DQI Case Study : Supplemental Material
Real-Time Visual Feedback to Guide Benchmark Creation: A Human-and-Metric-in-the-Loop Workflow

Anonymous EACL submission

References

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

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

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

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

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

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

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

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

A Artifact Case Study

A.1 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 (random sampling done 10 times for a fair comparison)

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 2, 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.

We use DQI to identify artifacts over four datasets: SNLI (Bowman et al., 2015), MNLI(Williams et al., 2017), SQUAD 2.0 (Rajpurkar et al., 2018), and Story CLOZE Task (Schwartz et al., 2017). 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 (Gururangan et al., 2018) and Story CLOZE Task (Schwartz et al., 2017). MNLI and SQUAD 2.0 have been shown to have a relatively smaller number of artifacts (Gururangan et al., 2018; Kaushik and Lipton, 2018), and therefore ensure adversarial evaluation of VAIDA. We evaluate each dataset using its test sets, or if unavailable, on its dev sets.

For fair comparison, we have taken illustrative samples from the AFLite paper (Bras et al., 2020)

Dataset	Sample ID	Split	Label	DQI Color
SNLI	S7	Good	Entailment	🟩
	S8			🟧
	S9		Neutral	🟩
	S10			🟧
	S11		Contradiction	🟩
	S12			🟥
	S5	Bad	Entailment	🟧
	S6			🟥
	S3		Neutral	🟥
	S4			🟥
	S1		Contradiction	🟧
	S2			🟩
Story CLOZE	S1	Good	True	🟥
	S2			🟧
	S3		False	🟧
	S4			🟧
	S5	Bad	True	🟩
	S6			🟧
	S7		False	🟥
	S8			🟩
SQUAD 2.0	S1	Good	True	🟥
	S2			🟩
	S3		False	🟩
	S4			🟩
	S5	Bad	True	🟥
	S6			🟥
	S7		False	🟩
	S8			🟩
MNLI	S1	Good	Entailment	🟥
	S2			🟩
	S3		Neutral	🟥
	S4			🟩
	S5		Contradiction	🟧
	S6			🟥
	S7	Bad	Entailment	🟩
	S8			🟥
	S9		Neutral	🟥
	S10			🟩
	S11		Contradiction	🟥
	S12			🟥

Table 1: Evaluating VAIDA over the most sensitive DQI component, Intra-Sample Word Similarity. Successes: green/orange for good split, red/orange for bad split. Failures: red for good split, green for bad split.

for SNLI (Table 2). We randomly sample for other datasets (Tables 3, 4 5, 6) as corresponding examples were not illustrated in those papers. 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 (Mishra et al., 2020). This is analogous to how humans learn quickly from few samples.

Results: DQI component colors across settings are correctly predicted according to AFLite categorization of good and bad splits on an average of 10/12 times in SNLI, 5/8 times in SQUAD 2.0 and Story CLOZE, and 7/12 times in MNLI $^{??}$ as illustrated in Table 1. 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.

Analysis: False positives and false negatives can be attributed to the limitation of AFLite in incorrectly classifying samples (Mishra et al., 2020). 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 (Mishra et al., 2020).

A.1.1 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 (shown for SNLI samples from Table 2) 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 7.

Sample	Terms T1	T2	T3	DQI C1
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 7: DQI_{C1} for Case (i)

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 15 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 18.

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 2: SNLI Samples used for Test Cases

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

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 8 - 19. These bounds therefore need to be reset at cold start particular to the sample's language.

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

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 3: MNLI Samples used for Test Cases

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 10: $DQI_{C2}$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 11: $DQI_{C2}$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 12: $DQI_{C2}$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 13: $DQI_{C2}$and DQI_{C6} (entailment) for S6, Case (i)

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 4: SQUAD 2.0 Test Cases - Dev Good

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 14: $DQI_{C2}$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 15: $DQI_{C2}$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 16: $DQI_{C2}$and DQI_{C6} (neutral) for S9, Case (i)

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 5: SQUAD 2.0 Test Cases - Dev Bad

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 17: $DQI_{C2}$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 18: $DQI_{C2}$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 19: $DQI_{C2}$and DQI_{C6} (contradiction) for S12, Case (i)

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 20, 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.

