# LT2326 Final project report

## Introduction and background

Natural language inference (NLI) is a staple natural language understanding benchmark task, as shown in part in its multiple inclusions in, for example, GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019). The task allows for some variety of formulation, but briefly put involves the truth-assessment of a hypothesis *h* given the assumed truth of a premise *p* (the input, collectively called an *inference pair* going forward), and an output label (typically *entailment*, *neutral*, or *contradiction*) to classify the relationship between the inference pair. Quick progress has been made on many NLI (and other NLU) benchmarks, shown for example in the quite brief window between GLUE and SuperGLUE. However, some researchers (see for example Schlangen 2021, Bender & Koller 2020, and Paullada et al. 2020) have raised concerns that the representations models often derive to solve a task might not correspond to the reasoning humans use to solve that same task. More specifically for NLI, Gururangan et al. (2018) trained a model on only the hypothesis half of inference pairs to perform well over chance on some prominent datasets: 67% on Stanford NLI (Bowman et al. 2015, henceforth SNLI) and ~54/~52 on MutliNLI matched/mismatched (Williams et al. 2018, MNLI), datasets with 3 well-balanced classes. Since NLI classification is of a cross-pair relation, this model's relative success cannot be attributed to what humans would use to solve the task. To combat problems of models using spurious patterns to solve the task, Adversarial NLI (Nie et al. 2020a, ANLI) was developed iteratively across rounds, each round with updated data in order to produce a challenging dataset. Relatedly, while the task has to be formulated with discrete labels to be able to interface with machine learning classification, human labels on the task typically have higher inter-annotator disagreement than can be attributed to error. This is acknowledged by Bowman et al. For SNLI (2015, p. 635) before high-disagreement pairs are nonetheless left unused for practical purposes, and further explored in e.g., Pavlick and Kwiatkowski (2019) and Nie et al. (2020b).

This project seeks to explore the relationship between model uncertainty and human judgment variation, and compare this relationship to typical classification metrics by answering these questions for one particular model:

- How does the model perform on some typical metrics (accuracy and macro F1)?
- How does the model's prediction uncertainty compare to human judgment distribution variation (both measured in entropy)?

Both lines of questioning are assessed both across epochs and across ANLI-style rounds.

## Model and data

All data size references reflect pre-preprocessing/data selection sizes. All datasets use the typical 3-class labelling system unless otherwise noted.

### Model

The assessed model is a pretrained BERT model (Devlin at al. 2019), finetuned for NLI on data selected for each round with the following hyperparameters: 6 epochs, 32 batch size, 5e-5 (constant, no scheduling) learning rate. Due to the computational load of full runs, these were not experimentally tuned except for the need for one full rerun due to an originally too high learning rate; it became clear just how strict BERT's aversion for a too-high learning rate actually is when the original run not only came back always predicting the same class, but even nearly the same logits.

### Training data

*SNLI (Bowman et al. 2015)*

This dataset contains ~570,000 inference pairs, collected through crowd-work (Amazon Mechanical Turk). Annotators are presented with a photo caption without the photo (used as premise) must provide three alternate descriptions: one that is definitely true (entailment), one that may be true (neutral) and one that is definitely false, thus providing three inference pairs. The datasets have some problems already outlined in the previous section, but more specifically: annotators tend to use premise-agnostic strategies for producing a given label, such as general language for entailments (e.g., existential quantifiers, likely hypernyms: dog --> animal), adjectival modifiers and purpose clauses for neutral (extra information that usually is not precluded by the premise), and negations and activity-precluding verbs for contradiction (e.g., *sleeping*) (Gururangan et al. 2018). This should not be a problem for the present project since its purpose is to investigate some effects of variables like these, not to train the best model possible. See note at top of references for data. Not included in submission due to size.

