# On the Effects of Transformer Size on In- and Out-of-Domain Calibration

**Soham Dan**
University of Pennsylvania
sohamdan@seas.upenn.edu

**Dan Roth**
University of Pennsylvania
danroth@seas.upenn.edu

## Abstract

Large, pre-trained transformer language models, which are pervasive in natural language processing tasks, are notoriously expensive to train. To reduce the cost of training such large models, prior work has developed smaller, more compact models which achieves a significant speedup in training time while maintaining competitive accuracy to the original model on downstream tasks. Though these smaller pre-trained models have been widely adopted by the community, it is not known how well are they calibrated compared to their larger counterparts. In this paper, focusing on a wide range of tasks, we thoroughly investigate the calibration properties of pre-trained transformers, as a function of their size. We demonstrate that when evaluated in-domain, smaller models are able to achieve competitive, and often better, calibration compared to larger models, while achieving significant speedup in training time. Post-hoc calibration techniques further reduce calibration error for all models in-domain. However, when evaluated out-of-domain, larger models tend to be better calibrated, and label-smoothing instead is an effective strategy to calibrate models in this setting.
.

## 1 Introduction

Large pre-trained transformer language models like BERT (Devlin et al., 2019; Liu et al., 2019) have revolutionized natural language processing, achieving state-of-the-art results in several tasks. The process of applying these models on a downstream task consists of two components: (1) Self-supervised pre-training on a large amount of text corpora and (2) Supervised fine-tuning on the downstream task. Due to the very large number of parameters of such transformer based architectures, the high downstream accuracies comes at a large computational cost (Sharir et al., 2020; Bender et al., 2021) during the pre-training stage and also to a lesser extent, while fine-tuning. To alleviate this computational cost, several models with fewer parameters have been proposed that significantly speed-up both the pre-training and the fine-tuning stages (Turc et al., 2019; Lan et al., 2020; Sanh et al., 2019; Sun et al., 2020). For example, the smallest model in (Turc et al., 2019) consists of only 4 million parameters compared to BERT-base which has 110 million parameters; this leads to a 65x speedup for pre-training time. It has been widely observed (Turc et al., 2019; Lan et al., 2020) that smaller models achieve comparable downstream task performance with a very significant speedup in training time.

A second issue with pre-trained models with a massive number of parameters, is their lack of calibration, which measures how well the model confidences (posterior probabilities) are aligned with the empirical likelihoods. In other words, for a calibrated model the probability associated with the predicted class label should reflect its ground truth correctness likelihood. Importantly, in the seminal work of (Guo et al., 2017), the authors demonstrate that for deep neural architectures *increasing model size negatively affects its calibration*, even though classification accuracy increases. In this paper, we extend this to investigate *the dependence of calibration on model size for pre-trained transformer models*. Since miscalibrated models can make very confident predictions even when they make errors, especially on out-of-distribution data (Gupta et al.), it is crucial to carefully study model calibration.

Recently, there has been some progress on studying the calibration of deep neural networks and specifically, pre-trained transformers (Guo et al., 2017; Desai and Durrett, 2020; Kong et al., 2020; Jagannatha et al., 2020). However, a careful study of how the size of the pre-trained model influences calibration is lacking. With the computational constraints of training large transformers like BERT and the increasingly wide adoption of smaller mod-

http://cogcomp.org/page/publication_view/953

els, it becomes essential to study the calibration of these variants. In this work, we make a thorough empirical study of the calibration properties of smaller transformer architectures of the BERT family, for a wide set of classification tasks. The set of models have rich variations over number of layers, number of hidden neurons and embedding representation. Additionally, we analyze the effects of techniques designed to help calibrate models: during training (eg: label smoothing) and post-hoc (eg: temperature scaling), on the smaller models, for both in- and out-of-domain datasets.

We establish the following results in this paper:

1. When evaluated in-domain, smaller models are as well calibrated as BERT-base, both with and without temperature scaling.
2. When evaluated out-of-domain, smaller models are worse calibrated than BERT-base. This persists, to a lesser extent, even after temperature scaling.
3. Label Smoothing, on the other hand, is not effective in-domain, but helps smaller models attain better calibration than BERT-base out-of-domain. It also helps improve accuracy as compared to the non-smoothed models, on out-of-domain data.

