A Further Discussion of Significance Testing, Power Analysis, and Post-Hoc Analysis

Null hypothesis significance testing: In this paper, we work within the framework of null hypothesis significance testing (NHST). NHST is not free from problems, in that certain systematic processes within the practice of scientific research and publishing can undermine its advantages, many of which have been explored in the literature (Gelman and Loken, 2013; Ioannidis, 2019; McShane et al., 2019). Nevertheless, it would be premature to discard the entire paradigm, and we believe there is still some value in considering power within NHST for several reasons.

First, despite its flaws, NHST remains a commonly used experimental framework in NLP research. Whether implicit of explicit, most experimental comparisons in the NLP literature have the structure of an experiment in the NHST framework, where having equivalent performance to an existing baseline is treated as a null hypothesis and the new model is argued to be significantly better (the typical case) or significantly worse (far rarer). But, whereas many fields that run experiments have standardized procedures for assessing statistical significance, NLP papers vary as to how formally they use a hypothesis testing framework to evaluate their results (Berg-Kirkpatrick et al., 2012; van der Lee et al., 2019; Azer et al., 2020).

Second, when done properly, NHST does provide a convenient way of summarizing results. Improvements in overall methdology, such as sharing code and data, sensitivity analyses, greater interest in null findings, and even pre-registration can vastly improve the validity of this paradigm, and we are seeing adoption of some of these practices within NLP.

Finally, there is also a great need for additional clarity with respect to precisely what claims are being made by NLP papers. In this work, we are primarily focused on claims made about trained models (i.e. in testing whether one particular instantiation of a model is significantly better than a particular instantiation of another model). It is, of course, also important to consider broader claims that might be made, such as about expected performance or computational budget (Dodge et al., 2019; Schwartz et al., 2019), and everything we have to say can be extended to incorporate such considerations. For the purpose of clarity, however, we restrict ourselves to the simplest sort of statistical claim.

Power and power analyses: The probability that a statistical test will reject the null hypothesis in an experiment is a function of several parameters, some of which are typically known or controllable, such as the sample size and significance threshold, and some of which are unknown, such as the details about exactly how models differ. Power tells us what this probability would be, if we knew the true values for these unknown parameters. Conditional on a particular difference existing (e.g. an expected difference in accuracy between two models for a particular data distribution), along with a statistical test, a significance threshold, power is the probability that the test will reject the null hypothesis and find the observed difference to be significant. In common statistical terminology, power is one minus the probability of false negatives in rejecting the null hypothesis or type II error.

While we will not, in general, know what the true power of an experiment is, by making reasonable assumptions, we can try to choose appropriate values for those parameters that we can control. By making assumptions about what we expect to observe, we can obtain estimates of how much power a test is likely to have, which may lead us to modify our experimental design, such as by increasing the sample size.

Importantly, proper experiment design requires specifying these parameters in advance of data collection, or otherwise using a valid stopping rule. One can *always* obtain a significant result by progressively collecting data until a significant result is found ("sampling to a foregone conclusion"), but this is not a valid procedure (Anscombe, 1954; Wagenmakers, 2007). Similarly, *post-hoc* power analysis, using estimates derived from the experiment itself, provides no additional information beyond a transformation of the observed p-value, and is thus not recommended (though see below).

Expanding on the algorithm in Figure 2, a simulation-based power analysis involves the following:

1. First, determine the statistical test, T, which will be used. For the example of comparing models depicted in Figure 1, we will use the binomial test to compare the systems (Dror et al., 2018).

2. Come up with a generative process which

could be used to generate data like that which we will collect. In this step, we need to make assumptions about the comparison of interest. Since the binomial test requires only the counts of how many people prefer each system, we need to specify a prior on generating those counts. For example, we might assume that 60% of people will prefer system B, so the generative process will be $c_B \sim \text{Binomial}(p = 0.6, n)$, where n is the total number of people to be sampled.

3. Choose a value of n for which we want to calculate power. Repeatedly (e.g., 10,000 times) draw many samples from our assumed generative process for that size of n .

4. For each simulated dataset of size n, run the chosen statistical test to check if difference between the observed counts is significant, and compute the proportion that are found to be significant. This is our estimate of power.

Note that more direct solutions for power analysis do exist for some settings, such as this one (see Appendix E.5 below).

Post-Hoc Power Analysis: Post-hoc power analysis is an issue when the true population effect has variance to it (O'Keefe, 2007; Hoenig and Heisey, 2001; Gelman, 2019). In the case of NLP models, there are several perspectives on the comparisons which can lead to differences regarding how we perceive post-hoc power analysis: (1) we are comparing one model vs. another on a particular test set, the effect we see is the true population effect, post-hoc power analysis is okay because it is deterministic; (2) we are comparing one model vs. another on a data distribution from which the test and dev set are drawn, post-hoc power is not okay; (3) we are comparing one training algorithm vs. another (including variance from both training procedures and test/dev set draws), post-hoc power analysis is still not okay. We specifically look at the case of (2). While (3) is interesting on its own, this is not the typical comparison done (yet) in NLP research and thus we do not have enough information on reported training variance to investigate this thoroughly here. The case of (1) is also atypical as the authors of a study typically wish to draw inferences about how well a model does on the true data distribution (hence, why a dev and test set are used).

B Type-M and Type-S errors

Although the most obvious risk of using underpowered experiments is that there is a greater chance of failing to detect a true effect, there is an additional harm of using an underpowered design, which has emerged in light of the replication crisis in science. This can be most easily understood through the idea of Type-M and Type-S error (Gelman and Carlin, 2014).

Type-M error is the extent to which an observed difference exaggerates the true effect, conditional on a finding being significant. Type-S error is the probability that an observed difference has the opposite sign of the true difference, again conditional on a finding being significant. Even in a low-powered experiment, there is some probability of finding an effect to be significant; the lower the power, however, the more likely it is that the observed significant difference has the opposite sign of the true effect, and the larger the degree to which the magnitude of the observed effect will tend to exaggerate the true effect.

Intuitively, if power is low, this means that the sample size is small relative to the effect size. As such, the difference will *only* be significant if an atypically large effect is observed. Assuming the use of a two-sided test, many of these significant findings will also have the wrong sign, as they will be nearly as likely to fall on either side of zero for a symmetric distribution.

Type-M and Type-S error rates can be estimated using the exact same process for power analysis as described in Figure 2. To do so, we need only augment the algorithm with these two additional steps:

3. Type-S error $\approx \sum_{i:p_i \leq \alpha} \frac{\mathbb{I}[\text{sign}(e_i) \neq \text{sign}(e*)]}{|j:p_j \leq \alpha|}$

4. Type-M error $\approx \sum_{i:p_i \leq \alpha} \frac{\text{abs}(e_i)/\text{abs}(e^*)}{|j:p_j \leq \alpha|}$

Figures 7 and 8 show scenarios for comparing classifiers on accuracy, corresponding to Figure 3 in the main text, but showing expected Type-M and Type-S error instead of power. As can be seen, Type-M and Type-S error increase with smaller sample sizes, smaller differences between models, and lower agreement rates, all corresponding to lower power.

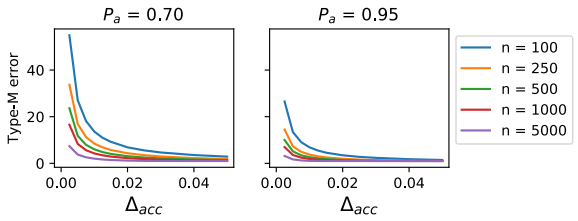


Figure 7: Type-M error (the factor by which observed significant effects are likely to exaggerate the true effect) for comparing classifiers on accuracy increases with smaller test sets (n), smaller differences between models (Δ_{acc}), and smaller agreement rates (P_a). Severe exaggerations of differences between models are likely with underpowered designs.

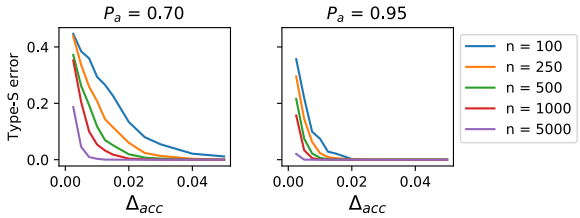


Figure 8: Type-S error (the probability that significant differences observed between models will have the opposite sign of the true difference) for comparing classifiers increases with smaller test sets (n), smaller differences between models (Δ_{acc}), and smaller agreement rates (P_a). Sign errors become resaonably likely with underpowered experiments.

C Numerical Example of a McNemar's Test Simulation

To provide a concrete example of comparing classifiers on accuracy, imagine that a test set for a benchmark task has 500 instances. Based on prior knowledge (see main paper), we might assume that our proposed model will achieve, at most, an absolute improvement of 2 percentage points over the state of the art ($\Delta_{acc} = 0.02$), and that the models are likely to agree on 90% of examples ($P_a = 0.9$). We can convert these assumptions into a distribution over outcomes which will define our generative process. In particular, for a random unseen instance, these assumptions imply that there is a 10% chance of a disagreement; the probability that our model is correct and the old model is incorrect is therefore 6%, and the opposite outcome has a probability of 4% (giving us the assumed net difference of 2%). Note that, because McNemar's test does not consider the on-diagonal elements, it is not necessary that we explicitly define the baseline accuracy. Thus, a valid probability distribution for use in this simulations could be that shown in Table 4.

	M1 correct	M1 incorrect
M2 correct	0.6	0.06
M2 incorrect	0.04	0.3

Table 4: A possible distribution corresponding to the case where models M1 and M2 will agree on 90% of examples (P_a) and M2 achieves a 2% improvement over M1 (Δ_{acc}). Note that the on-diagonal terms here will be dictated by the accuracy of M1 (or equivalently, by M2), but for our purposes, only need to be non-negative and sum to P_a for the sake of McNemar's test, which only looks at the off-diagonal elements.