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 6: Story CLOZE Test Cases

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 20.

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

Intra-sample STS: The first term, in Table 21, 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 24 - 27.

	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 21: T1 for DQI_{C5}, 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)

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 21 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.

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 23: DQI Terms, Case (i)

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 24: 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 25: 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 26: 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 27: Word Similarity Without Stop Words, Red: $>$ 6.8188, Yellow: 5.2483-6.8188 Green: $<$ 5.2483

A.1.2 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 28 - 43, indicate the need for hyperparameter scaling.

What requires Scaling? From tables 35 and 32-38, 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.

A.1.3 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 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 28: DQI_{C1} 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						
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 30: 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 31: 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 32: 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 33: 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 34: 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 38: 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 39: 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 40: Case (ii), Trigram Granularity Terms in DQI_{C6}

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 35: 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 36: DQI_{C3} for Case (ii)

			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 41: 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 42: DQI_{C6} 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 37: DQI_{C5} 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 43: DQI_{C7} for Case (ii)

A.1.4 Results Across Datasets

The following tables contain DQI component values across the sets of samples from Tables 2-6 in SNLI, MNLI, SQUAD 2.0, and Story CLOZE. Here, 'Good' denotes samples present in the 'Good' split of AFLite and 'Bad' denotes samples present in the 'Bad' Split of AFLite respectively.

Parameter 1: The following tables contain values for Parameter 1 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Term	T1	T2	T3	DQI C1
Good	**1.8996**	**6.0409**	**0.9532**	**7.6578**
Bad	0.6416	5.8135	0.9494	6.1609

Table 44: SNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	**1.8996**	**6.0409**	**0.9532**	**7.6578**
Bad	0.6416	5.8135	0.9494	6.1609

Table 45: SNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	1.6177	**104.6542**	**0.7550**	**80.6316**
Bad	**7.4100**	14.1068	0.6020	15.9023

Table 46: MNLI Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	1.7715	71.3947	-0.0023	1.6073
Bad	**11.1550**	**73.3092**	**-0.001**	**11.1476**

Table 47: SQUAD 2.0 Sub-Component and Overall Values for DQI_{c1}

.

Term	T1	T2	T3	DQI C1
Good	3.3010	13.4569	0.2772	7.0313
Bad	**4.7675**	**13.4895**	**0.2839**	**8.5972**

Table 48: Story-CLOZE Sub-Component and Overall Values for DQI_{c1}

.

Granularity	Split	T1	T2	Contribution
Words	Good	**121.9512**	**0.7269**	**88.6463**
	Bad	52.3560	0.6500	34.0314
Adjectives	Good	**31.7460**	**0.2966**	**9.4159**
	Bad	16.9205	0.3590	6.0745
Adverbs	Good	**21.0970**	**0.1847**	**3.8966**
	Bad	10.7875	0.1732	1.8684
Verbs	Good	**43.6681**	**0.2349**	**10.2576**
	Bad	16.5289	0.1893	3.1289
Nouns	Good	**49.2611**	**0.4351**	**21.4335**
	Bad	21.0084	0.3685	7.7416
Bigrams	Good	**1296.3443**	**0.9374**	**1215.1931**
	Bad	873.2862	0.9355	816.9592
Trigrams	Good	**7686.3951**	**0.9546**	**7337.4328**
	Bad	6119.9510	0.9422	5766.2178
Sentences	Good	9070.7819	**0.6607**	**5993.0656**
	Bad	**14537.0541**	0.2705	3932.2731
Sentences (Not Normalized)	Good	**3.0656**	**0.6607**	**3.7263**
	Bad	1.2655	0.2705	1.0607
DQIC2	Good	-	-	**8668.3012**
	Bad	-	-	6636.3641

