Comparing Test Sets with Item Response Theory

Clara Vania[♠*†] **Phu Mon Htut**[♣*] **William Huang**[♣*]
Dhara Mungra[♣] **Richard Yuanzhe Pang**[♣] **Jason Phang**[♣]
Haokun Liu[♢†] **Kyunghyun Cho**[♣] **Samuel R. Bowman**[♣]
[♠]Amazon [♣]New York University
[♢]Allen Institute for AI
vaniclar@amazon.co.uk, bowman@nyu.edu

Abstract

Recent years have seen numerous NLP datasets introduced to evaluate the performance of fine-tuned models on natural language understanding tasks. Recent results from large pretrained models, though, show that many of these datasets are largely saturated and unlikely to be able to detect further progress. What kind of datasets are still effective at discriminating among strong models, and what kind of datasets should we expect to be able to detect future improvements? To measure this uniformly across datasets, we draw on Item Response Theory and evaluate 29 datasets using predictions from 18 pretrained Transformer models on individual test examples. We find that Quoref, HellaSwag, and MC-TACO are best suited for distinguishing among state-of-the-art models, while SNLI, MNLI, and CommitmentBank seem to be saturated for current strong models. We also observe span selection task format, which is used for QA datasets like QAMR or SQuAD2.0, is effective in differentiating between strong and weak models.

1 Introduction

Many datasets have been created to evaluate various aspects of natural language understanding (NLU) in English. These datasets are useful to measure progress; however, it is evident from various leaderboards (Wang et al., 2018, 2019b; Rajpurkar et al., 2016; Zellers et al., 2018) that many of them are no longer challenging or discriminative enough to differentiate strong models such as those based on Transformers (Vaswani et al., 2017).[1] Even if these benchmarks are sound tests of important (and potentially unsolved) tasks, their usefulness is limited if they cannot measure further progress. In this paper, we ask: Which datasets are best in distinguishing current and possible future strong models?

We aim to compare datasets using a single metric that accounts for their effectiveness in separating current stronger and weaker models. To that end, we use Item Response Theory (IRT; Baker and Kim, 1993), a statistical framework from psychometrics that is widely used for the evaluation of test items in educational assessment. IRT assumes that the probability that a model will correctly handle an example in a test set depends on the model's latent ability parameter and three example-specific parameters, typically measuring example difficulty (how strong does a model have to be to get it right), discrimination (how effective the example is for differentiating between similar models), and guessing (how likely a weak model is to get the example right for spurious reasons).

This paper presents a large-scale IRT analysis of existing English NLU datasets. Unlike previous work which focuses on example-level analysis *within* individual datasets (Lalor et al., 2016, 2018), here we analyze example characteristics from a larger perspective by comparing individual examples *across* datasets. We evaluate test sets from 29 datasets in different formats—classification, multiple-choice QA, and span-selection QA. As responses, we use model predictions from 18 Transformer-based models, including some limited-capacity models chosen to expose better the dataset's ability to discriminate weaker from stronger predictors. We then fit a single IRT model on these responses using a variational inference method.[2]

[*]Equal contribution.
[†]Work done while at New York University.
[1]For example, the recent DeBERTa model (He et al., 2020) achieves parity with human annotators on the SuperGLUE benchmark score: https://super.gluebenchmark.com/leaderboard.
[2]Our data and code can be found at https://github.com/nyu-mll/nlu-test-sets.

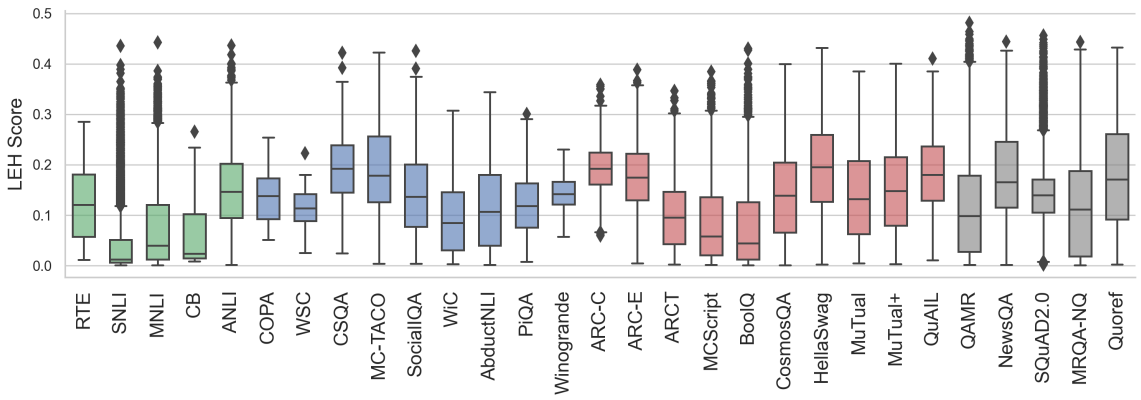

Figure 1: Distribution of test examples according to our proposed *locally estimated headroom* (LEH) scores (§ 4.1.1), which measure the **local slope** of the Item Characteristic Curve (ICC) for an example at the ability level corresponding to the **best model**, and thus reflect the effectiveness of that single example at distinguishing between near-state-of-the-art models. Datasets are grouped by task format: classification (green), sentence-level multiple-choice (blue), paragraph-level multiple-choice (red), and span selection (grey). Within each format, the datasets are sorted by their release date. More details on the datasets are given in Table 1.

We find:

- Quoref, HellaSwag, and MC-TACO contain the highest number of examples that can differentiate between near-state-of-the-art models, making them very likely to be effective at tracking near-future progress on the skills that they actually test (Figure 1).
- SQuAD2.0, NewsQA, QuAIL, MC-TACO, and ARC-Challenge have the most difficult examples.
- Span-based QA is an effective task format for discriminating between strong and weak models.
- CosmosQA, MC-TACO, Winogrande, and ARC-Challenge consist mostly of hard examples, while for most datasets, the example difficulty levels are more widely distributed.

2 Item Response Theory

Baker and Kim (1993) introduce Item Response Theory (IRT), a statistical framework to measure the probability of a responder (human or AI system) predicting a correct answer for a given item (test example). The probability of a responder i answering an item j correctly is estimated as a function of the responder's latent ability θ_i and the item characteristics, referred to as the item characteristic curve (ICC).

We use the 3-parameter (3PL) IRT model, where item behavior is governed by discrimination, difficulty, and guessing parameters. The discrimination parameter (α) defines how effective an item is for distinguishing predictors along the ability axis. The difficulty parameter (β) defines a minimum level of ability at which we expect to see high responder performance. The guessing parameter (γ) defines the probability of correctly answering an item by random guessing. Figure 2 shows example ICCs with different parameter values.

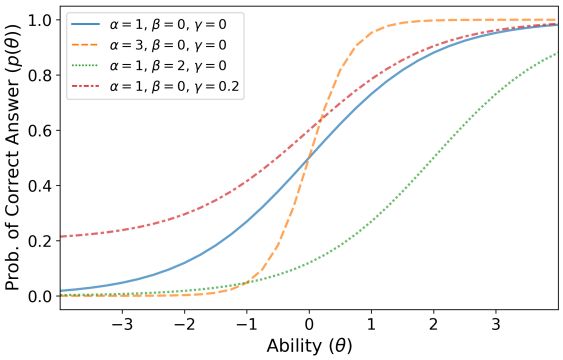

Figure 2: An example of item characteristic curves (ICCs) with different values for discrimination (α), difficulty (β), and guessing (γ) parameters. $p(\theta)$ is the probability of a correct answer for a given θ. θ measures a model's ability level (higher is better). α governs the steepness of the function, β determines the θ value at which the curve is the steepest, while γ defines the baseline likelihood that an arbitrarily weak model can guess correctly.