*MNLI (Williams et al. 2018)*

MNLI contains ~433,000 inference pairs. Hypothesis were collected in the same way as for SNLI, but premises are instead drawn from various genres, such as face-to-face conversations, letters, and government websites. Evaluation sets come in two kinds: from genres matching the training set (MNLI-m) and mismatched genres (MNLI-mm). As noted in the background, MNLI has some of the same issues as SNLI, but to a lesser degree. See note at top of references for data. Not included in submission due to size.

*ANLI (Nie et al. 2020a)*

ANLI was produced iteratively in-loop with a model. A BERT-large was trained on a concatenation of SNLI and MNLI (SMNLI, henceforth), and the best checkpoint served as a starting point for ANLI collection. For each round, an annotator is presented with a premise (sampled from Wikipedia for all rounds of collection, with some additional source genres for round 3) and a target label, for which the annotator produces a hypothesis. The candidate inference pair is input to the model, and the process is repeated until the model fails the prediction or an attempt threshold is reached. The resulting dataset of adversarial inference pairs is incorporated into the training of a new model (RoBERTa for rounds 2 and 3), which will be the model against which the next round of data collection is measured. In their respective training sets, ANLI1 has ~17,000 pairs, ANLI2 ~ 45,000, and ANLI3 ~100,000. The ANLI procedure is the inspiration for the round-based methodology of the present project. See references for data. Not included in submission due to size.

**Evaluation data**

*ChaosNLI (Nie at al. 2020b, CNLI)*

CNLI incorporates 100 reannotations of parts of popular NLI sets. The present project uses two out of three subsets: one based on SNLI (~1500, henceforth CNLI-S) and one on MNLI-m (~1600, CNLI-M). The third subset uses a binary classification NLI formulation, so was left out due interfacing and comparability issues. CNLI comes with traditional discrete labels (so can be used for computation of traditional performance metrics) but also include the label distributions from the reannotation effort. None of the inference pairs are from the training data of their respective original sets, and can thus be used to evaluate a model trained on that data. See references for data.

*NLI variation*[1] *(Pavlick and Kwiatkowski 2019, NLIvar)*

This dataset reannotates a mixed dataset from several original sources, including SNLI and MNLI, for human variation. Data comparison revealed that some of the inference pairs from these sets are from the training data of their respective original sets, which is addressed in the code. NLIvar annotators are instructed to grade each pair on a scale from –50 (definitely not true) to 50 (definitely true), with 50 annotators per pair. To interface with the three-class uncertainty distribution the model outputs and for a comparable entropy computation with the model and CNLI, these labels are descretised in a similar manner and with similar thresholds as the authors use for an experiment in the original paper (p. 685): < –16.7 = contradiction, > 16.7 = entailment, and in-between = neutral. This approach loses the benefit of the grading scale, but retains the inter-annotator distribution data in a more comparable form. These makeshift labels are only used for entropy comparison with humans and not for the computation of the traditional metrics. See references for data.

## Methodology

A new BERT model (accessed from *huggingface*, link in references) is finetuned four sequential rounds on updated data, somewhat (but not exactly) mirroring the round data update of ANLI. For computational reasons, round 1 data consists of a reduced sample of SMNLI of length equal to the total number of inference pairs in ANLI across all its training rounds (~163,000). Each round a new ANLI subset is added, taking the place of a number of pairs equal to its length (always from the remaining SMNLI data, never from ANLI data added in a previous round). This way the amount of data per round is always the same and not a variable to account for in the cross-round comparison. The data input across rounds is as in table 1.

**Table 1. Per-round training data**

| ROUND | DATA | SIZE (N INFERENCE PAIRS) |
|---|---|---:|
| 1 | SMNLI sample | 162,865 |
| 2 | Reduced SMNLI + ANL1 | 162,865 |
| 3 | Further reduced SMNLI + ANLI1 + ANLI2 | 162,865 |
| 4 | Further reduced SMNLI + ANLI1 + ANLI2 + ANLI3 | 162,865 |

A model checkpoint is saved for each of the 6 epochs for each round. After training, accuracy, macro f1, and model-human entropy correlation is computed and tracked across epochs and rounds. Entropy is Shannon entropy as per equation 1 per inference pair, where C is the label set, $p(c_i)$ is the probability of each label for the pair. The probability of each p(c) is assumed to be normalised between 0 and 1.[2] Correlation is Pearson, computed per epoch.[3]

[1] NLI variation is the name of the dataset on GitHub, but this is not an otherwise named dataset.

