, Denver, Colorado. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 252–256, Vancouver, Canada. Association for Computational Linguistics.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. *arXiv preprint arXiv:2206.11083*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.

Stefano Teso and Kristian Kersting. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Zhao Xu, Kai Yu, Volker Tresp, Xiaowei Xu, and Jizhi Wang. 2003. Representative sampling for text classification using support vector machines. In *European conference on information retrieval*, pages 393–407. Springer.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023. Are human explanations always helpful? towards objective evaluation of human natural language explanations. *arXiv preprint arXiv:2305.03117*.

Bingsheng Yao, Dakuo Wang, Tongshuang Wu, Zheng Zhang, Toby Li, Mo Yu, and Ying Xu. 2022. It is AI's turn to ask humans a question: Question-answer pair generation for children's story books. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 731–744, Dublin, Ireland. Association for Computational Linguistics.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. 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)*, pages 4094–4103, Hong Kong, China. Association for Computational Linguistics.

Eric Zelikman, Jesse Mu, Noah D Goodman, and Yuhuai Tony Wu. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning.

Shujian Zhang, Chengyue Gong, Xingchao Liu, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. 2022. ALLSH: Active learning guided by local sensitivity and hardness. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1328–1342, Seattle, United States. Association for Computational Linguistics.

Appendix

A e-SNLI Examples

Table 3 illustrates an example data of each category in the e-SNLI dataset. Every data instance contains a premise and hypothesis along with a human annotated label and free-form explanation.

Premise: *This church choir sings to the masses as they sing joyous songs from the book at a church.* **Hypothesis:** *The church is filled with song.* **Label:** *entailment* **Human-annotated explanation:** *“Filled with song” is a rephrasing of the "choir sings to the masses.*
Premise: *A man playing an electric guitar on stage.* **Hypothesis:** *A man is performing for cash.* **Label:** *neutral* **Human-annotated explanation:** *It is unknown if the man is performing for cash.*
Premise: *A couple walk hand in hand down a street.* **Hypothesis:** *A couple is sitting on a bench.* **Label:** *contradiction* **Human-annotated explanation:** *The couple cannot be walking and sitting a the same time.*

Table 3: Sample data of each category in e-SNLI (Camburu et al., 2018) dataset.

B Transfer Learning Ablation Study Diagrams

Figure 4 shows the results of our Ablation Study results described in Section 4.4. The explanation-generation model is fine-tuned from AL on e-SNLI dataset with two different AL settings, then we freeze the explanation-generation model to train the prediction model in AL simulation for Multi-NLI dataset under two settings. Setting 1/2 refers to the settings for Active Learning Simulation in Section 4.2.

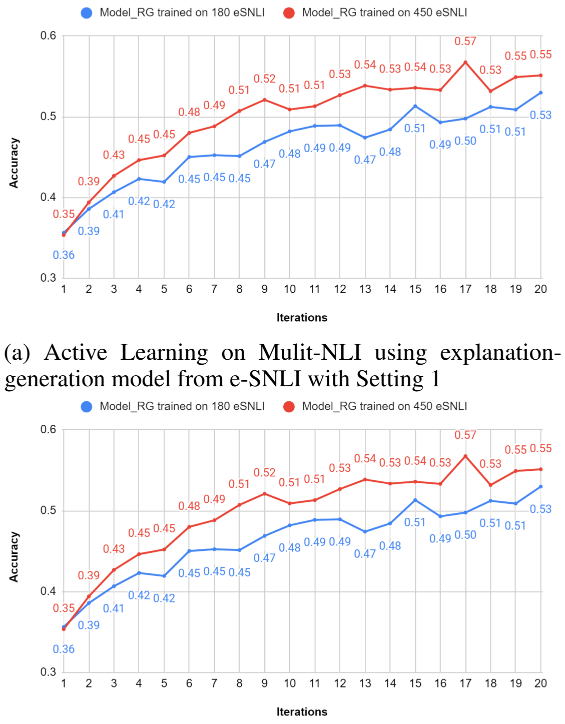
C System Environment and Hyper-Parameters

The computing resource of all the experiments we conducted in this paper has 128 Gigabytes of RAM. In addition, we use 2 NVIDIA Tesla V100 GPU for the preliminary experiment and 8 NVIDIA Tesla V100 GPU for the AL simulation experiment.

C.1 Preliminary Experiment

For the Preliminary experiment described in Section 4.1, we leverage the same set of fine-tuning hyper-parameters other than the number of fine-tuning epochs for the explanation-generation model (denotes as M_{EG}) and the prediction model (denotes as M_P). The same set of hyper-parameters is: $batch\_size\_per\_GPU = 2; learning\_rate = 1e^{-4}; input\_max\_length = 512; target\_max\_length = 64$

We conduct a hyper-parameter search for the number of fine-tuning epochs for each amount of sampled examples, details are shown in Table 4.



(a) Active Learning on Mulit-NLI using explanation-generation model from e-SNLI with Setting 1

(b) Active Learning on Mulit-NLI using explanation-generation model from e-SNLI with Setting 2

Figure 4: Results of Transfer Learning Ablation Study of AL Simulation experiment on our Dual-model system from e-SNLI to Multi-NLI. Setting 1/2 refers to the settings for Active Learning Simulation in Section 4.2.

# of train data per category / total	**epoch for M_{RG}**	**epoch for M_P**
10 / 30	25	100
50 / 150	25	250
100 / 300	10	250
500 / 1500	5	50
1500 / 4500	5	50
3000 / 9000	5	25
5000 / 15000	5	25
Full	1	1

Table 4: Fine-tuning epochs of each model in our dual-model system with different data amount settings.

C.2 AL Simulation Experiment

For both of the AL Simulation settings we experimented in Section 4.2, we leverage the same set of hyper-parameters for fine-tuning our dual-model AL system: $batch\_size\_per\_GPU =$

$2; learning\_rate = 1e^{-4}; M_{EG}\_train\_epoch = 20, M_P\_train\_epoch = 250; input\_max\_length = 512; target\_max\_length = 64$

D Proposal for an Interactive System

Our proposed dual-model system can be easily implemented as an interactive human-centered AI system for supporting domain experts and human annotators' task, when they have the needs to label both labels and explanations.

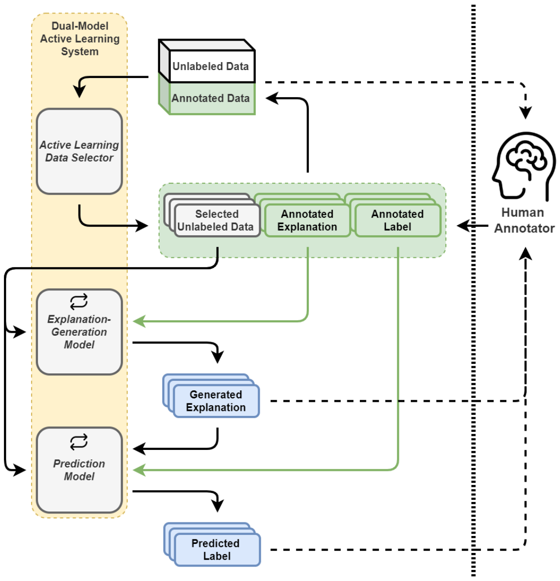


Figure 5: Our proposed dual-model system can be implemented as an interactive AL-based data annotation system to speed up users' annotation productivity. Such system can simply have an interface with four output functions (i.e., display unlabeled data, display AL selected data, display generated-explanation, and display predicted labeled) and one input function (i.e., annotate label and explanation for the unlabeled data.