Formally, the probability of individual i answering item j correctly is modeled as:

$$p_j(\theta_i) = \gamma_j + \frac{1 - \gamma_j}{1 + e^{-\alpha_j(\theta_i - \beta_j)}}. \quad (1)$$

2.1 IRT with Variational Inference

We use variational inference to infer IRT parameters from model response patterns using Pyro (Ranganath et al., 2014; Bingham et al., 2019). Lalor et al. (2019) found this method effective when fitting IRT models to responses on SNLI. Let n be the number of items and let m be the number of responders. The response patterns is $\mathbf{Y} \in \mathbb{R}^{n \times m}$, where the i-th row corresponds to responder i and the j-th column corresponds to item j. We define $y_{ij} \in [0, 1]$ as the response of model i to item j, where $y_{ij} = 1$ indicates a correct response and $y_{ij} = 0$ indicates an incorrect response. We approximate the joint probability of the parameters $\pi(\theta, \alpha, \beta, \gamma \mid \mathbf{Y})$ with a variational posterior:

$$q(\theta, \alpha, \beta, \gamma) = \prod_{i=1}^{I} \pi_i^{\theta}(\theta_i) \prod_{j=1}^{J} \pi_j^{\alpha}(\alpha_i) \pi_j^{\beta}(\beta_i) \pi_j^{\gamma}(\gamma_i) \tag{2}$$

where $\pi^{\rho}(\cdot)$ denotes the density for parameter ρ. For each parameter, we choose the following distributions:

$$\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \tag{3}$$

$$\log \alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \tag{4}$$

$$\beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \tag{5}$$

$$\text{sigmoid}^{-1}(\gamma) \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2) \tag{6}$$

We fit the posterior parameters by minimizing the evidence lower bound (ELBO). When calculating the ELBO, we weight the log-likelihoods of each item's parameter by the inverse of the item's dataset size to control for test set size.

Following Lalor et al. (2019), we use a prior of $\mathcal{N}(0, 1)$ for θ, β, and $\text{sigmoid}^{-1}(\gamma)$. While Lalor et al. (2019) uses $\mathcal{N}(0, 10^3)$ for item parameter priors, we encountered degenerate runs and instead use $\mathcal{N}(0, 1)$. For $\log \alpha$, we use $\mathcal{N}(0, \sigma_\alpha^2)$ where we set σ_α by searching $[0.25, 0.5]$ by increments of 0.05 and use the value yielding the highest ELBO after excluding degenerate runs. We use a sigmoid transformation for γ to constrain the guessing probability to $(0, 1)$.

3 Experiments

3.1 Datasets

Our goal is to perform a fine-grained evaluation of English NLU datasets that appear to discriminate among widely used Transformer-based models. To that end, we choose datasets based on the following criteria:

- They are plausibly unsolved, in that the best-reported model performance does not exceed estimated human performance (if available) by more than three metric points.
- They are relatively easy to use with current large pretrained models, and in particular, their inputs fit within a typical pretrained Transformer's 512-token limits. (This rules out tasks with full-document contexts or retrieval components.)
- They are evaluated at example-level, i.e., we focus our analysis on QA and other classification datasets, where each example corresponds to one item in the IRT. (This rules out structured prediction and sequence tagging tasks.)
- They have simple and reliable automatic metrics at the example level. (This rules out generation-based tasks.)

Table 1 lists the datasets we evaluate. For MNLI, we combine the matched and mismatched portions of the development and custom test sets for our analysis. For ANLI, we train models on SNLI, MNLI, and ANLI training examples. Similar to MNLI, we combine ANLI's three evaluation rounds of the development and the test sets for our analysis.

Custom Test Splits Some of our selected datasets do not have publicly available labeled test examples. For such cases, we create a new custom split by randomly sampling 50% of the validation examples as a new test set and keeping the rest for validation ("Cust." column in Table 1). For Natural Questions, we use the MRQA 2019 version (Fisch et al., 2019), as the original version includes some examples with very long contexts.[3] For MCTACO, the original dataset does not come with a training set. For our experiment, we use 80% of the validation set as our training set and the rest as a our validation set while leaving the original test set untouched.

[3] https://github.com/mrqa/MRQA-Shared-Task-2019

		|Train|	|Dev|	|Test|	Cust.	Metric	RoBERTa	Human
Classification	RTE (Dagan et al., 2005, et seq.)	2,490	138	139	✓	Acc.	87.6	93.6
	SNLI (Bowman et al., 2015)	550,152	10,000	10,000		Acc.	92.7	–
	MNLI (Williams et al., 2018)	392,702	9,823	9,824	✓	Acc.	89.7	92.0
	CommitmentBank (CB; De Marneffe et al., 2019)	250	28	28	✓	Acc.	90.5	95.8
	ANLI (Nie et al., 2020)	1,105,719	3,200	3,200		Acc.	50.8	–
Sentence-Level Multiple Choice	COPA (Roemmele et al., 2011)	400	50	50	✓	Acc.	86.0	100.0
	WSC (Levesque et al., 2012)	554	52	52	✓	Acc.	78.8	100.0
	CommonsenseQA (CSQA; Talmor et al., 2019)	9,741	610	611	✓	Acc.	74.6	88.9
	MC-TACO (Zhou et al., 2019)	3,026	757	9,442	✓	EM	55.9	75.8
	SocialIQA (Sap et al., 2019)	33,410	977	977	✓	Acc.	79.9	88.1
	WiC (Pilehvar and Camacho-Collados, 2019)	5,428	319	319	✓	Acc.	71.5	80.0
	Abductive NLI (AbductNLI; Bhagavatula et al., 2020)	169,654	766	766	✓	Acc.	85.0	92.9
	PIQA (Bisk et al., 2020)	16,113	919	919	✓	Acc.	77.6	94.9
	WinoGrande (Sakaguchi et al., 2020)	40,398	633	634	✓	Acc.	77.3	94.0
Paragraph-Level Multiple Choice	ARC-Easy (Clark et al., 2018)	2,251	570	2,376		Acc.	62.5	–
	ARC-Challenge (Clark et al., 2018)	1,119	299	1,172		Acc.	37.5	–
	ARCT (Habernal et al., 2018)	1,211	317	445		Acc.	86.7	79.8
	MCScript (Ostermann et al., 2018)	14,191	2,020	3,610		Acc.	92.8	98.2
	BoolQ (Clark et al., 2019)	9,427	1,635	1,635	✓	Acc.	85.7	89.0
	Cosmos QA (Huang et al., 2019)	25,262	1,492	1,493	✓	Acc.	79.4	94.0
	HellaSwag (Zellers et al., 2019)	39,905	5,021	5,021	✓	Acc.	84.1	95.6
	MuTual (Cui et al., 2020)	7,088	443	443	✓	Acc.	87.8	93.8
	MuTual+ (Cui et al., 2020)	7,088	443	443	✓	Acc.	77.9	93.0
	QuAIL (Rogers et al., 2020)	10,246	2,164	556		Acc.	73.3	–
Span Selection	QAMR (Michael et al., 2018)	50,615	18,908	18,770		EM	79.6	–
	NewsQA (Trischler et al., 2017)	76,568	4,343	4,293		EM	57.8	46.5
	SQuAD2.0 (Rajpurkar et al., 2018)	130,319	5,675	6,198	✓	EM	91.5	86.8
	MRQA-NQ (Kwiatkowski et al., 2019)	104,071	6,418	6,418	✓	EM	69.9	–
	Quoref (Dasigi et al., 2019)	19,399	1,209	1,209	✓	EM	78.7	93.0

Table 1: Datasets grouped by their task format and ordered by release year. **Cust.** denotes cases when we use our own custom split. **Metric**: evaluation metric used in this study. **RoBERTa**: model performance using $\text{RoBERTa}_{\text{Large}}$. **Human**: human performance.

