Real-Time Visual Feedback for Educative Benchmark Creation : A Human-and-Metric-in-the-Loop Workflow

Anjana Arunkumar aarunku5@asu.edu

Swaroop Mishra srmishr1@asu.edu

Bhavdeep Sachdeva bssachde@asu.edu

Chitta Baral chitta@asu.edu

Chris Bryan cbryan16@asu.edu

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University

Abstract

Recent research has shown that language models exploit 'artifacts' in benchmarks to solve tasks, rather than truly learning them. Considering that this behavior inflates model performance, shouldn't the creation of *better benchmarks* be our priority? In pursuit of this, we focus on *guiding crowdworkers*, an under-explored facet of addressing benchmark idiosyncrasies. We propose VAIDA, a novel benchmark creation paradigm for NLP. VAIDA provides realtime visual feedback to both crowdworkers and backend analysts on both sample and dataset quality, and aims to educate them on the same. VAIDA also facilitates sample correction to improve quality via recommendations. VAIDA is domain, model, task, and metric agnostic, and constitutes a paradigm shift for robust, validated, and dynamic benchmark creation via human-and-metric-in-the-loop workflows. We demonstrate VAIDA's effectiveness by leveraging a state-of-the-art data quality metric DQI over four datasets. We further evaluate via expert review and a user study with NASA TLX. We find that VAIDA decreases effort, frustration, mental, and temporal demand of crowdworkers and analysts, while simultaneously increasing the performance of both user groups.

1 Introduction

Researchers invest significant effort to create benchmarks in AI, including ImageNet [4], SQUAD [18], and SNLI [1], as well as to create, tune, and tweak models that solve these benchmarks. *Can we rely on these benchmarks?* A growing body of recent research [24, 16, 11] is revealing that models exploit spurious bias– unintended correlations between input and output [25] (e.g. the word 'not' is associated with the label 'contradiction' in Natural Language Inference (NLI) [6])– instead of the actual underlying features, to solve many popular benchmarks. Models therefore fail to generalize, and experience drastic performance drops when testing with out of distribution (OOD) data or adversarial examples [2, 14, 28]. These biases[1] have led to the overestimation of AI's true advancement [21], and limit its deployment in safety-critical domains [8]. This begs the question: *Shouldn't ML researchers consequently focus on creating 'better' datasets rather than developing increasingly complex models on bias-laden benchmarks?*

Benchmark creators report bias baselines– hypothesis-only baseline in NLI [5])– and if the baseline

[1] Henceforth 'bias' implies spurious bias and also artifacts

Preprint. Under review.

performance is high, they might have to delete all the data created or can leverage adversarial filtering algorithms like AFLite [21] to delete targeted subsets of the data. This, along with other bias mitigation approaches [3, 10] has the following limitations: (i) data deletion/augmentation and residual learning do not justify the original investment in data creation, and (ii) crowdworkers are not provided continuous feedback to learn what constitutes high quality data– and so have additional overhead due to the manual effort involved in sample creation/validation. One potential solution to these problems is *in situ* feedback about artifacts while benchmark data is being created. *To our knowledge, there are no approaches which provide realtime artifact identification, feedback, and reconciliation opportunities to data creators, nor guide them on data quality*.

Contributions: (i) We propose *VAIDA* (Visual Analytics for Interactively Discouraging Artifacts), a novel system for benchmark creation that provides continuous visual feedback to data creators as the benchmark is being created. VAIDA supports both artifact identification and resolution, implicitly educating two classes of users – *crowdworkers* and *analysts*– on data quality. (ii) We design a *crowdworker* workflow and interface to create and submit new data samples for benchmark inclusion. Feedback from VAIDA guides crowdworkers on why a sample likely constitutes an artifact. To assist with sample modification, we propose an AutoFix module, that allows for machine-assisted sample modification to achieve higher quality (i.e., lower bias and higher generalizability). (iii) We develop a series of visualizations for *analysts* to review and verify submitted samples, as well as analyze and resample train-test splits to build an optimal dataset. VAIDA allows visual exploration of the effect of a sample's addition to a dataset in both cold-start and pre-existing data scenarios. We also propose the use of TextFooler for adversarial transformation to increase benchmark robustness using model-in-the-loop. (iv) We leverage DQI [15], a data quality metric that identifies artifacts by decomposing samples according to their language properties, as part of a metric-in-the-loop approach to demonstrate VAIDA's effectiveness over a set of four benchmarks. (v) We further evaluate VAIDA empirically through expert review and a user study to understand the cognitive workload it imposes. The results[2] indicate that VAIDA decreases mental demand, temporal demand, effort, and frustration of crowdworkers (29.7%) and analysts(12.1%); it increases performance by 30.8% and 26% respectively, and also educates crowdworkers on how to create *high quality* samples. VAIDA represents a novel, and to our knowledge, substantial shift in how benchmarks can be developed and validated, as it enables dynamic identification and resolution of artifacts during benchmark creation. By allowing crowdworkers and analysts to intuitively work in sync, spurious bias can be minimized, in turn reducing the overestimation of AI systems' capabilities; this enables their deployment in safety-critical domains.

2 Task Selection and Controlled Dataset Creation

In this work, we apply VAIDA for a natural language inference task (though it is task-independent), and mimic the SNLI dataset creation and validation processes. Elicited annotation has been found to lead to social bias in SNLI using probablistic mutual information (PMI) [20]. Visual feedback is provided based on DQI (which takes PMI into account) to explicitly correct this bias, and discourage the creation of such samples. Also, human annotation of machine-generated sentences/sentences pulled from existing texts instead of elicitation has been suggested to reduce such bias [27]. However, machine-generated text might look artificial, and work has shown that text generation has its own set of quality issues [13]. While we use AutoFix and TextFooler as modules to automatically transform samples, they are designed to be used in parallel with human sample creation. Their results can also be further modified by humans prior to submission. We see less reliance on these tools over the course of our user study (Subsection 5.3).

Additionally, previous work [19] in controlled dataset creation trains crowdworkers, and selects a subset of the best-performing crowdworkers for actual corpus creation. Each crowdworker's work is reviewed by another crowdworker, who acts as an analyst (as per our framework) of their samples. However, in real-world dataset creation, such training and selection phases might not be possible. Additionally, the absence of a metric-in-the-loop basis for feedback provided during training can potentially bias (through trainers) the created samples.

3 Workflow and Modules

In this section, we describe VAIDA's high-level workflow and important backend processes.

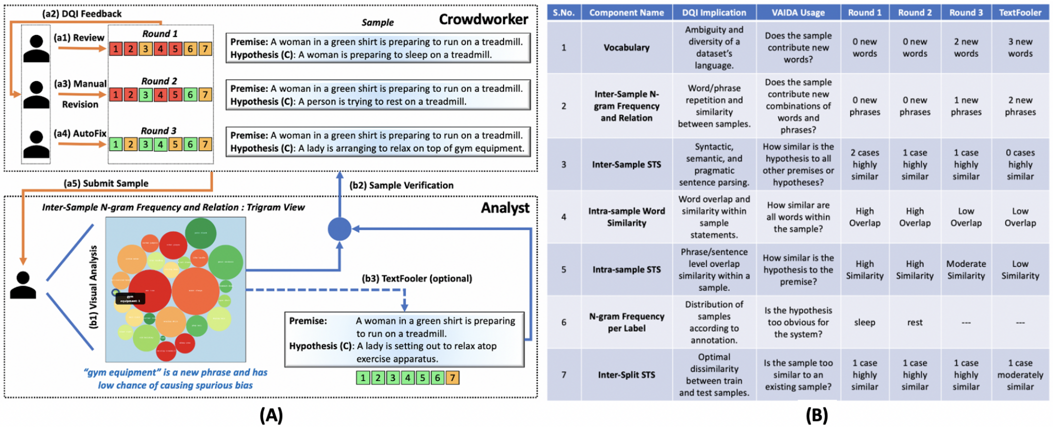
[2] Henceforth, red: decrease, and green: increase

3.1 Crowdworker and Analyst Workflows

VAIDA's high-level workflow is shown in Figure 1(A). Both crowdworkers and analysts work in parallel to create benchmark data points.

For crowdworkers, (a1) newly created samples are evaluated by DQI and (a2) realtime feedback is given to the user about potential biases. To fix an artifact, users can (a3) manually revise the sample, (a4) run AutoFix to automatically update it, or simply discard the sample and create a new one. After review (and potentially iterative DQI evaluations/revisions), (a5) the sample can be submitted for benchmark inclusion.

For analysts, (b1) VAIDA provides several visual interfaces to support detailed analysis and review of submitted samples, and to assess overall benchmark quality. Submitted samples enter a *pending* state until reviewed by the analyst, who *accepts*, *rejects*, or *modifies* the sample. (b2) Sample decisions are communicated back to crowdworkers to provide continuous feedback about performance and allow them to correct such samples. (b3) Analysts also have the option to submit low quality samples to TextFooler for adversarial transformation and augument with high quality samples to improve robustness of dataset, thereby ensuring minimal data loss.



Premise: A woman in a green shirt is preparing to run on a treadmill.
Hypothesis (C): A woman is preparing to sleep on a treadmill.

Premise: A woman in a green shirt is preparing to run on a treadmill.
Hypothesis (C): A person is trying to rest on a treadmill.

Premise: A woman in a green shirt is preparing to run on a treadmill.
Hypothesis (C): A lady is arranging to relax on top of gym equipment.

Premise: A woman in a green shirt is preparing to run on a treadmill.
Hypothesis (C): A lady is setting out to relax atop exercise apparatus.

S.No.	Component Name	DQI Implication	VAIDA Usage	Round 1	Round 2	Round 3	TextFooler
1	**Vocabulary**	Ambiguity and diversity of a dataset's language.	Does the sample contribute new words?	0 new words	0 new words	2 new words	3 new words
2	**Inter-Sample N-gram Frequency and Relation**	Word/phrase repetition and similarity between samples.	Does the sample contribute new combinations of words and phrases?	0 new phrases	0 new phrases	1 new phrases	2 new phrases
3	**Inter-Sample STS**	Syntactic, semantic, and pragmatic sentence parsing.	How similar is the hypothesis to all other premises or hypotheses?	2 cases highly similar	1 case highly similar	1 case highly similar	0 cases highly similar
4	**Intra-sample Word Similarity**	Word overlap and similarity within sample statements.	How similar are all words within the sample?	High Overlap	High Overlap	Low Overlap	Low Overlap
5	**Intra-sample STS**	Phrase/sentence level overlap similarity within a sample.	How similar is the hypothesis to the premise?	High Similarity	High Similarity	Moderate Similarity	Low Similarity
6	**N-gram Frequency per Label**	Distribution of samples according to annotation.	Is the hypothesis too obvious for the system?	sleep	rest	---	---
7	**Inter-Split STS**	Optimal dissimilarity between train and test samples.	Is the sample too similar to an existing sample?	1 case highly similar	1 case highly similar	1 case highly similar	1 case moderately similar

Figure 1: **(A)** VAIDA workflow– branches *(a)*: crowdworker, *(b)*: analyst functions. **(B)** Language properties considered in DQI, interpretation in VAIDA, and statistics for each feedback shown in **(A)** given 100 pre-existing dataset samples; STS: semantic textual similarity. C:"contradiction".

3.2 Modules

DQI and Traffic Signal Scheme: VAIDA communicates sample quality using an intuitive traffic signal color coding (red, yellow, green) to indicate if samples might lead to bias. The quality of individual features (aspects) of samples are evaluated based on decreasing presence of artifacts and increasing generalization capability. Based on overall sample quality, VAIDA computes the probability the sample will be accepted/rejected.

To demonstrate this, we leverage DQI [15], which can: (i) compute the overall data quality for a benchmark with n data samples, and (ii) compute the impact of a new $(n+1)^{th}$ data sample. When a crowdworker creates a new sample, DQI estimates its quality by calculating seven component values corresponding to a set of seven language properties; these are defined in Figure 1(B), along with their interpretation in VAIDA[3].

AutoFix: We propose AutoFix as a module to help crowdworkers avoid creating bad samples by recommending changes to a sample to improve its quality. The AutoFix algorithm is explained in Figure 2(a). Given a premise, hypothesis, and the DQI values for the hypothesis, AutoFix sequentially masks each word in the hypothesis and ranks words based on their impact on model output, i.e. their importance. DQI bins values into three classes: *high*, *acceptable*, and *low* quality. Hypothesis words are replaced in the order of importance to achieve *acceptable* quality. DQI hence controls the amount and aspect of changes made by AutoFix.

VAIDA employs AutoFix in an incremental manner to facilitate human-in-the-loop continuous feedback. By incrementally changing the sample one word at a time, users can understand how and why their sample is being modified and how DQI values are affected. Figure 1(A) shows AutoFix

[3] Hyperparameters depend on the application type. See Supplemental Material:Hyperparameters.

results after being used three times on a data sample, while Table 46 shows examples of AutoFix being applied to SNLI samples.

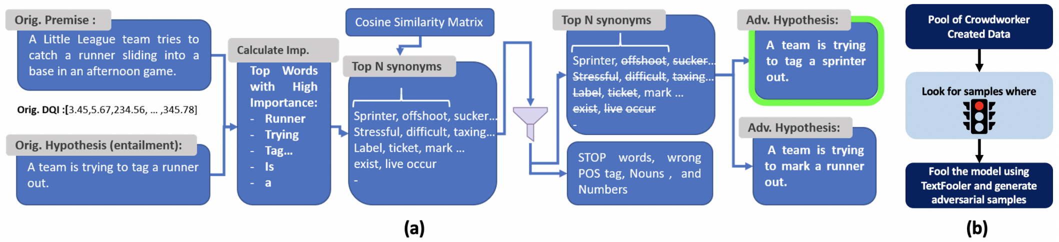

Orig. Premise :
A Little League team tries to catch a runner sliding into a base in an afternoon game.
Orig. DQI :[3.45,5.67,234.56, ... ,345.78]
Orig. Hypothesis (entailment):
A team is trying to tag a runner out.
Calculate Imp.
Top Words with High Importance:
- Runner
- Trying
- Tag...
- Is
- a
Cosine Similarity Matrix
Top N synonyms
Sprinter, offshoot, sucker...
Stressful, difficult, taxing...
Label, ticket, mark ...
exist, live occur
-
Top N synonyms
Sprinter, ~~offshoot~~, ~~sucker~~...
~~Stressful~~, ~~difficult~~, ~~taxing~~...
~~Label~~, ~~ticket~~, mark ...
~~exist~~, ~~live occur~~
-
STOP words, wrong POS tag, Nouns , and Numbers
Adv. Hypothesis:
A team is trying to tag a sprinter out.
Adv. Hypothesis:
A team is trying to mark a runner out.
(a)
Pool of Crowdworker Created Data
Look for samples where
Fool the model using TextFooler and generate adversarial samples
(b)

Figure 2: (a) AutoFix Algorithm. (b) TextFooler Algorithm.

Module	Premise	Orig. Hypothesis	DQI	Suggested Words	New Hypothesis	New DQI	Label
AutoFix	A man in a green apron smiles behind a food stand	A man smiles	3.324	smiles	A person is grinning	6.304	Entailment
TextFooler	One blond girl offers another blond girl some food	One girl offering another girl food	2.378	N/A	One girl tenders another food.	3.116	Entailment

Table 1: Examples for Autofix and TextFooler, with DQI's Intra-sample STS values for SNLI samples.

TextFooler: From an analyst's perspective, the quality of a submitted sample might be "too low" because (i) the crowdworker might not employ AutoFix, or (ii) there is a narrow acceptability range due to the criticality of the application domain, such as in BioNLP [12]. We therefore implement a module TextFooler [9] for adversarial sample transformation of low quality samples (instead of discarding them) to improve benchmark robustness (another key aspect in benchmark construction), and ensure that the crowdsourcing effort is not wasted.

We initially use AFLite, a recent adversarial filtering approach [2], to bin samples into *good* (retained samples) and *bad* (filtered samples) splits. Using TextFooler, we adversarially transform bad split data to flip the label; we revert back to the original label and evaluate this sample using DQI, as shown in Figure 1(A) and Table 46.

4 Interface Design Choices

VAIDA provides customized interfaces for both crowdworkers and analysts.

4.1 Crowdworker Interface

In addition to workflow functionalities, the crowdworker interface (Figure 3(a)) provides interface navigation, data creation, and feedback interpretation instructions (A). Sample creation (B) mimics the original SNLI crowdsourcing interface– examples (b1) are given, and the premise field (b2) autopopulates with captions from the Flickr30 corpus; three hypotheses (for entailment, neutral, contradiction labels) are to be entered at a time, though they are reviewed individually. DQI feedback (C) is shown for each component (c1), and hovering on these displays a tooltip that suggests sample fixes to improve quality (c2). (c3) Overall sample quality and (c4) estimated probability it will be accepted provide additional feedback. AutoFix (b3) can be used for automatic fixes. Samples enter a pending state (d1), until analyst review, upon which the count (d2) and pie chart (d3) update. Historical quality of samples submitted by the user (e1) , and (e2) current rank of the user are shown to help crowdworkers gauge their performance. Communication links for FAQs, and error reporting are also provided (F).

4.2 Analyst Interfaces

While crowdworkers work within a single tightly-coordinated interface to create, submit, and review samples, analysts can navigate between a set of nine interfaces (Figure 3(b)) to review samples in detail to make accept, reject, and modification decisions, and to assess overall benchmark quality.

(UI) The *single crowdworker view* provides a view similar to the crowdworker interface, and allows the analyst to review the work of a single crowdworker. The data creation panel is modified to allow the analyst to iterate over and review submitted samples. For low quality samples, the TextFooler module can be invoked (via a 'Generate Adversarial Example' button). (C1–C7) Other interfaces available to the analyst support detailed review of specific DQI components and allow the analyst to simulate how adding one or more submitted samples affects the benchmark's quality. Several visualization techniques are employed (treemap, node-link diagram, bubble chart, heatmap, bar chart, etc.) tailored to the specific DQI component of interest, but all interfaces consistently utilize the traffic signal color scheme to represent quality[4].

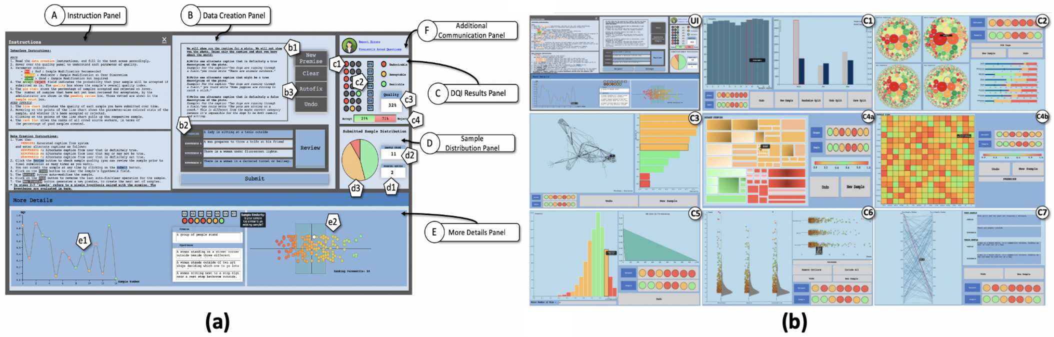


Figure 3: **(a)** VAIDA's crowdworker interface consists of six linked panels: (A) Instructions, (B) Data creation, (C) DQI results, (D) Sample distribution, (E) More details, and (F) Additional communication. **(b)** VAIDA's provides a collection of interfaces for the analyst supporting detailed analysis, review, and investigation of submitted samples and the overall state of the benchmark.

5 Evaluation

We evaluate VAIDA's efficacy at providing real time feedback to crowdworkers during benchmark creation using case studies, expert reviews, and user studies. First, we evaluate by assuming gold scores corresponding to the results for a recent adversarial filtering approach, AFLite [2]. Since this assumption has potential risks, as AFLite may have some limitations (outlined in [15]), expert review and the subsequent user study provide additional evaluation and feedback.