Table 49: SNLI Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	Good	299.2489	0.9223	275.9972
	Bad	**1026.2828**	**1.0000**	**1026.2828**
Adjectives	Good	147.7382	**1.0000**	147.7382
	Bad	**333.8001**	**1.0000**	**333.8001**
Adverbs	Good	14.9467	0.5166	7.7214
	Bad	**54.2488**	**0.7318**	**39.6992**
Verbs	Good	76.0906	0.6893	52.4492
	Bad	**182.7695**	**0.7130**	**130.3146**
Nouns	Good	225.1162	**0.9726**	218.9480
	Bad	**477.5051**	0.9704	**463.3709**
Bigrams	Good	4394.8945	**1.0000**	4394.8945
	Bad	**5615.4581**	**1.0000**	**5615.4581**
Trigrams	Good	16628.8816	0.9907	16474.2330
	Bad	**35285.2261**	**0.9735**	**34350.1676**
Sentences	Good	**15197.5684**	0.0049	74.4680
	Bad	11085.6756	**0.9680**	**10730.9339**
Sentences (Not Normalized)	Good	1.2314	0.0049	0.0060
	Bad	**11.1732**	**0.9680**	**10.8156**
DQIC2	Good	-	-	21646.4558
	Bad	-	-	**52700.84312**

Table 50: MNLI Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	Good	138.6878	**0.6744**	93.5310
	Bad	**615.0626**	0.6224	**382.8149**
Adjectives	Good	37.0775	**1.0000**	37.0775
	Bad	**161.0191**	**1.0000**	**161.0191**
Adverbs	Good	4.0080	0.7473	2.9951
	Bad	**18.7378**	**0.7610**	**14.2594**
Verbs	Good	30.1469	0.9051	27.2859
	Bad	**152.9500**	**0.9372**	**143.3447**
Nouns	Good	58.5576	**1.0000**	58.5576
	Bad	**255.8677**	1.0000	**255.8677**
Bigrams	Good	1665.8142	**0.9763**	1626.3344
	Bad	**4563.8191**	0.9755	**4452.0055**
Trigrams	Good	20526.6346	**1.0000**	20526.6346
	Bad	**39155.8925**	0.9821	**38455.0020**
Sentences	Good	**4811.1347**	-0.0013	-6.2544
	Bad	1996.9248	**0.2460**	**491.2435**
Sentences (Not Normalized)	Good	0.3991	-0.0013	-0.0005
	Bad	**1.3043**	**0.2460**	**0.3208**
DQIC2	Good	-	-	22366.1613
	Bad	-	-	**44355.87788**

Table 51: SQUAD 2.0 Sub-Component and Overall Values for DQI_{c2}, Good Split

Granularity	Split	T1	T2	Contribution
Words	Good	**396.9190**	**0.3661**	**145.3120**
	Bad	52.3560	0.3239	16.9581
Adjectives	Good	**77.3987**	**0.8307**	**64.2951**
	Bad	70.2610	0.8020	56.3493
Adverbs	Good	17.3230	0.4292	7.4350
	Bad	**27.8482**	**0.6178**	**17.2046**
Verbs	Good	59.4638	**0.5936**	**35.2977**
	Bad	**63.3871**	0.5511	34.9326
Nouns	Good	**270.8688**	0.8953	**242.5088**
	Bad	250.9358	**0.9289**	233.0942
Bigrams	Good	**4116.6448**	**1.0000**	**4116.6448**
	Bad	2991.6306	**1.0000**	2991.6306
Trigrams	Good	**30424.4890**	**1.0000**	**30424.4890**
	Bad	17757.2356	0.9383	16661.6141
Sentences	Good	**8161.7926**	-0.0015	-12.2426
	Bad	2544.5235	**0.0000**	**0.0000**
Sentences (Not Normalized)	Good	2.1199	-0.0015	-0.0031
	Bad	**2.1204**	**0.0000**	**0.0000**
DQIC2	Good	-	-	**35023.73666**
	Bad	-	-	20011.78371

Table 52: Story CLOZE Sub-Component and Overall Values for DQI_{c2}, Good Split

Parameter 2: Tables 52-55 contain values for Parameter 2 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Parameter 3: The following tables contain values for Parameter 3 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	9.1320	11.3955	14.3267
Bad	**10.3842**	**13.1062**	**16.6390**

Table 53: SNLI Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0468**	**0.0244**	**0.0103**
Bad	0.0404	0.0216	0.0094

Table 54: SNLI Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	9.4123	11.4508	14.3370
Bad	**10.3936**	**13.1156**	**16.7024**

Table 55: SNLI DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**334.2154**	**695.0772**	**1040.5142**
Bad	312.4684	643.3308	953.5445