3.2 Models

We aim to understand how examples from different datasets contribute to the evaluations of models with near-state-of-the-art abilities, so we include several pretrained Transformer-based models to approximate this. However, using only high-performing models could result in a poor IRT model fit (Martínez-Plumed et al., 2019) To avoid this, we add both weaker models and under-trained versions of our original models. We use ALBERT-XXL-v2 (Lan et al., 2020), $\text{RoBERTa}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Base}}$ (Liu et al., 2019), $\text{BERT}_{\text{Large}}$ and $\text{BERT}_{\text{Base}}$ (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and 12 MiniBERTas (Zhang et al., 2021b). [4] For each of the 18 Transformer-based models, we evaluate five different checkpoints—at 1%, 10%, 25%, and 50% of the maximum steps of the maximum epochs (Section 3.3), as well as the best checkpoint on the validation set, which need not be one of the other four. This yields a total of 90 model predictions for each test example.

3.3 Experimental Setup

Optimization We perform a hyperparameter sweep on each dataset, varying the learning rate $\in \{1e-5, 3e-5, 5e-6\}$. We tune the maximum epochs $\in \{10, 40\}$ for small datasets ($<$ 5k training examples), and $\in \{3, 10\}$ for other datasets (Zhang et al., 2021a). We use the `jiant` (Pruksachatkun et al., 2020b) library which is based on PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020).

We only perform hyperparameter tuning with the $\text{RoBERTa}_{\text{Large}}$ model and apply the best configuration to train all the other Transformer models. We use NVIDIA V100 Tensor Core GPUs for our experiments. On average, it takes approximately four hours to train RoBERTa on small datasets ($<$ 3k training examples), one day for medium-

[4] The MiniBERTas are RoBERTa models pretrained on 1M, 10M, 100M, or 1B words of raw text, and varying slightly in model size. There are three pretrained models for each pretraining data quantity, which are pretrained using different near-optimal hyperparameter values. We use all three variants in producing responses for IRT.

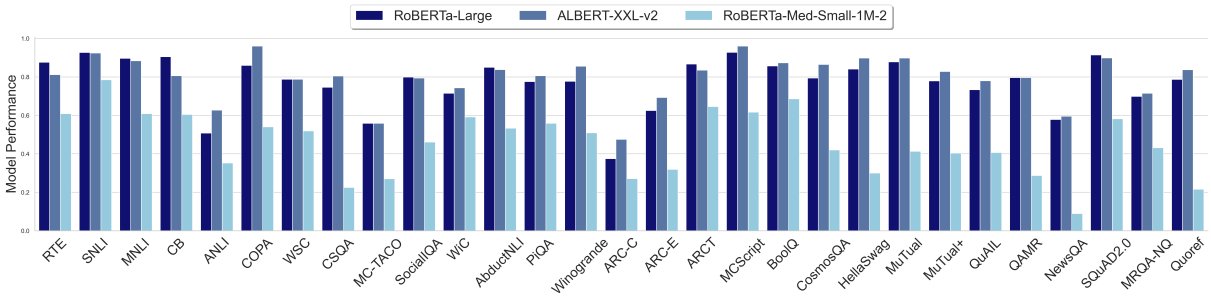

Figure 3: The best validation performance of ALBERT-XXL-v2, RoBERTa$_{\text{Large}}$, and the smallest MiniBERTa (RoBERTa-Med-Small-1M-2) on each dataset. The full results table with performance of all models is reported in the Appendix (Table 3)

sized datasets ($<$ 10k), and four days for large datasets ($> 10k$).

4 Results and Analysis

Figure 3 shows the performance of RoBERTa$_{\text{Large}}$, ALBERT-XXL-v2, and one of the low performing MiniBERTas (RoBERTa-Med-Small-1M-2) on all validation sets. Unsurprisingly, ALBERT-XXL-v2 and RoBERTa$_{\text{Large}}$ are the best-performing models, while the small MiniBERTa model achieves much lower performance. Full results using all 18 models can be found in the Appendix (Table 3).

4.1 IRT Analysis

4.1.1 Item Characteristics

Metric As our primary metric, we introduce *Locally Estimated Headroom* (LEH) score, which measures the ability of each test example to contribute to the evaluation of near-future progress. We calculate it as the derivative of the example's ICC (Figure 2) with respect to the highest latent ability score, which corresponds to ALBERT-XXL-v2. A high LEH score indicates that the best-performing model is still far from the example's saturation points—the flat sections of ICC inferred by our model. There is enough space along the curve that the IRT model expects the example to be able to differentiate future state-of-the-art models. Typically, different near-state-of-the-art models both succeed and fail on this kind of example, while weaker models mostly fail. A high LEH score implies that there is still enough room for potentially stronger models to perform better on this dataset.

To validate the use of LEH scores for detecting near-future improvements, we compare two IRT models. The first is fitted using responses from all models, while the second is fitted based on responses from BERT and other weaker models (excluding RoBERTa$_{\text{Large}}$, RoBERTa$_{\text{Base}}$, XLM-R, and ALBERT-XXL-v2). After that, we compute the correlation between the two sets of LEH scores, focusing on the 75$^{\text{th}}$ percentile for each dataset. The Pearson correlation is 95.5% with a median absolute difference of 0.007 and a standard deviation of 0.011. Out of the 29 datasets, only SQuAD2.0, CommensenseQA, MuTual, Quoref, and HellaSwag have more than 0.02 absolute difference in LEH scores. This strong correlation suggests that our ICCs fits are not overly sensitive to the exact characteristics of current state of the art models.

Analysis by LEH Scores Figure 1 shows the distribution of test examples for each dataset based on their LEH scores. For our analysis, we focus on the 75$^{\text{th}}$ percentile examples in each dataset as a rough proxy for how likely a dataset is to have a significant number of examples that are difficult or discriminative for near-future models.

We observe that Quoref, HellaSwag, and MC-TACO have examples with the highest LEH scores, suggesting sufficient headroom for future state-of-the-art models with a higher ability to achieve better performance on these datasets. SNLI, CommitmentBank, and MNLI have relatively low LEH scores, indicating that performance on these datasets is largely saturated. Additionally, we also measure how the 75$^{\text{th}}$ percentile LEH scores correlate with human-RoBERTa gap. Using 22 datasets that have human performance numbers (Table 1), we find that the Pearson correlation between the two is weakly positive (0.21).