By drawing many samples from this distribution of size $n = 500$ and computing a p-value using McNemar's test for each, we obtain an estimate that the power of this test is approximately 0.25 for a significance threshold of $\alpha = 0.05$, which is severely underpowered. This would also imply a Type-M error factor of 1.9; we would expect that a typical experiment that found the observed difference between models to be significant would exaggerate the true difference of 0.02 by a factor of 1.9, producing observed significant differences between models on the order of 0.04, on average. (See supplementary notebooks for calculations and interactive demonstration). As such, we conclude that this test set is too small to be able to reliably evaluate whether or not our model is significantly different from the state of the art, and should distrust any observed differences that are significant, unless we have poorly estimated the relevant parameters.

By contrast, if the test set contained 2000 examples, we would estimate the test to have nearly 80% power, with a Type-M factor of only 1.1, and would feel comfortable proceeding with and reporting on this evaluation. Similarly, if we had reason to think that our model represented a game-changing advance, and would achieve an improvement of 4 percentage points, or if we had reason to believe that the models would agree on 97.5% of examples, then we would have the power to evaluate this, even with only 500 examples.

D SQuAD 2.0 Analysis and Results

From the authors of SQuAD 2.0, we obtained pairwise agreement statistics on the SQuAD 2.0 development and test sets for all models that were submitted to the SQuAD 2.0 leaderboard and had

publicly visible development set predictions on the CodaLab platform. We removed six submissions whose exact match (EM) scores on test data were less than 50%; EM scores below 50% suggest a bug or misconfiguration of the model for predicting on the test set, as the majority baseline gets roughly 50% accuracy (by always predicting no-answer). We also removed one submission whose development set EM score was more than 20 points higher than its test EM score, as it seemed likely that the model had been trained on the development set. After this filtering, we were left with 144 models.

Figure 9 shows the correlation between validation and test data for both pairwise accuracy differences (Δ_{acc}) and agreement rates (P_a) on the SQuAD 2.0 leaderboard. As can be seen, these correlate well, suggesting that measuring these quantities on validation data can serve as a reasonable guide when doing a power analysis for a new model, though lower agreement rates on dev data to tend to slightly underestimate agreement on test. If the validation results are available for both models, these can be used to compute estimates of P_a and Δ_{acc}, and these can be used to compute the approximate power of the test set.

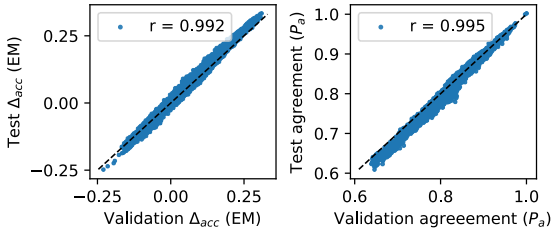


Figure 9: Correlation between validation and test data among all models submitted to the SQuAD 2.0 leaderboard for both pairwise accuracy differences (Δ_{acc} using exact match (EM); left), and agreement rates (P_a; right). In both cases, Pearson correlation (r) is over 0.99. Dashed lines show $y = x$.

To verify that using these estimates provide a reliable guide to power, we make use the predictions made by SQuAD 2.0 submissions on both validation and test data. In particular, if we assume that each submission is being compared to the previous model to demonstrate a significant and well-powered improvement over the previous baseline, we find that 19 out of 143 submissions showed sufficient improvement on the validation set to have at least 80% power (see Figure 10). Of these, 14 (74%) attain a significant improvement over the baseline on the test data (consistent with the expected value of 80%). Of the remaining 124 submissions, 3 (2.5%) would show a significant improvement over the baseline, but did not have sufficient power based on validation performance. Interestingly, while all other significant improvements were generally well-spaced over time, these three underpowered submissions were all beaten by a new submissions within 5 days. As an aside, we also note that the vast majority of submissions are significantly worse than the current SOTA, reinforcing the notion that real improvements are rare, and most improvements will be small.

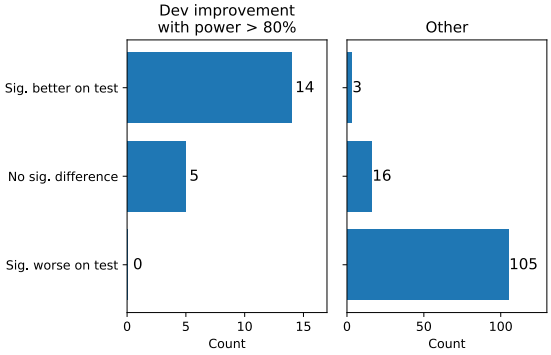


Figure 10: SQuAD 2.0 leaderboard submissions compared to previous SOTA, where we require for SOTA that submissions have 80% power (based on validation improvement and agreement), and a significant improvement on test data.

Caveats: Correlation between the effect size on the validation and test sets may not always be so high. Overconfidence in the power of your experiment may thus occur if the validation performance is greater than the test performance (as would be the case if no regularization was used and extensive hyperparameter tuning caused a model to overfit to the validation set). Alternatively, if comparing to a baseline with inflated performance on validation data (for the same reasons as above), running power analyses based purely on estimates from validation data would underestimate power. As such, combining validation estimates with reasonable priors is recommended.

E Accuracy

E.1 Data Collection

E.1.1 Model Predictions on Test Set and Model Prediction Agreement

From the authors of the GLUE benchmark – as well as authors of individual models – we obtain

the model test-set predictions on all tasks from a set of 10 high-performing models, which allows us to measure the extent to which their predictions overlap with each other. We select GLUE tasks which use accuracy as an evaluation metric. The relevant tasks are MNLI (Williams et al., 2018), MRPC (Dolan and Brockett, 2005), RTE (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SST-2 (Socher et al., 2013), QQP (Iyer et al., 2017), QNLI (Rajpurkar et al., 2016), and WNLI (Levesque et al., 2012). For consideration of other metrics, see Appendix F.

We use model predictions for: ELECTRA (small, base, large, large with tricks) (Clark et al., 2019b), XLNet (large) (Yang et al., 2019), T5 (Raffel et al., 2019), ALBERT (large) (Lan et al., 2020), BAM(large) (Clark et al., 2019a), RoBERTa (large) (Liu et al., 2019), and BERT (Devlin et al., 2019). We only had the model predictions available and extrapolated overlap from that, we did not have access to the models themselves, ground truth test set labels, nor dev set predictions for the models.

E.1.2 Comparisons and Claims

We gather data from GLUE papers regarding the accuracy tasks and manually label 119 comparisons and 57 claims of improvement (as denoted within a work by bolding of a new model's number and a claim of SOTA in the main text) across 14 papers (selected as being at or above the BERT score on the GLUE leaderboard with an accompanying publication). For each paper we examine if a specific comparison is made against a baseline that isn't claiming state of the art performance. For example, the STILTs approach (Phang et al., 2018) makes comparisons against non-SOTA baselines, which we add to our labeling scheme but filter out when fitting regressions to likely SOTA improvements. We mark this as **SOTA Comparison** = N. For claims of SOTA improvement, we examine this as some textual basis for the claim (e.g., "we drive state of the art performance on GLUE") coupled with bolding of values in a table reporting baselines against the model under test. We mark datapoints as **Claim of Improvement** = Y if they are an improvement claim. We mark effect size as the improvement from the best previous baseline (the current SOTA) on the test set on a per-dataset basis. We note that in several cases, worse results on the new model were bolded. We treated this as no claim of improvement. If results were not bolded but still higher for the new model we also treated this as no claim for improvement.

E.2 Regression-based approach to modeling power and MDEs

E.2.1 Predicting overlap

There are several versions of McNemar's test, each with their own unique method for calculating power, sample size, or minimum effect size. See, for example, discussions in Schlesselman (1982), Duffy (1984) Suissa and Shuster (1991), Connett et al. (1987), Fagerland et al. (2013), and Lachenbruch (1992).

The methods for calculating sample size or power by Connett et al. (1987); Schlesselman (1982); Suissa and Shuster (1991) require making an assumption about the odds ratio $\Phi = p_{10}/p_{01}$ as well as an estimate of the fraction of discordant pairs (disagreements between two models).

Fagerland et al. (2013) suggest that the exact unconditional version of the test by Suissa and Shuster (1991) has desirable properties. Thus, we use the implementation of the power calculations for this test from the `https://github.com/ekstroem/MESS` package.

How do we make an assumption about the odds ratio and fraction of discordant pairs? We first fit an OLS regression to the existing models on the GLUE leaderboard for all binary choice accuracy tasks using the aforementioned predictions provided by the leaderboard creators and individual authors of models,

$$\text{overlap}_i = \beta_0 + \beta_1 \text{min\_acc}_i + \beta_2 \text{acc\_diff}_i, \quad (1)$$

for all i that are a pairwise comparison between any two models, min_acc_i is the minimum accuracy between the two models under comparison, acc_diff_i is the gap between the two models, and overlap_i is the fraction of overlapping predictions. We end up with the model shown in Table 5.

We note that outcomes are biased toward a higher range of accuracy values and may not be a perfect prior. However, this does give us a fairly good linear fit for top-of-the-leaderboard results. We then can predict the expected overlap for a given model as:

$$\begin{aligned} \text{exp\_overlap} =& 0.41 + 0.58 \cdot \text{ min\_acc} \\ & - 0.47 \cdot \text{exp\_acc\_dif} \end{aligned} \quad (2)$$

Note now we can make an assumption on the expected fraction of discordant values and the odds

Dep. Variable:	y	**R-squared:**	0.966
Model:	OLS	**Adj. R-squared:**	0.966
Method:	Least Squares	**F-statistic:**	3820.
Date:	Thu, 14 May 2020	**Prob (F-statistic):**	3.62e-197
Time:	07:03:28	**Log-Likelihood:**	818.14
No. Observations:	270	**AIC:**	-1630.
Df Residuals:	267	**BIC:**	-1619.
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.4142	0.019	21.694	0.000	0.377	0.452
min_acc	0.5819	0.021	27.999	0.000	0.541	0.623
acc_diff	-0.4662	0.028	-16.625	0.000	-0.521	-0.411

Omnibus:	6.121	**Durbin-Watson:**	1.040
Prob(Omnibus):	0.047	**Jarque-Bera (JB):**	8.647
Skew:	-0.108	**Prob(JB):**	0.0133
Kurtosis:	3.850	**Cond. No.**	71.5

Table 5: OLS Regression Results for predicting GLUE model overlap from baseline accuracy and effect size.