[2] Tried some different computations for this, including computing it manually, and it is not strictly necessary for the SciPy function I end up using, but I cannot pass logits into it, so if I'm normalising the model output I might as well normalise the human distribution as well.

$$H(C) = -\sum_{c_i \in C}\left(p(c_i) * \log_2\left(p(c_i)\right)\right) \quad (1)$$

## Results

All results refer to refer to those of the creatively named model in models/bert-base-uncased-bs32-eps6-lr5e-05. See readme for filesystem navigation information.

Figures[4] 1-4 (for rounds 1-4, respectively) show accuracy, m-f1, and model/human entropy correlations across epochs. As seen in the figures, the model's predictions are quite stable across epochs. Given this, only the final epoch's cross-round comparison is included here (figure 5). Also mostly left out are the scatterplots of the kind in figure 6; showing the distribution the entropy correlations are based on. These exist for each round of each epoch for each evaluation set, and thus have a lot of information at the cost of being less interpretable in a generalisable manner. The one inclusion here is just to showcase its existence.

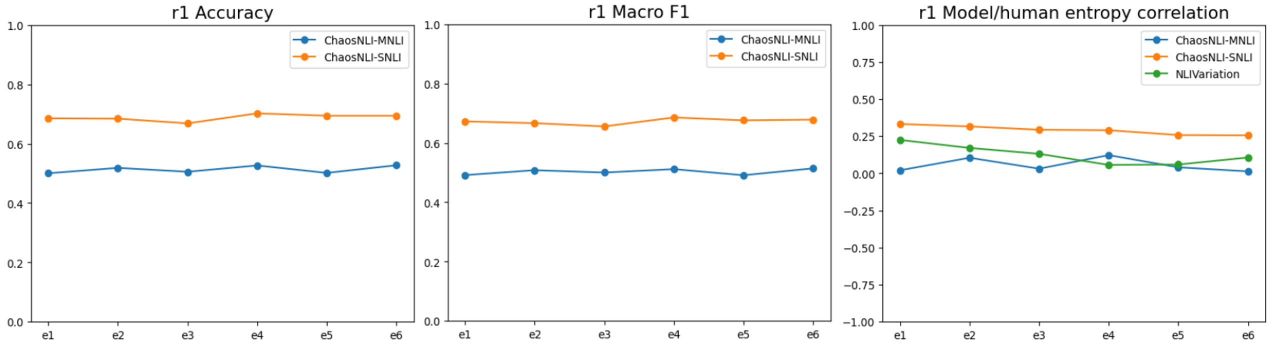

**Figure 1. Round 1 accuracy, macro F1, Model/human entropy correlation across epochs.**

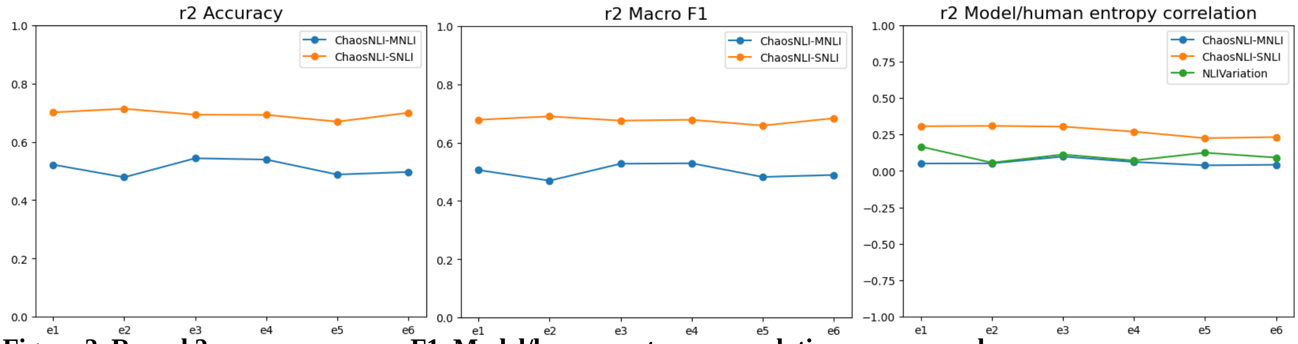

**Figure 2. Round 2 accuracy, macro F1, Model/human entropy correlation across epochs.**

[4] The figures themselves were done properly in Python, but note that the grids are a bit of a hackjob, as I could not justify spending the time to do it in properly at this point.