Analysis by Item Parameters Next, we analyze the distribution of test examples according to their discrimination and difficulty parameters (Figure 4). We observe that datasets with span selection for-

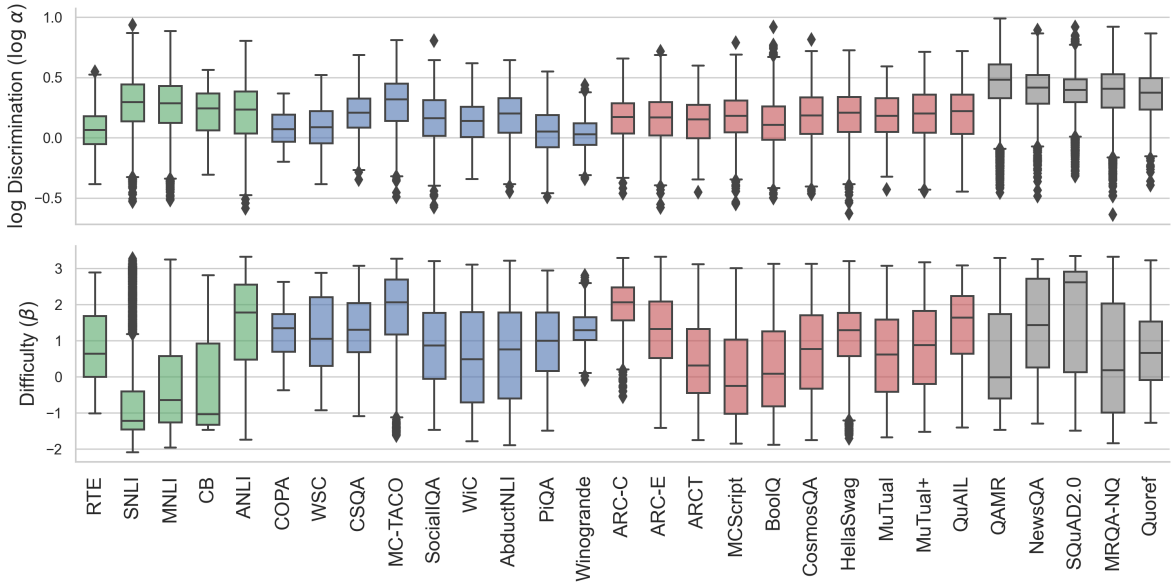

Figure 4: Distribution of test examples for each dataset based on the log discrimination ($\log \alpha$) parameter (top) and the difficulty (β) parameter (bottom).

mat (QAMR, NewsQA, SQuAD, MRQA-NQ, and Quoref) have the highest discrimination scores than other datasets, highlighting span selection as an effective task format for discriminating among strong and weak models. However, this might be because this task format typically features a much larger space of possible model outputs than the other formats we consider. It does not necessarily mean that span selection is the most suitable to test models' ability to understand language. As the span-based format restricts answers to be text spans in the given passage, there are concerns that it rarely requires reasoning ability which often involves answers not mentioned in the passage, and thus not reflecting comprehension ability of humans (Lai et al., 2017; Sugawara et al., 2018).

For the difficulty parameter, we do not observe a narrow task format that is superior to the others. However, we notice that the highest difficulty scores are obtained by QA datasets such as SQuAD2.0, NewsQA, QuAIL, ARC-Challenge, and MC-TACO. ANLI, which is created with adversarial model-in-the-loop crowdsourcing, also has of many hard examples. Impressionistically, training set size and creation date do not seem to correlate with either example's difficulty or discrimination parameters.

Figure 5 shows the distribution of examples jointly according to their difficulty and log discrimination parameters. We notice a half-moon shape pattern in most datasets, which indicates that most of the discriminative examples are either very easy or very difficult. Referring to the ICC curve (Figure 2), this indicates that there is high agreement among strong models or weak models, which corresponds to one of the saturation points in the ICC curve (upper or lower). The only dataset that does not have this pattern is Winogrande, which is difficult for all models.

ARC-Challenge, QuAIL, HellaSwag, CommonsenseQA, and MC-TACO show clusters with high density on the top right regions, indicating a large number of examples with high discrimination and difficulty scores. Other datasets have more scattered distributions. SNLI, MNLI, and MCScript show higher density on the bottom right regions, while NewsQA, SQuAD2.0, and MRQA-NQ show higher density on both the top and bottom right regions. Further analysis of the guessing parameters can be found in Appendix A.

4.2 Examples with Unanimous Responses

When fitting ICC on examples that have only correct responses or only incorrect responses, the discrimination parameter is unconstrained. We find that these examples make up 4% of our data. 13 of the 29 datasets contain at least one such example. Roughly 16% of NewsQA examples are incorrectly answered by all models, while the remaining 12 datasets have less than 10% of all correct or incorrect examples. To study the effect of examples with all correct or incorrect responses, we fit an

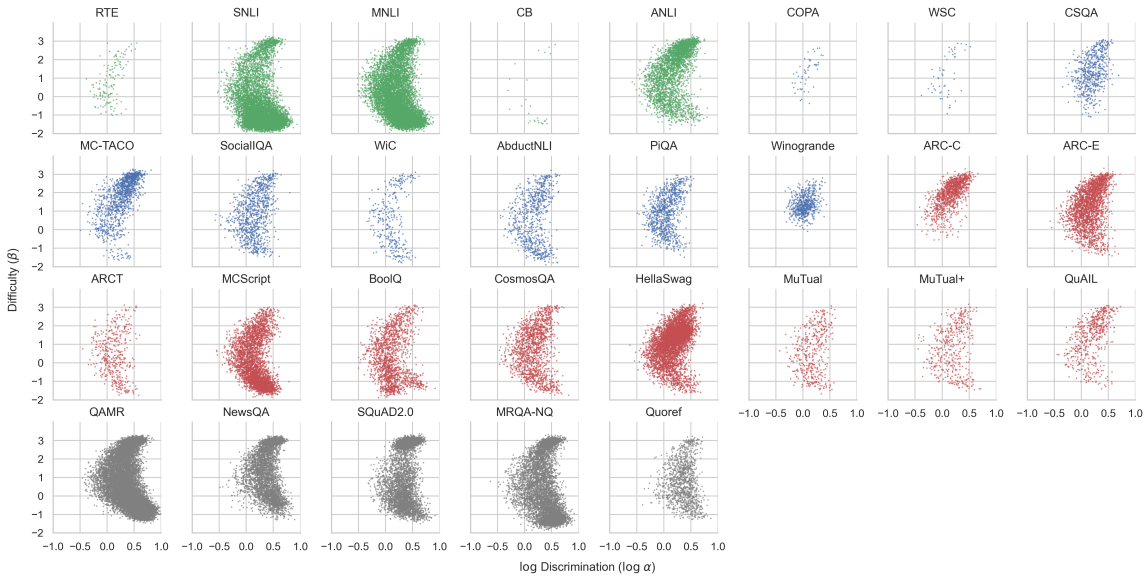

Figure 5: Distributions of log discrimination ($\log \alpha$) versus the difficulty (β) parameters for each dataset..

IRT model on responses excluding such examples and compare against parameters from the full set of responses. We find that the Pearson correlation for the discrimination at the 75$^{\text{th}}$ percentile is 97.2%, with a median absolute difference of 0.016 and standard deviation of 0.015. MC-TACO, CommitmentBank, and WSC differ by more than 0.04. Further, we find that the Pearson correlation for the LEH score at the 75$^{\text{th}}$ percentile is 98.9%, with a median absolute difference of 0.006 and standard deviation of 0.005. RTE, WiC, WinoGrande, QAMR, NewsQA, MRQA-NQ, MC-TACO, and BoolQ differ by 0.01. Given these high correlations, we do not exclude these examples when reporting our main results.