Dep. Variable:	y	**R-squared:**	0.944
Model:	OLS	**Adj. R-squared:**	0.933
Method:	Least Squares	**F-statistic:**	91.87
Date:	Tue, 26 May 2020	**Prob (F-statistic):**	1.37e-07
Time:	06:05:23	**Log-Likelihood:**	36.368
No. Observations:	14	**AIC:**	-66.74
Df Residuals:	11	**BIC:**	-64.82
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.4339	0.091	4.786	0.001	0.234	0.633
min_acc	0.5932	0.101	5.874	0.000	0.371	0.816
acc_diff	-1.2849	0.588	-2.186	0.051	-2.578	0.009

Omnibus:	0.299	**Durbin-Watson:**	2.022
Prob(Omnibus):	0.861	**Jarque-Bera (JB):**	0.163
Skew:	0.214	**Prob(JB):**	0.922
Kurtosis:	2.691	**Cond. No.**	140.

Table 6: OLS Regression Results for predicting SQuAD 2.0 model overlap.

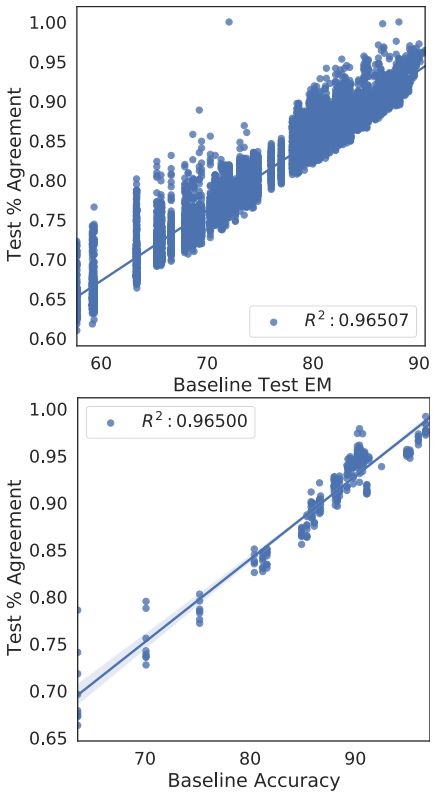


Figure 11: SQuAD 2.0 (top) and GLUE (bottom) % agreement of new model vs. the accuracy of the baseline in the comparison (assuming improvement in the new model).

ratio, the latter being:

$$\Phi = \frac{1 - \text{exp\_overlap} + \text{exp\_acc\_diff}}{1 - \text{exp\_overlap} - \text{exp\_acc\_diff}} \quad (3)$$

This is all that is necessary for McNemar's test and thus we can then simply solve for the minimum expect treatment effect for the given sample size of the dataset and a power of 80%. Note that for QQP we use the normal approximation rather than exact unconditional test as the large sample size makes the exact test intractable. See Duffy (1984).

We fit such a regression to GLUE tasks and achieve an R^2 of 0.97. Repeating this for SQuAD 2.0, we get an R^2 of 0.94, with fit shown in Table 6. See Figure 11 for a plot indicating the level of agreement plotted against baseline accuracy. See also additional model comparisons for overlap in Appendix I.

E.2.2 Predicting Effect Size

A similar regression can be run to predict the expected effect size given the baseline accuracy: how much do models typically improve given the current SOTA. To fit an OLS regression predicting this value, we gather data from GLUE papers regarding the accuracy tasks and manually label 119 comparisons and 57 claims of improvement (as denoted within a work by bolding of a new model's number and a claim of SOTA in the main text) across 14 papers (selected as being at or above the BERT score on the GLUE leaderboard with an accompanying publication). We fit the regression:

$$\widehat{\Delta}_i = \beta_0 + \beta_1 \text{baseline}_i + \hat{\beta}_2 \text{task}_i, \quad (4)$$

to see how predictable the expected effect size is, where $\widehat{\Delta}_i$ is the predicted effect size, baseline_i is the baseline model's accuracy, and task_i is a categorical variable (in the regression this ends up being a set of dummy variables for each category so we denote $\hat{\beta}$ to emphasize this). Note that for SQuAD 2.0, we use a separate regression without the task variable since it is a single-task leaderboard.

We achieve an $R^2 = 0.69$ which is not a perfect fit, but still provides a prior on likely effect size. Similarly, we achieve an $R^2 = .67$ when fitting a regression to SOTA improvements on the SQuAD 2.0 leaderboard (selected as being a significant improvement in time-ordered submissions).

See Table 7 and Table 8 for regression coefficients and model fits. Figure 13 shows the per-task distribution of effect sizes against baseline accuracies in GLUE papers for SOTA improvements. Figure 12 shows the effect size distribution as a histogram.

E.2.3 Caveats for Regression-based Approach

Fitting a regression to predict overlap between a baseline and a new model has a good linear fit. However, this may not be the case for every dataset. Additionally, predicting effect sizes via a linear fit is not a perfect prior. The measurements of power in this case are meant to simulate estimating power before running evaluation on a test set, as running power analysis using only the observed effect may lead to the issues of post-hoc power estimation.

E.3 No Prior Approach (Lachenbruch, 1992)

What do you do if there is no prior data available (as in a new task) and so you cannot make assumptions about discordant pairs or odds ratio? Lachenbruch (1992) discusses this exact problem in the context of clinical trials, and proposes an alternative method based on the work of (Connett et al., 1987) which allows you to make

	Dependent variable:
	effect.size
Previous.Best	-0.264^{***}
	(0.032)
TaskMNLI-mm	0.150
	(0.621)
TaskMRPC	0.023
	(0.622)
TaskQNLI	2.139^{***}
	(0.639)
TaskQQP	-0.195
	(0.719)
TaskRTE	1.018
	(0.628)
TaskSST-2	1.536^{**}
	(0.686)
TaskWNLI	-0.520
	(0.789)
Constant	24.342^{***}
	(2.837)
Observations	61
R^2	0.690
Adjusted R^2	0.642
Residual Std. Error	1.309 (df = 52)
F Statistic	14.455^{***} (df = 8; 52)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: OLS regression for predicting effect size for GLUE tasks.

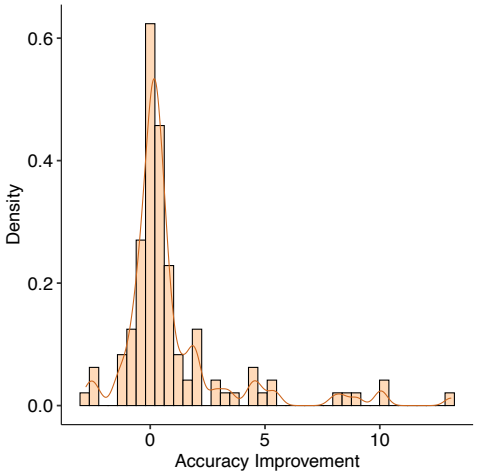


Figure 12: The reported difference from the best performing new model to the best performing baseline in accuracy across all accuracy datasets in the GLUE Benchmark. Note: unlike Table 10, we do not limit these to claims of improvement, but only to papers which introduce a new model and compare against some baseline. Mean: +0.959 Std.Err.: 0.23

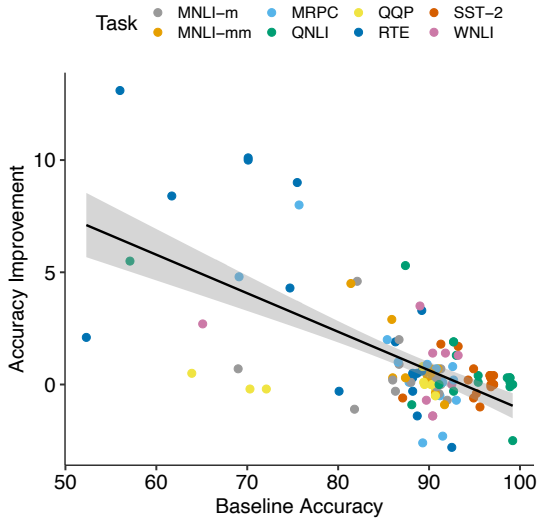


Figure 13: The effect size given the baseline model accuracy observed across GLUE tasks. As the baseline model moves toward the range of current GLUE submissions, reported model gains decrease toward 0. Fitting a regression yields an $R^2 = 0.69$.

assumptions about potential marginal probabilities, providing a midpoint value, as well as an upper and lower bound. We use an implementation of this from: `https://rdrr.io/rforge/biostatUZH/man/sampleSizeMcNemar.html` and solve for the expected accuracy minimum given a fixed dataset sample size and baseline accuracy for each of the lower bound, midpoint, and upper bound. In practice, we find the Lachenbruch (1992) prior to be very close to the values we obtain from the above regression (see Table 9). Importantly this method requires no assumptions and is meant to give an idea for whether it is worth pursuing a study for the given size of the test set.

E.4 Extended Results

Table 9 contains additional MDE estimates using a two-sample proportion test as in Appendix E.5, the Lachenbruch (1992) methodology. We also provide the standard errors and n for each average effect size, the OLS regression predicting the next effect size for a new SOTA $\widehat{\Delta}$, and the current difference from SOTA and next on the leaderboard. We note that MDE calculations are roughly similar except for the upper and lower bounds provided in the Lachenbruch (1992) calculation. We also note that predicted SOTA results are far lower than past averages since the average includes early large results like those of Devlin et al. (2019). We can see that in some cases the predicted effect size

Dep. Variable:	y	**R-squared:**	0.672
Model:	OLS	**Adj. R-squared:**	0.644
Method:	Least Squares	**F-statistic:**	24.55
Date:	Tue, 26 May 2020	**Prob (F-statistic):**	0.000334
Time:	06:05:23	**Log-Likelihood:**	45.711
No. Observations:	14	**AIC:**	-87.42
Df Residuals:	12	**BIC:**	-86.14
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.1331	0.023	5.910	0.000	0.084	0.182
x1	-0.1408	0.028	-4.955	0.000	-0.203	-0.079

Omnibus:	19.911	**Durbin-Watson:**	2.643
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	18.487
Skew:	1.995	**Prob(JB):**	9.68e-05
Kurtosis:	6.971	**Cond. No.**	17.3

Table 8: OLS Regression Results for predicting effect size from baseline accuracy for SQuAD 2.0 improvements.

is even smaller than the lowest bound MDE and we may wish to consider the usefulness of further comparisons on individual datasets in such cases.