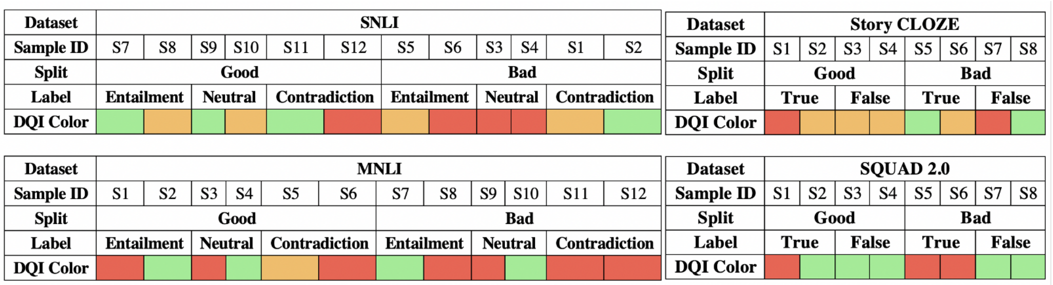
5.1 Case Studies

We evaluate our traffic signal scheme (based on DQI) over four datasets: SNLI [1], MNLI[26], SQUAD 2.0 [17], and Story CLOZE Task [24]. In the case of SQUAD 2.0 and Story CLOZE, we split each sample into multiple samples– for e.g., in Story CLOZE there are two ending choices per sample and so we make two samples, with label *True* for the sample with the correct ending and *False* for the sample with the incorrect ending.
The presence of a large number of artifacts has been shown in several studies on SNLI [6] and Story CLOZE Task [24]. MNLI and SQUAD 2.0 have been shown to have a relatively smaller number of artifacts [6, 11], and therefore ensure adversarial evaluation of VAIDA. We evaluate each dataset using its test sets, or if unavailable, on its dev sets.
Setup: For each dataset, we filter using AFLite and divide it into two categories: *good* and *bad*, where each category respectively refers to the set of samples retained and removed after adversarial filtering. Evaluation with VAIDA involves providing feedback in two different settings: (i) no preexisting samples, and (ii) 100 preexisting samples corresponding to the good category. For (ii), random sampling of 100 pre-existing samples is done 10 times, for a fair comparison.
In (ii), we: (a) compute DQI for the existing sample set as x_1, (b) recompute DQI for the sample set after a new sample is added as x_2, and (c) calculate $\Delta x = x_1 - x_2$. The crowdworker interface shows the DQI components corresponding to Δx. In the analyst interface, both Δx as 'sample' and x_2 as the 'dataset' quality are shown component-wise in each view. For fair comparison, we have taken illustrative samples from the AFLite paper [2] for SNLI. We randomly sample for other datasets[5], as corresponding examples were not illustrated in those papers.
Configuring the Boundary Separating Red, Yellow and Green Flags There exist two hyperparameters separating the boundary between red, yellow, and green flags. We tune hyperparameters on 0.01% of data manually in a supervised manner [15]. This is analogous to how humans learn quickly from few samples. Hyperparameters depend on the application task [3]. On the other hand, they help in controlling the hardness of a benchmark, which can be leveraged in an active learning setting to develop dynamic benchmarks.
Results: DQI component colors across settings are correctly predicted according to AFLite catego-

[4] See Supplemental Material: Interface Design for interface intuitions and description

[5] See Supplemental Material: Evaluation, for details across all components and analyses.

rization of good and bad splits on an average [6] of 10/12 times in SNLI, 5/8 times in SQUAD 2.0 and Story CLOZE, and 7/12 times in MNLI [5] as illustrated in Table 29. We convert SQUAD 2.0 and Story CLOZE into NLI format, with *answer* and *ending* corresponding to *hypothesis*, and *context* and *story* corresponding to *premise*, respectively.



Dataset	**SNLI**											
Sample ID	S7	S8	S9	S10	S11	S12	S5	S6	S3	S4	S1	S2
Split	**Good**						**Bad**					
Label	**Entailment**		**Neutral**		**Contradiction**		**Entailment**		**Neutral**		**Contradiction**	
DQI Color	🟩	🟧	🟩	🟧	🟩	🟥	🟧	🟥	🟥	🟥	🟧	🟩

Dataset	**Story CLOZE**							
Sample ID	S1	S2	S3	S4	S5	S6	S7	S8
Split	**Good**				**Bad**			
Label	**True**		**False**		**True**		**False**	
DQI Color	🟥	🟧	🟧	🟧	🟩	🟧	🟥	🟩

Dataset	**MNLI**											
Sample ID	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Split	**Good**						**Bad**					
Label	**Entailment**		**Neutral**		**Contradiction**		**Entailment**		**Neutral**		**Contradiction**	
DQI Color	🟥	🟩	🟥	🟩	🟧	🟥	🟩	🟥	🟥	🟩	🟥	🟥

Dataset	**SQUAD 2.0**							
Sample ID	S1	S2	S3	S4	S5	S6	S7	S8
Split	**Good**				**Bad**			
Label	**True**		**False**		**True**		**False**	
DQI Color	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩

Table 2: Evaluating VAIDA over the most sensitive DQI component, Intra-Sample Word Similarity. Successes: green/orange for good, red/orange for bad split. Failures: red for good, green for bad split.

Analysis: False positives and false negatives can be attributed to the limitation of AFLite in incorrectly classifying samples [15]. Additionally, we have two observations: (i) VAIDA's prediction accuracy decreases as the artifact level in a dataset decreases. (ii) The values of most DQI sub-components do not change significantly (<25% of the time) after adding samples in both categories. However, it changes considerably (>60% of the time) across two sub-components: Intra-sample word overlap and word similarity, both of which belong to the fifth component of DQI. This can again be explained by AFLite's sensitivity towards word overlap [15].

5.2 Expert Review

We present an initial prototype of our tool, to a set of three researchers with expertise in NLP and knowledge of data visualization, in order to judge the interface design. For each expert, the crowdworker interface and then analyst interfaces were demoed. Participants could ask questions and make interaction/navigation decisions to facilitate a natural user experience. All the experts appreciated the easily interpretable traffic-signal color scheme and found the organization of the interfaces—providing separate detailed views within the analyst workflow– a way to prevent cognitive overload (too much information on one screen) while allowing multi-granular analysis; this would help in classifying samples of middling quality as benchmark size increases with relative ease.

5.3 User Study

Setup: We approach several software developers, testing managers, and undergraduate/ graduate students. Based on their domain familiarity (in NLP and visualization), we split them into 23 crowdworkers and 8 analysts for constructing NLI samples, given premises. There are 100 high quality samples in the system at the time each participant participates in each round. Their experience is evaluated using NASA Task Load Index[7][7], where each task is scored in a 100-points range, with 5-point steps. To conduct an ablation study, we introduce modules one at a time (and finally the complete system) to all user classes as follows: (i) Crowdworkers— conventional crowdsourcing, traffic signal feedback, AutoFix, all, and (ii) Analysts— conventional analysis, traffic signal feedback, visualizations, TextFooler, all. For both user categories, a preliminary walkthrough of panels using 2 fixed samples– chosen randomly from the set used for the case study with SNLI– is conducted for each round of the study (Figure 4(a)).

Analysis: Figure 4(b) summarizes study results, averaged over all user responses. The users are presented with system modules in the order listed, and are asked to report scores relative to the original score they assign the conventional crowdsourcing/analysis approaches; at the end of each round, they are also asked for their comments[8].

Crowdworkers: Traffic signal feedback initially increases time (29.2%) and effort (65%) required to create high quality samples, as users have to correct them. However they are more confident (performance– 23.1%) of sample quality. AutoFix usage causes an unexpected increase in effort (10%) and frustration (77.8%), as users do not fully trust recommendations without visual feedback.

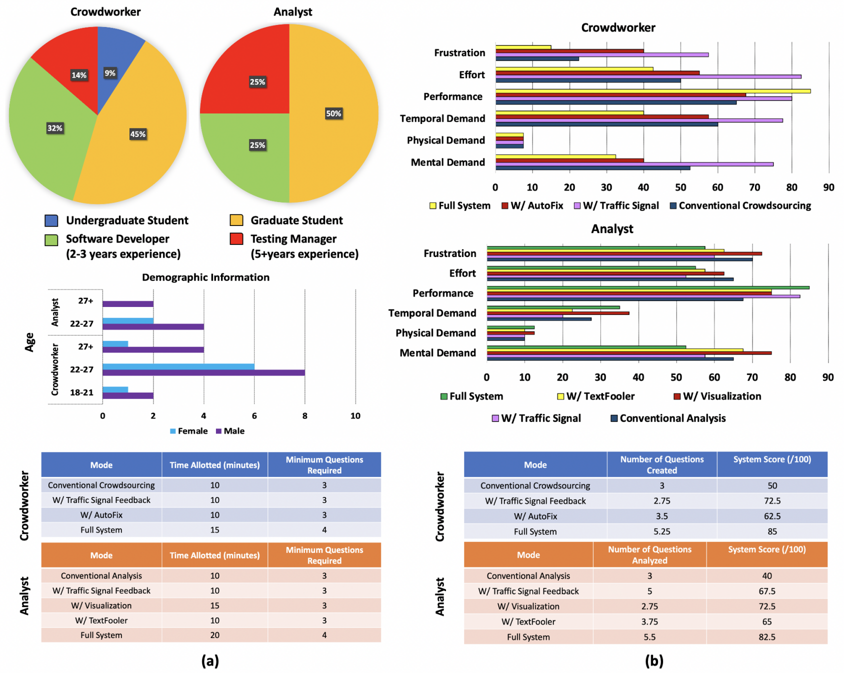
[6] We run (i) once and (ii) 10 times.

[7] See Supplemental Material: User Study for more details.

[8] We aggregate comment analysis here, see Supplemental Material: Expert and User Comments for quotes.

The drastic improvement over all aspects (frustration– 33.3%, mental demand– 38.1%, temporal demand– 33.3%, effort– 15%, average decrease in difficulty– 29.7%, performance– 30.8%) in the case of using the full system is in line with this observation. The number of questions created per round (traffic signal– 8.3%, AutoFix– 16.7%, full system– 75%) as well as system scores (traffic signal– 45%, AutoFix– 25%, full system– 70%) also follows this trend, across all types of crowdworkers.

Analysts: Analysts find the task easier (effort– 19.3%, performance– 22.2%) with traffic signal feedback, as quality is clearly marked.



Mode	Time Allotted (minutes)	Minimum Questions Required
Conventional Crowdsourcing	10	3
W/ Traffic Signal Feedback	10	3
W/ AutoFix	10	3
Full System	15	4

Mode	Time Allotted (minutes)	Minimum Questions Required
Conventional Analysis	10	3
W/ Traffic Signal Feedback	10	3
W/ Visualization	15	3
W/ TextFooler	10	3
Full System	20	4

Mode	Number of Questions Created	System Score (/100)
Conventional Crowdsourcing	3	50
W/ Traffic Signal Feedback	2.75	72.5
W/ AutoFix	3.5	62.5
Full System	5.25	85

Mode	Number of Questions Analyzed	System Score (/100)
Conventional Analysis	3	40
W/ Traffic Signal Feedback	5	67.5
W/ Visualization	2.75	72.5
W/ TextFooler	3.75	65
Full System	5.5	82.5

Figure 4: (a) User Study Setup, (b) User Study Results (averaged over all responses)

When analysts are shown the visualization interfaces, they are explicitly taught to differentiate the traffic signal colors in the visualizations as being indicative of how the sample affects the overall dataset quality, i.e., the colors in different component views represent individual terms of the components calculated over the whole dataset (analysts can toggle between the states of original dataset and new sample addition). We find that users initially find this more difficult to do (mental demand– 15.4%, temporal demand– 36.4%, frustration– 3.5%), though they agree that it improves their judgement of quality (performance– 11.1%). Analysts averaged behavior on TextFooler models the conventional approach quite closely, as analysts are seen to have a tendency to send all samples that are unclear to TextFooler immediately. With the full system, analysts also report improvement in all aspects (average decrease in difficulty– 12.1%), particularly mental demand (19.2%) and performance (26%), considering that the system increases the likelihood of a low hypothesis baseline. The visualization usage also improves, as analysts learn component relationships. Altogether, sample evaluation by analysts increases (full system– 83.3%), following this trend, and analysts are more assured of their performance (full system score– 106.25%).

Learning Curve: At the end of the study, all users are asked the following: *"What do you think high quality means?"* We find that users are able to distinguish certain patterns that promote higher quality, such as keeping sentence length appropriate and uniform across labels (not too long/short), using complex phrasing ('not bad')/gender information/modifiers across labels, and decreasing premise-hypothesis word overlap; they also do not display undesirable behavior like tweaking previously submitted samples just to create more.

User Education: We also conduct a variation of the study where a subset of participants (7 crowdworkers and 2 analysts) agreed to create/ analyze samples, for varying numbers of pre-accepted samples (Figure 5), in only the full system condition. In general, as the number of samples increases, the proportion of red or mixed samples also increases, and those green decreases. We find that when beginning from the cold start condition, as the sample number increases, due to their familiarity with the system, both crowdworkers and analysts are able to leverage the system better to avoid red samples. However, when participants are directly started in situations with > 500 samples in the system, their unfamiliarity with the system initially causes a steepening of the learning curve compared to the cold start condition; this also tapers and saturates more slowly than cold start as the users gain experience. In the case of cold start, we find that users who create $\sim$50 samples report lesser reliance on AutoFix as they get better at creating higher quality samples; those who analyze $\sim$75 samples use TextFooler more efficiently as they understand how to deal with samples of middling quality better.

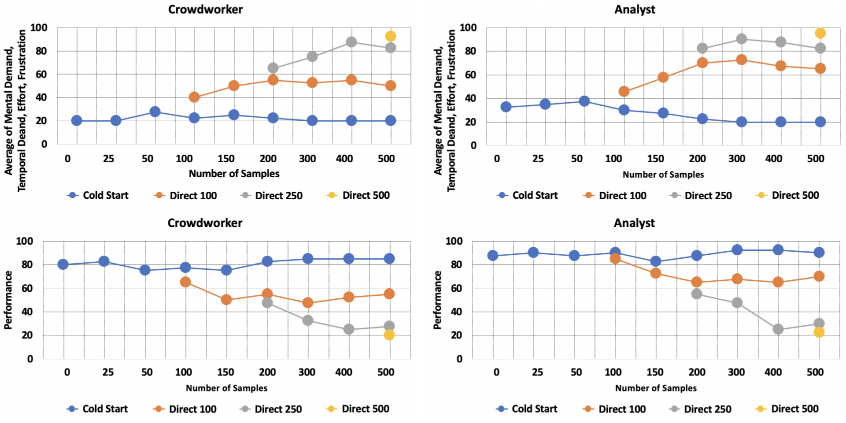


Figure 5: User education curves. Cold start has no pre-existing samples, and direct-n has n pre-existing samples. Mental Demand, Temporal Demand, Frustration, and Effort are averaged, Physical Demand is ignored. Performance is plotted separately as it shows differing behavior than the others.

6 Discussions and Future Work

VAIDA decreases mental demand, as well as the pressure (temporal demand, frustration) of creation and validation. Users report greater satisfaction with their work (performance); we also observe that users process more than the required samples with the full system in our user study due to the lower effort expended. This implies possible higher crowdworker retention and engagement. The use of AutoFix and TextFooler initially compensates for learning/decision fatigue; participants become less reliant on these over time and report less fatigue, indicating an understanding of data quality and its relevant features. We intend to integrate VAIDA with an actual crowdsourcing framework, and create a high quality benchmark, which we will evaluate using DQI and other metrics[6]. In our ablation study, we introduce modules in a fixed order to users, as per the patterns of usage preferred by the experts. We will compare this directly with the effectiveness of explicit user training [19] on patterns that decrease quality, as well as user training using an alternate workflow to see if/how user strategy changes. While both AutoFix and TextFooler can potentially result in the production of artificial sentences, as human intervention is allowed, we expect the final sample to be natural for a larger benchmark; we observe this in our user study. The most sensitive DQI component is found to involve word overlap; AFLite's removed samples also exhibit larger word overlap among other artifacts [6]. We will modify the n-gram, and sentence related DQI subcomponents to increase the range of bias captured.

7 Conclusion

We propose VAIDA, a paradigm to address benchmark bias, by integrating human-in-the-loop sensemaking with continuous feedback from a data quality metric, DQI. We design complementary workflows for both crowdworkers and analysts, to create new samples, evaluate them for the existence of artifacts, and review/repair samples to ensure the overall benchmark quality. We also develop AutoFix for automated data repair, and design a mechanism for adversarial tranformation to improve data quality by leveraging TextFooler. We also construct several visualization interfaces to analyze quality considerations at multiple granularities. VAIDA is evaluated with several case studies, a set of expert reviews– which provide qualitative feedback about the overall workflow experience– and a user study with NASA TLX. We find that usage of VAIDA decreases mental demand, temporal demand, effort, and frustration of crowdworkers (29.7%) and analysts(12.1%); it increases performance by 30.8% and 26% respectively, and educates users on data quality. VAIDA demonstrates a novel, dynamic approach for building benchmarks and mitigating bias, and serves as a starting point for the next generation of benchmarks in AI.

8 Broader Impact

- **Greater Accessibility:** The process of creating big datasets involves heavy resource investment. Model development to solve such datasets and top leaderboards necessitates further resource utilization. This skews deep learning research to favor those communities with high resources. Using VAIDA to create smaller, high quality datasets, can hence lessen resource requirement, and increase accessibility to low resource communities.
- **Environmental Impact:** The heavy computation involved in training models on large datasets adversely affects the environment on a broader scale [23]. Reduced dataset size can reduce the magnitude of this effect.
- **White Box Benchmarks:** Current trends in AI are model transparency, interpretability, and explainability. VAIDA facilitates the explanationa dn corection of sample artifacts, to build robust benchmarks. This approach can be extended to other AI domains, opening up a new developmental paradigm for white-box benchmarks and models. This can serve to boost the trust of communities that apply and rely on AI, such as health care, and develop safer, reliable AI.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their thoughtful feedback. The support of DARPA SAIL-ON program (W911NF2020006) is gratefully acknowledged.

References

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[2] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*, 2020.

[3] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

[6] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018.

[7] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[9] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, 2019.

[10] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019.

[11] Divyansh Kaushik and Zachary C Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *arXiv preprint arXiv:1808.04926*, 2018.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] Nitika Mathur, Tim Baldwin, and Trevor Cohn. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*, 2020.

[14] R Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*, 2019.

[15] Swaroop Mishra, Anjana Arunkumar, Bhavdeep Singh Sachdeva, Chris Bryan, and Chitta Baral. Dqi: A guide to benchmark evaluation. *arXiv: Computation and Language*, 2020.

[16] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*, 2018.

[17] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[18] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[19] Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. Crowdsourcing a high-quality gold standard for qa-srl. *arXiv preprint arXiv:1911.03243*, 2019.

[20] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 2017.

[21] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[22] David Schuff, Karen Corral, and Ozgur Turetken. Comparing the understandability of alternative data warehouse schemas: An empirical study. *Decision support systems*, 52(1):9–20, 2011.

[23] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *arXiv preprint arXiv:1907.10597*, 2019.

[24] Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A Smith. The effect of different writing tasks on linguistic style: A case study of the roc story cloze task. *arXiv preprint arXiv:1702.01841*, 2017.

[25] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[26] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.

[27] Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395, 2017.

[28] Yuan Zhang, Jason Baldridge, and Luheng He. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*, 2019.

9 Supplemental Material

9.1 Infrastructure Used

In Section 2, we describe VAIDA's flow by high level workflow and back-end processes(DQI, AutoFix, and TextFooler). Further, as discussed in Subsection 2.2 DQI can be used for the: i) overall benchmark, and ii) impact of new samples. Depending on the task at hand we run our experiments in different hardware settings. The DQI calculations run mostly using CPU, for new samples as well as overall samples. The AutoFix procedure, as explained in Subsection 2.2, gives user assistance for improving quality on a per submission basis. Therefore that does not require high GPU intensive systems, one can use that to speed up the process; we have provision to shift it to a GPU as well if necessary. For the TextFooler the fine tuning of the model is run on "TeslaV100-SXM2-16GB"; CPU cores per node 20; CPU memory per node: 95,142 MB; CPU memory per core: 4,757 MB– this is not a necessity as code has been tested on lower configuration GPUs as well but we have run our experiments in this setting. The attack part of the TextFooler requires more memory and we run that code on "Tesla V100-SXM2-32GB" com-pute Capability: 7.0 core Clock: 1.53GHz, coreCount: 80, device Memory Size: 31.75GiB device Memory Bandwidth: 836.37GiB/s.

9.2 Run-time estimations

The DQI calculation run on CPU(for real life setting purposes), for the approximate estimate for the time taken, we run experiments for fixed data size of 10K samples. If the DQI calculations are done to calculate the impact of individual new samples it take a couple of seconds. On the other hand, If we take the whole 10k size dataset it takes around 48 hours to complete the process on CPU. This whole process can be run in parallel to reduce the time taken to 16 hours. The Textfooler part consists of two steps the fine tuning part and attack part for generating adversaries. For fine tuning models we use "TeslaV100-SXM2-16GB" and it takes 20-30 minutes to complete the process. For the attack part we use "Tesla V100-SXM2-32GB", which takes 2-3 hrs for completing 20k data samples. This estimate requires the cosine similarity matrix for word embeddings to be calculated before hand which takes around 1-2 hrs, but this step has to be done only if the word embeddings are modified. This is a rare task so we have kept this separated.

9.3 Hyper Parameter

The main focus for this study in all is to look at the estimations of DQI and its variations. Keeping that in mind we have kept basic hyper-parameters fixed in the experiments. We keep the learning rate to 1e-5, the number of epoch during the experiments have varied from 2-3, per gpu train batch size and eval batch size varies from 8-64 samples, the results shown are with respect to 8 batch size in this paper, adam epsilon is set to 1e-8, weight decay is set to 0, maximum gradient normalisation is set to 1, and maximum sequence length is set to 128. The variations and range in the DQI parameters are dataset specific.

For TextFooler the the semantic similarity is fixed to 0.5 uniformly for all the experiments shown in this paper.

Hyperparameters depend on the application task: [15] design DQI as a generic metric to evaluate diverse benchmarks. However, the definitions of what constitutes high and low quality will vary depending on the application. For example, BiomedicaNLP might have lower tolerance levels for spurious bias than General NLP. Another case is in water quality– cited as an inspiration for DQI by [15]– where the quality of water needed for irrigation is different than that of drinking or medicine. We can therefore say that the hyper-parameters in the form of boundaries separating high and low quality data (i.e., inductive and spurious bias) are dependent on applications.