Table 56: MNLI Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0148**	**0.0108**	**0.0067**
Bad	**0.0111**	**0.0084**	**0.0056**

Table 57: MNLI Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**334.2221**	**695.0839**	**1040.5209**
Bad	312.474	643.3364	953.5501

Table 58: MNLI DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**129.8631**	**171.7117**	**228.9109**
Bad	88.9812	110.6097	141.2737

Table 59: SQUAD 2.0 Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	0.0051	0.0039	0.0026
Bad	**0.0055**	**0.0042**	**0.0094**

Table 60: SQUAD 2.0 Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**129.8657**	**171.7143**	**228.9135**
Bad	88.984	110.6125	141.2765

Table 61: SQUAD 2.0 DQI_{C3}

Split	SIML=0.3	SIML=0.35	SIML=0.4
Good	**285.1348**	**513.1720**	**820.2516**
Bad	209.0823	368.5646	594.0969

Table 62: Story CLOZE Term 1 for DQI_{c3}

Split	e=0.25	e=0.33	e=0.5
Good	**0.0069**	**0.0053**	**0.0036**
Bad	**0.0069**	**0.0053**	**0.0036**

Table 63: Story CLOZE Term 2 for DQI_{c3}, with SIML=0.4

Sample Set	DQI C3 (e=0.5)		
	SIM=0.5	SIM=0.6	SIM=0.7
Good	**285.1384**	**513.1756**	**820.2552**
Bad	209.0859	368.5682	594.1005

Table 64: Story CLOZE DQI_{C3}

Parameter 4: The following tables contain values for Parameter 4 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	DQIC4
Good	**0.0004**
Bad	0.0001

Table 65: SNLI DQI_{c4}

Split	DQIC4
Good	**0.0197**
Bad	0.0011

Table 66: MNLI DQI_{c4}

Split	DQIC4
Good	**5.2208**
Bad	0.4577

Table 67: SQUAD 2.0 DQI_{c4}

Split	DQIC4
Good	**0.0025**
Bad	0.0008

Table 68: Story CLOZE DQI_{c4}

Parameter 5: The following tables contain values for Parameter 5 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	**2.2349**	**2.8763**	**4.0125**	**6.3065**
Bad	2.2215	2.8558	3.9784	6.2237

Table 69: SNLI Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.1439**	**0.0038**	**6.4064e-05**	**20.3518**	**0.0903**
Bad	0.1430	0.0007	1.2711e-05	19.9288	0.0900

Table 70: SNLI Terms 2,3,4,5,6 for DQI_{c5}

Split	DQI C5
Good	**24.6024**
Bad	24.1409

Table 71: SNLI DQI_{c5}, with ISIM=0.5

Split	ISIM=0.3	ISIM=0.4	ISIM=0.5	ISIM=0.6
Good	**2.2233**	**2.8585**	**3.9884**	**6.3364**
Bad	2.1256	2.6986	3.6843	5.5845

Table 72: MNLI Term 1 for DQI_{c5}

Split	T2	T3	T4	T5	T6
Good	**0.0791**	0.0162	1.1073E-05	**15.3835**	14.7547
Bad	0.0741	**0.0307**	**20.9407E-05**	12.3932	**17.6181**

Table 73: MNLI Terms 2,3,4,5,6 for DQI_{c5}

Split	**DQI C5**
Good	**34.2219**
Bad	33.8006

Table 74: MNLI DQI_{c5}, with ISIM=0.5

Split	**ISIM=0.3**	**ISIM=0.4**	**ISIM=0.5**	**ISIM=0.6**
Good	2.5073	3.3460	5.0031	9.1300
Bad	**2.5379**	**3.4012**	**5.1352**	**9.6189**

Table 75: SQUAD 2.0 Term 1 for DQI_{c5}

Split	**T2**	**T3**	**T4**	**T5**	**T6**
Good	**0.0085**	0.0052	7.3081E-06	22.9314	**102.9990**
Bad	0.0079	**0.0524**	**7.4403E-05**	**27.0966**	88.8872