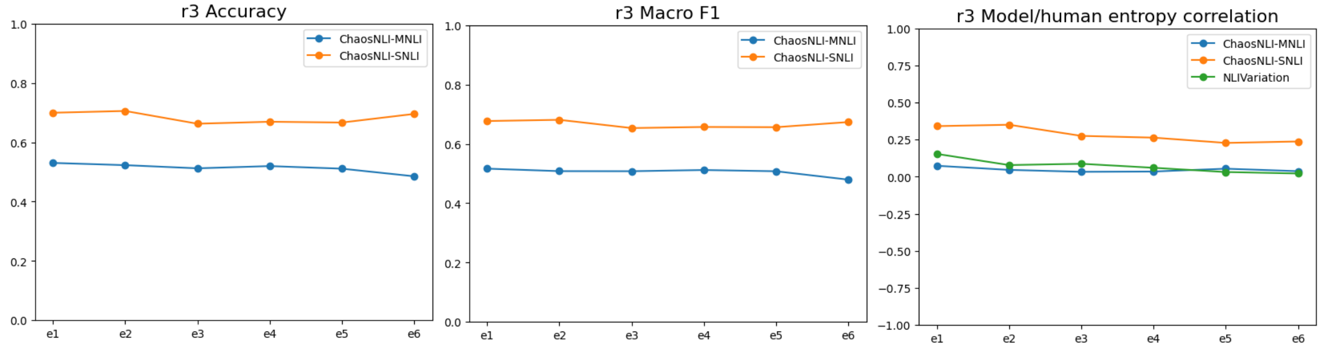

**Figure 3. Round 3 accuracy, macro F1, Model/human entropy correlation across epochs.**

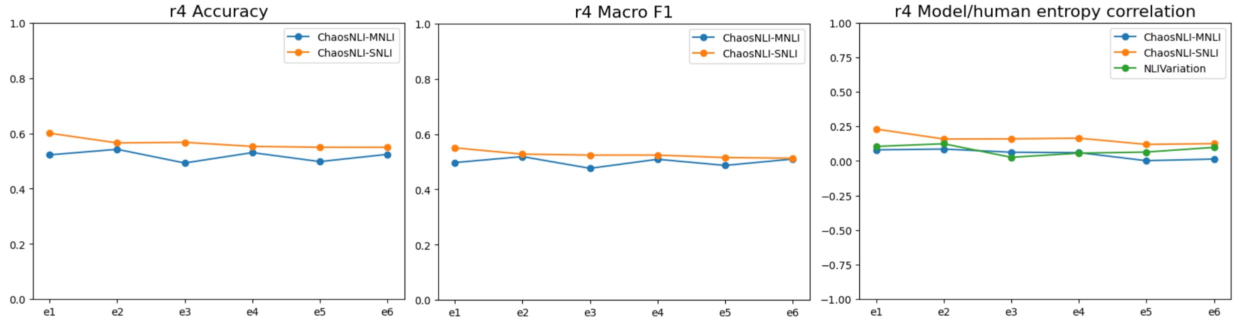

**Figure 4. Round 3 accuracy, macro F1, Model/human entropy correlation across epochs.**

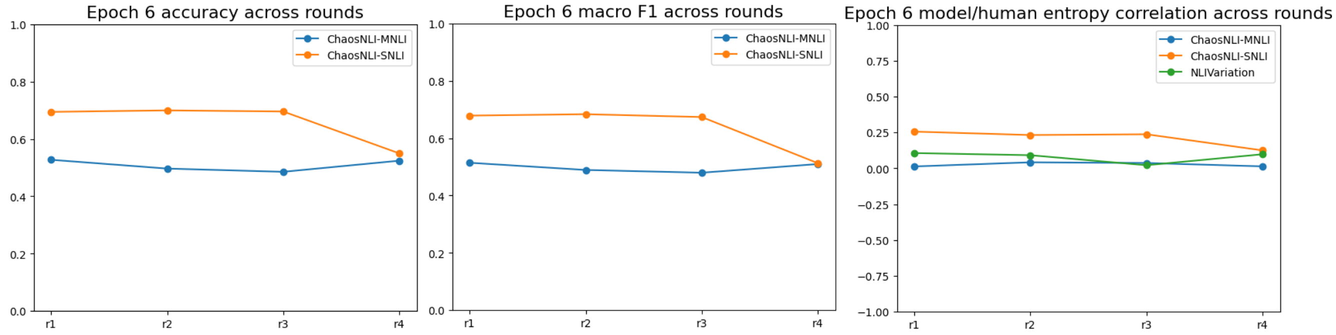

**Figure 5. Final epoch accuracy, macro F1, Model/human entropy correlation across rounds.**

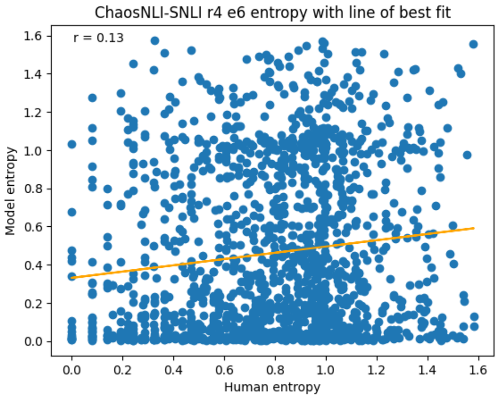

**Figure 6. Example of entropy data for one epoch checkpoint, one evaluation set (CNLI-S). Top left is correlation for distribution.**

## Analysis and discussion

As noted, the model does not improve or even change much across epochs, suggesting it either learned most of what it was going to learn, or it needed different hyperparameters. Since one training run already had to be discarded due to too high learning rate, it could be that the one used here (5e-05) is only on the upper end of what BERT wants. Training loss (not included here5) was typically very low, which, alongside the stagnant metrics, is another indicator that more epochs would not have solved the issue. The accuracies (and m-f1) that the model hovers at across both epochs and *most* rounds (a bit over 50% for CNLI-M and a bit under 70% for the CNLI-S) is interestingly more or less the same as the results Gururangan et al. (2018) got on these respective original datasets after having trained only on hypotheses. Whether the spurious pattern recognition is the same in both cases is difficult to say without further investigation.

Accuracy and m-f1 mirror each other very well, indicating that the model does not get to its above-chance performance by neglecting any classes.

Model/human entropy correlation barely (if at all) ever drops below 0, which is not necessarily a given if we assume model strategies for solving the task differ from humans. There is basically no effect size for CNLI-M, and NLIvar fluctuates at a weak correlation. Interestingly, the correlation is strongest (by a decent margin in some rounds/epochs) for CNLI-S. At face value, this is a good sign for the model; the set on which is scores the best in the traditional metrics is also where its uncertainty best aligns with human variation. This could be interpreted to validate the traditional score. However, SNLI is an easy set, both generally (as can be seen by the performance of Gururangan et al.'s hypothesis-only model), but also relative to the data fed to the model in the later rounds, since ANLI is adversarially constructed. While what passes as difficult for a model *can* be different from what passes for the same for a human, a generally easy dataset where annotator disagreement is avoided is likely to yield little human-interpreted ambiguity. The model only needs to have confident predictions to match this, regardless of how well a proxy the underlying reasoning is to that of a human.