9.4 Interface Design

Careful Selection of Visualizations Prior to the design of test cases and a user interface, data visualizations highlighting the effects of sample addition are built. Considering the complexity of the formulas for the components of empirical DQI, we carefully select visualizations to help illustrate and analyze the effect to which individual text properties are affected.

All DQI Component Values are Shown for Each Visualization: We show all DQI component values for each visualization, since the user needs to optimize across several dependent components while selecting the best quality data. All DQI component values are tracked across different visualizations using two separate panels present at the bottom of the screen. The first panel shows the component-wise values as colored circles for the overall dataset prior to adding the sample. The second panel is initially a set of grayscale circles. Once the new sample is added, both the panels are updated. The first panel may not show any color changes, as it represents the overall dataset. The second however, will now display colored circles based on the DQI component values of the individual new sample. The values of the components can be viewed with a tooltip.

Traffic Signal Color Scheme: The color combination of Red-Yellow-Green used in all the visualizations represents the quality of the component/property being observed/analyzed. Here, red represents an undesirable quality value, yellow a permissible value, and green an ideal value. The color scale follows a pattern of red-yellow-green-yellow-red unless otherwise specified, centered around the ideal value of a component.

9.4.1 Vocabulary

Which Characteristics of Data are Visualized? The contribution of samples to the size of the vocabulary is tracked using a dual axis bar chart. This displays the vocabulary size, along with the vocabulary magnitude, across the train, dev, and test splits for the dataset. Also, the distribution of sentence lengths is plotted as a histogram. Each sample contributes two sentences, i.e., the premise and hypothesis statements. Figure 6 illustrates this.

Interactions: Interactions are supported through a tooltip and buttons. The tooltip displays the quantities in both charts on mouseover, and the buttons are used to update the chart. There are five “interactions” supported:

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the train split by default. A script to calculate the new words this sample contributes to the vocabulary set is run, and the bar chart is accordingly updated. The sentence lengths of the premise and hypothesis statements are used to update the histogram. The updated portions of both the charts are highlighted, as shown in Figure 7. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of ’addition of a new sample’ by using the saved state variables to restore the visualizations back to their original state.
- **Randomization of Split *(Randomize Split)*:** The samples are distributed randomly between the train, dev, and test splits, using a 70:10:20 split ratio. Once the split is randomized, the new sample cannot be removed from the split anymore, as it is not necessarily a part of the train set. In order to account for annotator bias, the annotator id of dataset samples is used to create mutually exclusive annotator sets across splits. Additionally, the split is designed such that if a premise has multiple hypothesis statements and is therefore repeated across samples, then all samples containing that premise belong to the same split. This split operation can be performed multiple times, as an attempt to understand the effect of data ordering on the DQI component values for the overall dataset. The previous state of the visualization is saved in a set of variables.
- **Undo Split *(Undo Split)*:** This reverses the operations of ’randomization of split’ by using the saved state variables to restore the visualizations back to their original state. Only the latest randomization operation is reversed.
- **Save Split *(Save Split)*:** Once the split is satisfactory, this button can be used to freeze this split state for the remainder of the analysis. On addition of the next sample, this frozen state is used for the initialization of the visualizations.

9.4.2 Inter-sample N-gram Frequency and Relation

Which Characteristics of Data are Visualized? There are different granularities of samples that are used to calculate the values of this component, namely: words, POS tags, sentences, bigrams, and trigrams. The granularities’ respective frequency distributions and standard deviations are utilized for this calculation.

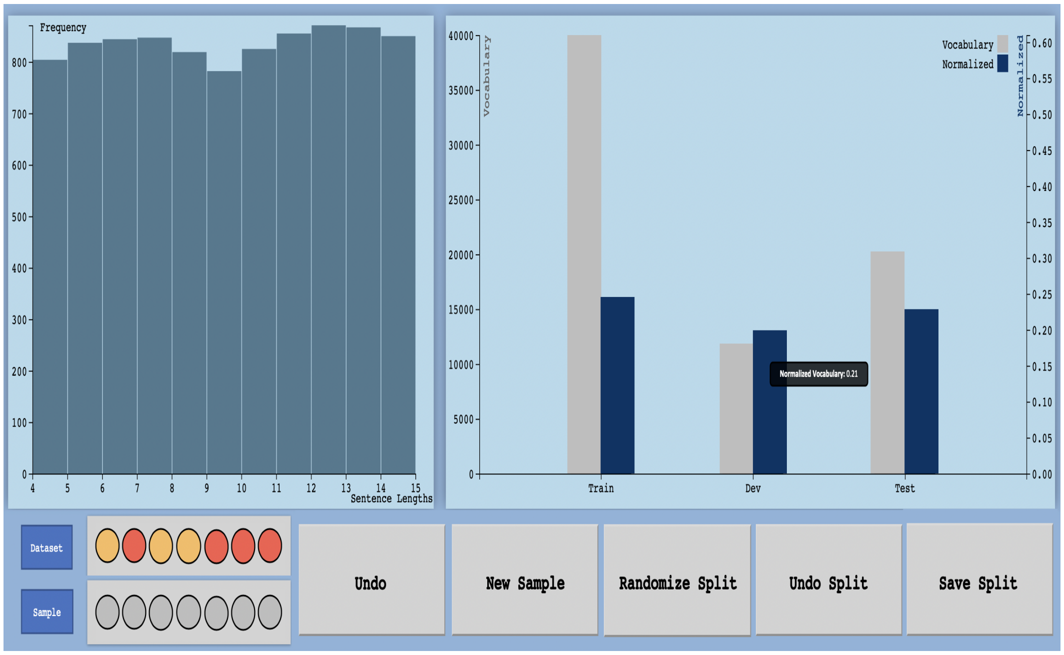


Figure 6: DQI_{c1} Visualization Prior to New Sample Addition

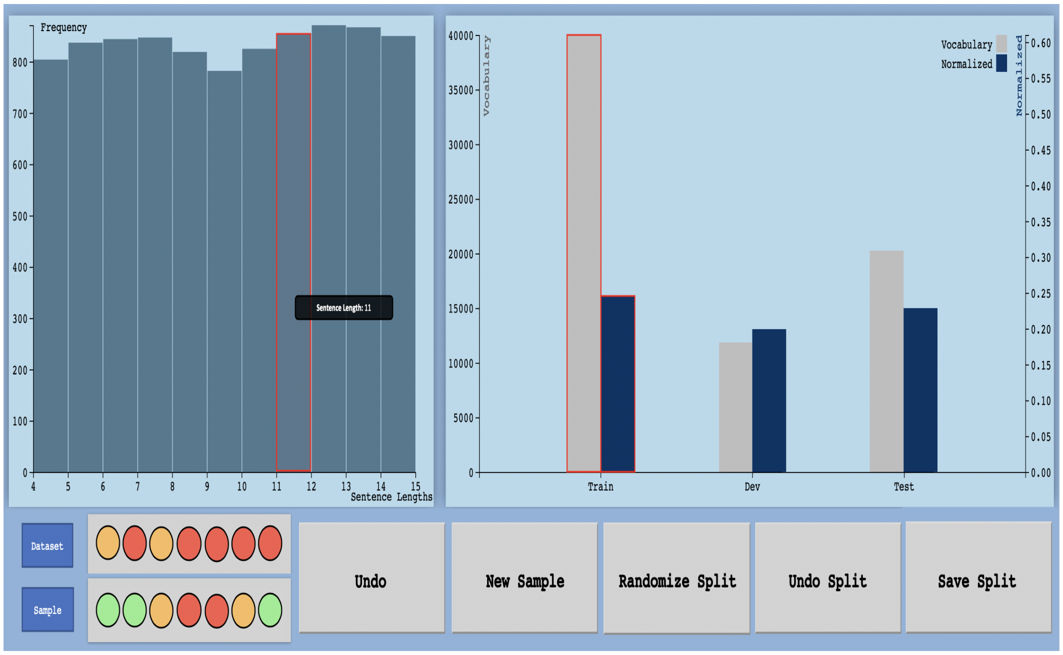


Figure 7: DQI_{c1} Visualization On New Sample Addition

Bubble Chart for visualizing the frequency distribution: A bubble chart is used to visualize the frequency distribution of the respective granularity. This design choice is made in order to clearly view the contribution made by a new sample when added to the existing dataset in terms of different granularities. The bubbles are colored according to the bounds set for frequencies by the hyperparameters, and sized based on the frequency of the elements they represent. Additionally, some insight into variance can be obtained from this chart, by observing the variation in bubble size.

Bullet Chart for impact of new sample: The impact of sample addition on standard deviation can be viewed using the bullet chart. The red-yellow-green color bands for each granularity represent the

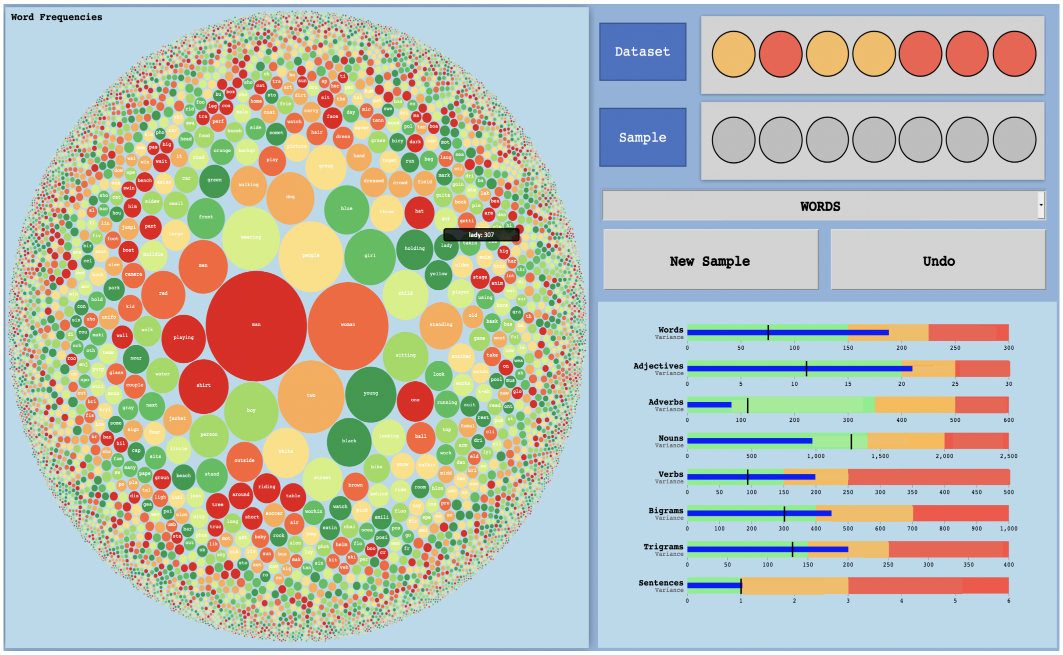


Figure 8: DQI_{c2} Visualization Prior to New Sample Addition

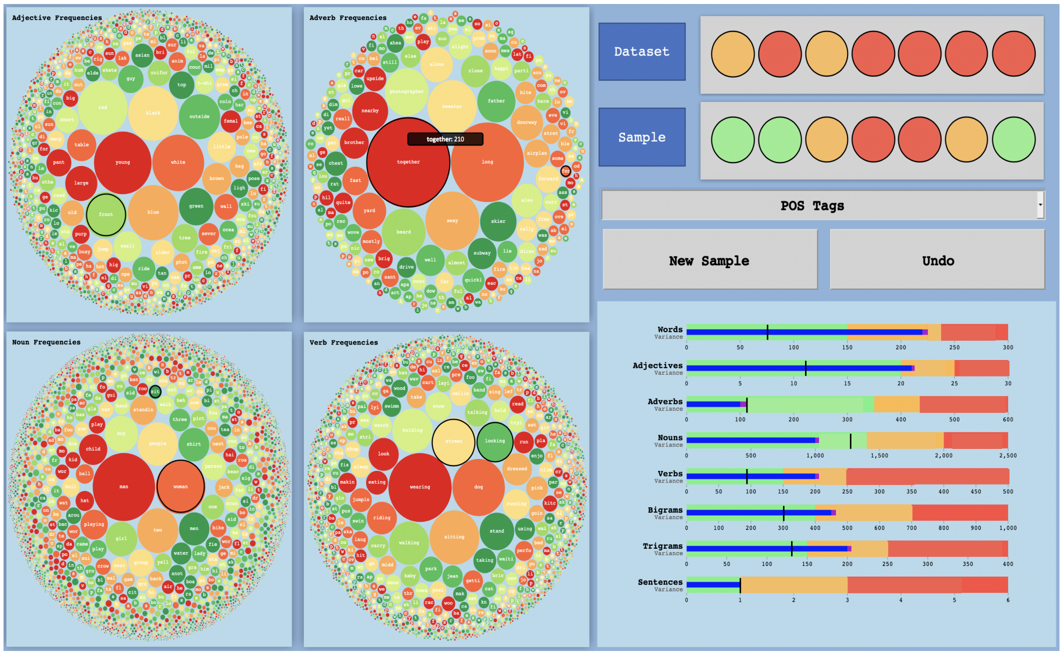


Figure 9: DQI_{c2} Visualization On New Sample Addition

standard deviation bounds of that granularity. The vertical black line represents the ideal value of the standard deviation of that granularity. The two horizontal bars represent the value of standard deviation before and after the new sample's addition. Figure 8 illustrates the visualization.

Interactions: A tooltip, buttons, and a drop down are used for interactions. The tooltip displays the quantities in both charts on mouseover, and the buttons/drop down are used to update the chart. The following tasks are supported by the latter.

- **Changing Granularity *(Drop Down)*:** The drop down menu is used to select the granularity of the bubble chart displayed, as shown in Figure 8.
- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and an updated bubble chart of the word frequency distribution is generated. The new words that are added/ existing words that are updated are highlighted with thick black outlines in the chart. The granularity of the view can be changed using the drop down. The additions/modifications in the frequency distribution are similarly highlighted across all granularities, as illustrated in Figure 9. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

9.4.3 Inter-sample STS

Which Characteristics of Data are Visualized? The main units used in this DQI component are the similarity values between sentences across the dataset. This refers to either premise or hypothesis statements, relative to all other premise/hypothesis statements. In order to understand the similarity relations of sentences, a force layout and horizontal bar chart are used. This is illustrated in Figure 10.

Force Layout for Similar Sentence Pairs In the force layout, those sentence pairs with a similarity value that meets the minimum threshold are connected. Each node represents a sentence. The thickness of the connecting line depends on how close the similarity value is to the threshold.

Horizontal Bar Chart for Most Similar Sentences In the horizontal bar chart, the sentences that are most similar to the given sentence are ordered in terms of their similarity value. The bar colors are centered around the threshold.

Interactions: Interactions via tooltip display the sentence id- i.e., the sample id, and whether the sentence is a premise/hypothesis of that sample- and similarity value in case of both the charts. The two charts are also linked on click of a node in the force layout. Other interactions are fuelled by buttons. The complete set of tasks is as follows:

- **Displaying Horizontal Bar Chart *(on node click)*:** By selecting a node in the force layout, a horizontal bar chart is produced, that displays the ten most similar sentences to the sentence represented by the node. The benefits of the bar chart are two-fold. First, the bar chart accounts for sentence links not present in the force layout. It displays those sentences whose similarity value is below the minimum threshold. This can help if certain sentences are isolated without links in the force layout. Second, it enhances the readability of information present in the force layout by drilling down on a subset, if the dataset size is very large.
- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and two new nodes are created in the force layout. The outline of these two nodes is in black, and by default, the premise is auto-selected to generate the bar chart. If the new sample's sentences appear in the bar chart for any other sample, then the outline of those bars is in black, as illustrated in Figure 11. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

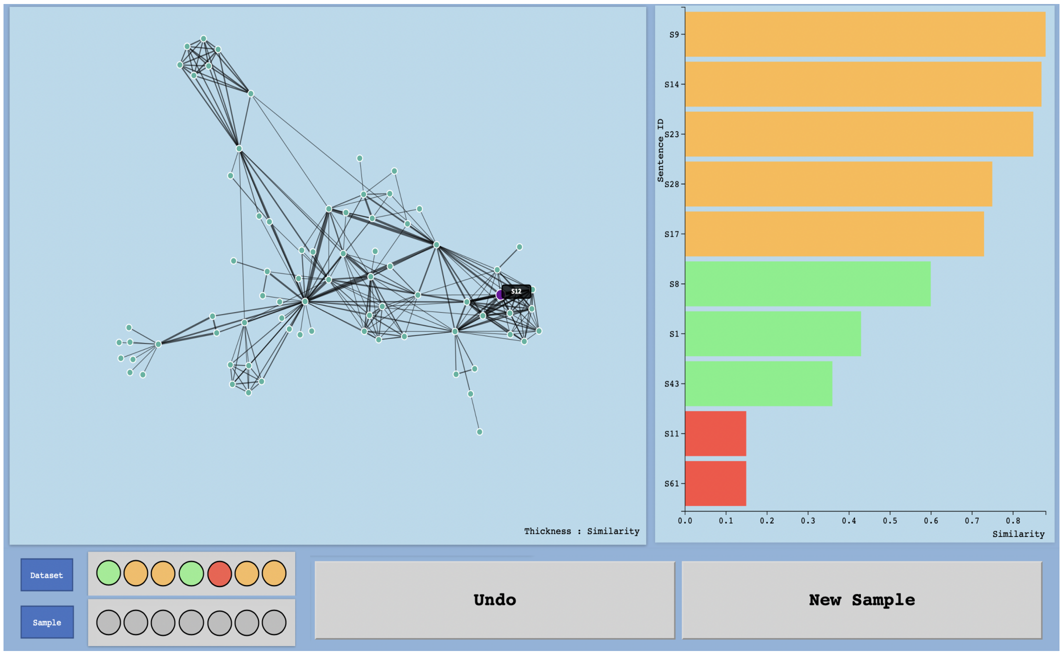


Figure 10: DQI_{c3} Visualization Prior to New Sample Addition

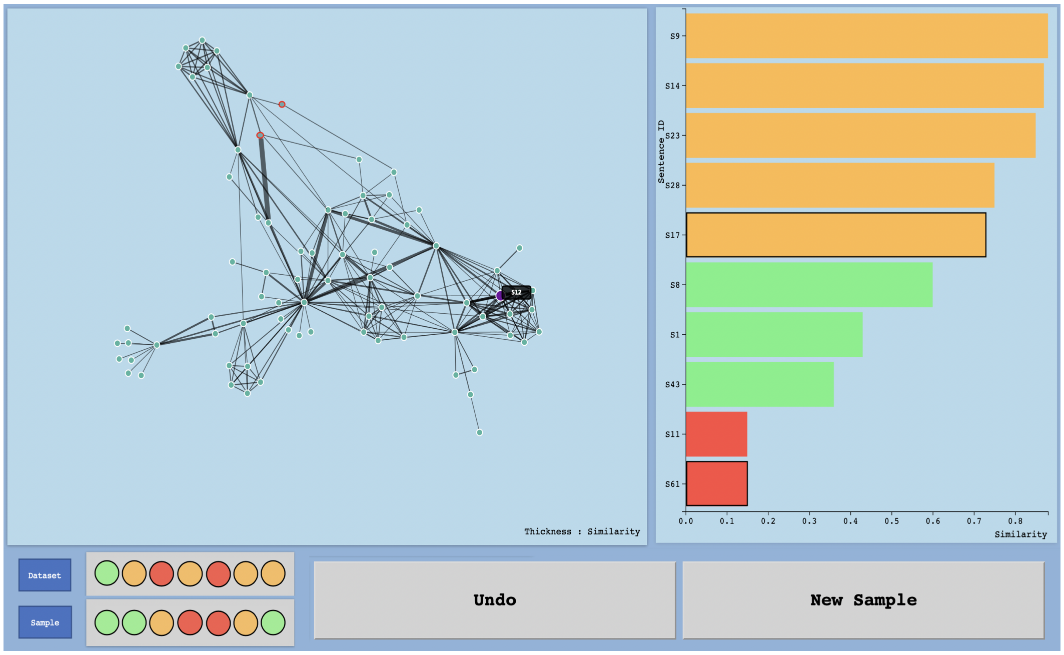


Figure 11: DQI_{c3} Visualization On New Sample Addition

9.4.4 Intra-sample Word Similarity

Which Characteristics of Data are Visualized? In this section, A sample's word similarity is viewed in terms of premise-only, hypothesis-only, and both. The relationship between non-adjacent words in the sample's sentences is analyzed specifically.