Table 76: SQUAD 2.0 Terms 2,3,4,5,6 for DQI_{c5}

Split	**DQI C5**
Good	**130.9472**
Bad	121.1793

Table 77: SQUAD 2.0 DQI_{c5}, with ISIM=0.5

Split	**ISIM=0.3**	**ISIM=0.4**	**ISIM=0.5**	**ISIM=0.6**
Good	**3.1103**	**4.5013**	**7.7337**	14.4898
Bad	3.0639	4.4163	7.5943	**14.7772**

Table 78: Story CLOZE Term 1 for DQI_{c5}

Split	**T2**	**T3**	**T4**	**T5**	**T6**
Good	**0.0400**	0.0027	3.1939E-05	**0.0400**	2.6196e-06
Bad	0.0398	**0.0084**	**9.7664E-05**	0.0398	**7.6306e-06**

Table 79: Story CLOZE Terms 2,3,4,5,6 for DQI_{c5}

Split	**DQI C5**
Good	**7.8164**
Bad	7.6824

Table 80: Story CLOZE DQI_{c5}, with ISIM=0.5

Parameter 6: The following tables contain values for Parameter 6 across SNLI, MNLI, SQUAD 2.0, and Story CLOZE.

Split/Label	**Entailment**	**Neutral**	**Contradiction**
Good	1110	1430	708
Bad	5626	5008	6118

Table 81: SNLI Sample counts for Splits across Labels

Split-Label	**T1**	**T2**
Good-Entailment	8829.2425	**0.9387**
Bad-Entailment	**21655.2868**	0.8571
Good-Neutral	7467.5349	0.8699
Bad-Neutral	**31616.2545**	**0.9141**
Good-Contradiction	4932.7421	**0.9210**
Bad-Contradiction	**29145.0957**	0.8783

Table 82: SNLI Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	**T1**	**T2**
Good-Entailment	**142.8571**	**0.7277**
Bad-Entailment	81.9672	0.6110
Good-Neutral	**153.8462**	**0.9118**
Bad-Neutral	117.6471	0.7071
Good-Contradiction	**163.9344**	**0.6764**
Bad-Contradiction	101.0101	0.6088

Table 83: SNLI Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	**T1**	**T2**
Good-Entailment	**42.1230**	**0.34114**
Bad-Entailment	26.4201	0.30551
Good-Neutral	**48.8998**	0.46865
Bad-Neutral	38.1534	**0.47497**
Good-Contradiction	**43.1593**	0.31019
Bad-Contradiction	29.2826	**0.32385**

Table 84: SNLI Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	**T1**	**T2**
Good-Entailment	**18.4128**	0.056911
Bad-Entailment	11.0963	**0.05816**
Good-Neutral	8.6798	0.09709
Bad-Neutral	**14.6135**	**0.43124**
Good-Contradiction	**37.9795**	**0.34286**
Bad-Contradiction	23.7192	0.21583

Table 85: SNLI Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	**T1**	**T2**
Good-Entailment	**41.7885**	**0.16091**
Bad-Entailment	22.9410	0.05348
Good-Neutral	**48.9476**	0.17946
Bad-Neutral	38.9105	**0.20192**
Good-Contradiction	**53.5045**	**0.20000**
Bad-Contradiction	34.6380	0.13589

Table 86: SNLI Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	**T1**	**T2**
Good-Entailment	**59.2768**	**0.49650**
Bad-Entailment	34.3643	0.38238
Good-Neutral	**62.7353**	**0.44534**
Bad-Neutral	46.4253	0.40586
Good-Contradiction	**66.3570**	**0.45653**
Bad-Contradiction	39.9202	0.37431

Table 87: SNLI Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-Entailment	1131.7133	**0.93307**
Bad-Entailment	**1173.5409**	0.93206
Good-Neutral	1261.2663	0.93783
Bad-Neutral	**1598.1514**	**0.94117**
Good-Contradiction	1100.8597	**0.94325**
Bad-Contradiction	**1369.0528**	0.93387

Table 88: SNLI Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-Entailment	5921.2942	**0.94672**
Bad-Entailment	**7757.5306**	0.93496
Good-Neutral	6414.8208	0.94517
Bad-Neutral	**10229.7186**	**0.95015**
Good-Contradiction	5478.1014	**0.95359**
Bad-Contradiction	**8984.3224**	0.94430