E.5 Calculating Power or Sample Size with Binomial Test

If we assume that samples are *unpaired* – the new model and baseline evaluation samples are drawn from the same data distribution but aren't necessarily the same samples – we can use a binomial test for significance.

In this case, we assume that we have two models and each draw brings a 1 if the model is correct or 0 if incorrect. We would like to use the two-sample proportion test, and have two binomial distributions with p_1 and p_2 as the mean probabilities. Our null hypothesis is $H_0 : p_1 = p_2$. We have an alternative hypothesis (two sided) is $H_1 : p_1 \neq p_2$. Note, in R we can use the function **power.prop.test()** to calculate power, the MDE, or the sample size of the tests. See also a tutorial here: https://imai.fas.harvard.edu/teaching/files/Handout9.pdf.

F Additional Metrics

In this appendix, we provide guidance on how we might apply power analysis to metrics beyond what is covered in the main paper.

Recall, Precision, F1, Matthew's correlation: While accuracy is the most commonly used metric in the GLUE benchmark, other tasks make use of other metrics such as F1 and Matthew's correlation. F1 is particularly relevant in cases of binary classification where there is strong class imbalance, such that even the baseline of predicting the most common class will achieve high accuracy.

If we have good prior information, we can use an approach akin to that recommended for accuracy, but replacing McNemar's test with a randomization test (as used for machine translation, see §4 in main paper). In particular, given an evaluation on paired data (as is the case for all benchmark datasets), one can test for a significant difference between models in terms of F1 (or any other metric) using a randomization test. That is, on each iteration, we randomize the assignment of which model each prediction came from for every instance with probability 0.5, and compute the resulting overall difference in F1. Repeating this thousands of times gives us the null distribution, and we can then check to see whether the observe difference in F1 is in the tails of this distribution, which can thereby be converted into a p-value (see Dror et al. (2018) for more details).

Because F1 (and related metrics) cannot be represented as a simple sum over individual instances, in order to completely specify a hypothetical data generating process, we need to assume values for all cells in the confusion matrix, per class. That is for each class we would need to assume values for the cells as shown in Table 11, where the relevant distribution of predictions are for the instances with the corresponding label, and the values for each class sum to one.

Dataset	Size	SOTA	MDE Binomial	MDE (Lachenbruch, 1992)	MDE regression	$\widehat{\Delta}$	$\lvert\Delta\rvert$ (std.err.,n)	Δ_{SOTA}
WNLI	147	94.5%	+5.38%	+5.42%(5.36%, 5.45%)	+5.26%	-1.17%	1.72 (0.917, 4)	0.0%
MRPC	1725	92.0%	+2.40%	+1.91% (0.45%, 2.48%)	+1.62%	+0.03%	+0.625 (0.234, 8)	+0.6%
SST-2	1821	97.2%	+1.34%	+1.10% (0.43%,1.35%)	+1.02%	+0.18%	+0.571 (0.197, 7)	-0.3%
RTE	3000	91.7%	+1.89%	+1.48% (0.26%, 1.96%)	+1.23%	+1.11%	+3.89 (1.23, 10)	+0.8%
QNLI	5463	97.5%	+0.77%	+0.60% (0.14%, 0.78%)	+0.55%	+ 0.69%	+1.31 (0.552, 9)	+0.9%
MNLI-m	9796	91.6%	+1.08%	+0.82% (0.08%, 1.12%)	+0.67%	+0.12%	+ 0.97 (0.442, 10)	+0.2%
MNLI-mm	9847	91.3%	+1.09%	+0.84% (0.08%, 1.14%)	+0.68%	+0.34%	+ 1.29 (0.550, 8)	+0.3%
QQP	390965	91.0%	+0.18%	+ 0.13% (8.45×10^{-5}%, 0.19%)	+0.11%	+0.08%	0.36 (0.121, 5)	+0.1%
SQuAD 2.0	8862	90.724%	+1.18%	+0.91% (0.09%, 1.23%)	+0.556%	+0.528%	+2.23% (0.431,14) †	+0.146%

Table 9: The minimum detectable effect (MDE) for various datasets given the current top accuracy on the leaderboard on May 6th, 2020. See Appendix E for expanded details. How to use this table? Suppose you are building a model to get SOTA on any of these datasets. If you don't have a reasonable expectation that your model will exceed the MDE, then it is not worth proceeding with the study on a dataset of this size and instead either more data should be collected or a different (larger) dataset used. MDE (Lachenbruch, 1992) provides a mid-point and upper/lower bound assumptions using the most conservative and generous estimates of model agreement. MDE Binomial uses the binomial test as the assumed statistical test and calculates the MDE using the exact mechanism from Appendix E.5. See also discussion by Arend and Schäfer (2019). $\widehat{\Delta}$ is the expected effect by fitting a regression to all SOTA improvement claims found in reviewed papers. $\lvert\Delta\rvert$ (std.err., n) is the average improvement in surveyed papers that claimed SOTA and had a positive effect size reported for the dataset (with standard error and the number of papers in parentheses). † indicates that the SQuAD 2.0 average improvement was based on improvements to the SQuAD leaderboard, but weren't necessarily reported as improvements in a publication. Δ_{SOTA} is the gap between the SOTA model (ALBERT + DAAF + NAS) on GLUE and the next best model (ERNIE) – this was not included in the regression.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Power	57	0.698	0.352	0.034	0.407	1.000	1.000
P	57	0.220	0.283	0.000	0.00000	0.348	1.000
Statistic	N	Percentage			-		
% Powered	57	0.456%	–	–	–	–	–
% Significant	57	0.509%	–	–	–	–	–
% significant and Powered	57	0.368%	–	–	–	–	–

Table 10: We examine the claims of SOTA improvement in surveyed GLUE papers and use a leave-one-out regression-based estimate of effect size and overlap to simulate how many authors would have found their study to be well-powered. We also examine how many of the observed effects were likely significant based on predicted model overlap. We note that if we use the *observed* effect in a post-hoc analysis, the proportion of studies falling below the MDE is even higher.

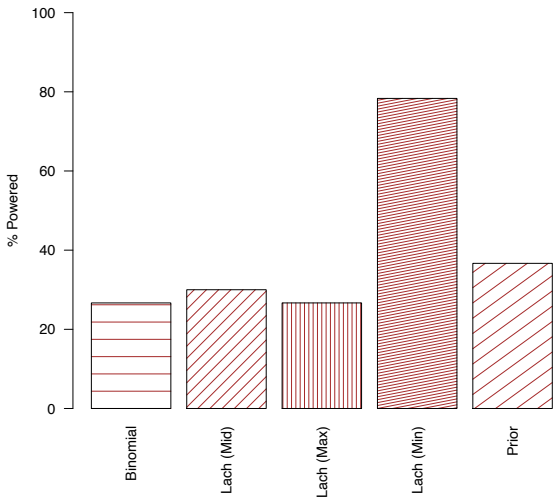


Figure 14: Of the claims of improvement over a given baseline (indicated in text and via bolded values in tables) across 14 papers on the GLUE leaderboard (also seen in Table 10). We find only 26.7% of observed effects met the MDE to the binomial power calculation, 30% met the MDE according to the midpoint calculation of (Lachenbruch, 1992), 26.7% met the MDE when using the upper bound from the (Lachenbruch, 1992) calculation, 78.3% met the MDE when using the most generous (unlikely) assumptions for power according to the MDE (Lachenbruch, 1992) calculation, and 36.7% met the MDE when using the regression-fitted prior of model overlap. Note: this assumes the true population effect *is* the test set effect size. While this is post-hoc power analysis, we felt it may be useful to consider in the context that for a given model comparison on a given test set there is no variance and thus post-hoc power analysis is acceptable. However, for claims that include the entire data distribution this no longer holds and we refer back to the main text.

	M1 negative	M1 positive
M2 negative	p(both neg.)	p(only M1 pos.)
M2 positive	p(only M2 pos.)	p(both pos.)

Table 11: A contingency table representing the distribution of possible outcomes for two models (M1 and M2) on the instances of a single class of labels. The cells of this table should sum to 1.0 for each class

In addition, we need to assume the true distribution of labels in the data distribution of interest, $p(c)$ for c in $\{1, \ldots, C\}$. Given these assumptions, we could then simulate an arbitrary number of datasets from this process. For each instance, we would first sample a true label (c), and then sample the model predictions from the corresponding contingency table. For each simulated dataset, we could then apply the randomization test (using thousands of randomizations). By repeating this process many times, we can directly estimate power for the corresponding assumptions and sample size n.

This process is not particularly efficient, but can still be run relatively quickly on a laptop. The more difficult part is choosing good values for the necessary probabilities. However, such an approach can still be used to test for how sensitive power is to variations in assumptions. It is also possible to make simplifying assumptions, such as that the rate of false positives and false negatives will be the same across classes, or to estimate some parameters from training data, such as the underlying distribution of labels. The same technique can easily be extended to other metrics that depend on the contingency table, such as Matthew's correlation.

G Additional Details for the BLEU Scores Power Analysis

In this section, we provide further details for the machine translation (MT) data generation procedure as well as an analysis of how power varies for a range of values of P_0 and b_0, the parameters estimated from the empirical observations.