Overview Chart for Average Word Similarities and Heatmap for Single Sample The overview chart that is used is a one-level tree map, which uses the average value of all word similarities per sample- i.e., concatenated premise and hypothesis- to color and group its components. This is

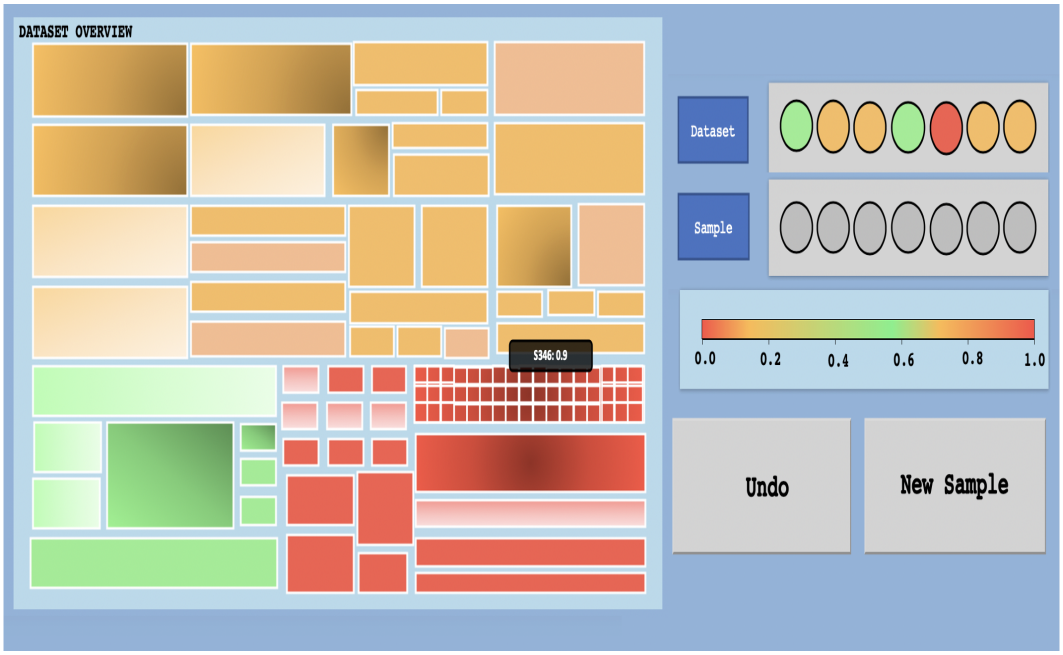


Figure 12: DQI_{c4} Visualization Prior to New Sample Addition

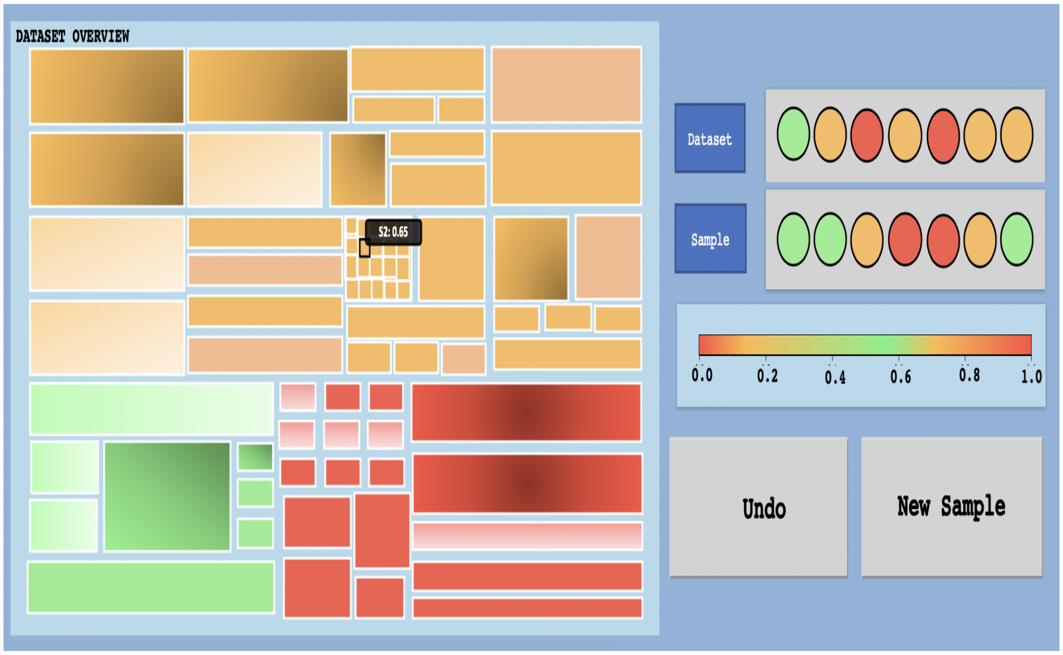


Figure 13: DQI_{c4} Visualization On New Sample Addition: Dataset View

illustrated in Figure 12 The detailed view is a heat map of all the words in a single sample, ass shown in Figure 14.

Interactions: Tooltips display the sample id for the tree map, and the similarity value between words for the heat map. Other interactions include a drop down used to select the sentence to be viewed in the heat map, linking the heat map to the tree map on click, and buttons to modify the visualizations. The tasks are as follows:

- **Displaying Heat Map *(on Tree Map click)*:** By clicking on a box of the tree map, the user is shown the heat map of the clicked on sample.

- **Displaying the Tree Map *(on Heat Map click)*:** By clicking anywhere on the heat map, the user is taken back to the tree map view.
- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and a new box is added to the tree map, with a black outline to highlight it, as illustrated in Figure 13. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.
- **Change Heat Map View *(Drop Down)*:** Using the drop down, the heatmap can be changed to show word similarities for the (a) premise, (b) hypothesis, or (c) both sentences.

9.4.5 Intra-sample STS

Which Characteristics of Data are Visualized? Premise-Hypothesis similarity is analyzed on the basis of length variation, meeting a minimum threshold, and similarity distribution across the dataset. The first is addressed already in the vocabulary property by viewing the sentence length distribution. The other two are visualized using a histogram and kernel density estimation curve, as shown in Figure 15.

Histogram and Kernel Density Curve for Sample Distribution The histogram represents the distribution of the samples, and is colored by centering around the threshold as the ideal value. The number of bins can be changed, and therefore multi-level analysis can be conducted. The kernel density curve is used to check for the overall skew of the distribution.

Interactions: Tooltips on the histogram display the number of samples per bin. Buttons and a text box are used for implementing other interactions:

- **Re-binning Histogram *(textbox)*:** By filling a new value in the textbox, the number of bins in the histogram changes to that value.
- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, the histogram and density plot are updated accordingly. The bar in the histogram to which the sample contributes is outlined in black across all histogram binnings, as illustrated in Figure 16. The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

9.4.6 N-Gram Frequency per Label

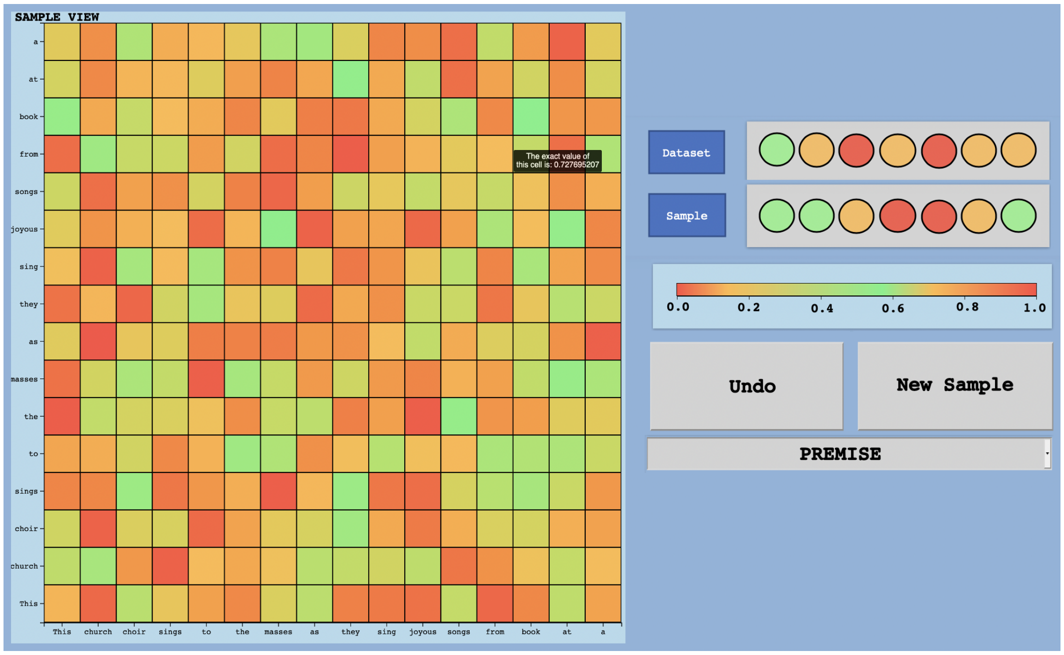
Which Characteristics of Data are Visualized? This component drills down on the second component, to view the patterns seen in granularities per label. There are two small multiples charts, divided based on label, used in this view- a violin plot and a box plot.

Violin plot and Kernel Density Curve for Skew of Distribution: The violin plots are structured to display both jittered points, according to their frequency distribution, as well as a kernel density curve to judge the skew of the distribution. The points each represent an element of the granularity.

Box Plots for More Information The box plots are used to garner more information about the distribution, in terms of its min, max, median, mean, and inter quartile range. These help further characterize the distribution, as well as provide a quantitative definition of the skew seen using density curves. Jittered points representing elements are present in this plot as well.

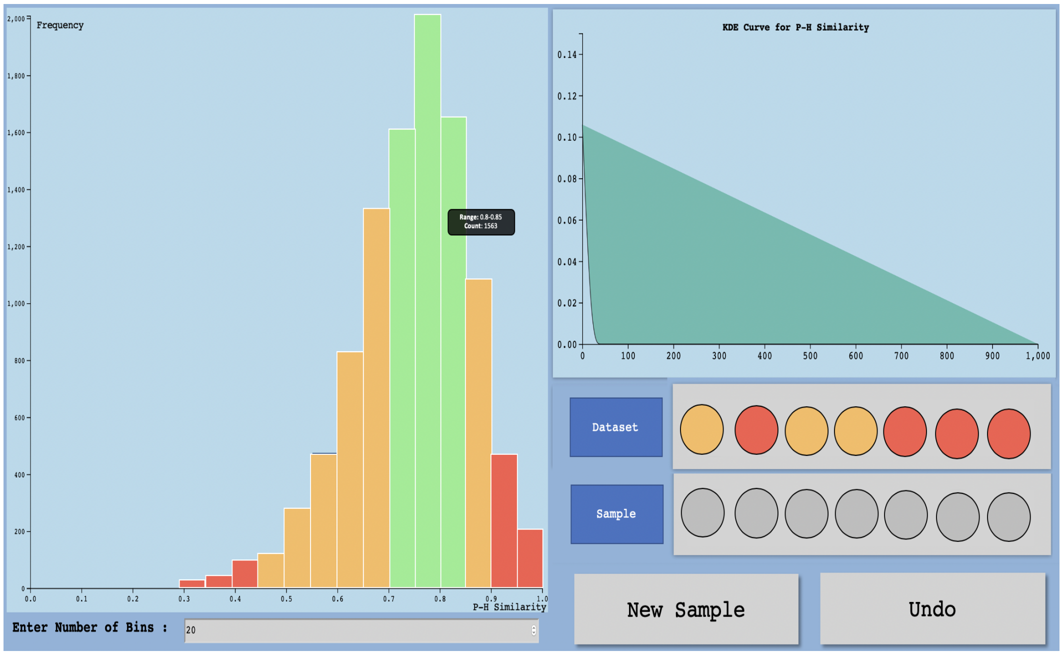
Interactions: On mouseover of a point in both visualizations, the element and its frequency are displayed in a tooltip. Other interactions are based on a dropdown and buttons as follows:

- **Changing Granularity *(Drop Down)*:** The drop down menu is used to select the granularity of the plots displayed, as shown in Figure 17. This granularity can be in terms of words, POS tags, bigrams, trigrams, or sentences.



The exact value of this cell is: 0.727695207

Figure 14: DQI_{c4} Visualization On New Sample Addition: Sample View



Range: 0.8-0.85
Count: 1563

Enter Number of Bins : 20

Figure 15: DQI_{c5} Visualization Prior to New Sample Addition

- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and updated plots of the word frequency distribution are generated. The new words that are added/ existing words that are updated are highlighted with thick white outlines in the chart. The granularity of the view can be changed using the drop down. The additions/modifications in the frequency distribution are similarly highlighted across all granularities. This is shown in Figure 19 and 20 .The component value panels are updated as well. The previous state of the visualization is saved in a set of variables.

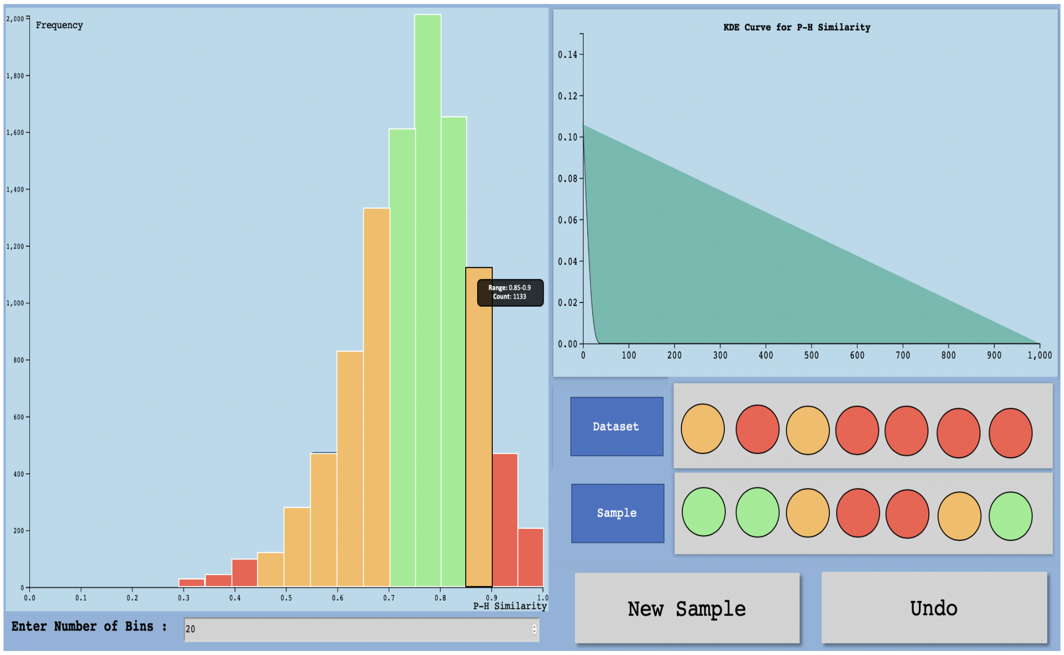


Figure 16: DQI_{c5} Visualization On New Sample Addition

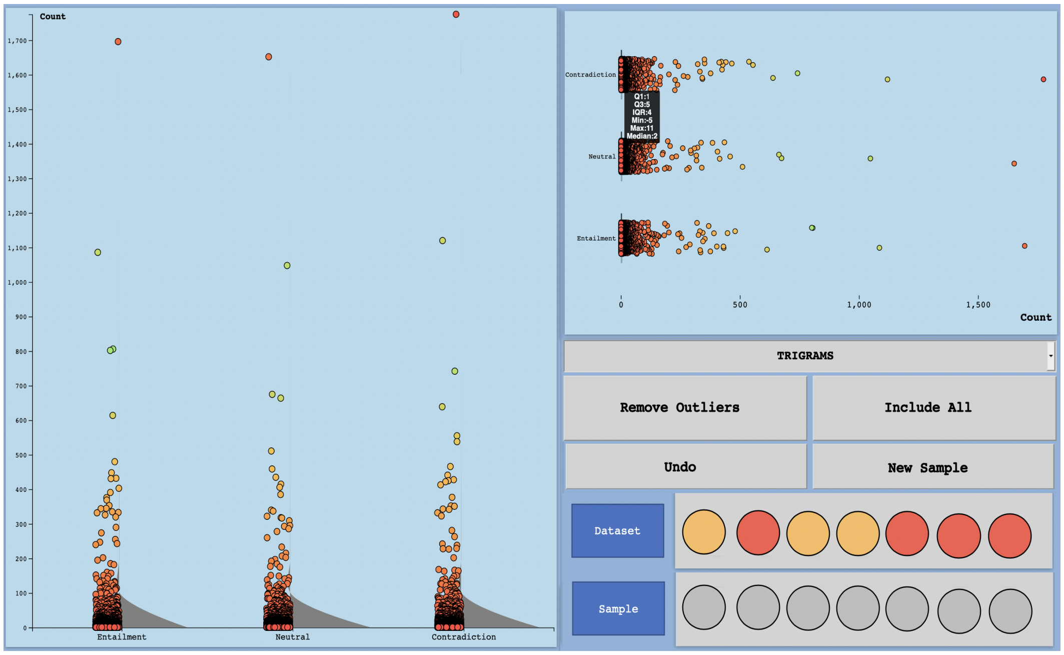


Figure 17: DQI_{c6} Visualization Prior to New Sample Addition

- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.
- **Outlier Handling *(Remove Outliers)*:** This removes elements with frequency counts less than the median to get a less skewed picture of the remainder of the distribution. The component value panels are updated as well, as illustrated in Figure 18. The previous state of the visualization is saved in a set of variables.

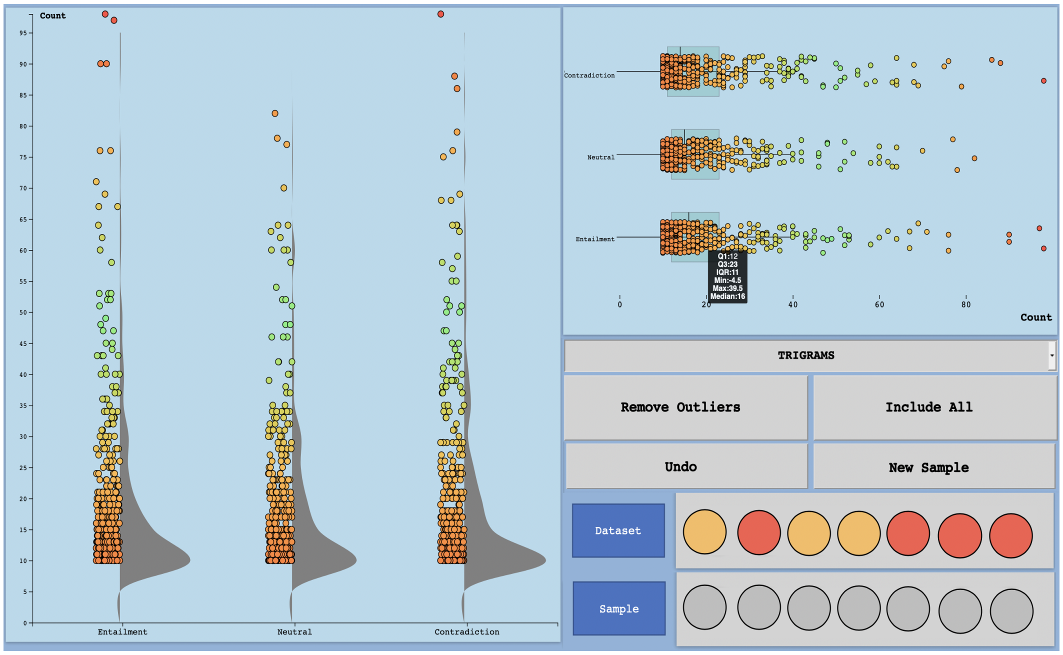


Figure 18: DQI_{c6} Visualization after removing outliers Prior to New Sample Addition

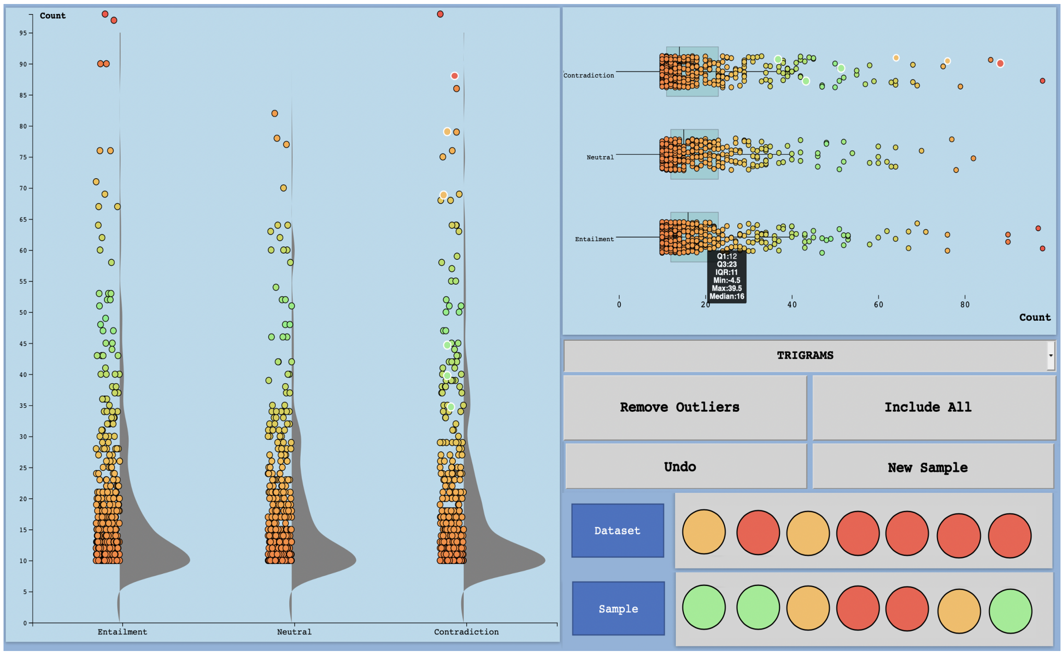


Figure 19: DQI_{c6} Visualization On New Sample Addition

- **Full Distribution View *(Include All Samples)*:** This reverses the operations of ’outlier handling’ by using the saved state variables to restore the visualizations back to their original state.