4.3 Analysis by Task Group

Next, we analyze each task-type group in more detail, focusing on the example's scores around the 75$^{\text{th}}$ percentile.

Classification We observe that all datasets have moderate discrimination scores. Most ANLI examples have relatively high difficulty scores, while SNLI, MNLI, and CommitmentBank have the lowest difficulty scores.

Sentence-Level Multiple Choice All of the datasets in this group have relatively low discrimination scores compared to span selection datasets. Figure 5 shows that MC-TACO, Winogrande, and CommonsenseQA all have a higher density of difficult examples, while for other datasets the distribution is more spread.

Paragraph-Level Multiple Choice QuAIL and ARC-Challenge examples have high difficulty but moderate discrimination scores. As seen in Figure 5, these datasets have a higher density in the top right regions, showing a large proportion of difficult examples. ARCT shows moderate difficulty despite its known artifacts (Niven and Kao, 2019), indicating that it can still be challenging for models. Compared to other datasets, BoolQ has the highest number of easy examples. However, as it is a binary classification task, the random baseline performance is already high.

To investigate this, we calculate the number of examples in each test set that have γ parameter below 0.5. In general, we find that 88% of the test examples have $\gamma < 0.5$, implying that most of the examples contributed to the inferences of α, β, and θ. BoolQ was the only exception in which approximately 56% of examples were assigned $\gamma > 0.5$. After filtering out these guessable examples in BoolQ, we find that its test examples have slightly higher discrimination scores with little change in difficulty scores.

Span Selection We observe that span selection datasets are the most discriminative. However, in terms of difficulty, only SQuAD2.0 and NewsQA are among the top five.

4.3.1 Analysis on Model Ability

For a sanity check, we further analyze how each model scores according to our fitted IRT parame-

Name	Example	Difficulty (β)
MNLI	*Premise*: And, you know, with this, you know, it wasn't many opportunities for kids to be special, because kids weren't, you know, you were pushed out of adult conversation, and just really pushed to the side. *Hypothesis*: Children were pushed out of adult conversation, and really just pushed to the side in general. *Label*: entailment	3.27
MNLI	*Premise*: Look, it's your skin, but you're going to be in trouble if you don't get busy. *Hypothesis*: The boss will fire you if he sees you slacking off. *Label*: neutral	-1.87
MC-TACO	The Beatles are giving a press conference about their new film, Magical Mystery Tour .What time of day was the press conference? (1) 4:00 PM ✓ (2) 12:00 PM ✓ (3) 3 p.m ✓ (4) 6:00 AM ✗	2.86
MC-TACO	Because then they feel like they are forced to stay in that situation."On average, how often do they feel stuck in the situation? (1) 54 months ✗ (2) 6 centuries ✗ (3) once every 6 years ✗ (4) every few seconds ✗ (5) once every 2 seconds ✗ (6) once every 18 years ✗	-1.67

Table 2: Hardest and easiest examples along with their estimated difficulty score for MNLI and MC-TACO.

ters. We observe a positive correlation between ability and average model accuracy (Appendix B). Generally, within a model, the best validation checkpoint obtains the highest average model accuracy and/or ability score. Across models, ALBERT-XXL-v2 performs typically best.

4.4 Qualitative Analysis

To better understand what kinds of examples are difficult or discriminating, we analyze the 20 examples with the lowest and highest scores for the discrimination and the difficulty parameters from five datasets: SQuAD2.0, MC-TACO, QuAIL, MNLI, and BoolQ. The first three are datasets with high discrimination and/or difficulty scores. MNLI and BoolQ have moderate discrimination and difficulty scores and low label entropy (three-class classification for MNLI and binary choice for BoolQ).

We observe that the 20 most difficult BoolQ examples are labeled *False* (the minority class), while 19 of the 20 easiest examples are labeled *True*. For MNLI, we find that the 20 easiest MNLI examples are labeled *neutral* while the 20 hardest examples are a mixture of *entailment* and *contradiction*.

In MC-TACO, each example contains a varying number of answer choices. For each choice, a model needs to predict whether the answer is *True* or *False*. We find that all answer choices in top 20 easiest examples are labeled *False* (the majority class), whereas for difficult examples the answer choices are either all *True* or a mix of *True* and *False* (Table 2). For SQuAD2.0 and QuAIL, we analyze the context length, the answerability of a question, and the lexical overlap between context and questions. However, we do not find any clear evidence that any of them might indicate the difficulty level of test examples.

For BoolQ, we observe that the 20 most discriminating examples are all labeled *False* while 13 of the 20 least discriminating examples are labeled *True*. Table 2 shows the hardest and the easiest examples of MNLI and MC-TACO.

5 Related Work

Prior work on using IRT to evaluate NLP systems mostly relies on human responses. Hopkins and May (2013) use IRT to estimate the relative ability of a set of machine translation systems using responses from pairwise comparison of system outputs by human judges. Otani et al. (2016) extend this work by including a baseline translation to the pairwise comparison. Lalor et al. (2016, 2018) use IRT to identify hard examples in natural language inference data based on human responses. In a follow-up study, Lalor et al. (2019) compare human versus model responses and find that both are positively correlated and demonstrate the use cases of IRT parameters in training set filtering. Sedoc and Ungar (2020) use IRT to evaluate chatbot systems.

The work by Martínez-Plumed et al. (2019) is the first to study the idea of using model responses (as opposed to human responses) for IRT in machine learning research. For NLU, Lalor and Yu

(2020) use model responses to estimate difficulty parameters of several GLUE datasets for dynamic data selection in curriculum learning. In concurrent work, Rodriguez et al. (2021) study how IRT can be used for more nuanced leaderboard evaluations. Their experiments demonstrate that IRT can produce a more reliable ranking of models than the traditional metrics. They also show that IRT is not only useful for better understanding of individual examples in the dataset and task, but also effective in identifying annotation errors.

For other dataset evaluations, in addition to providing a benchmark, the SuperGLUE paper also compares a set of candidate datasets using a fixed pool of machine learning models and human annotators (Nangia and Bowman, 2019). Wang et al. (2019a) investigate pretraining tasks and paradigms for effective transfer learning methods. Pruksachatkun et al. (2020a) study when and why intermediate-task training is useful for a given target task. Vu et al. (2020) introduce task embeddings to predict the most beneficial source task for a given target task. Schlegel et al. (2020) propose an evaluation framework for machine reading comprehension (MRC) datasets and reveal some concerns regarding factual correctness and the presence of linguistic cues in existing MRC gold datasets.

6 Conclusion

Given the large number of NLU datasets introduced in recent years, what kinds of datasets are effective to measure near-future progress? Our analysis on 29 test sets using IRT gives us reason to believe that, among the datasets we evaluate, Quoref, HellaSwag, and MC-TACO are best able to discriminate among current (and likely future) strong models. Meanwhile, SNLI, MNLI, and CommitmentBank seem to be saturated and ineffective for measuring future progress.

Our analysis of examples' difficulty and discrimination parameters shows that datasets with many hard examples do not always contain examples that can discriminate between strong and weak models. We find that QA datasets are more difficult than other datasets. We also find span selection as the most effective task format for discriminating between strong and weak models.

According to our LEH score, datasets that seem to be solved are unlikely to see improvements with future pretrained models. Therefore, the skills they intend to test are either largely solved, to the extent that they are solvable, or not well isolated (e.g., due to data artifacts). Focusing on the skills for which these solved test sets are originally designed to evaluate would most likely require a new dataset that better isolates the reasoning ability of interest.