Table 89: SNLI Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Repetition	1	2	3	4	5	6
Good-Entailment	0.9844	0.0155	0	0	0	0
Bad-Entailment	0.9659	0.0308	0.001849	0	0.0007	0.0005
Good-Neutral	0.9667	0.0325	0.0007	0	0	0
Bad-Neutral	0.9785	0.0204	0.0010	0	0	0
Good-Contradiction	0.9798	0.0201	0	0	0	0
Bad-Contradiction	0.9785	0.0204	0.0010	0	0	0

Table 90: SNLI Sentence Granularity Repetitions

Split-Label	T3
Good-Entailment	**0.1457**
Bad-Entailment	0.1330
Good-Neutral	0.1496
Bad-Neutral	**0.1571**
Good-Contradiction	0.1313
Bad-Contradiction	**0.1434**

Table 91: SNLI T3 for DQI_{c6}

Split-Label	T4
Good-Entailment	**0.0100**
Bad-Entailment	0.0021
Good-Neutral	**0.0084**
Bad-Neutral	0.0022
Good-Contradiction	0.0197
Bad-Contradiction	**0.0020**

Table 92: SNLI T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	**15.3475**	11.6614
Words	**0.9313**	0.6596
Adjectives	**1.2190**	0.9185
Adverbs	**1.5708**	1.1850
Verbs	**0.9667**	0.7001
Nouns	**1.0623**	0.7358
Bigrams	0.3646	**0.4893**
Trigrams	0.1860	**0.2760**

Table 93: SNLI T5 for DQI_{c6}

Split-Label	DQI C6
Good	**556.6914**
Bad	320.2893

Table 94: SNLI DQI_{c6}

Split/Label	Entailment	Neutral	Contradiction
Good	6150	6098	6082
Bad	700	60	240

Table 95: MNLI Sample counts for Splits across Labels

Split-Label	T1	T2
Good-Entailment	2.69E+04	0.8133
Bad-Entailment	6.47E+03	0.9542
Good-Neutral	2.78E+04	0.8465
Bad-Neutral	4.76E+16	1.0000
Good-Contradiction	4.62E+04	0.9378
Bad-Contradiction	1.05E+17	1.0000

Table 96: MNLI Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-Entailment	5.67E+02	0.970607701
Bad-Entailment	9.48E+02	0.957116548
Good-Neutral	8.70E+02	0.488048002
Bad-Neutral	6.74E+02	0.794573643
Good-Contradiction	9.40E+02	0.965482191
Bad-Contradiction	7.01E+02	0.885763001

Table 97: MNLI Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-Entailment	1.16E+02	0.7834
Bad-Entailment	2.83E+02	1.0000
Good-Neutral	2.86E+02	1.0000
Bad-Neutral	1.92E+02	0.8771
Good-Contradiction	3.47E+02	1.0000
Bad-Contradiction	2.67E+02	1.0000

Table 98: MNLI Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-Entailment	2.56E+01	0.4803
Bad-Entailment	5.20E+01	0.6531
Good-Neutral	3.61E+01	0.6091
Bad-Neutral	7.15E+01	0.6521
Good-Contradiction	3.43E+01	0.5017
Bad-Contradiction	5.19E+01	0.3939

Table 99: MNLI Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-Entailment	1.71E+02	0.7901
Bad-Entailment	1.61E+02	0.6620
Good-Neutral	1.43E+02	0.5911
Bad-Neutral	1.69E+02	0.3061
Good-Contradiction	1.79E+02	0.7271
Bad-Contradiction	1.30E+02	0.5636

Table 100: MNLI Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-Entailment	2.61E+02	0.8994
Bad-Entailment	4.52E+02	0.9447
Good-Neutral	4.68E+02	1.0000
Bad-Neutral	2.61E+02	0.7235
Good-Contradiction	4.84E+02	1.0000
Bad-Contradiction	2.80E+02	0.9287

Table 101: MNLI Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-Entailment	3.38E+03	0.9361
Bad-Entailment	4.83E+03	1.0000
Good-Neutral	9.21E+03	1.0000
Bad-Neutral	1.91E+03	1.0000
Good-Contradiction	1.04E+04	1.0000
Bad-Contradiction	2.97E+03	1.0000