G.1 Data Generation Procedure

Recall that using the randomization test to determine whether two MT systems are statistically different gives rise to the null distribution of differences in BLEU.[14] If we had access to large amounts

[14]The bootstrap is another valid approach to testing for differences between models (Koehn, 2004; Graham et al., 2014; Dror et al., 2018), though note the concerns highlighted by Riezler and Maxwell (2005).

of parallel text, we could instead sample many subsets of real sentences and evaluate the difference between models on those subsets, which allow us to characterize the mean and variance of the difference in model performance. Such estimates could then be used to estimate power directly. Because we do not have access to such data, however, we instead rely on the randomization approach, in which we run several thousand trials where the paired output translations for a subset of the test set samples are swapped. In order to estimate power, we would like to be able to generate many datasets from a data generating procedure, which we can parameterize by various parameters, such as the difference between models. Rather than generating raw text, however, and computing BLEU scores on that, we instead attempt to generate only the data necessary for the randomization test. How can we do this?

In our case, the answer to this question lies in establishing a relationship between individual samples and the permuted set within each trial of the randomization test. This relationship is as follows: *the sum of individual changes to the difference in BLEU, from swapping single samples at a time, closely approximates the net change to the difference in BLEU, from swapping those samples all at once.*[15] Let S be the set of test set samples swapped during a single trial of the randomization test and $R_B(S)$ be the difference in BLEU between the paired outputs after swapping the examples in S. Δ_B is the original difference in BLEU and δ_i is the change to the difference in BLEU from swapping test sample i and leaving all other samples unswapped. Then, we find that,

$$\sum_{i \in S} \delta_i \approx R_B(S) - \Delta_B$$

This relationship is illustrated in Figure 15: Figure 15a shows the difference between two models evaluated on the 2019 test set, and Figure 15b shows the difference between a different pair of models evaluated on the 2018 test set. We found the same relationship is true for the 2017 and 2016 test sets, as well.

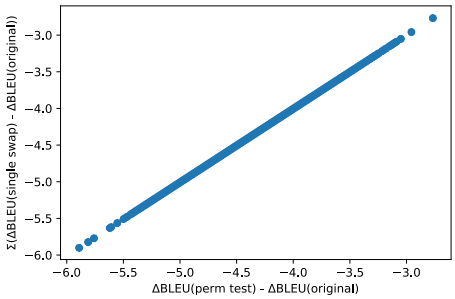
Now that we have established a relationship to closely approximate the outcome of each randomization trial, all that remains is to define a distribution from which the individual changes to the difference in BLEU can be sampled. This distribution is a mixture of a Delta distribution at zero and a Laplace distribution. The Delta distribution accounts for the proportion of samples (P_0) such that swapping any of them individually results in no change to the difference in BLEU, i.e. the effect is zero. For the remaining samples, we fit a Laplace distribution, as shown in Figure 16. This Laplace is parametrized by two parameters: location (μ) and scale (b). By fitting this mixture to the individual effects computed from evaluating BLEU differences on many pairs of models, we discover that the variance parameter scales inversely proportional to the size of the dataset. Thus, we report an overall b_0 value for each dataset, such that $b_0 = b_k * n_k$, where b_k is the Laplace scale parameter obtained from dataset k containing n_k samples.

For generating synthetic data, we need to specify μ and b, as well as P_0. However, because we want the effect of swapping half the non-zero samples from this distribution to equal the difference in BLEU between models, we only use the above fits to estimate b_0. We thus complete the generative process by assuming values for Δ_B, n, P_0, b_0, and setting $\mu = -2 \cdot \Delta_B / (n \cdot (1 - P_0))$ such that the average effect of a random subset of $n/2$ instances is equal to $-\Delta_B$. Table 3 in the main paper shows a range of observed values for P_0 and b_0.

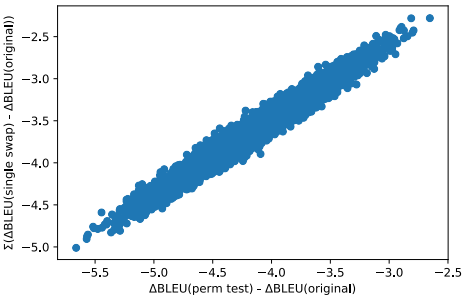
G.2 Variation in Power Estimates for a Range of Parameter Values

Now that we have defined the data generation procedure, and have estimates for the two parameters, P_0 and b_0, that are needed to simulate datasets, we can estimate power for a range of values for sample size n and difference in BLEU Δ_B, and see how these estimates vary as P_0 and b_0 change. To provide a concrete example, suppose that we have two machine translation models that we expect will differ by $\Delta_B = 1$ BLEU point. For a dataset of $n = 2{,}000$ sentences, we assume that the models will perform equally for $P_0 = 0.2$, i.e. 20% of sentences, and will assume a base scale parameter of $b_0 = 26$. To compute power, we would follow the process in Algorithm 1, with the following modifications. On each iteration, we would draw individual changes to the difference in BLEU from the distribution specified above, with $P_0 = 0.2$, $\Delta_B = 1$, $b_0 = 26$, and $n = 2000$. For each such draw, we would apply the randomization test to compute a null distribution, using the sum of in-

[15] Note that this does not directly solve the problem of computing BLEU at the sentence level (Chen and Cherry, 2014), as it still mimicking the process of evaluating BLEU on a corpus.

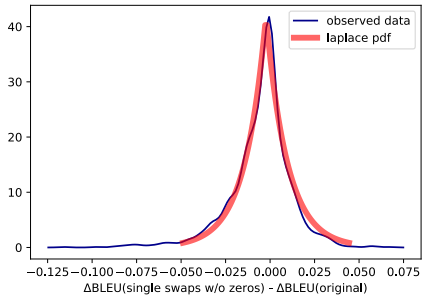


(a) Model trained on WMT19 data versus model trained on WMT18 data, evaluated on the 2019 test set.

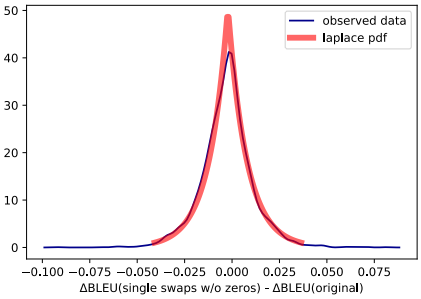


(b) Model trained on WMT18 data versus model trained on WMT16 data, evaluated on the 2018 test set.

Figure 15: Correlation between individual changes to Δ_B and the net effect.



(a) Model trained on WMT19 data versus model trained on WMT18 data, evaluated on the 2019 test set.



(b) Model trained on WMT18 data versus model trained on WMT16 data, evaluated on the 2018 test set.

Figure 16: Fitting a Laplace distribution to individual non-zero effects.

dividual amounts as the total effect of flipping a random subset of pairs. Based on the null distribution, we compute if the difference is significant for this trial. Repeating this many times and observing the proportion of trials that are found to be significant gives us the approximate power.

Figure 17 shows power for a range of values for Δ_B, n, P_0 and b_0. When P_0 is low, as is true for the observed data in Table 3, effect sizes and sample sizes need to be larger in order for an experiment to be well-powered. But as P_0 gets higher, a given effect size can be detected by a smaller sample size. On the other hand, as b_0 increases and consequently the scale parameter b for the Laplace grows, even large effect sizes cannot be detected by test sets containing 5,000 samples.

H Details of Human Evaluation Section

H.1 Meta-analysis of human ratings for EMNLP 2019

To assess the state of statistical power in a typical NLP study using human evaluation, we sampled papers from the mean EMNLP 2019 workshop that contained the phrase "human eval". This first pass returned 117 papers, of which 86 had relevant human evaluations (in which models were compared), with the remainder either referencing human evaluation, or containing some other type of evaluation, such as comparing the agreement between automated metrics and human performance. Because some papers had more than one such evaluation, we had 97 experiments for analysis. Of these 51 were Likert experiments (as discussed in the main text), 38 were some form of direct model comparison, and 8 were other.

Significance testing was rare and was reported, in some form, in only 24% of experiments. Bolding or starring the best results in a table was more common, occurring in 63% of human rating experiments in our set. Whether bold results implies that the author is claiming a meaningful difference is not always clear. We did find one single case of authors performing a power analysis to estimate sample size among the papers we surveyed (Garbacea et al., 2019). However, because that paper did not involve a comparison of models to a baseline, it

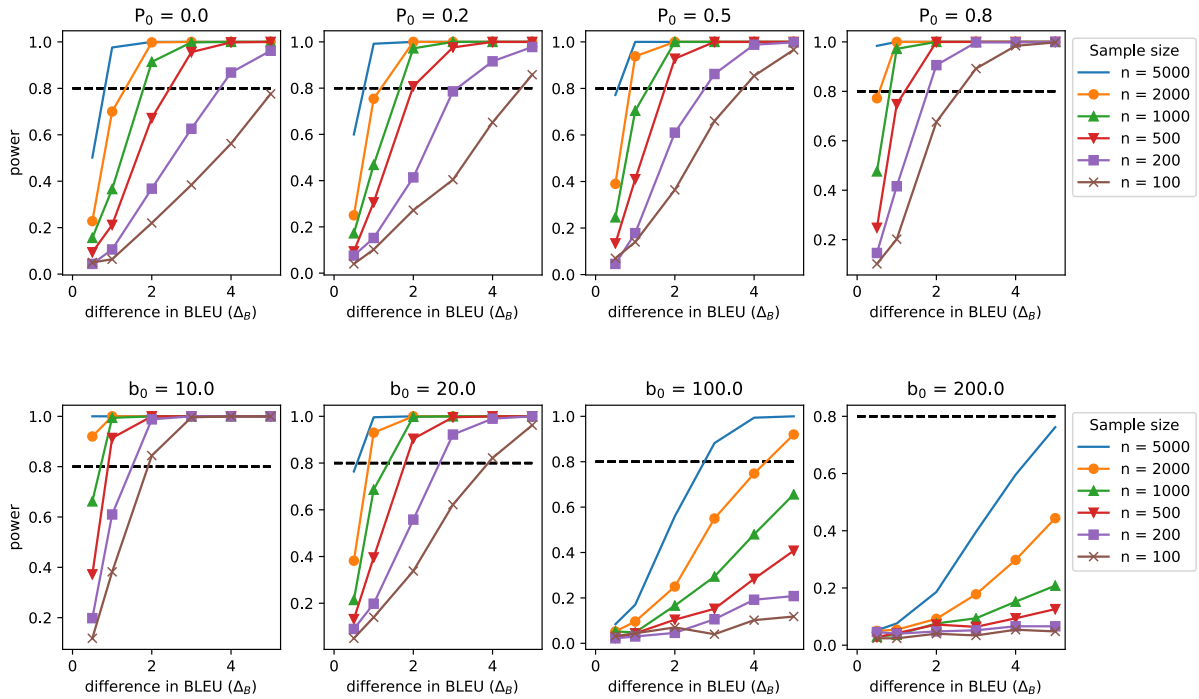


Figure 17: Power Analysis for BLEU scores: Variation in estimates of power for different values of P_0 (top) and b_0 (bottom). For the top row, b_0 = 25.8, and for the bottom row, P_0 = 0.13.

was not included in our analysis. In addition, we note that few details were provided, such that we were unable to ascertain precisely how the power analysis was done.