On the other hand, datasets that perform well according to our LEH metric show the best signs of being amenable to future hill-climbing. This *does not* entail that we should focus future research on these benchmarks, since we do not evaluate whether they test the skills they mean to test, or whether these skills are important for scientific or practical progress on natural language understanding. Finally, we argue that this evaluation should be done periodically, as datasets and models improve over time.

For future work, one can study multi-dimensional variables for both model ability and item parameters, which could reveal a factorization of datasets by skills. Other potential directions include expanding our analysis to a broader range of tasks and analyzing the relationship between the estimated IRT parameters and the human-model gap.

Acknowledgments

We thank John Lalor, João Sedoc, Nikita Nangia, Sebastian Schuster, Iacer Calixto, and the anonymous reviewers for feedback. This work has benefited from financial support to SB by Eric and Wendy Schmidt (made by recommendation of the Schmidt Futures program), Samsung Research (under the project *Improving Deep Learning using Latent Structure*), Apple, and Intuit, and from in-kind support by the NYU High-Performance Computing Center and by NVIDIA Corporation (with the donation of a Titan V GPU). This material is based upon work supported by the National Science Foundation under Grant No. 1922658. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Ethical Considerations

We present an objective approach for comparing the difficulty of test sets examples across datasets and demonstrate it on a large set of established datasets. We expect this to contribute to the development of more challenging benchmarks for NLP datasets and potentially to develop more challenging mod-

els. One concern worth noting is that most of the evaluation datasets we study are crowdsourced or drawn from naturally occurring data. Thus, they likely demonstrate harmful stereotypes to some degree or even score models more highly for demonstrating them. In general, models that perform well on these datasets should not be deployed directly without additional measures to measure and eliminate any harms that stereotypes like these could cause in the target application settings.

References

Frank B. Baker and Seock-Ho Kim. 1993. Item response theory : parameter estimation techniques. *Journal of the American Statistical Association*, 88:707.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.*, 20:28:1–28:6.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, Lisbon, Portugal. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5925–5932, Hong Kong, China. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The CommitmentBank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. 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 1930–1940, New Orleans, Louisiana. Association for Computational Linguistics.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

John P. Lalor and Hong Yu. 2020. Dynamic data selection for curriculum learning via ability estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 545–555, Online. Association for Computational Linguistics.

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. In *8th International Conference on Learning Representations, ICLR 2020*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

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. Unpublished manuscript available on arXiv.

Fernando Martínez-Plumed, Ricardo B.C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18 – 42.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. 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 560–568, New Orleans, Louisiana. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. 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, Online. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, Austin, Texas. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020a. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020b. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. 2014. Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, volume 33 of *JMLR Workshop and Conference Proceedings*, pages 814–822. JMLR.org.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Boyd-Graber Jordan. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, Berlin Chen, Benjamin Van Durme, Edouard Grave, Ellie Pavlick, and Samuel R. Bowman. 2019a. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476, Florence, Italy. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019b. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

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, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021a. Revisiting few-sample BERT fine-tuning. In *International Conference on Learning Representations*.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2021b. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Discrimination vs. Guessing

In addition to the analysis of discrimination versus difficulty parameters, we also look at the distribution of the guessing (γ) parameters. From Figure 6, we observe that all QA datasets with span selection format generally have low guessing parameters, meaning that they are difficult to predict correctly by random guessing. This makes sense as span selection has higher label entropy than classification or multiple-choice task. We find that several datasets have examples with varying guessing parameters: For SNLI we see a high density of examples that can be predicted easily by random guessing while for MNLI, HellaSwag, and MCScript, there are more examples with low guessing parameters.

B Additional Analysis on Model Ability

Figure 7 plots model abilities θ against their average accuracy over all test examples, where each point represents a model checkpoint (Section 3.2). We use different colors for different models (e.g., dark blue for ALBERT-XXL-v2), and different shapes to mark different checkpoints.

Since we only perform tuning on $\text{RoBERTa}_{\text{Large}}$, some of these models might have worse performance than if they were individually tuned.

C Task Descriptions

In this section, we provide a short description for each dataset.

RTE The series of Recognizing Textual Entailment datasets (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009) correspond to a two-class textual entailment classification task. Given a premise sentence and a hypothesis sentence, the task is to decide whether the premise entails the hypothesis.

SNLI The Stanford Natural Language Inference corpus (Bowman et al., 2015) is a textual entailment dataset, formulated as a three-class classification task. Given a premise sentence and a hypothesis sentence, the task is to determine if the premise entails the hypothesis, contradicts it, or neither. The SNLI dataset is created using premises taken from image captions.

MNLI The Multi-Genre Natural Language Inference corpus (Williams et al., 2018) is also a textual entailment dataset, similar to that of SNLI. The MNLI dataset is built to cover a broad range of genres, including written and spoken text. Half of its test set is created from text that is out of domain relative to the training set.

CommitmentBank CommitmentBank (De Marneffe et al., 2019) is a dataset formulated as a three-class textual entailment classification task. Given a piece of text and an embedded clause, models must decide whether the embedded clause is entailed by the text.

ARCT The Argument Reasoning Comprehension Task (Habernal et al., 2018) is a multiple-choice question answering dataset. Given an argument, a claim, and a premise, the task is to select the correct implicit warrant (which explains why the premise implies the claim) from two choices.

ARC-Easy ARC (Clark et al., 2018) is a multiple-choice QA dataset composed of real multiple-choice science questions in grade schools. ARC-Easy is composed of the easier questions that do not satisfy the criteria used to built ARC-Challenge (described below).

ARC-Challenge ARC-Challenge (Clark et al., 2018) is the subset of ARC that contains questions that are incorrectly answered by both a retrieval-based algorithm and a word co-occurrence algorithm.

MCScript The MCScript (Ostermann et al., 2018) is a QA dataset with multiple-choice format. The dataset tests models' commonsense knowledge, in particular, script knowledge which corresponds to the sequence of actions people do in a particular situation.

Cosmos QA Cosmos QA (Huang et al., 2019) is a multiple-choice reading comprehension dataset, and it is intended to require extensive abstractive commonsense reasoning. Unlike CommonsenseQA, Cosmos QA requires comprehension over an auxiliary article, instead of simply responding to a free-standing question.

HellaSwag HellaSwag (Zellers et al., 2019) is a commonsense reasoning multiple-choice dataset. It is built using adversarial filtering with BERT. Given a story, the task is to select the most plausible continuation.