## 2 Background

In this section, we describe how we measure calibration and two techniques that help calibrate models: *Temperature Scaling* and *Label Smoothing*.

**Calibration Metric**: Let us define the following notation: $K$ is the number of classes, $z_i$ denotes the raw logits from the model for the $i^{th}$ example and $\sigma^{(k)}$ denotes the $k^{th}$ value of the softmax layer $\sigma$, corresponding to the probability for the $k^{th}$ class (for $k \in [1, ..., K]$). Then, the confidence on the $i^{th}$ example is $p_i = max_k \sigma(z_i)^{(k)}$.

A model is well calibrated if the confidence on a prediction is aligned with the accuracy on that prediction, in expectation. The widely adopted **Expected Calibration Error (ECE)** metric (Guo et al., 2017) measures exactly this: difference in expectation between confidence and accuracy. Empirically this is approximated by dividing the data into $M$ confidence based bins, i.e., $B_m$ (where $m \in \{1, 2, ..., M\}$) contains all datapoints $i$ for which $p_i$ lies in $(\frac{m-1}{M}, \frac{m}{M}]$. If $acc(B_m)$ and $conf(B_m)$ denotes the average accuracy and prediction confidence for the points in $B_m$, ECE is defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|,$$

where, $|B_m|$ denotes the number of datapoints in $B_m$ and $n$ is the total number of samples ($\sum_{m=1}^{M} B_m$). In our experiments we set $M = 10$.

**Reliability diagrams** are a popular graphical representation of calibration. It plots the bucket-wise accuracies $acc(B_m)$ versus the confidences $conf(B_m)$. The identity line denotes perfect calibration. The greater the deviation from the identity line, higher is the mis-calibration of the model.

**Post-hoc calibration**: The calibration properties of a model can be evaluated directly *out-of-box* (OOB) based on the softmax scores of the model's predictions. *Temperature scaling* is designed to improve the calibration of a model after training. It rescales the logits $z_i$ by a factor of $T$, before applying softmax $\sigma$. On the $i^{th}$ example, the new confidence prediction is $q_i = \max_k \sigma(\frac{z_i}{T})^{(k)}$ Thus, as $T \to \infty$, $q_i \to \frac{1}{K}$, $\forall i$, which is the uniform distribution with maximum uncertainty. As $T \to 0$, the probability collapses to a point mass ($q_i = 1$) and if $T = 1$, $p_i = q_i$. The optimal temperature $T$ is tuned on the dev-set by a line search algorithm.

**Label Smoothing** (Szegedy et al., 2016) leads to a modified fine-tuning procedure to address overconfident predictions. While Maximum Likelihood Estimation (MLE), sharpens the model's posterior distribution around the target labels, label smoothing introduces uncertainty to smoothen the posterior over the labels. Label smoothing constructs a new target vector from the one-hot target vector, with a probability of $1 - \alpha$ on the target label and $\frac{\alpha}{K-1}$ on all the other labels. Then, in the standard manner, the cross entropy loss is minimized between the model predictions and the modified target vectors. Label smoothing has been shown to implicitly calibrate neural networks (Müller et al., 2019) and (Desai and Durrett, 2020) have shown it is effective for calibrating models on out-of-distribution data.

## 3 Experiments

### 3.1 Models

We consider a family of smaller pre-trained transformer models from (Turc et al., 2019) with the number of layers (L) ranging from 2 to 12 and the number of hidden neurons (H) ranging from 128 to 768. This family of models allows us to carefully study calibration as a function of L and H, since the other parameters like training data and architecture type are constant across them. We focus on 5 models: Tiny (L=2, H=128), Mini (L=4, H=256), Small (L=4, H=512), Medium