Because we chose to focus on ordinal ratings, we further annotated those in order to record the mean ratings and experimental characteristics (number of annotators, number of items, number of annotators per item), as well as all differences for all metrics between the model being proposed and the best performing baseline evaluated in the paper, as discussed in the main text.

H.2 Human evaluation datasets

For our analyses, we make use of the following datasets:

- From Hashimoto et al. (2019) we use the evaluation data for Reddit, language modeling, and summarization. The data is available at https://worksheets.codalab.org/worksheets/0x88644b5ee189402eb19d39d721d1005c
- From Dathathri et al. (2020) we use the available ratings. The data is available at https://github.com/uber-research/PPLM
- For WMT19 (http://statmt.org/wmt19/translation-task.html), the data is available at https://www.computing.dcu.ie/~ygraham/newstest2019-humaneval.tar.gz
- For Holtzman et al. (2020), we obtain the human evaluation data directly from the authors.

H.3 Linear Mixed Effect Models

To assess power in the human ratings framework, we used linear mixed effect models with random intercepts and slopes for worker and item, as in Barr et al. (2013). Following best practices, we use the following structure, where w is a particular worker and i is a particular item. There are seven parameters, corresponding to the parameters needed for running a power analysis: fixed effects β_0 (the intercept) and β_1 (the model effect), and variance parameters for the worker intercept (σ_{0w}), the item intercept (σ_{0i}) and their respective slope variance parameters (σ_{1w} and σ_{1i}). There is also a variance parameter for the overall error (σ_{wi}). We transform the Likert ratings to be on a [0, 1] scale and treat them as normally distributed (which we note is an imperfect assumption). We give fit parameters for these values, on a few datasets, in Tables 13, 14, and 15.

$$Y_{wi} = \beta_0 + W_{0w} + I_{0i} + (\beta_1 + W_{1w} + I_{1i})X_i + e_{wi} \quad (5)$$

$$I_{0i} \sim N(0, \sigma_{0i}) \quad (6)$$

$$W_{0i} \sim N(0, \sigma_{0w}) \quad (7)$$

$$I_{1i} \sim N(0, \sigma_{1i}) \quad (8)$$

$$W_{1i} \sim N(0, \sigma_{1w}) \quad (9)$$

$$e_{wi} \sim N(0, \sigma_{wi}) \quad (10)$$

For simplicity and convergence issues, we do not include a correlation parameter in the random effect structure.

To assess power, we use two possible variance settings derived from the model fits ("high variance" and "low variance" settings, in the main text) and show these in Table 16. We systematically vary the number of annotators (always assuming each annotator annotates each item, which is not always true in typical experiments), the number of items, and the effect size. We note that simulations can be customized to the planned analysis, including aspects such as how many items will be annotated by each annotator.

To compute power, we use each setting of the parameters to simulate 200 experiments and compute the proportion that detect a significant positive effect ($t > 1.96$). Significant effects in the opposite direction ($t < -1.96$) do not count as detections. Code for these model fits and simulations is included with the online materials. However, we note that these should be used as a starting point, rather than being blindly copied, as details may differ in each experimental setting.

H.4 Head to head human evaluations

Another commonly used form of human evaluation is head to head comparison, where raters are shown a pair of outputs (one from each model), and asked to choose which they prefer, sometimes with "neither" as a third option. Head to head comparisons offer some advantages over ratings-basd approaches (Yannakakis and Martínez, 2015; van der Lee et al., 2019), but do not scale as well when comparing many models.

As with ordinal judgements, there are multiple ways of analyzing such data. If we treat annotator judgements as independent and identically distributed (such as if we only collect one judgement from each annotator), we could model this simply in terms of the underlying probabilities that a random annotator will prefer each model (as in the opening example in the main paper). In that case, running a power analysis would be a simple as assuming values for the underlying probabilities of each category (win, lose, draw), as usual based on pilot data or prior assumptions, and simulating many draws from that prior, checking in each sample to see if there is a statistically significant difference between win and lose.

On the other hand, if multiple judgements will be collected from each annotator and/or for each pair of outputs, then it makes sense to use a richer model to account for all sources of variation, as described above (see §H.3). In particular, the mixed effects framework can be adopted, potentially by modeling the outcome as a logistic model (in the case of win or lose), with ties either excluded or split.

Dataset	Number of Workers	Number of Items
Hashimoto et al. (2019) (LM)	124	50
Hashimoto et al. (2019) (summarization)	96	99
Hashimoto et al. (2019) (Reddit)	123	99
WMT19	176	1997
Dathathri et al. (2020)	15	1358
Holtzman et al. (2020)	140	1399

Table 12: Number of workers and items in each of our convenience sampled datasets.

Dataset	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\sigma}_{wi}$
Hashimoto et al. (2019) (LM)	0.55	-0.03						0.25
Hashimoto et al. (2019) (summarization)	0.58	0.06						0.26
Hashimoto et al. (2019) (Reddit)	0.55	0.05	0.03	0.01				0.23
WMT19	0.86	0.04						0.12
Dathathri et al. (2020)	0.62	0.04	-0.05	-0.03				0.16
Holtzman et al. (2020)	0.59	0.02	0.04	0.02	0.01	0	-0.04	0.16

Table 13: Fit fixed effect coefficients for each model along with the residual model variance. If only one model is compared to a baseline, there is a value for intercept and β_1. If more than one model, there is an additional parameter for each model. Because we use contrast coding, each coefficient can be interpreted as the difference from the grand mean.

Dataset	$\hat{\sigma}_{0w}$	$\hat{\sigma}_{1w}$	$\hat{\sigma}_{2w}$	$\hat{\sigma}_{3w}$	$\hat{\sigma}_{4w}$	$\hat{\sigma}_{5w}$	$\hat{\sigma}_{6w}$
Hashimoto et al. (2019) (LM)	0	0.11	0.11				
Hashimoto et al. (2019) (summarization)	0	0.13	0.11				
Hashimoto et al. (2019) (Reddit)	0.11	0.04	0.08	0.06	0.17		
WMT19	0.07	0.04	0.13				
Dathathri et al. (2020)	0	0.04	0.05	0.05	0.05		
Holtzman et al. (2020)	0.09	0.05	0.03	0.04	0.04	0.02	0.04

Table 14: Fit random effects standard deviations for worker. As in the equations above, $\hat{\sigma}_{0w}$ is the worker intercept and the rest of the parameters are worker slopes for each model.

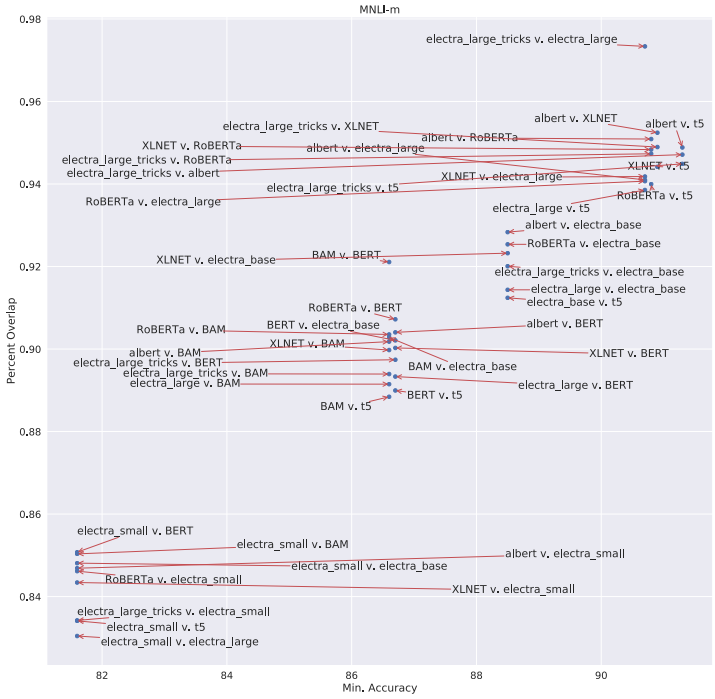
Dataset	$\hat{\sigma}_{0i}$	$\hat{\sigma}_{1i}$	$\hat{\sigma}_{2i}$	$\hat{\sigma}_{3i}$	$\hat{\sigma}_{4i}$	$\hat{\sigma}_{5i}$	$\hat{\sigma}_{6i}$
Hashimoto et al. (2019) (LM)	0.04	0.14	0.1				
Hashimoto et al. (2019) (summarization)	0.07	0	0.18				
Hashimoto et al. (2019) (Reddit)	0	0.13	0.11	0.14	0.14		
WMT19	0.05	0.03	0.15				
Dathathri et al. (2020)	0	0.16	0.19	0.16	0.16		
Holtzman et al. (2020)	0	0.13	0.1	0.12	0.11	0.13	0.13