BoolQ BoolQ (Clark et al., 2019) is a boolean (yes/no) reading comprehension QA dataset built

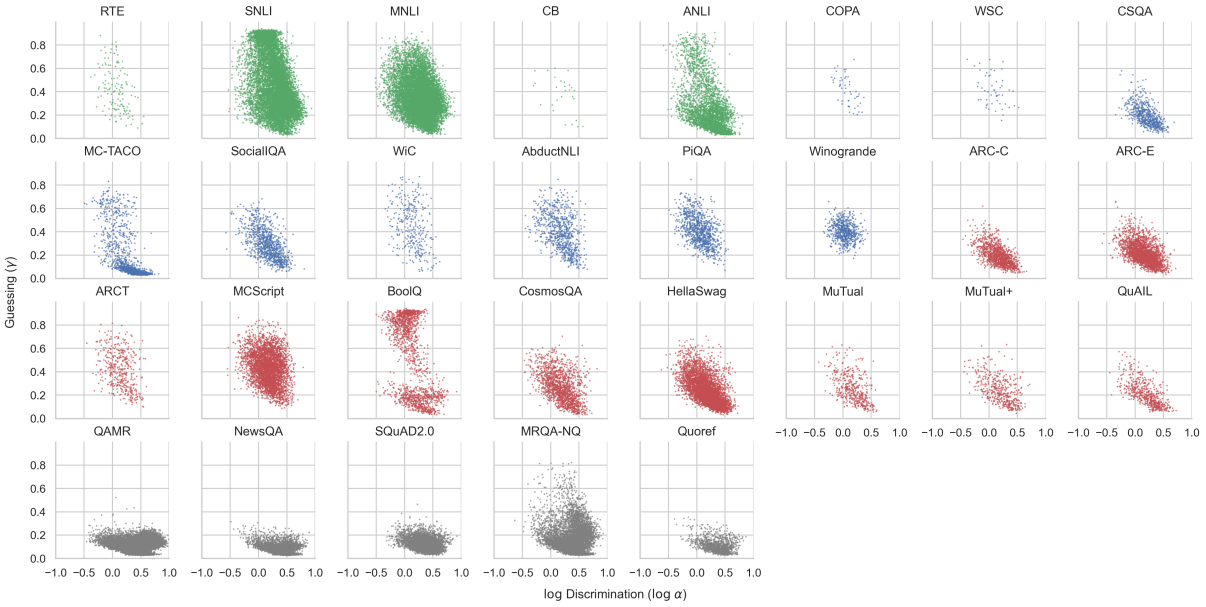

Figure 6: Plots of the log discrimination ($\log \alpha$) versus the guessing (γ) parameters for each dataset.

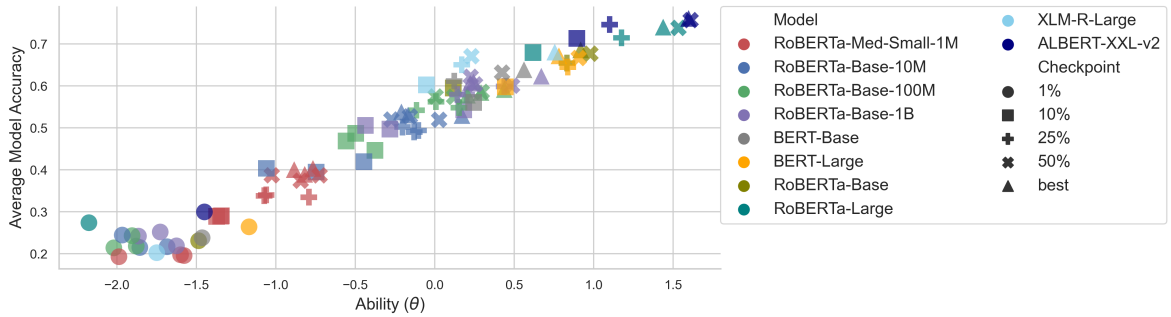

Figure 7: Average model accuracy over all datasets vs. ability (θ). The three different hyperparameter configurations of each MiniBERTa are represented by a single color for ease of readability. Best viewed in color.

using the same pipeline used to produce the (non-boolean) Natural Questions (Kwiatkowski et al., 2019).

MuTual MuTual (Cui et al., 2020) is a multiple-choice QA dataset for multi-turn dialogue reasoning. The dataset is created from Chinese students' English listening comprehension exams, and it is intended to require a variety of commonsense reasoning skills.

MuTual-Plus MuTual-Plus (Cui et al., 2020) is a variant of MuTual, in which one of the choices in each set of answers is replaced by a safe response (i.e., "could you repeat that"). If all other choices are incorrect, then the model is supposed to select the safe response. This variant of MuTual is built so that we can evaluate if the model can select the safe response when all other options are incorrect.

QuAIL QuAIL (Rogers et al., 2020) is a reading comprehension dataset formulated as a multiple choice task. One feature of QuAIL is that it combines "commonsense, text-based, and unanswerable questions." It is also designed such that it has a balanced distribution of genres and reasoning types.

COPA Choice of Plausible Alternatives (Roemmele et al., 2011) is a dataset for sentence-level multiple-choice task. Given a premise and a question that asks for the cause or effect of the premise, the task is to choose the most plausible hypothesis from two options.

WSC The Winograd Schema Challenge (Levesque et al., 2012) is a sentence-level multiple-choice commonsense reasoning dataset. Given a piece of text, a pronoun, and a list of possible noun phrases, the model must choose the correct

referent to the pronoun. The dataset is designed such that world knowledge is required to make the correct choices. We use the SuperGLUE (Wang et al., 2019b) version of the dataset.

CommonsenseQA CommonsenseQA (Talmor et al., 2019) is a multiple-choice QA dataset which is designed to test a range of commonsense knowledge.

SocialIQA SocialIQA (Sap et al., 2019) is a dataset that is specifically designed to test a models' capabilities related to emotional and social intelligence in everyday situations.

MC-TACO MC-TACO (Zhou et al., 2019) is a multiple-choice QA dataset that is designed to test temporal commonsense reasoning, in particular: duration, temporal ordering, typical time, frequency, and stationarity. Each question consists of a varying number of choices, and for each answer choice, a model needs to predict whether the answer is correct or incorrect.

WiC The Word-in-Context (Pilehvar and Camacho-Collados, 2019) dataset which is designed to test the word sense disambiguation skill of a model. Given two pieces of text (a phrase or a sentence) with a polysemous word in both, a model needs to predict whether the two words are used in the same sense.

PIQA The Physical Interaction Question Answering dataset (Bisk et al., 2020) is a multiple-choice QA dataset that is designed to test the physical commonsense reasoning skill. Given a physical task expressed in text, a model needs to select the most sensible solution.

WinoGrande The WinoGrande dataset (Sakaguchi et al., 2020) is built through a crowdsourcing procedure that incorporates adversarial filtering. Given a sentence with a blank (where the blank corresponds to a noun phrase), the task is to select the correct filler. The dataset is designed to test the commonsense reasoning skill.

Abductive NLI The Abductive Natural Language Inference dataset (Bhagavatula et al., 2020) is a multiple-choice dataset. Given a premise, the task is to select the most likely explanation from the given hypotheses.

QAMR The Question-Answer Meaning Representations (Michael et al., 2018) is a QA dataset where the question-answer pairs are created from sentences' predicate-argument relationships.

NewsQA NewsQA (Trischler et al., 2017) is a QA dataset formulated as span selection task. The dataset is built by crowdworkers using passages taken from CNN news articles.

SQuAD2.0 SQuAD2.0 (Rajpurkar et al., 2018) is a QA dataset that combines the span-selection reading-comprehension questions in SQuAD 1.1 (Rajpurkar et al., 2016) with over 50,000 unanswerable questions. The unanswerable questions were written by crowdworkers to look like the answerable ones. A model must either select an answer span or decline to answer.

Quoref Quoref (Dasigi et al., 2019) is a QA dataset that is designed to test coreferential reasoning ability. The dataset is formulated as a span selection QA task.

MRQA Natural Questions The Natural Questions dataset (Kwiatkowski et al., 2019) is a dataset designed to test a model's ability in reading comprehension. The questions are taken from real-word queries, while the context passages are taken from Wikipedia articles. We use the MRQA version of it which contains a preprocessed version of a subset of questions in Natural Questions.