| Model (L/H) | SNLI | | | MNLI | | | COLA | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc.(↑) | OOB (↓) | TS(↓) | Acc.(↑) | OOB(↓) | TS (↓) | Acc.(↑) | OOB(↓) | TS (↓) |
| 2/128 | 82.05 | 2.61 | 1.14 | 69.72 | 3.61 | 1.80 | 69.39 | **2.25** | **0.95** |
| 4/256 | 86.67 | 3.64 | 1.23 | 76.05 | 4.75 | 1.95 | 70.54 | 4.25 | 2.37 |
| 4/512 | 87.24 | 3.63 | **0.80** | 78.01 | 4.28 | 0.95 | 74.38 | 7.42 | 2.06 |
| 8/512 | 88.72 | 4.46 | 1.41 | 80.15 | 4.79 | 1.35 | 76.58 | 4.78 | 3.03 |
| *Albert* | 89.07 | **0.86** | 0.91 | 83.62 | **3.29** | 0.94 | 79.08 | 4.90 | 2.47 |
| $BERT_{base}$ | **89.29** | 2.70 | 1.30 | **84.02** | 4.72 | **0.82** | **80.80** | 4.31 | 2.08 |
| Model (L/H) | SST-2 | | | QQP | | | TwitterPPDB | | |
| | Acc.(↑) | OOB (↓) | TS(↓) | Acc.(↑) | OOB(↓) | TS (↓) | Acc.(↑) | OOB(↓) | TS (↓) |
| 2/128 | 80.04 | **4.49** | 2.46 | 84.21 | 3.06 | 1.44 | 84.62 | 6.27 | 3.99 |
| 4/256 | 85.55 | 7.07 | 1.67 | 88.28 | 2.79 | 1.47 | 88.99 | 5.06 | 2.29 |
| 4/512 | 88.53 | 7.64 | 4.61 | 88.56 | 3.87 | **0.68** | 88.36 | 5.74 | 2.73 |
| 8/512 | 89.22 | 7.83 | 4.14 | **89.51** | 3.08 | 1.14 | 87.85 | 6.66 | 3.14 |
| *Albert* | **91.97** | 4.73 | **1.49** | 89.03 | 1.03 | 0.70 | **90.21** | **3.17** | **2.14** |
| $BERT_{base}$ | 90.60 | 8.07 | 4.45 | 89.47 | **1.54** | 0.74 | 88.77 | 5.73 | 3.40 |

Table 1: Variation of Acc. (Accuracy) and ECE (defined in Sec. 2) as a function of model size (L/H denotes the number of layers/number of hidden neurons) across 6 different tasks. Acc. is in % (↑ denotes higher is better ) and OOB, TS are in ECE (↓ denotes lower is better). The results are average over 5 iterations with random initialization. The best results in each column are bolded. $BERT_{base}$ and Albert (uses parameter-sharing) have L=12 and H=768.

(L=8, H=512), and Base (L=12, H=768). Note that the first 4 models have far fewer parameters than BERT; the Tiny model has only 4m parameters compared to the 110m parameters in BERT-Base. To investigate the effect of other types of parameter reduction techniques beyond reducing the number of neurons or layers, we also experiment with Albert (Lan et al., 2020). Albert uses factorized embeddings and cross layer parameter sharing to reduce the number of parameters to only 12 million. We use the 12 layer Albert Base model which is architecturally comparable to BERT Base. For all models, we experiment with three settings:

**Out-of-box (OOB) Calibration**: We directly use the confidences $p_i$ (on the $i^{th}$ example) from the model to compute ECE. No specialized techniques are used to explicitly calibrate the model.

**Temperature Scaling (TS)** (Guo et al., 2017): We use this post-hoc (does not require model-retraining) calibration technique that finds the optimal temperature $T$ as that which achieves the lowest ECE on the dev-set, using line-search.

**Label Smoothing (LS)**: We train a label-smoothed model with hyper-parameter $\alpha = 0.1$. This model can be used out-of-box or with temperature scaling.

### 3.2 Tasks

We perform experiments on various NLP tasks:

**Natural Language Inference**: The Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and the Multi-Genre Natural Language Inference (MNLI) (Williams et al., 2017) datasets are used. Each of them have three classes corresponding to the relations between the hypothesis and the premise: entailment, contradiction and neutral.

**Paraphrase Detection**: The Quora Question Pairs (QQP) (Iyer et al., 2017) and the TwitterPPDB (Lan et al., 2017) datasets are used, where the former contains semantically equivalent questions from Quora and the latter contains semantically equivalent tweets from Twitter. Both datasets have two classes corresponding to similar/ dis-similar pairs.