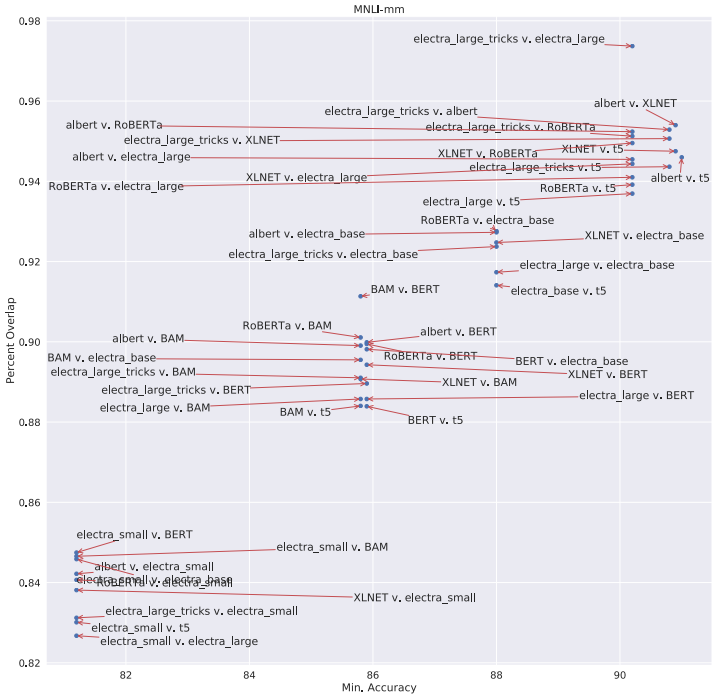
Table 15: Fit random effects standard deviations for item. As in the equations above, $\hat{\sigma}_{0i}$ is the item intercept and the rest of the parameters are item slopes for each model.

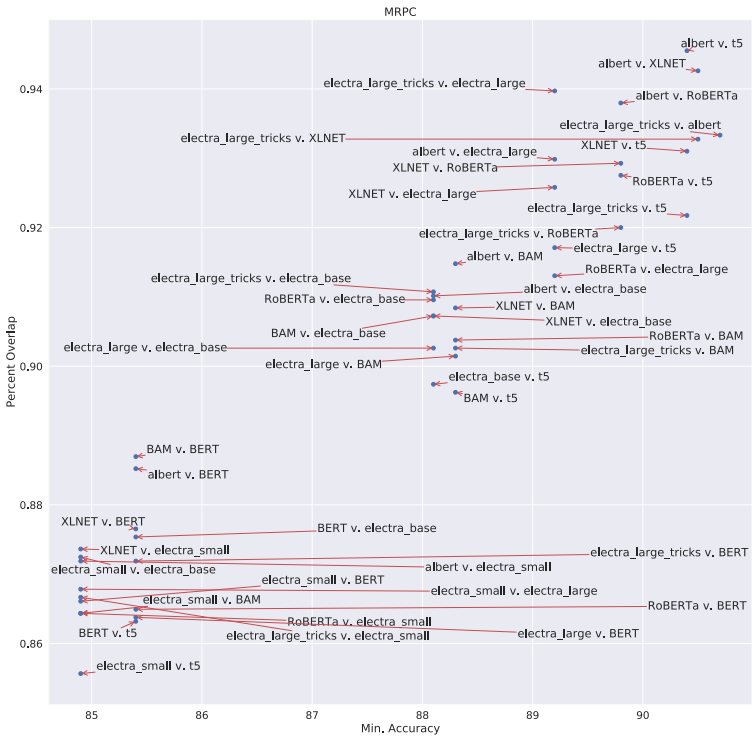
Scenario	σ_{w0}	σ_{w1}	σ_{i0}	σ_{i1}	σ_{wi}
Low variance	0.01	0.04	0.01	0.13	0.16
High variance	0.01	0.11	0.04	0.14	0.26

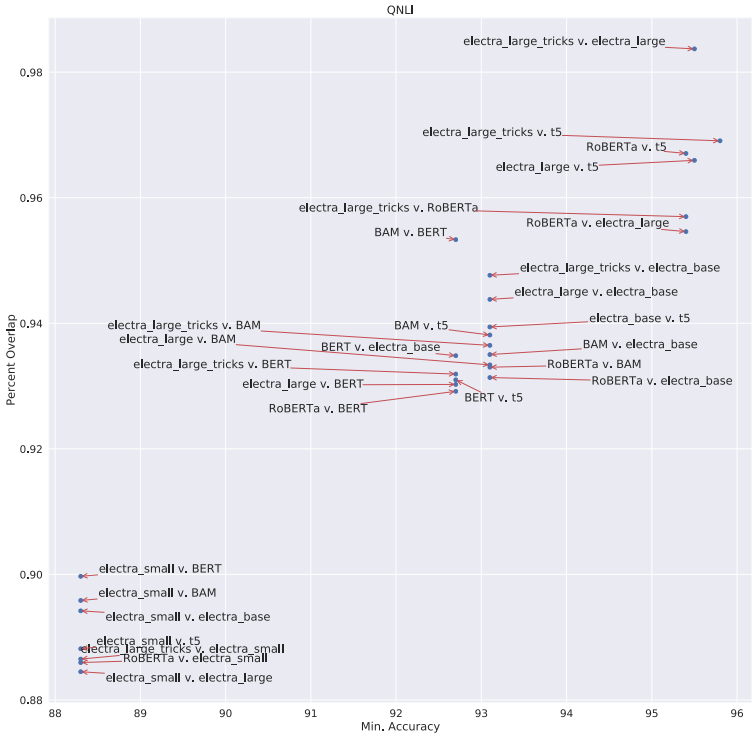
Table 16: An example of high variance and low variance settings. The standard deviations correspond to the variance parameters for worker intercept, worker slope, item intercept, item slope, and sigma, respectively.