Table 102: MNLI Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-Entailment	9.27E+03	0.9573
Bad-Entailment	2.93E+04	1.0000
Good-Neutral	4.54E+04	0.9913
Bad-Neutral	4.61E+03	0.8822
Good-Contradiction	1.04E+05	1.0000
Bad-Contradiction	6.96E+03	0.9937

Table 103: MNLI Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Repetition	1	2	3
Good-Entailment	0.9512	0.0484	0.0003
Bad-Entailment	0.9884	0.0115	0.0000
Good-Neutral	0.9612	0.0363	0.0024
Bad-Neutral	1.0000	0.0000	0.0000
Good-Contradiction	0.9844	0.0150	0.0005
Bad-Contradiction	1.0000	0.0000	0.0000

Table 104: MNLI Sentence Granularity Repetitions

Split-Label	T3
Good-Entailment	0.0647
Bad-Entailment	0.0860
Good-Neutral	0.0926
Bad-Neutral	0.0590
Good-Contradiction	0.1000
Bad-Contradiction	0.2290

Table 105: MNLI T3 for DQI_{c6}

Split-Label	T4
Good-Entailment	0.0803
Bad-Entailment	0.0202
Good-Neutral	0.0041
Bad-Neutral	0.0484
Good-Contradiction	0.2018
Bad-Contradiction	0.0326

Table 106: MNLI T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	14.6049	72.1687
Words	1.2571	0.8533
Adjectives	1.4557	1.7959
Adverbs	0.7319	0.9429
Verbs	1.0382	1.0345
Nouns	1.7755	1.5836
Bigrams	0.4008	0.3561
Trigrams	0.6547	0.9724

Table 107: MNLI T5 for DQI_{c6}

Split-Label	DQI C6
Good	2.74E+05
Bad	1.53E+17

Table 108: MNLI DQI_{c6}

Split/Label	True	False
Good	10946	10770
Bad	914	1086

Table 109: SQUAD 2.0 Sample counts for Splits across Labels

Split-Label	T1	T2
Good-True	4431.2159	0.0007
Bad-True	1921.2260	0.5448
Good-False	4412.2037	0.0014
Bad-False	1853.6963	0.5009

Table 110: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-True	263.6776	1.0000
Bad-True	954.5225	1.0000
Good-False	259.3381	0.3105
Bad-False	776.2031	1.0000

Table 111: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-True	75.3820	1.0000
Bad-True	244.8719	1.0000
Good-False	70.8210	1.0000
Bad-False	222.5754	1.0000

Table 112: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-True	6.31677	0.6666
Bad-True	27.6740	0.6494
Good-False	6.4805	0.6632
Bad-False	24.6482	0.6878

Table 113: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-True	58.2850	0.8789
Bad-True	219.8726	0.8851
Good-False	59.0344	0.9066
Bad-False	208.3846	0.9113

Table 114: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-True	110.8118	1.0000
Bad-True	415.9473	1.0000
Good-False	109.7139	1.0000
Bad-False	307.1137	1.0000

Table 115: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-True	2923.9305	0.9768
Bad-True	5800.9793	0.9762
Good-False	2834.7978	0.9758
Bad-False	5157.4516	0.9749

Table 116: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-True	35363.3144	1.0000
Bad-True	49074.7258	1.0000
Good-False	34076.1381	1.0000
Bad-False	40854.1931	1.0000

Table 117: SQUAD 2.0 Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Label	T3
Good-True	0.0085
Bad-True	0.00852
Good-False	0.0079
Bad-False	0.0078

Table 118: SQUAD 2.0 T3 for DQI_{c6}

Split-Label	T4
Good-True	0.0104
Bad-True	0.0106
Good-False	0.1165
Bad-False	0.0954

Table 119: SQUAD 2.0 T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	20.5287	9.6533
Words	0.0711	0.0682
Adjectives	0.6497	1.1487
Adverbs	0.4012	0.6832
Verbs	0.4918	0.8153
Nouns	0.5183	0.9957
Bigrams	0.1262	0.05600
Trigrams	0.1366	0.09422