Unlike across epochs, there is a notable development across rounds: in round 4. CNLI-S drops in the traditional metrics to match those of CNLI-M. Entropy correlation drops as well to a somewhat lesser degree, which strengthens the previous point (more difficult data --> more disagreement). Round 4

[5] By the time I realised I likely should plot training and validation loss, I could not justify the time. However, training loss does exist in the model filesystem at epoch-level.

adds ANLI3, which is the largest infusion of adversarial data of the project, as well as the portion of ANLI data produced last in the process against the most robust model version. Given this, it is perhaps more surprising that CNLI-M does not suffer than that CNLI-S does. Round 4 is the only one whose training data retains no SMNLI data, and the consistent cross-round outcome on CNLI-M might indicate that MNLI's data facilitates more generalisable evaluation. However, given that the model only scored 50% on the traditional metrics, the case might just be that there exists some exploitable pattern that this model just did not learn. Since Gururangan et al. still achieved 50% accuracy on only hypothesis, it is not unlikely.

Finally, there is an intuition that was not borne out in the results. If we assume more exploitable data in early rounds and less so in later rounds (particularly round 4), then later rounds should require more 'correct' reasoning, which we hope to be more human-like. Ideally, this should lead to uncertainty where humans vary and confidence when they are more uniform, which we can translate to high entropy correlation. There is a slight increase for NLIvar in round 4, but that is it. However, it is fair to say that this intuition was not actually tested; if the hypothesis[6] is that a model which learns non-spurious patterns (i.e., the targeted patterns) will have a level of uncertainty that matches human variation, then a test of the hypothesis requires that the model actually performs well. It does not tell us much that a model that does not perform terribly well also does not correlate with humans.

### Conclusion

The project was carried out largely according to the original intent, albeit with some changes. Its breadth somewhat overshadowed its depth; perhaps an additional metric could have been used instead of doing comparisons both across epochs and rounds. Relatedly, the rounds could have been designed more distinctly. The ANLI rounds are quite mismatched in size, so the difference between project rounds is uneven as a result. A better split might be round 1 --> sample of SNLIof len(ANLI), round 2 --> sample of MNLI of len(ANLI), round 3 --> ANLI (all rounds). This would also make the effects of a particular dataset clearer. The intuition behind the carried-out approach was to differentiate between different ANLI subsets, but in round 2, ANLI1 is likely too small a portion of the data to have a discernible impact.

Given the mediocre accuracy and m-f1, it is difficult to interpret the entropy correlations, as noted. The hypothesis might also be too simple; even if there had been positive results, it seems difficult to tie this strongly to human reasoning. For example, to reason like a human is not just to be certain and uncertain in parallel with humans. A near 100% confident entailment is as entropic as a similar contradiction, so entropy might not be exactly what should be sought. A more granular metric might be interesting for future work.

## References

**GitHub links are provided if the data from the present project used those repositories to get the data. SNLI and MNLI data files can be generated from the ANLI repository, which is where I got them from.**

[6] I use the word because it fits, not claiming I did proper hypothesis testing here

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5185-5198).

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Data accessed from ANLI GitHub

Devlin, J., Chang, M-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Accessed from huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020a). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Data accessed from github.com/facebookresearch/anli/tree/main

Nie, Y., Zhou, X., and Bansal, M. 2020b. What Can We Learn from Collective Human Opinions on Natural Language Inference Data?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9131–9143. Data accessed from github.com/easonnie/ChaosNLI

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. In *Patterns*, 2(11).

Pavlick, E., & Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. In *Transactions of the Association for Computational Linguistics*, 7, 677-694. Data accessed from github.com/epavlick/NLI-variation-data

Schlangen, D. (2021). Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674

Williams, A., Nangia, N., & Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Data accessed from ANLI GitHub

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in neural information processing systems*, 32.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355,