9.4.7 Inter-split STS

Which Characteristics of Data are Visualized? Train-Test similarity must be kept minimal to prevent data leakage. This component’s main feature is finding the train split sample that is most similar to a given test split sample.

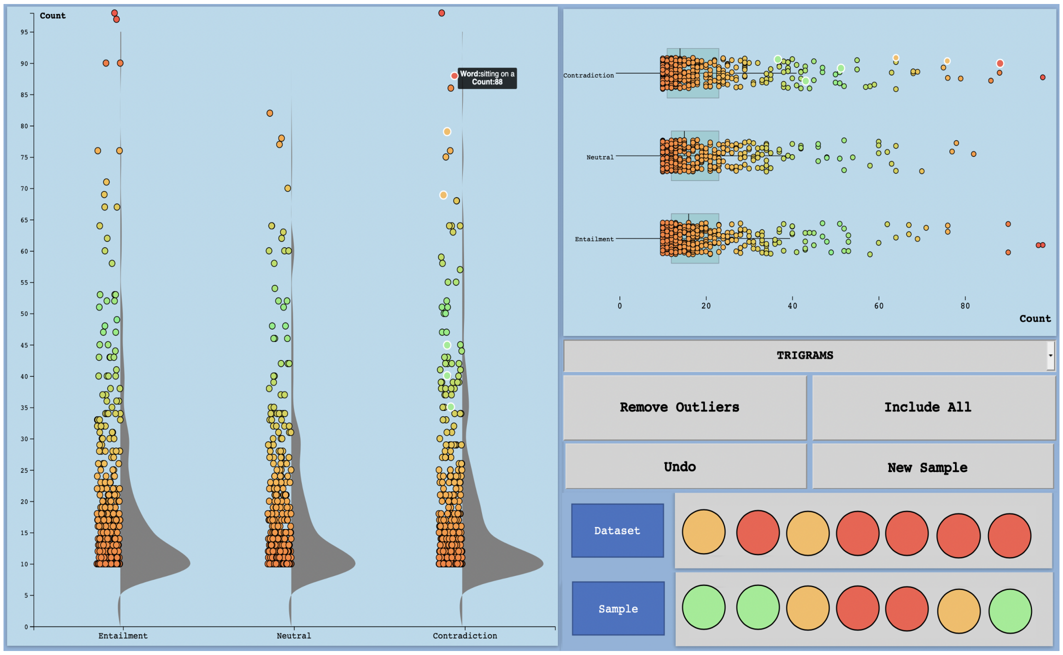


Figure 20: DQI_{c6} Visualization with mouseover On New Sample Addition

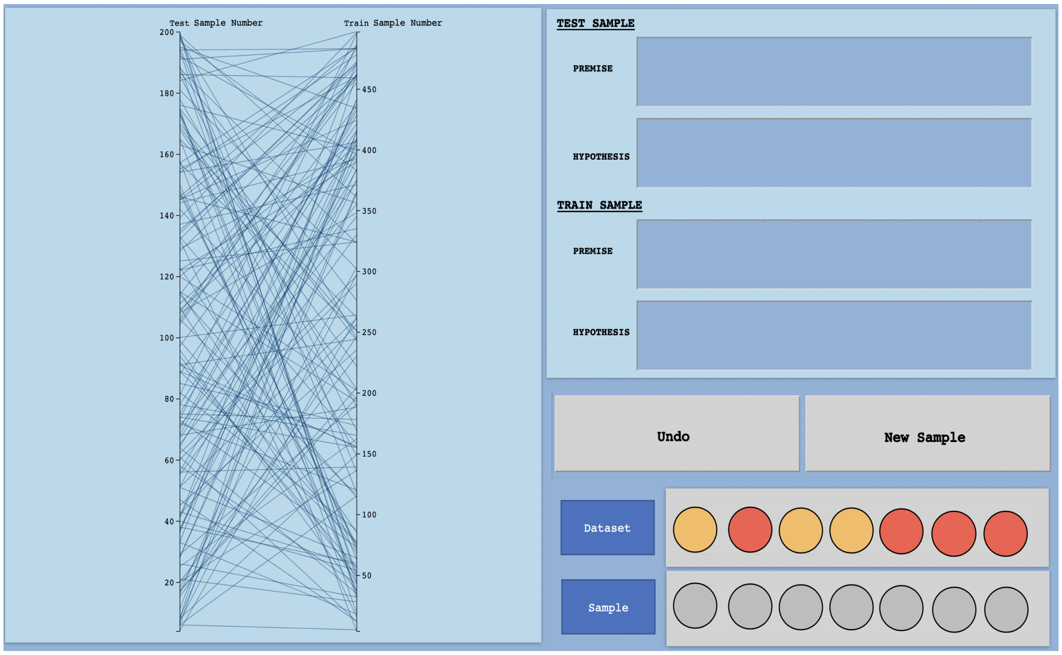
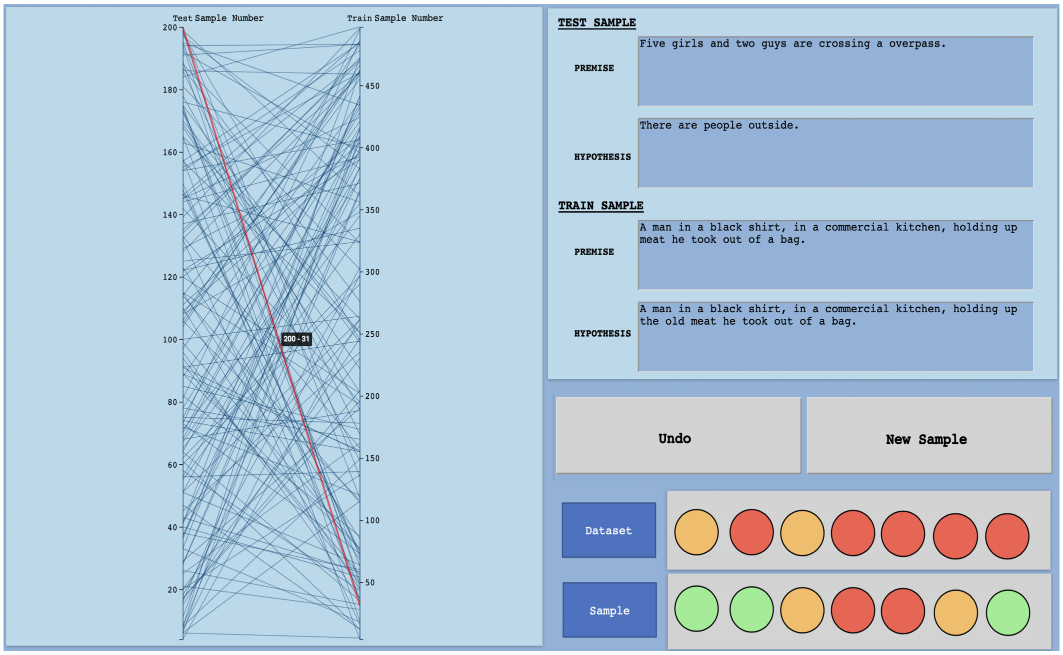


Figure 21: DQI_{c7} Visualization Prior to New Sample Addition

Parallel Coordinate Graph for Train-Test Similarity: A subset of test and train samples, all found to have close similarity within their respective splits, and significant similarity across the splits are plotted as a one step parallel coordinate graph, with test samples along one axis, and train samples along the other. This subset is seeded with those samples closest in similarity to the new sample to be introduced, based on the third component's visualization. The links connecting points on the two axes are drawn between the most similar matches across the split, as shown in Figure 21.

Interactions: Interactions include a tooltip that displays the sample ids connected on mouseover of a link, text boxes filled on click of a link, and other tasks by buttons:



TEST SAMPLE

PREMISE: Five girls and two guys are crossing a overpass.

HYPOTHESIS: There are people outside.

TRAIN SAMPLE

PREMISE: A man in a black shirt, in a commercial kitchen, holding up meat he took out of a bag.

HYPOTHESIS: A man in a black shirt, in a commercial kitchen, holding up the old meat he took out of a bag.

Dataset 🟠 🔴 🟠 🔴 🔴 🔴 🔴

Sample 🟢 🟢 🟠 🔴 🔴 🟠 🟢

Figure 22: DQI_{c7} Visualization On New Sample Addition

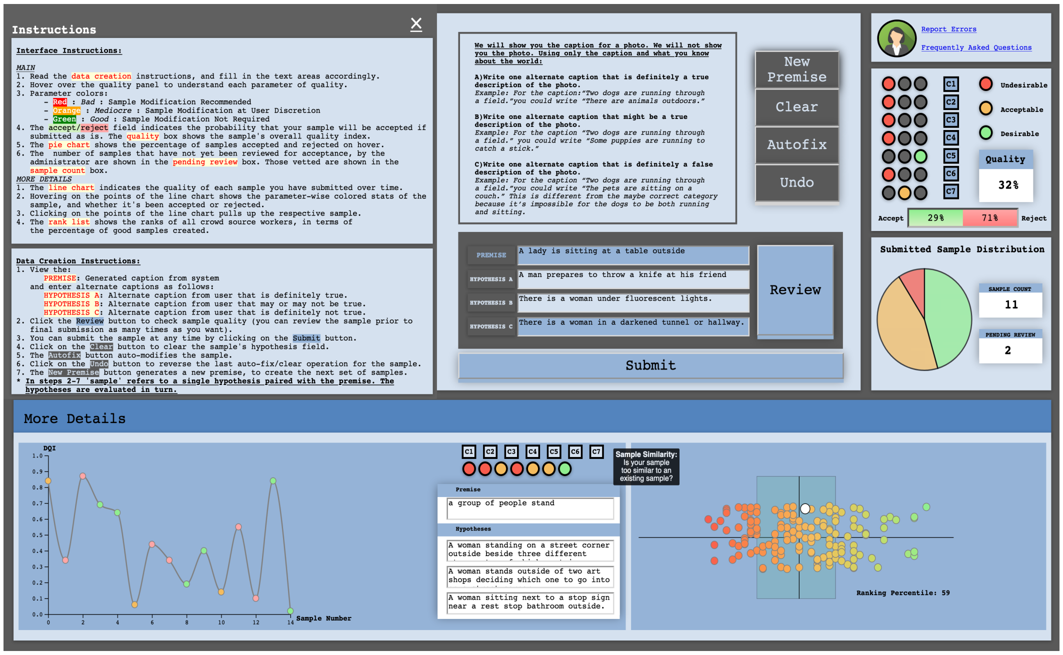
- **Details of Linked Pair *(on click of link)*:** Clicking on a link causes the link to turn red, and the premises and hypotheses of the two samples are displayed in the text boxes on the screen. Clicking on another link changes the values of the textboxes, and highlights only the new link.
- **Addition of a New Sample *(New Sample)*:** The new sample is added to the dataset, and the sample is added to the axis of the parallel coordinates plot depending on the split that it belongs to, as determined by the component one visualization. The sample's link is auto-selected and the textboxes are accordingly updated. The component value panels are updated as well, as illustrated in Figure 22. The previous state of the visualization is saved in a set of variables.
- **Removal of a New Sample *(Undo)*:** This reverses the operations of 'addition of a new sample' by using the saved state variables to restore the visualizations back to their original state.

UI for Data Creation and Valiation: The UI design is two-fold. It targets two aspects of data creation- crowd source worker creation, and analyst review. The first phase uses colored flags to provide feedback to a crowd source worker about the quality of the sample they have created, so that they can fix it manually/with autofix assistance before submitting for higher return. The second phase uses the data visualizations to help the analyst determine if the sample should be added, rejected, or fixed.

9.4.8 Crowd-Source Worker:

The design choices made are heavily focused on the notion of providing simple, yet critical feedback to the crowd source worker, to enhance the quality of data created by means of minimizing spurious bias. The methods and principles used in building the interface used for SNLI's [1] data collection process are the basis of our interface design. There are two types of feedback given in the UI, pre-submission and post-submission of the sample.

Instructions A sliding panel instruction tab is on the left corner of the screen. It consists of two sets of instructions. The first set goes over all general interface functionality descriptions, including post-submission user feedback. The second set specifically focuses on the pre-submission feedback loop.



Instructions

Interface Instructions:

MAIN
1. Read the data creation instructions, and fill in the text areas accordingly.
2. Hover over the quality panel to understand each parameter of quality.
3. Parameter colors:
   - Red : *Bad* : Sample Modification Recommended
   - Orange : *Mediocre* : Sample Modification at User Discretion
   - Green : *Good* : Sample Modification Not Required
4. The accept/reject field indicates the probability that your sample will be accepted if submitted as is. The quality box shows the sample's overall quality index.
5. The pie chart shows the percentage of samples accepted and rejected on hover.
6. The number of samples that have not yet been reviewed for acceptance, by the administrator are shown in the pending review box. Those vetted are shown in the sample count box.

MORE DETAILS
1. The line chart indicates the quality of each sample you have submitted over time.
2. Hovering on the points of the line chart shows the parameter-wise colored stats of the sample, and whether it's been accepted or rejected.
3. Clicking on the points of the line chart pulls up the respective sample.
4. The rank list shows the ranks of all crowd source workers, in terms of the percentage of good samples created.

Data Creation Instructions:
1. View the:
   PREMISE: Generated caption from system
   and enter alternate captions as follows:
   HYPOTHESIS A: Alternate caption from user that is definitely true.
   HYPOTHESIS B: Alternate caption from user that may or may not be true.
   HYPOTHESIS C: Alternate caption from user that is definitely not true.
2. Click the Review button to check sample quality (you can review the sample prior to final submission as many times as you want).
3. You can submit the sample at any time by clicking on the Submit button.
4. Click on the Clear button to clear the sample's hypothesis field.
5. The Autofix button auto-modifies the sample.
6. Click on the Undo button to reverse the last auto-fix/clear operation for the sample.
7. The New Premise button generates a new premise, to create the next set of samples.

* **In steps 2-7 'sample' refers to a single hypothesis paired with the premise. The hypotheses are evaluated in turn.**

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

A)Write one alternate caption that is definitely a true description of the photo.
Example: For the caption:"Two dogs are running through a field."you could write "There are animals outdoors."

B)Write one alternate caption that might be a true description of the photo.
Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."

C)Write one alternate caption that is definitely a false description of the photo.
Example: For the caption "Two dogs are running through a field."you could write "The pets are sitting on a couch." This is different from the maybe correct category because it's impossible for the dogs to be both running and sitting.

PREMISE: A lady is sitting at a table outside
HYPOTHESIS A: A man prepares to throw a knife at his friend
HYPOTHESIS B: There is a woman under fluorescent lights.
HYPOTHESIS C: There is a woman in a darkened tunnel or hallway.

Accept 29% 71% Reject

Quality 32%

Submitted Sample Distribution

SAMPLE COUNT 11
PENDING REVIEW 2

More Details

Premise
a group of people stand

Hypotheses
A woman standing on a street corner outside beside three different
A woman stands outside of two art shops deciding which one to go into
A woman sitting next to a stop sign near a rest stop bathroom outside.

Sample Similarity: Is your sample too similar to an existing sample?

Ranking Percentile: 59

Figure 23: Crowd Source Worker View

Pre-Submission Feedback Loop: After reviewing the main instruction panel, the user can begin data creation. There is an instructions box displayed at all times on the main creation panel, which gives examples used in the original SNLI interface design, to make users understand the nature of the samples they are required to create. The premise field is auto-filled with captions from the Flickr30k corpus. This field can be changed to a fresh premise at any time by clicking on the 'new premise' button. The 3 types of hypothesis (entailment, neutral, and contradiction) must be entered in their respective fields.

DQI based on past history Following this, each hypothesis is evaluated individually with the premise. Henceforth, the use of the term sample denotes premise and only the hypothesis under consideration. The hypothesis under consideration can be cleared at any time by clicking the 'clear' button. The user must click the 'Review' button at least once before submitting. The 'Review' button populates the DQI indication panel, which displays the values of the DQI components with respect to both the newly created sample and the existing set of accepted samples. The general aspect of data that is being analyzed by a component can be viewed on a tooltip, on mouseover of the component label. The messages displayed are as follows:

- Vocabulary: Does your sample contribute new words?
- Combinations: Does your sample contribute new combinations of words and phrases?
- Sentence Similarity: How similar is your hypothesis to all other premises or hypotheses?
- Word Similarity: How similar are all the words within your sample?
- PH Score: How similar is your hypothesis to the premise?
- Label Giveaway: Is your hypothesis too obvious for our system?
- Sample Similarity: Is your sample too similar to an existing sample?

Feedback Flags The values of the DQI components are indicated using a traffic signal analogy (red, yellow, and green), thereby indicating if a particular aspect of the data created might lead to bias. The colors respectively advise the user to stop, revise, and proceed in their sample creation tactics. The probability of the newly created sample being accepted/rejected is also displayed. Based on this feedback, the user can choose to: (i) manually fix their sample and review it again, (ii) 'auto-fix' the sample by paraphrasing it using concept net, (iii) submit the sample as is. Once the user is satisfied with the sample created, they can submit the sample. Once the sample has been submitted, the ’pending review’ box is accordingly updated, as is the ’count’ box for total number of submitted samples.

Post-Submission Feedback Loop: We retain the notion of a background expert reviewing samples to ensure that the sentences use appropriate ideas and language. Once the analyst reviews the sample and marks it as accepted/rejected (see section 8.2), the following updates occur on the crowdsource worker’s UI [9] :

- The line chart on the secondary panel indicates the quality of the user’s submitted samples over time. It is color coded according to whether the sample was accepted or rejected. On hovering over any one sample, the quality level of that sample are displayed on a tooltip. On click the sample appears in a text box.
- The ’pending review’ box count on the main panel is decremented by one.
- The ranks are displayed using a box plot that calibrates ranks based on the percentage of accepted samples created by each user.
- The pie chart on the main panel is updated according to the accept/reject percentages.

Additional Communication Links: There are additional FAQ and Reporting Problem links present in the interface. The FAQs deal with data creation guidelines, and the Reporting Problems form is intended for technical issues only. This is in accordance with similar functionalities from the original SNLI interface. Figure 23 illustrates the crowdsource worker’s UI.

9.4.9 Analyst:

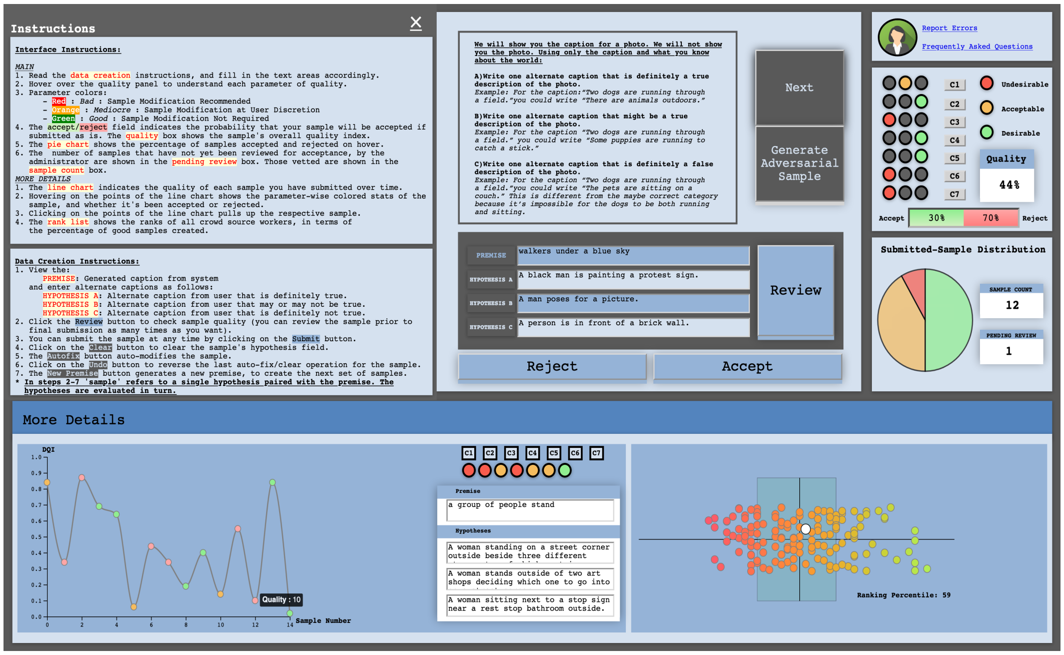
Analysts’ basic interface similar to crowdsource workers’: The analyst interface is focused on the data validation process. The layout of the interface follows the same pattern as that of the crowd source workers interface. This is done so that the analyst understands the environment presented to the crowd source worker for data creation. The sliding panel for instructions, data entry boxes, DQI indication panel, and communication links are retained as is. The piechart, count box, pending review box, line chart, and rank box plot change depending on the annotator id associated with the sample being evaluated, as they represent the performance of that particular annotator.