Table 120: SQUAD 2.0 T5 for DQI_{c6}

Split-Label	DQI C6
Good	75918.2760
Bad	105949.3404

Table 121: SQUAD 2.0 DQI_{c6}

Split/Label	True	False
Good	2568	2568
Bad	800	800

Table 122: Story CLOZE Sample counts for Splits across Labels

Split-Label	T1	T2
Good-True	1.30E+05	0.9984
Bad-True	5.06E+16	1.0000
Good-False	1.30E+05	0.9984
Bad-False	5.06E+16	1.0000

Table 123: Story CLOZE Terms 1 and 2 for DQI_{c6}, Sentence Granularity

Split-Label	T1	T2
Good-True	5.47E+05	0.9792
Bad-True	5.22E+05	0.8618
Good-False	5.47E+05	0.5316
Bad-False	4.96E+05	0.8537

Table 124: Story CLOZE Terms 1 and 2 for DQI_{c6}, Word Granularity

Split-Label	T1	T2
Good-True	129.1883	0.7800
Bad-True	133.5904	0.7711
Good-False	121.0435	0.7459
Bad-False	128.3632	0.8014

Table 125: Story CLOZE Terms 1 and 2 for DQI_{c6}, Adjective Granularity

Split-Label	T1	T2
Good-True	41.1600	0.5959
Bad-True	49.9482	0.5368
Good-False	36.9653	0.6145
Bad-False	54.7544	0.6194

Table 126: Story CLOZE Terms 1 and 2 for DQI_{c6}, Adverb Granularity

Split-Label	T1	T2
Good-True	103.8261	0.5285
Bad-True	115.6968	0.5828
Good-False	112.3307	0.5946
Bad-False	113.4481	0.5155

Table 127: Story CLOZE Terms 1 and 2 for DQI_{c6}, Verb Granularity

Split-Label	T1	T2
Good-True	551.3272	0.8898
Bad-True	458.9138	0.8862
Good-False	520.3204	0.9047
Bad-False	462.2876	0.9252

Table 128: Story CLOZE Terms 1 and 2 for DQI_{c6}, Noun Granularity

Split-Label	T1	T2
Good-True7139.05776	1.0000	
Bad-True5158.2473	1.0000	
Good-False6941.1989	1.0000	
Bad-False5006.1656	1.0000	

Table 129: Story CLOZE Terms 1 and 2 for DQI_{c6}, Bigram Granularity

Split-Label	T1	T2
Good-True54497.5504	1.0000	
Bad-True33876.9502	1.0000	
Good-False50906.0915	1.0000	
Bad-False33618.6103	1.0000	

Table 130: Story CLOZE Terms 1 and 2 for DQI_{c6}, Trigram Granularity

Split-Label	T3
Good-True	0.0085
Bad-True	0.0079
Good-False	0.0085
Bad-False	0.0078

Table 131: Story CLOZE 2.0 T3 for DQI_{c6}

Split-Label	T4
Good-True	0.0104
Bad-True	0.1165
Good-False	0.0106
Bad-False	0.0954

Table 132: Story CLOZE 2.0 T4 for DQI_{c6}

Granularity/Split	Good	Bad
Sentences	382.2842	2262.7417
Words	1.0447	1.0192
Adjectives	3.9910	5.0527
Adverbs	1.7714	3.1284
Verbs	2.2377	3.5188
Nouns	5.8841	7.3696
Bigrams	1.6522	1.9489
Trigrams	4.9660	6.8154

Table 133: Story CLOZE T5 for DQI_{c6}

Split-Label	DQI C6
Good	1.01E+17
Bad	1.01E+17

Table 134: Story CLOZE DQI_{c6}

Parameter 7: The following tables contain values for Parameter 7 across SNLI, MNLI, and SQUAD 2.0. Story CLOZE does not have a separate training set and is hence not evaluated.

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	**0.0031**	**0.0042**	**0.0063**
Bad	0.0029	0.0040	0.0057

Table 135: SNLI DQI_{c7}

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	0.0004	0.0005	0.0002
Bad	**0.0009**	**0.0011**	**0.0005**

Table 136: MNLI DQI_{c7}

Split	SSMIL=0.2	SSMIL=0.3	SSMIL=0.4
Good	**1.2500**	**1.4285**	**1.6666**
Bad	0.0029	0.0040	0.0057

Table 137: SQUAD 2.0 DQI_{c7}