**Grammaticality Detection**: The Corpus of Linguistic Acceptability (COLA) (Warstadt et al., 2018) is used. It contains two classes corresponding to whether sentences are grammatical or not.

**Sentiment Analysis**: The Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) is used in the binary classification setting, where movie reviews are assigned positive or negative labels.

For details on the datasets, refer to Appendix A.

### 3.3 In Domain Calibration

For each of the different datasets, we fine-tune [1] the various models on the train-set and evaluate their calibration error on the test-set. Additionally, we calibrate the model in-domain through temperature scaling, where the optimal $T$ is tuned on the dev-

[1] Refer to Appendix B for details on hyper-parameter choices: fine-tuning epochs, learning rate and batch size.

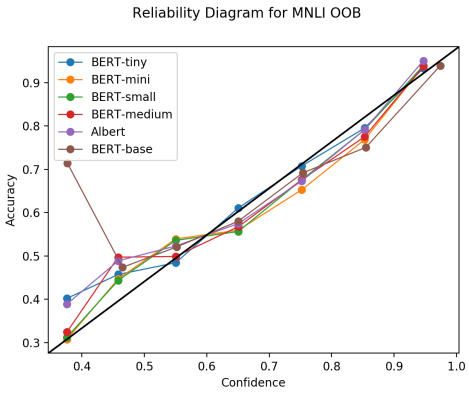


Figure 1: Reliability Diagram for In-Domain MNLI test-set, evaluated out-of-box. The closer the lines are to the identity line, better is the calibration. Note the first 3 bins with confidences from [0.0-0.3] do not contain any points and thus, we start from the 4$^{th}$ bin.

set. [2] Table 1 shows the accuracies and the ECE for the various models on the different datasets. We see that models with far fewer parameters than BERT-base, have competitive accuracy as well as competitive (and sometimes better) calibration as compared to BERT-base. This holds even after temperature scaling, which reduces the ECE for all the models. Fig. 1 shows the reliability diagram for MNLI, where we see that the different smaller models are as well calibrated as BERT-base.

### 3.4 Out-of-domain calibration

We further investigate the effect of model size on calibration for out-of-domain data. For Natural Language Inference, all models are fine-tuned on the SNLI train-set and evaluated on the MNLI test-set. For Paraphrase Detection, all models are fine-tuned on the QQP train-set and evaluated on the TwitterPPDB test-set. We also investigate the effect of (1) Temperature Scaling (where the optimal temperature is chosen based on performance on the dev-set for the source domain: SNLI or QQP) and (2) Label Smoothing with $\alpha = 0.1$, on calibration.

In the reliability diagram in Fig. 2 and in Table 2, we see that smaller models suffer from higher calibration error (ECE) on out-of-domain data, when evaluated out-of-box (OOB) or with temperature scaling (TS). The gap in ECE between smaller models and BERT-base is more severe for the SNLI to MNLI transfer. However, Label Smoothing is very effective in the out-of-domain setting. It significantly reduces calibration error of all models

[2]We also try label-smoothing, but it gives worse results than temperature scaling for in-domain data, across all models.

| Model (L/H) | **SNLI → MNLI** | | | | |
|---|---|---|---|---|---|
| | Acc. | OOB | TS | Acc. | LS |
| 2/128 | 47.73 | 19.64 | 18.34 | 56.57 | **3.65** |
| 4/256 | 56.57 | 15.61 | 12.92 | 61.83 | 6.17 |
| 4/512 | 57.61 | 14.55 | 11.16 | 63.91 | 6.82 |
| 8/512 | 63.13 | 15.43 | 9.38 | 66.76 | 6.91 |
| *Albert* | 67.09 | 8.36 | 8.13 | 68.59 | 4.18 |
| $BERT_{base}$ | **69.88** | **7.25** | **4.06** | **71.35** | 4.98 |
| Model (L/H) | **QQP → TwitterPPDB** | | | | |
| | Acc. | OOB | TS | Acc. | LS |
| 2/128 | 85.95 | 8.89 | 7.70 | 85.57 | **5.07** |
| 4/256 | 86.34 | 10.03 | 8.07 | 88.08 | 5.28 |
| 4/512 | 86.94 | 9.013 | 7.50 | 88.32 | 6.32 |
| 8/512 | 86.58 | 8.84 | 7.62 | 89.24 | 5.37 |
| *Albert* | 86.86 | 8.05 | 7.69 | 87.97 | 6.78 |
| $BERT_{base}$ | **87.35** | **7.59** | **7.09** | **90.22** | 7.06 |