Review Button The ’Next’ buttons loads the next created sample set that must be reviewed. The text fields are filled with the premise and all hypotheses statements matching that premise. On clicking ’Review’, the analyst reviews each hypothesis paired with the premise individually, as done in the crowdsource worker interface.

Buttons for Appropriate Visualizations: The DQI indication panel has buttons that link to each component’s respective visualization. There are buttons present instead of labels for each component in this panel that can be used to navigate to each visualization in turn. The sample considered in the visualizations as the ’new sample’ is the sample that is under review.

Data Validation The ’Accept’ button can be used to accept the sample as is, and causes the piechart, pending review box, count box, rank box plot, and line chart for the annotator of the sample to be updated. The ’Reject’ button is used mainly to discard samples that contain obscenities, have incoherent/ungrammatical hypothesis statements, and have hypothesis statements of length less than three words. If the sample has low quality, but can be converted to a higher quality adversarial sample with some modification and resubmitted, the ’Generate Adversarial Sample’ button sends the sample to Text-Fooler. Samples that are auto-fixed at the analyst end in this manner are displayed as the yellow slice of the pie chart. Crowdsource workers receive lesser rewards for these samples. Figure 24 illustrates this.

[9] these updates are only loaded at the start of each new user login session



Interface Instructions:

MAIN
1. Read the data creation instructions, and fill in the text areas accordingly.
2. Hover over the quality panel to understand each parameter of quality.
3. Parameter colors:
   - Red : *Bad* : Sample Modification Recommended
   - Orange : *Mediocre* : Sample Modification at User Discretion
   - Green : *Good* : Sample Modification Not Required
4. The accept/reject field indicates the probability that your sample will be accepted if submitted as is. The quality box shows the sample's overall quality index.
5. The pie chart shows the percentage of samples accepted and rejected on hover.
6. The number of samples that have not yet been reviewed for acceptance, by the administrator are shown in the pending review box. Those vetted are shown in the sample count box.

MORE DETAILS
1. The line chart indicates the quality of each sample you have submitted over time.
2. Hovering on the points of the line chart shows the parameter-wise colored stats of the sample, and whether it's been accepted or rejected.
3. Clicking on the points of the line chart pulls up the respective sample.
4. The rank list shows the ranks of all crowd source workers, in terms of the percentage of good samples created.

Data Creation Instructions:
1. View the:
   PREMISE: Generated caption from system
   and enter alternate captions as follows:
   HYPOTHESIS A: Alternate caption from user that is definitely true.
   HYPOTHESIS B: Alternate caption from user that may or may not be true.
   HYPOTHESIS C: Alternate caption from user that is definitely not true.
2. Click the Review button to check sample quality (you can review the sample prior to final submission as many times as you want).
3. You can submit the sample at any time by clicking on the Submit button.
4. Click on the Clear button to clear the sample's hypothesis field.
5. The Autofix button auto-modifies the sample.
6. Click on the Undo button to reverse the last auto-fix/clear operation for the sample.
7. The New Premise button generates a new premise, to create the next set of samples.

* **In steps 2-7 'sample' refers to a single hypothesis paired with the premise. The hypotheses are evaluated in turn.**

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

A)Write one alternate caption that is definitely a true description of the photo.
Example: For the caption:"Two dogs are running through a field."you could write "There are animals outdoors."

B)Write one alternate caption that might be a true description of the photo.
Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."

C)Write one alternate caption that is definitely a false description of the photo.
Example: For the caption "Two dogs are running through a field."you could write "The pets are sitting on a couch." This is different from the maybe correct category because it's impossible for the dogs to be both running and sitting.

PREMISE: walkers under a blue sky
HYPOTHESIS A: A black man is painting a protest sign.
HYPOTHESIS B: A man poses for a picture.
HYPOTHESIS C: A person is in front of a brick wall.

Premise
a group of people stand

Hypotheses
A woman standing on a street corner outside beside three different
A woman stands outside of two art shops deciding which one to go into
A woman sitting next to a stop sign near a rest stop bathroom outside.

Figure 24: Analyst View

Task	**Description**	Component
New Sample	Adds the sample under review to dataset and updates visualizations.	All
Undo	Removes sample under review from dataset and updates visualizations.	All
Randomize Split	Randomized re-sampling of data across splits in a 70:10:20 ratio.	Vocabulary
Undo Split	Reverses last random split generated.	Vocabulary
Save Split	Freezes split for the remainder of analysis.	Vocabulary
Changing Granularity	View granularity can be changed by selecting drop down option.	Inter-sample N-gram Frequency and Relation, N-Gram Frequency per Label
Change Heat Map View	Using the drop down, the heatmap shows word similarities for the (a) premise, (b) hypothesis, or (c) both sentences.	Intra-sample Word Similarity
Rebinning Histogram	By filling a new value in the textbox, the number of bins in the histogram changes to that value.	Intra-sample STS
Remove Outliers	Removes elements with frequency count less than median count of granularity being viewed.	N-Gram Frequency per Label
Include All Samples	Displays all elements for a granularity.	N-Gram Frequency per Label

Table 3: Task Descriptions for Visual Interfaces

9.5 Evaluation

Test cases have been developed to show the efficacy of DQI in our proposed data creation paradigm, with varying numbers of preexisting samples. We tune the hyperparameters proportionally, based on the dataset size. The value ranges for the DQI component colors are also set accordingly. DQI has been calculated for the following cases:

(i) No Preexisting Samples

(ii) 100 Preexisting Samples from the Good Split of the SNLI Test Set

In case (i), DQI of the new sample is calculated. In case (ii), first, DQI for the preexisting sample set is computed, as x_1. Then, the new sample is added and DQI is recalculated for the updated sample set, as x_2. The new samples, shown in Table 4, have been taken from a recent work on adversarial filtering, AFLite.

Then, the difference $\Delta x = x_1 - x_2$ is calculated. On the main interface, the crowd source worker views the colors of DQI components corresponding to Δx. The analyst views Δx as 'Sample' and x_2 as 'Dataset' component colors on the visualizations.

9.5.1 Case(i) - Addressing Cold Start

Case (i) addresses the situation of cold-start for DQI. Unlike adversarial filtering algorithms, DQI can be used even with low data levels. In the situation of cold start, the component initialization is as follows:

Vocabulary: The first term is scaled appropriately as it takes the size of the dataset into account. The second term returns the standard deviation between the premise and hypothesis lengths. Since the third term defines upper and lower bounds on sentence length, it takes a value of one as long as the lengths of both the premise and hypothesis statements exceed three words, and zero if it is three words or less, as seen for sample 5 in Table 9.

Inter-sample N-gram Frequency and Relation: Term 1 captures the inverse of standard deviation, and hence yields infinity in the case of POS tags, when a word with that POS tag does not occur at all, or only occurs once as standard deviation tends to zero. In some cases, the standard deviation can be zero, as seen in Table 17 for trigrams, as each trigram occurs an equal number of times. High non-infinite values for term one are seen for bigrams and trigrams due to their balanced distributions in a sample, as in Table 20.

Sentences are seen to differ across samples in terms of the language used, and their length. Therefore, when setting the upper and lower bounds of granularities for Term 2, standardizing the bounds for cold start fails in the case of POS tags, particularly adverbs, as in seen Tables 10 - 21. These bounds therefore need to be reset at cold start particular to the sample's language.

Inter-sample STS: The first term focuses on the standard deviation of similarity values that cross a threshold between all sentences. Since there is only one similarity value calculated, the value of Term 1, as in Table 24, is set to that similarity value to prevent it from becoming infinity. The second term is always taken to have a value of 2, as there is no definite set threshold for taking a maximum.

Intra-sample Word Simlarity: The fourth component scales appropriately, as it takes the size of the dataset into account and can therefore be directly computed, as in Table 24.

Intra-sample STS: The first term, in Table 23, deals with whether the Premise-Hypothesis similarity crosses a threshold. This scales as it takes dataset size into account, and can be calculated for different threshold values. The second and third terms, Table 22, involve the calculation of the mean and standard deviation of length difference between the premise and hypothesis. Therefore, the second term is directly computed, while the third is always zero, since only one value is present. The fourth term's value, in Table 22, also uses standard deviation and is directly taken to be the similarity between the premise and hypothesis, as only one value is calculated. The fifth and sixth terms look at word overlap and word similarity levels between the premise and hypothesis, and can be directly calculated. These are represented in Tables 42 - 45.

N-gram Frequency per Label: Since cold start only involves the text data of a single sample, the label of that sample is the only one with initialized values in DQI_{C6}. Table 23 has Terms 1 and 2 of DQI_{C6}, as they are equivalent to the terms of DQI_{C2} for the label of the new sample. These terms are set to zero for the other two labels. Table 22 has Terms 3 and 4, which are the same as terms 2 and 3 of DQI_{C5}, and are only computed for the label of the new sample. Also, since the counts of all granularities are only initialized for a single label, the fifth term is set to zero for all samples.

Inter-split STS: Since DQI_{C7} is calculated on the basis of the most similar training sample for every test set sample, it is not applicable to the case of cold start, as there is only one sample. Hence, its value is taken as zero.

9.5.2 Case(ii)-Adding to the Test Good Split

A 100 samples are taken at random 10 times from the good split of the SNLI Test set and x_1 is calculated. Then the new sample is added to the dataset. x_2 and Δx are calculated. For all components, DQI values are calculated using the same hyperparameter values as those used for the full test set. The results, shown in Tables 26 - 41, indicate the need for hyperparameter scaling.

What requires Scaling? From tables 27 and 33-36, we find that hyperparameters used to set upper and lower bounds for POS tag frequencies across and within labels require significant scaling. Additionally, we find that sentence, bigram, and trigram terms should be omitted when calculating the DQI until their overall frequencies and variance reach a certain threshold. This is because terms inversely proportional to the standard deviation of the distributions of those granularities are found to explode for lesser numbers of samples.

9.5.3 Assigning Colors

The new sample set has six samples removed by AFLite, that from the bad split of the Dev set, and six that are retained, i.e.,from the good split of the Dev set. In both case (i) and case (ii), we find that on adding samples to the existing dataset, there is no significant difference in the term/component values except in the cases of word overlap and word similarity, seen in T5 and T6 of DQI_{C5}. We observe that DQI component colors are correctly predicted 10/12 times on an average. Also, the change in DQI_{C5} corresponding to word overlap and word similarity is as expected as per the findings of AFLite.

Sample ID	Premise	Hypothesis	Label	Split
S1	A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill.	contradiction	Dev-Bad
S2	The dog is catching a treat.	The cat is not catching a treat.	contradiction	Dev-Bad
S3	Three young men are watching a tennis match on a large screen outdoors.	Three young men watching a tennis match on a screen outdoors, because their brother is playing.	neutral	Dev-Bad
S4	A girl dressed in a pink shirt, jeans, and flip-flops sitting down playing with a lollipop machine.	A funny person in a shirt.	neutral	Dev-Bad
S5	A man in a green apron smiles behind a food stand.	A man smiles.	entailment	Dev-Bad
S6	A little girl with a hat sits between a woman's feet in the sand in front of a pair of colorful tents.	The girl is wearing a hat.	entailment	Dev-Bad
S7	People are throwing tomatoes at each other.	The people are having a food fight.	entailment	Dev-Good
S8	A man poses for a photo in front of a Chinese building by jumping.	The man is prepared for his photo.	entailment	Dev-Good
S9	An older gentleman speaking at a podium.	A man giving a speech.	neutral	Dev-Good
S10	A man poses for a photo in front of a Chinese building by jumping.	The man has experience in taking photos.	neutral	Dev-Good
S11	People are waiting in line by a food vendor.	People sit and wait for their orders at a nice sit down restaurant.	contradiction	Dev-Good
S12	Number 13 kicks a soccer ball towards the goal during children's soccer game.	A player passing the ball in a soccer game.	contradiction	Dev-Good

Table 4: SNLI Samples used for Test Cases

Sample ID	**Premise**	**Hypothesis**	**Label**	**Split**
S1	To their good fortune, he's proving them right.	He is showing that they guessed correctly.	entailment	Dev-Good
S2	Strange as it may seem to the typical household, capital gains on its existing assets do not contribute to saving as measured in NIPA.	The increased equity of a house may not be considered as savings by NIPA.	entailment	Dev-Good
S3	Among runners-up is Boston solo Eleanor Newhoff.	Eleanor Newhoff had trained hard for the Olympic triathlon.	neutral	Dev-Good
S4	This was used for ceremonial purposes, allowing statues of the gods to be carried to the river for journeys to the west bank, or to the Luxor sanctuary.	Statues were moved to Luxor for funerals and other ceremonies.	neutral	Dev-Good
S5	Or just a philosophy of any weapon to hand?	They don't allow any weapon.	contradiction	Dev-Good
S6	Diets for men in their prime	A plan to keep men fat.	contradiction	Dev-Good
S7	Justice Kennedy does not care what law librarians across the country Reporters from 1790 through 1998.	Justice Kennedy doesn't care if do with all the Supreme Court the Supreme Court Reporters from 1790 to 1998 are thrown away.	entailment	Dev-Bad
S8	are you originally from uh Texas	You're originally from Texas?	entailment	Dev-Bad
S9	Click here for Finkelstein's explanation of why this logic is expedient.	Click here for Finkelstein's explanation of why this logic is expedient due to philosophical constraints.	neutral	Dev-Bad
S10	Two, most other productive operations are easier to study and understand, since few firms have 40,000 locations and a large proportion of their workforce working outdoors.	The productivity of the operations is directly related to the workforce that's based outdoors.	neutral	Dev-Bad
S11	Treat yourself and bill it to Si.	Don't treat yourself, Si has to pay for that.	contradiction	Dev-Bad
S12	Eh! Monsieur Lawrence, called Poirot.	Poirot did not call upon Monsieur Lawrence.	contradiction	Dev-Bad

Table 5: MNLI Samples used for Test Cases

[t]

Sample ID	Question	Context	Answer	impossible	Split
S1	By how many kilometers are shear waves separated when measuring the crust?	Seismologists can use the arrival times of seismic waves in reverse to image the interior of the Earth. Early advances in this field showed the existence of a liquid outer core (where shear waves were not able to propagate) and a dense solid inner core. These advances led to the development of a layered model of the Earth, with a crust and lithosphere on top, the mantle below (separated within itself by seismic discontinuities at 410 and 660 kilometers), and the outer core and inner core below that. More recently, seismologists have been able to create detailed images of wave speeds inside the earth in the same way a doctor images a body in a CT scan. These images have led to a much more detailed view of the interior of the Earth, and have replaced the simplified layered model with a much more dynamic model.	at 410 and 660 kilometers	True	Dev-Good
S2	Where is Geoffrey Parker from?	The plague repeatedly returned to haunt Europe and the Mediterranean throughout the 14th to 17th centuries. According to Biraben, the plague was present somewhere in Europe in every year between 1346 and 1671. The Second Pandemic was particularly widespread in the following years: 1360–63; 1374; 1400; 1438–39; 1456–57; 1464–66; 1481–85; 1500–03; 1518–31; 1544–48; 1563–66; 1573–88; 1596–99; 1602–11; 1623–40; 1644–54; and 1664–67. Subsequent outbreaks, though severe, marked the retreat from most of Europe (18th century) and northern Africa (19th century). According to Geoffrey Parker, "France alone lost almost a million people to the plague in the epidemic of 1628–31."	France	True	Dev-Good
S3	When was the European Convention on Human Rights established?	None of the original treaties establishing the European Union mention protection for fundamental rights. It was not envisaged for European Union measures, that is legislative and administrative actions by European Union institutions, to be subject to human rights. At the time the only concern was that member states should be prevented from violating human rights, hence the establishment of the European Convention on Human Rights in 1950 and the establishment of the European Court of Human Rights. The European Court of Justice recognised fundamental rights as general principle of European Union law as the need to ensure that European Union measures are compatible with the human rights enshrined in member states' constitution became ever more apparent. In 1999 the European Council set up a body tasked with drafting a European Charter of Human Rights, which could form the constitutional basis for the European Union and as such tailored specifically to apply to the European Union and its institutions. The Charter of Fundamental Rights of the European Union draws a list of fundamental rights from the European Convention on Human Rights and Fundamental Freedoms, the Declaration on Fundamental Rights produced by the European Parliament in 1989 and European Union Treaties.	1950	False	Dev-Good
S4	What did Lavoisier perceive the air had lost as much as the tin had gained?	In one experiment, Lavoisier observed that there was no overall increase in weight when tin and air were heated in a closed container. He noted that air rushed in when he opened the container, which indicated that part of the trapped air had been consumed. He also noted that the tin had increased in weight and that increase was the same as the weight of the air that rushed back in. This and other experiments on combustion were documented in his book Sur la combustion en général, which was published in 1777. In that work, he proved that air is a mixture of two gases; 'vital air', which is essential to combustion and respiration, and azote ("lifeless"), which did not support either. Azote later became nitrogen in English, although it has kept the name in French and several other European languages.	weight	False	Dev-Good

Table 6: SQUAD 2.0 Test Cases - Dev Good

Sample ID	Question	Context	Answer	impossible	Split
S5	Why are normal body cells attacked by NK cells?	Natural killer cells, or NK cells, are a component of the innate immune system which does not directly attack invading microbes. Rather, NK cells destroy compromised host cells, such as tumor cells or virus-infected cells, recognizing such cells by a condition known as "missing self." This term describes cells with low levels of a cell-surface marker called MHC I (major histocompatibility complex) – a situation that can arise in viral infections of host cells. They were named "natural killer" because of the initial notion that they do not require activation in order to kill cells that are "missing self." For many years it was unclear how NK cells recognize tumor cells and infected cells. It is now known that the MHC makeup on the surface of those cells is altered and the NK cells become activated through recognition of "missing self". Normal body cells are not recognized and attacked by NK cells because they express intact self MHC antigens. Those MHC antigens are recognized by killer cell immunoglobulin receptors (KIR) which essentially put the brakes on NK cells.	express intact self MHC antigens	True	Dev-Bad
S6	What did higher material living standards lead to for most of human history?	For most of human history higher material living standards – full stomachs, access to clean water and warmth from fuel – led to better health and longer lives. This pattern of higher incomes-longer lives still holds among poorer countries, where life expectancy increases rapidly as per capita income increases, but in recent decades it has slowed down among middle income countries and plateaued among the richest thirty or so countries in the world. Americans live no longer on average (about 77 years in 2004) than Greeks (78 years) or New Zealanders (78), though the USA has a higher GDP per capita. Life expectancy in Sweden (80 years) and Japan (82) – where income was more equally distributed – was longer.	better health and longer lives	True	Dev-Bad
S7	What happens as they build phase 1?	The owner produces a list of requirements for a project, giving an overall view of the project's goals. Several D&B contractors present different ideas about how to accomplish these goals. The owner selects the ideas he or she likes best and hires the appropriate contractor. Often, it is not just one contractor, but a consortium of several contractors working together. Once these have been hired, they begin building the first phase of the project. As they build phase 1, they design phase 2. This is in contrast to a design-bid-build contract, where the project is completely designed by the owner, then bid on, then completed.	they design phase 2	False	Dev-Bad
S8	When was the Third Assessment Report published?	Another example of scientific research which suggests that previous estimates by the IPCC, far from overstating dangers and risks, have actually understated them is a study on projected rises in sea levels. When the researchers' analysis was "applied to the possible scenarios outlined by the Intergovernmental Panel on Climate Change (IPCC), the researchers found that in 2100 sea levels would be 0.5–1.4 m [50–140 cm] above 1990 levels. These values are much greater than the 9–88 cm as projected by the IPCC itself in its Third Assessment Report, published in 2001". This may have been due, in part, to the expanding human understanding of climate.	2001	False	Dev-Bad

Table 7: SQUAD 2.0 Test Cases - Dev Bad

Sample ID	Story	Ending	Label	Split
S1	Fred receives a specialty coffee maker for Christmas. He finally opens it after leaving it in its box for a few weeks.Fred decides to make himself a cappuccino.To his surprise, it tastes just as good as the ones he buys outside.	Frank will save about $25 a week making coffee himself.	True	Dev-Good
S2	My family is sharing a bowl of popcorn.Mom is reading a book and eating one piece at a time.Dad and I are playing iPad games and eating handfuls at a time.We have played this game before!	Dad and I love popcorn.	True	Dev-Good
S3	I got a job as a shopping mall Santa last December. The hours were long.The pay was bad.But I found interacting with the kids to be completely amazing.	I found that playing Santa was not worth my time off.	False	Dev-Good
S4	Carry has been short her whole life.She could never reach the top shelf at the store.Greg saw her struggling to reach.He went over and helped her.	She refused his help and walked away.	False	Dev-Good
S5	Lou was on a diet.She was eating very little.But she still struggled to lose weight!Then she added an exercise regimen.	Lou was finally able to lose weight.	True	Dev-Bad
S6	Kim had been working extra hard for weeks.She learned of a promotion up for grabs at her company.It came with a new office and great benefits.Finally all her work paid off and she was offered the promotion.	She was happy to get the promotion.	True	Dev-Bad
S7	James has just started working at a company with a ping pong table.He has always wanted to play ping pong with a coworker.One day after work, his friend challenges him to a game.James plays very well, but eventually loses the game.	James was worried because he beat his boss at ping pong.	False	Dev-Bad
S8	Dan loves the sport of bowling.His dad taught him how to play when he was little.The use to compete in tournaments together.His dad has since passed away.	Dan never liked to bowl anyway.	False	Dev-Bad