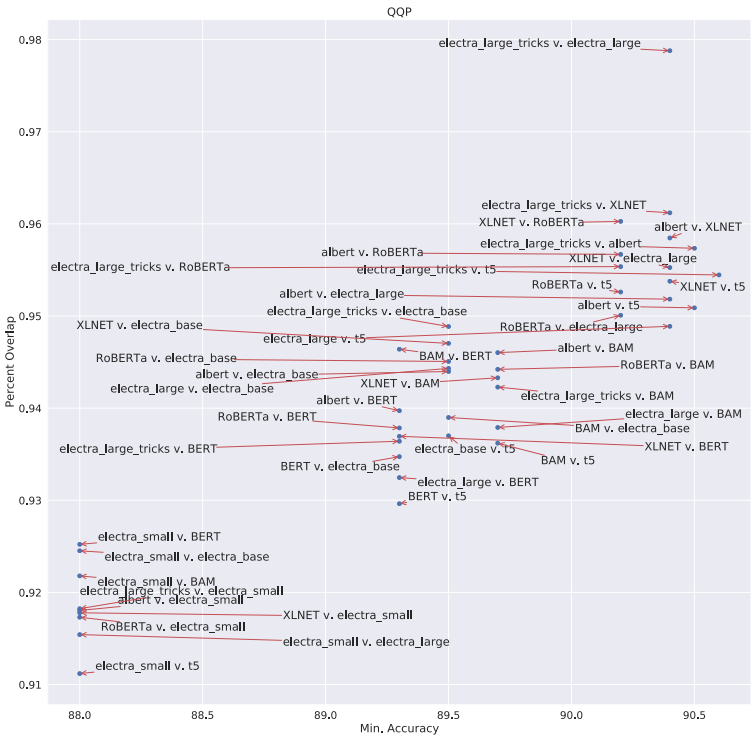
I Additional Plots of Model Overlap

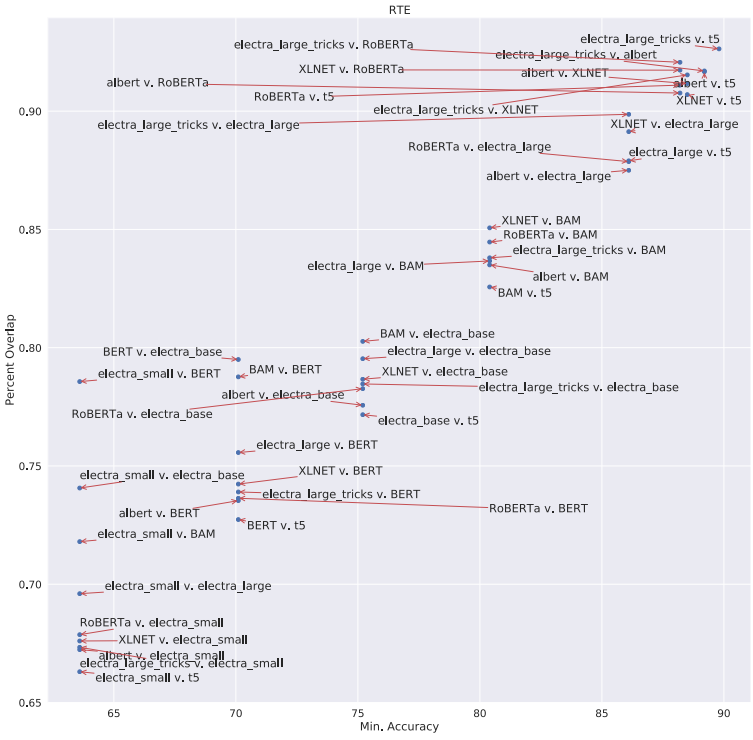


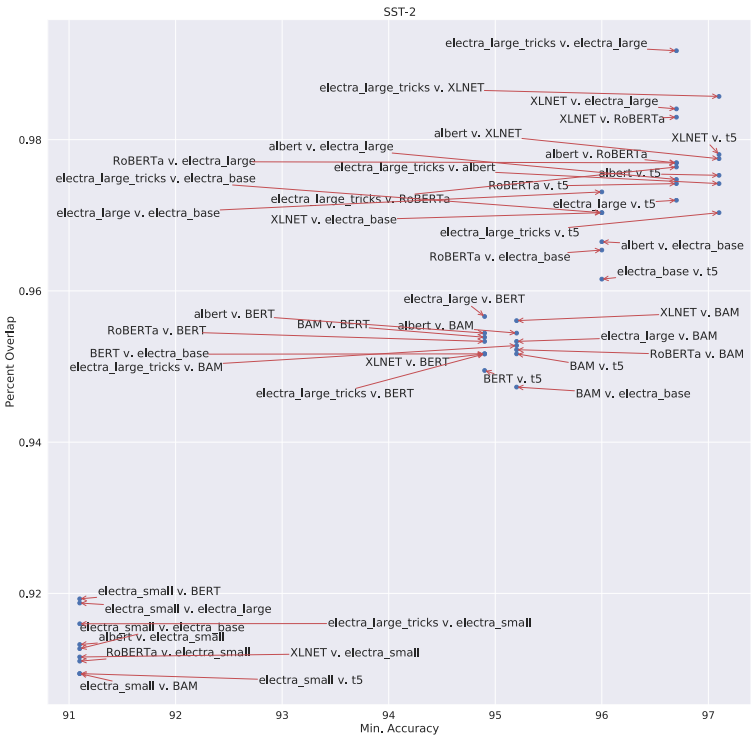


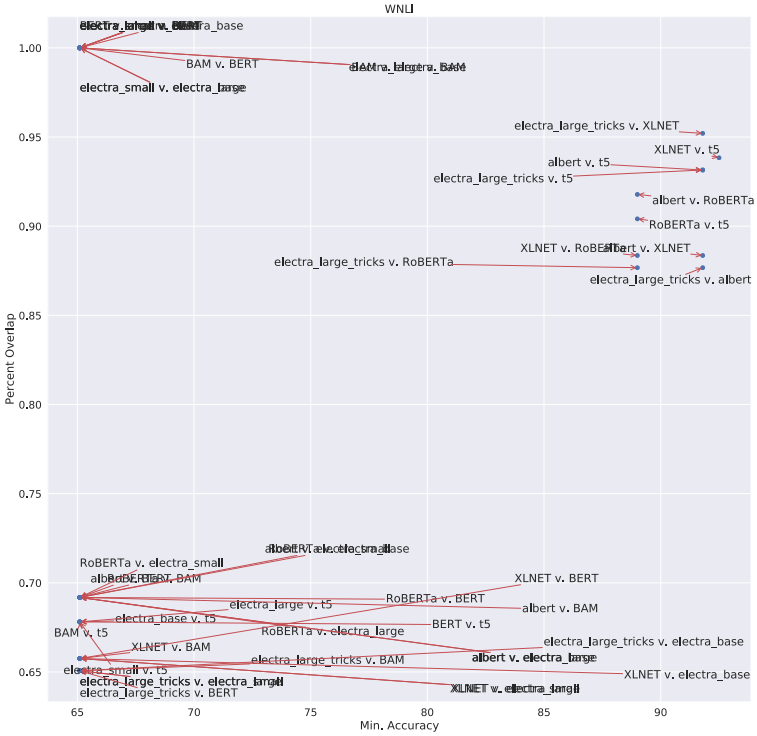

MRPC
albert v. t5
albert v. XLNET
electra_large_tricks v. electra_large
albert v. RoBERTa
electra_large_tricks v. albert
electra_large_tricks v. XLNET
XLNET v. t5
albert v. electra_large
XLNET v. RoBERTa
RoBERTa v. t5
XLNET v. electra_large
electra_large_tricks v. t5
electra_large_tricks v. RoBERTa
electra_large v. t5
albert v. BAM
RoBERTa v. electra_large
electra_large_tricks v. electra_base
albert v. electra_base
RoBERTa v. electra_base
XLNET v. BAM
XLNET v. electra_base
BAM v. electra_base
RoBERTa v. BAM
electra_large v. electra_base
electra_large_tricks v. BAM
electra_large v. BAM
electra_base v. t5
BAM v. t5
BAM v. BERT
albert v. BERT
XLNET v. BERT
BERT v. electra_base
XLNET v. electra_small
electra_large_tricks v. BERT
albert v. electra_small
electra_small v. electra_base
electra_small v. BERT
electra_small v. electra_large
electra_small v. BAM
RoBERTa v. BERT
RoBERTa v. electra_small
BERT v. t5
electra_large_tricks v. electra_small
electra_large v. BERT
electra_small v. t5
Percent Overlap
0.94
0.92
0.90
0.88
0.86
85
86
87
88
89
90
Min. Accuracy


QNLI
electra_large_tricks v. electra_large
electra_large_tricks v. t5
RoBERTa v. t5
electra_large v. t5
electra_large_tricks v. RoBERTa
RoBERTa v. electra_large
BAM v. BERT
electra_large_tricks v. electra_base
electra_large v. electra_base
electra_base v. t5
electra_large_tricks v. BAM
BAM v. t5
electra_large v. BAM
BERT v. electra_base
BAM v. electra_base
electra_large_tricks v. BERT
RoBERTa v. BAM
electra_large v. BERT
RoBERTa v. electra_base
BERT v. t5
RoBERTa v. BERT
electra_small v. BERT
electra_small v. BAM
electra_small v. electra_base
electra_small v. t5
electra_large_tricks v. electra_small
RoBERTa v. electra_small
electra_small v. electra_large
Percent Overlap
0.98
0.96
0.94
0.92
0.90
0.88
88
89
90
91
92
93
94
95
96
Min. Accuracy


QQP
Percent Overlap
Min. Accuracy
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
88.0
88.5
89.0
89.5
90.0
90.5
electra_large_tricks v. electra_large
electra_large_tricks v. XLNET
XLNET v. RoBERTa
albert v. XLNET
albert v. RoBERTa
electra_large_tricks v. albert
XLNET v. electra_large
electra_large_tricks v. RoBERTa
electra_large_tricks v. t5
RoBERTa v. t5
XLNET v. t5
albert v. electra_large
albert v. t5
electra_large_tricks v. electra_base
XLNET v. electra_base
RoBERTa v. electra_large
electra_large v. t5
BAM v. BERT
albert v. BAM
RoBERTa v. electra_base
RoBERTa v. BAM
albert v. electra_base
XLNET v. BAM
electra_large v. electra_base
electra_large_tricks v. BAM
albert v. BERT
RoBERTa v. BERT
electra_large v. BAM
BAM v. electra_base
electra_large_tricks v. BERT
electra_base v. t5
XLNET v. BERT
BAM v. t5
BERT v. electra_base
electra_large v. BERT
BERT v. t5
electra_small v. BERT
electra_small v. electra_base
electra_small v. BAM
electra_large_tricks v. electra_small
albert v. electra_small
XLNET v. electra_small
RoBERTa v. electra_small
electra_small v. electra_large
electra_small v. t5


RTE
Percent Overlap
Min. Accuracy
0.90
0.85
0.80
0.75
0.70
0.65
65
70
75
80
85
90
electra_large_tricks v. t5
electra_large_tricks v. RoBERTa
electra_large_tricks v. albert
XLNET v. RoBERTa
albert v. XLNET
albert v. t5
albert v. RoBERTa
RoBERTa v. t5
XLNET v. t5
electra_large_tricks v. XLNET
electra_large_tricks v. electra_large
XLNET v. electra_large
RoBERTa v. electra_large
electra_large v. t5
albert v. electra_large
XLNET v. BAM
RoBERTa v. BAM
electra_large_tricks v. BAM
electra_large v. BAM
albert v. BAM
BAM v. t5
BAM v. electra_base
BERT v. electra_base
electra_large v. electra_base
electra_small v. BERT
BAM v. BERT
XLNET v. electra_base
electra_large_tricks v. electra_base
albert v. electra_base
RoBERTa v. electra_base
electra_base v. t5
electra_large v. BERT
electra_small v. electra_base
XLNET v. BERT
electra_large_tricks v. BERT
albert v. BERT
RoBERTa v. BERT
BERT v. t5
electra_small v. BAM
electra_small v. electra_large
RoBERTa v. electra_small
XLNET v. electra_small
albert v. electra_small
electra_large_tricks v. electra_small
electra_small v. t5


SST-2
Percent Overlap
Min. Accuracy
electra_large_tricks v. electra_large
electra_large_tricks v. XLNET
XLNET v. electra_large
XLNET v. RoBERTa
albert v. XLNET
XLNET v. t5
albert v. electra_large
albert v. RoBERTa
RoBERTa v. electra_large
electra_large_tricks v. albert
albert v. t5
electra_large_tricks v. electra_base
RoBERTa v. t5
electra_large_tricks v. RoBERTa
electra_large v. t5
electra_large v. electra_base
XLNET v. electra_base
electra_large_tricks v. t5
albert v. electra_base
RoBERTa v. electra_base
electra_base v. t5
electra_large v. BERT
albert v. BERT
BAM v. BERT
albert v. BAM
XLNET v. BAM
RoBERTa v. BERT
electra_large v. BAM
BERT v. electra_base
RoBERTa v. BAM
electra_large_tricks v. BAM
XLNET v. BERT
BAM v. t5
BERT v. t5
electra_large_tricks v. BERT
BAM v. electra_base
electra_small v. BERT
electra_small v. electra_large
electra_small v. electra_base
electra_large_tricks v. electra_small
albert v. electra_small
RoBERTa v. electra_small
XLNET v. electra_small
electra_small v. t5
electra_small v. BAM


WNLI
Percent Overlap
Min. Accuracy
BAM v. BERT
electra_small v. electra_large
electra_large_tricks v. XLNET
XLNET v. t5
albert v. t5
electra_large_tricks v. t5
albert v. RoBERTa
RoBERTa v. t5
XLNET v. RoBERTa
albert v. XLNET
electra_large_tricks v. RoBERTa
electra_large_tricks v. albert
RoBERTa v. electra_small
XLNET v. BERT
albert v. BAM
electra_large v. t5
RoBERTa v. BERT
electra_base v. t5
BERT v. t5
BAM v. t5
RoBERTa v. electra_large
electra_large_tricks v. electra_base
XLNET v. BAM
electra_large_tricks v. BAM
XLNET v. electra_base
electra_small v. t5
electra_large_tricks v. BERT