Table 2: Variation of accuracy and ECE as a function of model size for the domain shift from SNLI to MNLI (above) and from QQP to TwitterPPDB (below). Acc. (Accuracy) is in % (higher is better) and OOB,TS,LS are in ECE (lower is better).

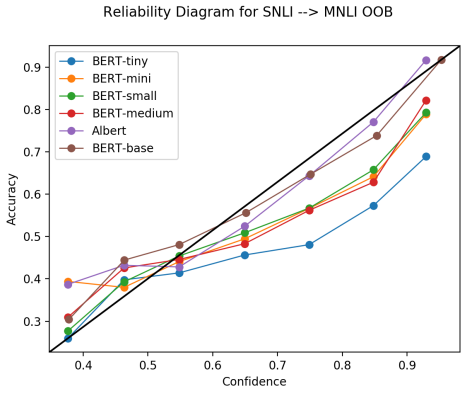


Figure 2: Reliability Diagram for Out-of-Domain MNLI test-set, evaluated out-of-box.

in general, but helps more for smaller models, as seen in both the transfer tasks. Additionally, label smoothing helps improve accuracy for all models when compared to their OOB counterparts.

## 4 Conclusion

We presented a thorough empirical study of the effects of model size (number of parameters) on calibration. Through experiments on a number of tasks, we demonstrated that smaller transformer models are as well, and sometimes better, calibrated compared to BERT-Base, when evaluated in-domain. On out-of-domain evaluation, larger models are better calibrated, out-of-box. Label-smoothed models are better calibrated and more accurate, on out-of-

domain data, with smaller models benefiting more from Label Smoothing.

## Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080. This work was supported by Contract FA8750-19-2-0201 with the US Defense Advanced Research Projects Agency (DARPA). The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense, the Army Research Office or the U.S. Government.

## References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. Bert & family eat word salad: Experiments with text understanding. *arXiv preprint arXiv:2101.03453*.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs.

Abhyuday Jagannatha et al. 2020. Calibrating structured output predictors for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2078–2092.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in-and out-of-distribution data. *arXiv preprint arXiv:2010.11506*.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1234.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When does label smoothing help? *arXiv preprint arXiv:1906.02629*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training nlp models: A concise overview. *arXiv e-prints*, pages arXiv–2004.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models.

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

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

## A Dataset Details

Since the GLUE tasks (Wang et al., 2018) do not have an annotated public test-set, we split the dev-set equally such that one half forms the new dev-set and the other half forms the test-set. The dev-set is used for hyper-parameter selection. Table 3 shows the details for each of the datasets considered.

## B Hyper-parameter Selection

All models are used from the HuggingFace Transformers Library (Wolf et al., 2019). All models are fine-tuned for 2 to 4 epochs with the best value chosen on the basis of the accuracy on the dev set. We set the batch size as 16 with a learning rate of 2e-5, gradient clip of 1.0, and no weight decay. All models are optimized using AdamW (Loshchilov and Hutter, 2018). All the experiments are performed on NVIDIA 24GB GPUs (although most models can be run on 11GB GPUs).

| Dataset | *Train* | *Dev* | *Test* |
|---|---|---|---|
| SNLI | 549,368 | 4,922 | 4,923 |
| MNLI | 392,702 | 4907 | 4908 |
| SST-2 | 67,349 | 910 | 911 |
| QQP | 363,871 | 20,216 | 20,217 |
| TwitterPPDB | 46,667 | 5,060 | 5,060 |
| COLA | 8,551 | 531 | 532 |

Table 3: Number of training, development and test examples for the various datasets we experiment with.