Table 8: Story CLOZE Test Cases

Sample	Terms			DQI C1
	T1	T2	T3	
S1	0.0693	2.121	1.0000	2.1906
S2	0.0396	0.7071	1.0000	0.7467
S3	0.1089	2.1213	1.0000	2.2302
S4	0.1188	7.7781	1.0000	7.8969
S5	0.06930	5.6568	0.0000	0.0693
S6	0.1188	11.3137	1.0000	11.4325
S7	0.0594	0.0000	1.0000	0.0594
S8	0.0792	4.9497	1.0000	5.0289
S9	0.0693	1.4142	1.0000	1.4835
S10	0.0891	4.9497	1.0000	5.0388
S11	0.0990	2.8284	1.0000	2.9274
S12	0.1089	2.8284	1.0000	2.9373

Table 9: DQI_{C1} for Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	13.0958	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	4.0000	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	15	32.7698	0.1578	0
Trigrams	16	64.0000	0.7647	0

Table 10: $DQI_{C2}$and DQI_{C6} (contradiction) for S1, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	4	6.9282	1.0000	0
Adjectives	0	nan	nan	0
Adverbs	0	nan	nan	0
Verbs	1	inf	1.0000	0
Nouns	3	6.3639	1.0000	0
Bigrams	9	20.4101	0.2727	0
Trigrams	8	22.6274	0.5555	0

Table 11: $DQI_{C2}$and DQI_{C6} (contradiction) for S2, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	11	23.5495	1.0000	0
Adjectives	3	6.3639	1.0000	0
Adverbs	0	6.3639	nan	0
Verbs	2	4.0000	1.0000	0
Nouns	5	12.5000	1.0000	0
Bigrams	19	37.4563	-0.1851	0
Trigrams	20	45.0185	0.2000	0

Table 12: $DQI_{C2}$and DQI_{C6} (neutral) for S3, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	12	41.5692	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	4	inf	1.0000	0
Nouns	5	12.5000	1.0000	0
Bigrams	20	89.4427	0.8095	0
Trigrams	19	4.6757e+16	1.0000	0

Table 13: $DQI_{C2}$and DQI_{C6} (neutral) for S4, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	14.3457	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	11	36.4828	0.6667	0
Trigrams	10	6.8359e+16	1.0000	0

Table 14: $DQI_{C2}$and DQI_{C6} (entailment) for S5, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	12	30.8285	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	7	20.0041	1.0000	0
Bigrams	25	125.0000	0.8461	0
Trigrams	24	7.0540e+16	1.0000	0

Table 15: $DQI_{C2}$and DQI_{C6} (entailment) for S6, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	6	14.6969	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	4	9.2376	1.0000	0
Bigrams	11	36.4828	0.6667	0
Trigrams	10	6.8359e+16	1.0000	0

Table 16: $DQI_{C2}$and DQI_{C6} (entailment) for S7, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	8	17.2819	1.0000	0
Adjectives	2	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	19	4.6757e+16	1.0000	0
Trigrams	17	inf	1.0000	0

Table 17: $DQI_{C2}$and DQI_{C6} (entailment) for S8, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	7	3.3356e+16	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	inf	1.0000	0
Bigrams	10	6.8359e+16	1.0000	0
Trigrams	8	inf	1.0000	0

Table 18: $DQI_{C2}$and DQI_{C6} (neutral) for S9, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	9	20.4100	1.0000	0
Adjectives	3	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	2	inf	1.0000	0
Nouns	4	8.0000	1.0000	0
Bigrams	19	4.6757e+16	1.0000	0
Trigrams	17	4.6757e+16	1.0000	0

Table 19: $DQI_{C2}$and DQI_{C6} (neutral) for S10, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	10	23.7170	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	8	18.4752	1.0000	0
Bigrams	20	1.4046e+17	1.0000	0
Trigrams	18	7.0027e+16	1.0000	0

Table 20: $DQI_{C2}$and DQI_{C6} (contradiction) for S11, Case (i)

Granularity	Count	DQI C2,C6 - T1	DQI C2,C6 - T2	DQI C6 - T5
Sentences	2	1.0000	1.0000	0
Words	11	16.3156	1.0000	0
Adjectives	1	inf	1.0000	0
Adverbs	0	inf	nan	0
Verbs	1	inf	1.0000	0
Nouns	8	11.3137	1.0000	0
Bigrams	18	55.6619	0.6000	0
Trigrams	18	7.0027e+16	1.0000	0

Table 21: $DQI_{C2}$and DQI_{C6} (contradiction) for S12, Case (i)

Sample	DQI C5 -T2,C6 - T3	DQI C5 - T3,C6 - T4	DQI C5 - T4
S1	0.2500	nan	0.8938
S2	0.5000	nan	0.9060
S3	0.2500	nan	0.8722
S4	0.0830	nan	0.6512
S5	0.1111	nan	0.6982
S6	0.0588	nan	0.6806
S7	1.0000	nan	0.7443
S8	0.1250	nan	0.7672
S9	0.3333	nan	0.8219
S10	0.1250	nan	0.7750
S11	0.2000	nan	0.7616
S12	0.2000	nan	0.8255

Table 22: T2/3 and T3/4 for DQI_{C5}/DQI_{C6}, T4 for DQI_{C5} , Case (i)

		Terms	
Sample Set		**T1**	
	ISIM=0.5	ISIM=0.6	ISIM=0.7
+S1	2.53901172	3.40305015	5.15852057
+S2	2.46282325	3.26756734	4.85347200
+S3	2.68605483	3.67251159	5.80405898
+S4	6.61292347	19.5239860	20.4998054
+S5	5.04523160	10.1825780	557.710874
+S6	5.53586344	12.4007484	51.6536766
+S7	4.09274400	6.92833358	22.5556185
+S8	3.74140198	5.97801932	14.8633715
+S9	3.10654715	4.50651832	8.20339191
+S10	3.6359872	5.71335622	13.3282739
+S11	3.8217013	6.18568557	16.2170311
+S12	3.0714259	4.43298421	7.96294530

Table 23: T1 for DQI_{C5}, Case (i)

Sample	**DQI C3 - T1**	**DQI C3 - T2**	**DQI C4**
S1	0.8938	2.0	0.9896
S2	0.9060	2.0	0.7779
S3	0.8722	2.0	1.3180
S4	0.6512	2.0	0.9093
S5	0.6982	2.0	0.0848
S6	0.6806	2.0	1.1088
S7	0.7443	2.0	0.6826
S8	0.7672	2.0	1.0860
S9	0.8219	2.0	0.5084
S10	0.7750	2.0	0.9601
S11	0.7616	2.0	1.1597
S12	0.8255	2.0	1.2076

Table 24: T1 and T2 for DQI_{C3}, DQI_{C4}, Case (i)

Sample	**DQI C1**	**DQI C2**	**DQI C3**	**DQI C4**	**DQI C5 (ISIM=0.5)**	**DQI C6**	**DQI C7**
S1	2.1906	80.2076	2.8938	0.9896	12.3961	80.4576	0
S2	0.7467	32.4274	2.9060	0.7779	9.7696	32.9274	0
S3	2.2302	49.4839	2.8722	1.3180	15.0742	49.7339	0
S4	7.8969	4.6757E+16	2.6512	0.9093	18.2884	4.6757E+16	0
S5	0.0693	6.8359E+16	2.6982	0.0848	16.3837	6.8359E+16	0
S6	11.4325	7.0540E+16	2.6806	1.1088	23.0456	7.054E+16	0
S7	0.0594	6.8359E+16	2.7443	0.6826	16.4604	6.8359E+16	0
S8	5.0289	4.6757E+16	2.7672	1.0860	15.8438	4.6757E+16	0
S9	1.4835	1.0171E+17	2.8219	0.5084	77.4403	1.01715E+17	0
S10	5.0388	9.3514E+16	2.7750	0.9601	16.2461	9.3514E+16	0
S11	2.9274	2.1048E+17	2.7616	1.1597	20.1601	2.10487E+17	0
S12	2.9373	7.0027E+16	2.8255	1.2076	16.6541	7.0027E+16	0

Table 25: DQI Terms, Case (i)

Sample Set		**Terms**		**DQI C1**
	T1	**T2**	**T3**	
Original	5.8200	6.6656	0.9300	12.0190
+S1	5.7921	6.6347	0.9307	11.9669
+S2	5.7822	6.6507	0.9307	11.9719
+S3	5.8020	6.6409	0.9307	11.9826
+S4	5.8119	6.6550	0.9307	12.0056
+S5	5.7723	6.6590	0.9208	11.9038
+S6	5.7822	6.6849	0.9307	12.0038
+S7	5.7822	6.6470	0.9307	11.9685
+S8	5.7921	6.6422	0.9307	11.9739
+S9	5.8020	6.6551	0.9307	11.9958
+S10	5.7921	6.6422	0.9307	11.9739
+S11	5.7921	6.6355	0.9307	11.9677
+S12	5.8317	6.6355	0.930	12.0073

Table 26: DQI_{C1} for Case (ii)

Sample Set	**Sentences**		**Words**		**Adjectives**		**Adverbs**		**Verbs**		**Nouns**		**Bigrams**		**Trigrams**		**DQI C2**
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	**T1**	**T2**	
Original	2807.2405	0.9800	137.2755	0.6371	52.0534	0.3111	20.0385	-0.04	46.8398	-0.025	54.2786	0.3888	707.8112	0.8852	2723.6406	0.8910	5927.1970
+S1	2849.6668	0.9802	137.0171	0.6368	55.6705	0.3065	21.7786	-0.1111	50.8642	-0.0356	49.5464	0.3452	697.9764	0.8815	2706.4317	0.8857	5922.7847
+S2	2849.6668	0.9802	137.0171	0.6368	55.6705	0.3065	21.7789	-0.1111	50.8642	-0.0356	49.5464	0.3452	697.9764	0.8815	2706.4317	0.8857	5922.7847
+S3	2849.6668	0.9802	137.9140	0.6393	52.6620	0.2414	17.4592	0.0833	43.8252	-0.0661	55.2815	0.3505	712.9377	0.8847	2763.8091	0.8924	6009.2173
+S4	2849.6668	0.9802	138.3361	0.6392	54.2001	0.2576	24.9929	0.1250	48.5320	-0.0313	50.1523	0.3498	706.9163	0.9043	2765.4396	0.8921	6021.0912
+S5	2849.6668	0.9802	135.4295	0.6365	49.2904	0.2619	23.3950	0.0000	49.0989	-0.0840	52.0959	0.3432	697.8102	0.9029	2649.2411	0.8895	5892.6612
+S6	2849.6668	0.9802	137.1086	0.6379	53.9239	0.3609	20.0385	-0.0400	48.0375	-0.0538	52.8044	0.3463	711.5407	0.9064	2723.0651	0.8903	5984.3517
+S7	2849.6668	0.9802	137.4205	0.6359	48.4367	0.2015	35.9211	0.1538	45.0502	-0.0361	54.6786	0.4303	710.2298	0.9058	2739.3807	0.8916	6003.5736
+S8	2849.6668	0.9802	136.2514	0.6368	49.6075	0.2268	57.0399	0.3846	49.9798	-0.0445	52.5582	0.3432	705.7911	0.9052	2693.8612	0.8888	5962.1966
+S9	2849.6668	0.9802	137.6593	0.6375	58.2917	0.3388	24.5189	-0.0244	52.4063	0.0041	50.5623	0.3237	707.6845	0.9048	2742.9126	0.8915	6002.3536
+S10	2849.6668	0.9802	136.2477	0.6371	56.5772	0.2511	29.8974	-0.1034	51.6379	-0.0206	51.8621	0.3484	708.3581	0.9052	2718.4279	0.8899	5968.5017
+S11	2849.6668	0.9802	137.7623	0.6373	49.6725	0.2197	20.5196	-0.0667	47.5031	-0.0370	54.6531	0.3741	717.2547	0.9062	2767.0664	0.8921	6027.7480
+S12	2849.6668	0.9802	139.5281	0.6413	59.9832	0.3101	15.2008	-0.2727	52.8410	0.0723	50.6446	0.3174	713.8007	0.9052	2763.0228	0.8920	6027.8220

Table 27: DQI_{C2} for Case (ii)

				Terms				DQI C3 (e=0.5)	
Sample Set		**T1**			**T2 (SIM=0.5)**				
	SIM=0.5	SIM=0.6	SIM=0.7	e=0.25	e=0.33	e=0.5	SIM=0.5	SIM=0.6	SIM=0.7
Original	14.1194	4.9647	4.2968	200.0000	200.0000	198.4692	212.5886	203.4339	202.766
+S1	14.0959	4.9880	4.2882	202.0000	202.0000	199.9066	214.0025	204.8946	204.1948
+S2	14.2729	4.8939	4.3000	202.0000	202.0000	200.9450	215.2179	205.8389	205.245
+S3	14.1055	4.9749	4.2710	202.0000	202.0000	199.9066	214.0121	204.8815	204.1776
+S4	14.1285	4.9797	4.3134	202.0000	202.0000	200.4539	214.5824	205.4336	204.7673
+S5	14.1522	4.9797	4.3072	202.0000	202.0000	200.4539	214.6061	205.4336	204.7611
+S6	14.1961	4.9827	4.3041	202.0000	202.0000	200.4539	214.65	205.4366	204.758
+S7	14.1656	4.9842	4.3197	202.0000	202.0000	200.4539	214.6195	205.4381	204.7736
+S8	14.2711	4.9873	4.3015	202.0000	202.0000	200.9450	215.2161	205.9323	205.2465
+S9	14.2321	4.9836	4.3214	202.0000	202.0000	200.9450	215.1771	205.9286	205.2664
+S10	14.2859	4.9888	4.2944	202.0000	202.0000	200.9450	215.2309	205.9338	205.2394
+S11	14.1403	4.9720	4.3122	202.0000	202.0000	200.4539	214.5942	205.4259	204.7661
+S12	14.1707	4.9874	4.3211	202.0000	202.0000	199.9066	214.0773	204.894	204.2277

Table 28: DQI_{C3} for Case (ii)

Sample Set	DQI C4
Original	0.00657581
+S1	0.00653241
+S2	0.00652070
+S3	0.00654317
+S4	0.00652860
+S5	0.00610259
+S6	0.00653705
+S7	0.00651307
+S8	0.00653624
+S9	0.00649185
+S10	0.00653108
+S11	0.00653874
+S12	0.00654020

Table 29: DQI_{C4} for Case (ii)

				Terms					DQI C5 (ISIM=0.5)
Sample Set		**T1**		**T2**	**T3**	**T4**	**T5**	**T6**	
	ISIM=0.5	ISIM=0.6	ISIM=0.7						
Original	3.79338794	5.79942751	9.64213607	0.13869626	0.06846071	0.00106449	19.2658	0.08669236	4.00160940
+S1	3.77492292	5.75927311	9.55986754	0.13950276	0.06756993	0.00105670	19.1081	0.08686184	3.98305231
+S2	3.77320467	5.75527455	9.54885537	0.13988920	0.06771915	0.00105824	19.1048	0.08711365	3.98187126
+S3	3.77796738	5.76636257	9.57941700	0.13950276	0.06756993	0.00105429	19.0986	0.08666733	3.98609436
+S4	3.80946946	5.84007436	9.69296631	0.13797814	0.06754694	0.00105432	19.2038	0.08661618	4.01604886
+S5	3.80273001	5.82425011	9.73687404	0.13854595	0.06744772	0.00105055	19.1196	0.08696758	4.00977423
+S6	3.80524680	5.83015604	9.72041244	0.13704206	0.06799806	0.00105172	19.1444	0.08642433	4.01133864
+S7	3.79613706	5.80879868	9.69710399	0.14008322	0.06781511	0.00104881	19.1444	0.08708462	4.00508420
+S8	3.79286615	5.80114342	9.67578885	0.13873626	0.06744340	0.00104868	19.1246	0.08673365	4.00009449
+S9	3.78510214	5.78300049	9.62542175	0.13969571	0.06763740	0.00105033	19.7681	0.08710369	3.99348558
+S10	3.79176275	5.79856261	9.66861134	0.13873626	0.06744340	0.00104875	19.1295	0.08675259	3.99899116
+S11	3.79366621	5.80301526	9.68099727	0.13931034	0.06751676	0.00104867	19.1840	0.08695819	4.00154198
+S12	3.78458008	5.78178193	9.62204642	0.13931034	0.06751676	0.00105054	19.1213	0.08674638	3.99245772

Table 30: DQI_{C5} for Case (ii)

				Terms			
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	7.1303e+16	1.0000	1045.3358	2.0833	7.1303e+16	1.0000	92.8203
+S1	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S2	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S3	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S4	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S5	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S6	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S7	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S8	1.4267e+17	1.0000	1045.3358	2.0000	7.1303e+16	0.9600	93.7485
+S9	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S10	7.1303e+16	1.0000	1075.9298	2.1250	7.1303e+16	1.0000	93.7485
+S11	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485
+S12	7.1303e+16	1.0000	1045.3358	2.0833	1.4267e+17	1.0417	93.7485

Table 31: Case (ii), Sentence Granularity Terms in DQI_{C6}

				Terms			
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	113.4748	0.5548	136.5557	0.6599	105.1059	0.5255	2.4416
+S1	113.4748	0.5548	136.5557	0.6599	103.7067	0.5219	2.4509
+S2	113.4748	0.5548	136.5557	0.6599	107.3208	0.5339	2.4325
+S3	113.4748	0.5548	137.7114	0.6182	105.1059	0.5255	2.3670
+S4	113.4748	0.5548	138.5993	0.6422	105.1059	0.5255	2.4336
+S5	109.7512	0.5298	136.5557	0.6599	105.1059	0.5255	2.4566
+S6	117.4812	0.5679	136.5557	0.6599	105.1059	0.5255	2.4518
+S7	115.2611	0.5520	136.5557	0.6599	105.1059	0.5255	2.4241
+S8	110.1518	0.5562	136.5557	0.6599	105.1059	0.5255	2.4491
+S9	113.4748	0.5548	136.5917	0.6604	105.1059	0.5255	2.4467
+S10	113.4748	0.5548	134.4891	0.6595	105.1059	0.5255	2.4267
+S11	113.4748	0.5548	136.5557	0.6599	110.1129	0.5304	2.4310
+S12	113.4748	0.5548	136.5557	0.6599	112.6038	0.5459	2.4524

Table 32: Case (ii), Word Granularity Terms in DQI_{C6}

				Terms			
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	65.4824	0.1935	48.9086	0.1130	44.8057	-0.2113	2.6514
+S1	74.6675	0.0909	50.8008	0.1500	57.0071	0.0164	2.8685
+S2	61.3138	-0.0588	52.7111	0.0815	51.3651	-0.1351	3.1961
+S3	76.2138	0.0588	46.8815	0.1339	60.6168	0.0476	3.0158
+S4	62.4955	-0.0423	58.8794	0.2480	52.4764	-0.1389	3.2262
+S5	71.8135	-0.0133	48.3257	0.1707	57.2251	0.0667	2.9149
+S6	71.5360	0.0571	50.7164	0.1897	49.4934	0.0000	2.5007
+S7	69.5736	0.1475	52.5575	0.0676	58.1186	0.0312	2.6028
+S8	73.1520	0.1250	45.2213	0.1000	51.0064	0.0149	2.7511
+S9	68.4000	0.0000	48.3109	0.0615	52.7210	0.0000	2.8224
+S10	72.3354	0.0684	48.7879	0.1147	53.0237	0.0667	3.0774
+S11	68.2115	-0.0410	47.9655	0.1355	50.9620	-0.0294	2.6320
+S12	74.7011	0.0000	51.4393	0.0518	45.1122	-0.1384	2.6840

Table 33: Case (ii), Adjective Granularity Terms in DQI_{C6}

	Terms						
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	18.4752	0.2000	21.4630	0.1765	6.3640	0.0000	5.1159
+S1	3.6029e+16	1.0000	16.4141	-0.0769	6.3640	0.0000	3.0036
+S2	10.0021	0.3333	13.4297	0.2632	9.2376	0.0000	2.9621
+S3	16.0997	0.4287	25.0000	0.3333	6.3640	0.0000	4.8231
+S4	inf	1.0000	20.8025	0.0000	9.2376	0.2000	3.4788
+S5	20.0042	0.5000	19.2428	0.1250	12.5	0.3333	4.2973
+S6	inf	1.0000	21.4630	0.1765	6.3639	0.0000	2.9468
+S7	28.6378	0.6000	19.0918	0.0000	6.3639	0.0000	3.5977
+S8	18.4752	0.2000	27.6955	0.4444	9.2376	0.2000	3.4223
+S9	21.6481	0.2727	28.6216	0.3000	6.3639	0.0000	5.3589
+S10	8.0632	-0.2307	19.2428	0.1250	9.6096	0.0000	4.3729
+S11	inf	1.0000	19.2428	0.1250	9.2376	0.2000	4.0262
+S12	inf	1.0000	23.7684	0.2222	6.3639	0.0000	4.1769