ANLI The Adversarial Natural Language Inference dataset (Nie et al., 2020) is a textual entailment dataset built using an iterative human-and-model-in-the-loop procedure in order to find hard examples.

Dataset	Best	ALBERT	RL	RB	XLM-R	BL	BB	RB- 100M-1	RB-100M-2	RB-100M-3	RB-10M-1	RB-10M-2	RB-10M-3	RB-1B-1	RB-1B-2	RB-1B-3	RMS-1M-1	RMS-1M-2	RMS-1M-3
RTE	86.6	81.2	87.6	80.4	57.2	81.9	73.9	60.9	66.7	66.7	58.7	59.4	57.2	67.4	64.5	71.7	60.9	60.9	58.0
SNLI	92.6	92.4	92.7	91.7	91.7	90.5	90.6	87.6	88.8	87.5	85.1	86.1	85.7	88.7	88.4	89.2	79.1	78.5	79.3
MNLI-m	90.2	88.3	89.8	86.5	87.5	85.3	80.4	75.5	76.6	74.6	69.3	70.6	68.6	77.8	77.9	79.8	61.3	62.6	60.9
MNLI-mm	90.2	88.5	89.5	86.2	87.8	85.0	81.1	77.1	77.4	75.5	70.6	71.2	70.8	79.7	78.9	80.8	62.6	62.3	62.3
CB	90.5	80.6	90.5	88.0	69.9	84.6	78.7	63.7	77.3	80.5	63.5	63.6	77.3	83.5	83.2	84.8	63.8	60.5	57.7
ANLI-R1	73.8	75.9	66.6	56.1	60.1	58.5	52.6	47.7	48.1	46.6	45.6	48.6	47.4	49.5	47.2	51.1	37.8	39.8	39.7
ANLI-R2	48.9	57.7	44.6	41.5	41.5	42.9	44.5	38.8	40.9	39.1	39.7	42.0	40.5	40.4	40.2	41.3	36.1	36.6	36.4
ANLI-R3	44.4	54.4	41.3	40.8	42.0	43.8	42.1	38.3	40.8	40.7	38.0	38.2	38.3	40.7	41.0	41.7	31.8	34.2	33.5
COPA	79.1	96.0	86.0	72.0	62.0	80.0	68.0	66.0	74.0	70.0	68.0	58.0	70.0	74.0	72.0	72.0	58.0	54.0	52.0
WSC	89.0	78.8	78.8	76.9	61.5	65.4	57.7	61.5	61.5	59.6	55.8	59.6	50.0	59.6	55.8	61.5	59.6	51.9	67.3
CommonsenseQA	72.1	80.5	74.6	58.5	23.8	60.8	57.4	33.8	36.1	32.5	26.1	26.1	23.3	41.3	38.7	43.0	21.6	22.5	25.1
MC-TACO	44.0	55.9	55.9	48.6	47.7	43.2	38.7	37.8	40.5	37.8	35.1	34.2	28.8	37.8	41.4	40.5	20.7	27.0	24.3
SocialIQA	78.5	79.4	79.9	70.5	38.9	66.1	59.9	54.4	55.2	56.0	51.8	49.9	50.8	58.5	56.9	55.6	43.5	46.1	44.1
WiC	70.5	74.3	71.5	71.2	71.5	69.6	68.7	63.3	64.9	68.0	61.4	63.3	62.4	65.8	66.5	68.3	60.2	59.2	59.9
Abductive NLI	83.9	83.8	85.0	72.2	76.4	66.4	62.4	56.9	57.4	56.7	55.0	54.8	55.0	57.7	56.1	60.3	52.9	53.3	53.9
PiQA	77.1	80.6	77.6	68.2	53.5	65.7	59.7	60.7	61.7	61.5	60.6	60.4	60.0	62.9	60.3	62.2	55.0	55.9	55.9
Winogrande	79.3	85.6	77.7	64.6	52.6	52.9	52.0	50.9	54.5	52.6	51.2	51.8	51.2	53.7	55.6	52.8	47.6	50.9	48.7
ARC-C	–	47.5	37.5	31.8	38.1	39.8	31.8	31.8	30.4	29.8	30.8	30.4	30.4	29.4	30.8	32.8	27.8	27.1	29.8
ARC-E	66.0	69.3	62.5	53.2	40.9	61.4	56.8	39.1	41.6	40.0	37.0	36.0	34.0	41.8	43.0	46.5	31.8	31.9	29.8
ARCT	70.1	83.5	86.7	76.3	76.6	74.4	72.2	66.5	68.0	66.5	64.6	66.1	66.5	70.3	68.7	70.3	66.1	64.6	64.6
MCScript	90.0	96.0	92.8	85.0	90.1	84.1	80.7	74.1	74.7	73.5	68.9	70.5	69.7	73.7	76.2	77.8	60.6	61.7	60.1
BoolQ	87.1	87.3	85.7	82.2	84.2	76.1	75.7	74.3	73.5	72.5	69.5	72.7	71.9	73.6	74.3	74.1	67.3	68.6	66.9
CosmosQA	81.9	86.5	79.4	66.8	71.7	64.9	56.0	54.7	53.4	51.5	44.0	46.8	48.9	55.9	54.3	54.0	39.9	42.0	40.6
HellaSwag	85.2	89.8	84.1	61.6	74.5	43.9	37.2	36.9	35.4	34.7	33.8	33.6	33.0	37.5	36.4	36.5	29.9	30.0	29.4
Mutual	71.3	89.8	87.8	73.3	26.9	72.5	65.5	56.2	58.7	57.8	50.8	52.8	53.3	63.9	59.1	62.5	43.3	41.3	42.4
Mutual+	62.6	82.8	77.9	63.9	65.5	65.9	54.2	51.9	49.7	50.6	47.2	43.8	45.8	54.9	53.5	53.5	35.9	40.4	40.2
QuAIL	47.9	78.0	73.3	67.0	63.9	52.9	54.4	52.2	53.6	53.6	47.8	50.0	48.5	54.2	54.1	57.3	40.7	38.7	41.8
QAMR	79.1	79.6	79.6	77.6	79.3	76.4	73.5	70.1	71.8	70.4	63.4	64.4	62.6	73.7	73.1	74.6	24.0	28.7	27.8
NewsQA	–	59.6	57.8	53.1	56.2	53.5	48.0	36.2	38.5	34.7	29.1	29.6	27.5	40.4	42.4	45.9	6.5	8.9	9.7
SQuAD2.0	86.8	89.9	91.5	88.6	87.0	88.6	86.3	82.0	82.9	80.7	75.8	77.2	75.8	82.7	83.5	84.8	57.1	58.2	58.2
MRQA-NQ	57.4	71.5	69.9	65.0	66.9	67.3	66.1	59.0	58.9	57.5	52.9	53.0	52.8	62.1	60.6	61.5	41.0	43.1	42.4
Quoref	74.9	83.7	78.7	66.7	76.3	69.8	62.9	48.0	46.7	43.9	34.9	35.6	34.5	51.0	50.1	56.0	14.6	21.6	16.8

Table 3: Results of all our models on each validation set. **RL**: RoBERTa$_{\text{Large}}$, **RB**: RoBERTa$_{\text{Base}}$, **BL**:BERT$_{\text{Large}}$, **BB**:BERT$_{\text{Base}}$, **RMS**: RoBERTa-Med-Small. **Best** denotes best known performance on the original dataset's validation set.