Table 34: Case (ii), Adverb Granularity Terms in DQI_{C6}

	Terms						
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	65.4824	0.1935	51.9736	-0.0598	35.1110	-0.1081	2.7836
+S1	40.3696	-0.2069	48.5430	-0.1525	29.9195	-0.2405	2.4728
+S2	43.9037	-0.2424	53.3506	-0.0093	30.1625	-0.0909	2.6133
+S3	37.4444	-0.3030	56.2047	-0.1057	27.3594	-0.2286	2.3308
+S4	42.1040	-0.3333	46.2161	-0.0973	31.2449	-0.1667	2.5586
+S5	38.3571	-0.3714	50.6384	-0.0182	24.4386	-0.2000	2.5610
+S6	41.7648	-0.2537	48.9552	-0.0280	28.8722	-0.1642	2.7063
+S7	46.5989	-0.2537	53.4887	-0.1260	31.1722	-0.2500	2.2977
+S8	35.4040	-0.3548	48.3655	-0.0990	26.0207	-0.2615	2.7680
+S9	40.6156	-0.2000	53.4014	-0.1056	32.0340	-0.2307	2.5957
+S10	41.3657	-0.3230	53.0775	-0.0847	29.1653	-0.2876	2.2606
+S11	42.3999	-0.2187	46.3814	-0.1452	33.3842	-0.1267	2.6794
+S12	37.5858	-0.2258	49.7109	-0.1071	26.0396	-0.0667	2.6669

Table 35: Case (ii), Verb Granularity Terms in DQI_{C6}

	Terms						
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	42.7808	-0.3056	53.6301	0.2841	38.7466	-0.2050	2.3372
+S1	38.3026	-0.3659	52.7785	0.2989	39.4878	-0.2601	2.4916
+S2	35.9868	-0.2752	51.9745	0.3097	41.0652	-0.2558	2.3264
+S3	36.7162	-0.3247	52.4598	0.2667	41.5999	-0.2485	2.3551
+S4	36.7565	-0.2617	53.2731	0.2570	37.4839	-0.2075	2.3918
+S5	33.0670	-0.2752	54.0598	0.3030	44.1367	-0.2817	2.3645
+S6	38.3611	-0.3250	54.9709	0.3040	42.2864	-0.2528	2.5035
+S7	37.7188	-0.3414	51.8644	0.2844	37.6200	-0.2327	2.6013
+S8	38.9773	-0.3254	55.4119	0.3028	41.6562	-0.2441	2.4018
+S9	35.4958	-0.3200	50.3967	0.3313	39.9118	-0.2121	2.4067
+S10	32.9868	-0.2765	52.1225	0.2954	38.6028	-0.2484	2.4450
+S11	36.0093	-0.3333	55.2239	0.3352	42.8904	-0.2402	2.4570
+S12	34.8526	-0.3509	50.4304	0.3113	51.0263	-0.2448	2.5026

Table 36: Case (ii), Noun Granularity Terms in DQI_{C6}

				Terms			
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	497.2044	0.8411	620.1037	0.9075	415.2737	0.8610	0.7924
+S1	497.2043	0.8411	620.1037	0.9075	403.4774	0.8206	0.7928
+S2	497.2043	0.8411	620.1037	0.9075	427.4754	0.8636	0.7917
+S3	497.2043	0.8411	625.7171	0.8873	415.2737	0.8610	0.7694
+S4	497.2043	0.8411	616.7056	0.9055	415.2737	0.8610	0.7864
+S5	473.5139	0.8528	620.1037	0.9075	415.2737	0.8610	0.8045
+S6	518.7792	0.8684	620.1037	0.9075	415.2737	0.8610	0.8088
+S7	503.1652	0.8648	620.1037	0.9075	415.2737	0.8610	0.7960
+S8	491.4631	0.8588	620.1037	0.9075	415.2737	0.8610	0.8069
+S9	497.2043	0.8411	617.3021	0.9064	415.2737	0.8610	0.7986
+S10	497.2043	0.8411	619.8558	0.9072	415.2737	0.8610	0.7936
+S11	497.2043	0.8411	620.1037	0.9075	437.4726	0.8657	0.8003
+S12	497.2043	0.8411	620.1037	0.9075	427.2611	0.8623	0.7915

Table 37: Case (ii), Bigram Granularity Terms in DQI_{C6}

				Terms			
Sample Set	**entailment**		**neutral**		**contradiction**		
	T1	**T2**	**T1**	**T2**	**T1**	**T2**	**T5**
Original	1567.0110	0.7652	2174.6543	0.7302	1135.1086	0.7193	1.7297
+S1	1567.0110	0.7652	2174.6543	0.7302	1154.0280	0.7094	1.7212
+S2	1567.0110	0.7652	2174.6543	0.7302	1157.8255	0.8636	1.7298
+S3	1567.0110	0.7652	2215.9640	0.7163	1135.1086	0.7193	1.6799
+S4	1567.0110	0.7652	2245.9485	0.7355	1135.1086	0.7193	1.7383
+S5	1517.6459	0.7571	2174.6543	0.7302	1135.1086	0.7193	1.7468
+S6	1642.3849	0.7601	2174.6543	0.7302	1135.1086	0.7193	1.7383
+S7	1593.6394	0.7615	2174.6543	0.7302	1135.1086	0.7193	1.7406
+S8	1529.5108	0.7521	2174.6543	0.7302	1135.1086	0.7193	1.7470
+S9	1567.0110	0.7652	2204.5792	0.7324	1135.1086	0.7193	1.7470
+S10	1567.0110	0.7652	2190.9585	0.7245	1135.1086	0.7193	1.7235
+S11	1567.0110	0.7652	2174.6543	0.7302	1199.7393	0.7288	1.7470
+S12	1567.0110	0.7652	2174.6543	0.7302	1199.7393	0.7288	1.7383

Table 38: Case (ii), Trigram Granularity Terms in DQI_{C6}

				Terms		
Sample Set	**entailment**		**neutral**		**contradiction**	
	T3	**T4**	**T3**	**T4**	**T3**	**T4**
Original	0.1846	0.2003	0.1465	0.1226	0.1008	0.3662
+S1	0.1846	0.2003	0.1465	0.1226	0.1037	0.3485
+S2	0.1846	0.2003	0.1465	0.1226	0.1046	0.3514
+S3	0.1846	0.2003	0.1480	0.1195	0.1008	0.3662
+S4	0.1846	0.2003	0.1448	0.1195	0.1008	0.3662
+S5	0.1811	0.1894	0.1465	0.1226	0.1008	0.3662
+S6	0.1712	0.2065	0.1465	0.1226	0.1008	0.3662
+S7	0.1923	0.1931	0.1465	0.1226	0.1008	0.3662
+S8	0.1824	0.1887	0.1465	0.1226	0.1008	0.3662
+S9	0.1846	0.2003	0.1484	0.1197	0.1008	0.3662
+S10	0.1846	0.2003	0.1464	0.1191	0.1008	0.3662
+S11	0.1846	0.2003	0.1465	0.1226	0.1033	0.3473
+S12	0.1846	0.2003	0.1465	0.1226	0.1033	0.3473

Table 39: Terms 3 and 4 in DQI_{C6} for Case (ii)

Sample Set	**DQI C6**
Original	228.3537
+S1	202.4647
+S2	197.6054
+S3	196.3454
+S4	196.1489
+S5	200.7986
+S6	213.8920
+S7	202.4102
+S8	202.2893
+S9	198.4766
+S10	202.7345
+S11	200.9509
+S12	197.8010

Table 40: DQI_{C6} for Case (ii)

Sample Set	**DQI C7**		
	SSIM=0.2	SSIM=0.3	SSIM=0.4
Original	0.00304989	0.00421324	0.00629840
+S1	0.00189475	0.00229266	0.00290212
+S2	0.00216703	0.00270372	0.00359374
+S3	0.00186796	0.00225356	0.00283975
+S4	0.00196072	0.00238996	0.00305981
+S5	0.00188903	0.00228429	0.00288872
+S6	0.00190351	0.00230549	0.00292271
+S7	0.00201427	0.00247000	0.00319224
+S8	0.00187124	0.00225832	0.00284732
+S9	0.00197442	0.00241034	0.00309330
+S10	0.001886216	0.00228017	0.00288214
+S11	0.002048964	0.00252237	0.00328026
+S12	0.002076182	0.00256374	0.00335058

Table 41: DQI_{C7} for Case (ii)

Sample	**Overlap Count**	**length(hypothesis) / Overlap Count**
S1	3	**2.0000**
S2	2	**1.5000**
S3	8	**1.1250**
S4	1	10.0000
S5	2	**3.5000**
S6	2	**5.5000**
S7	1	**4.0000**
S8	2	3.5000
S9	0	**40.0000**
S10	2	3.5000
S11	1	**5.0000**
S12	3	3.0000

Table 42: Word Overlap, Red: $<$ 3.9375, Yellow: 3.9375-9.8333 Green: $>$ 9.8333

Sample	**Overlap Count**	**length(hypothesis+premise) / Overlap Count**
S1	3	**3.3333**
S2	2	**3.0000**
S3	8	**2.3750**
S4	1	**13.0000**
S5	2	**4.5000**
S6	2	**7.0000**
S7	1	**7.0000**
S8	2	5.0000
S9	0	**70.0000**
S10	2	5.5000
S11	1	**11.0000**
S12	3	4.6667

Table 43: Word Overlap, Red: < 5.5347, Yellow: 5.5347-17.1944 Green: > 17.1944

Sample	Premise Word Count	Hypothesis Word Count	Sum of Word Similarities
S1	10	9	5.4753
S2	6	7	2.7865
S3	12	15	**8.9008**
S4	15	6	**9.8715**
S5	9	3	6.5202
S6	17	6	**29.0358**
S7	7	6	**3.6143**
S8	12	7	**6.5335**
S9	7	5	**3.6679**
S10	127	7	**6.0583**
S11	9	12	**4.3558**
S12	12	9	28.5806

Table 44: Word Similarity With Stop Words, Red: > 10.4317, Yellow: 8.8017-10.4317 Green: < 8.8017

Sample	Premise Word Count	Hypothesis Word Count	Sum of Word Similarities
S1	6	4	**5.3800**
S2	3	3	2.9008
S3	10	9	**8.8910**
S4	10	3	**7.9413**
S5	7	2	**6.0292**
S6	11	3	**9.7704**
S7	4	3	**3.6234**
S8	7	3	**6.2102**
S9	4	3	**3.1786**
S10	7	4	**6.2102**
S11	5	6	**4.3768**
S12	9	5	7.8905

Table 45: Word Similarity Without Stop Words, Red: > 6.8188, Yellow: 5.2483-6.8188 Green: < 5.2483

9.6 User Study

AutoFix Suggestions: See Table 46.

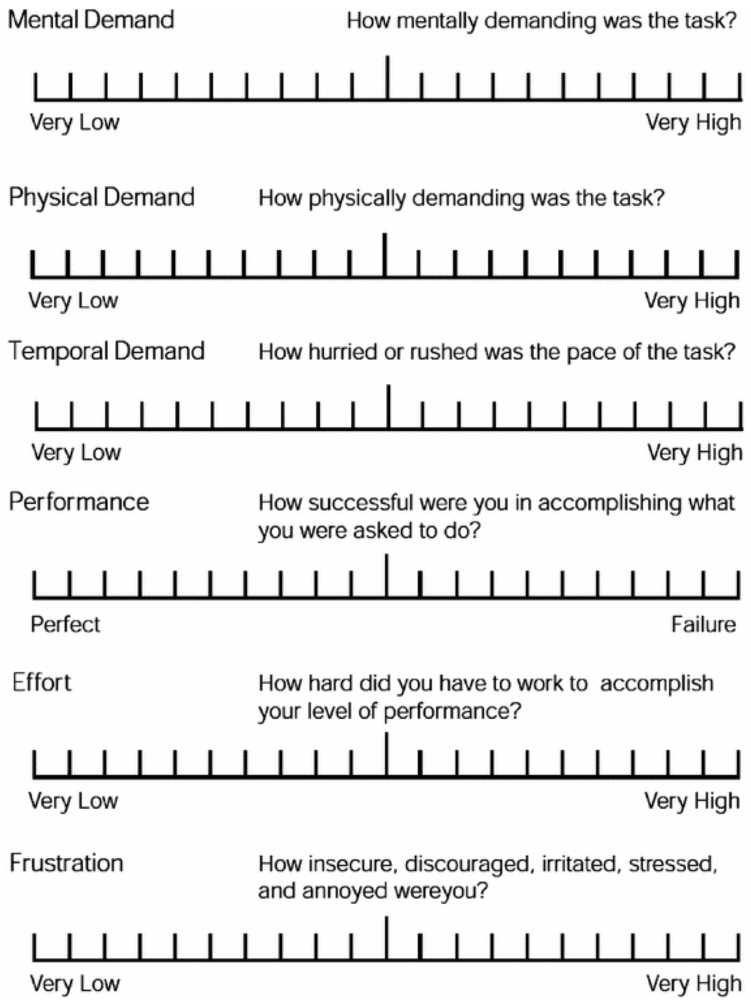
Premise	Orig. Hypothesis	DQI	Suggested Words	New Hypothesis based on suggestions	New DQI
A woman, in a green shirt, preparing to run on a treadmill.	A woman is preparing to sleep on a treadmill	2.4650170	preparing,sleep	A woman is organizing to rest on a treadmill	2.5275722
The dog is catching a treat	The cat is not catching a treat	2.752542	catching	the cat is not getting a treat	3.6909140
Three young men are watching a tennis match on a large screen outdoors	Three young men watching a tennis match on a screen outdoors, because their brother is playing	2.6435402 891414217	young,watching, playing	Three youthful men observing a tennis match on a screen outdoors, because their brother is performing.	2.6787982
A man in a green apron smiles behind a food stand	A man smiles	3.2367785	smiles	A person is grinning.	6.303777

Table 46: A few samples for Autofix with ISSTS in DQI

NASA TLX: The NASA Task Load Index (NASA-TLX) is a subjective, multidimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance [7].

NASA-TLX divides the total workload into six subjective subscales that are represented on a single page. There is a description for each of these subscales that the subject should read before rating. They rate each subscale within a 100-point range, with 5-point steps, as shown in Figure 25. Providing descriptions for each measurement can be found to help participants answer accurately [22]. The descriptions are as follows:

- **Mental Demand:** How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
- **Physical Demand:** How much physical activity was required? Was the task easy or demanding, slack or strenuous?
- **Temporal Demand:** How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
- **Performance:** How successful were you in performing the task? How satisfied were you with your performance?
- **Effort:** How hard did you have to work (mentally and physically) to accomplish your level of performance?
- **Frustration:** How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?



Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Figure 25: NASA TLX Form

We record participant demographics– age, gender, and occupation. Participants are asked to fill this form at the end of each round of the user study. We also record the number of questions participants successfully create, as well as a record of how often participants use each module in the full system round. At the end of the user study, participants are asked what their impression of data quality is, and their free response is recorded.

Subscale Wise Results: Individual results of the averaged subscales in Figure 5 are shown in Figures 26,27. Physical demand does not change significantly across user study rounds.

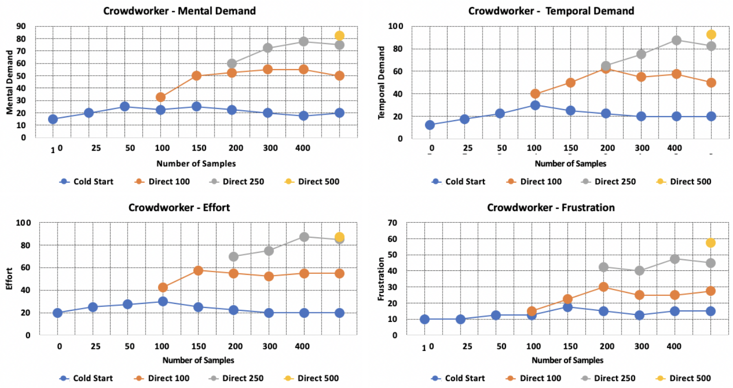


Figure 26: NASA TLX– Crowdworker Subscale Results

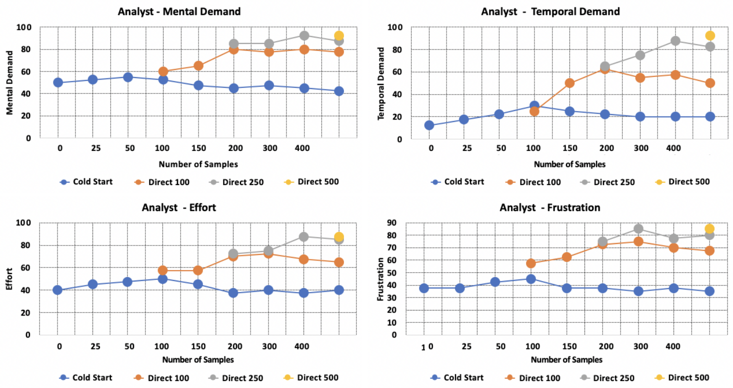


Figure 27: NASA TLX– Analyst Subscale Results

9.7 Expert and User Comments

Experts *(P)*: We present an initial prototype of our tool, to a set of three researchers with expertise in NLP and knowledge of data visualization, in order to judge the interface design. For each expert, the crowdworker interface and then analyst interfaces were demoed. Participants (P) could ask questions and make interaction/navigation decisions to facilitate a natural user experience. All the experts appreciated the easily interpretable traffic-signal color scheme and found the organization of the interfaces—providing separate detailed views within the analyst workflow– a way to prevent cognitive overload (too much information on one screen); P_2 said the latter *"... enhances readability for understanding the data at different granularities."*. P_1 suggested the inclusion of *"... a provenance module within the analyst views to show historical sample edits and overall data quality changes over time to understand how data quality evolves as the benchmark size increases... this would help with the bubble plot and tree map which will get more cluttered and complex as data size increases"*. Additionally P_3 remarked that *"The frequency of samples of middling quality should increase as benchmark size increases, but the initial exposure that analysts will have with higher or lower quality samples should lessen the learning curve as they are familiar enough with interface subtleties by the time they begin to encounter more challenging cases."*

Crowdworkers *(C)*: When presented with traffic signal feedback, crowdworkers report that the time and effort required to create high quality samples increases–*"You need to keep redoing the sample since when you see it's all red, you know it's probably not going to be accepted"*(C_3); however, they are more confident about their performance and sample quality *"...when there's green, I know I've done it right, and it cuts down on my having to create a lot of samples to get paid"* (C_{15}). We find that AutoFix usage [7] causes an unexpected increase in mental and temporal demand, as well as frustration; we attribute this to observed user behavior– *"I'm not sure how much I trust this recommendation without seeing the colors"*(C_{12}), and *"I'd prefer to change a couple of things since I can't see the feedback anymore*(C_{21}). The drastic improvement over all aspects (highest for frustration) in the case of using the full system is in line with this observation–*"This is so easy, I can create samples really fast, and I have a better chance of getting more accepted."*(C_8) and *"Now that I get the feedback along with the recommendation, I can see the quality improvement. So using the recommendation is now definitely faster."*(C_{12}). The number of questions created per round as well as system scores also follows this trend, across all types of crowdworkers.

Analysts *(A)*: In the case of direct quality feedback, i.e., traffic signals, analysts report an increased performance and find the task easier–*"... it's easier to directly choose based on quality... and it takes care of typos too, the typo samples are marked down so the work goes pretty fast"*(A_3). When analysts are shown the visualization interfaces, they are explicitly taught to differentiate the traffic signal colors in the visualizations as being indicative of how the sample affects the overall dataset quality, i.e., the colors in different component views represent individual terms of the components calculated over the whole dataset (analysts can toggle between the states of original dataset and new sample addition). We find that users initially find this more difficult to do– *"It takes a little time to figure out how to go through the views. I learned that in the samples I looked at, components three and seven seemed to be linked. So I'd look at those first the next time I used the system"* (A_6) and *"... it takes me some time to figure out how to read the interfaces effectively, but it does make me more secure in judging sample quality at multiple granularities and that would help if I was doing this for a particular application"*(A_1). Analysts averaged behavior on TextFooler models the conventional approach quite closely, as analysts are seen to have a tendency to either– *"... deciding to reject or repair is difficult when you don't have the sample or dataset feedback... and what if the repaired sample still isn't good enough?"*(A_4), or– *" I like having this option to repair... I don't need to waste time on analyzing something that isn't outright an accept or reject, I can send it to be repaired and come back to it later"*(A_8). When shown the full system, analysts also report improvement in all aspects, particularly mental demand and performance–*"I can be sure of not having to redo things since it's likely that I will be able to get a low hypothesis baseline using this system"*$(A_2$, $A_1)$. The visualization usage also improves– *"... I went to component three right off the bat this time, I knew that I could look at the linked components..."* (A_6). Altogether, sample evaluation by analysts increases, following this trend, and analysts are more assured of their performance.