How Reliable are Model Diagnostics?

Vamsi Aribandi[*]
Google Research
aribandi@google.com

Yi Tay
Google Research
yitay@google.com

Donald Metzler
Google Research
metzler@google.com

Abstract

In the pursuit of a deeper understanding of a model's behaviour, there is recent impetus for developing suites of probes aimed at diagnosing models beyond simple metrics like accuracy or BLEU. This paper takes a step back and asks an important and timely question: how *reliable* are these diagnostics in providing insight into models and training setups? We critically examine three recent diagnostic tests for pre-trained language models, and find that likelihood-based and representation-based model diagnostics are not yet as reliable as previously assumed. Based on our empirical findings, we also formulate recommendations for practitioners and researchers.

1 Introduction

Contemporary statistical models based on deep learning have made incredible progress towards solving complex language tasks (Radford et al., 2019; Devlin et al., 2019; Raffel et al., 2020). These models generally trade off the interpretability and simplicity of traditional models for powerful parameterizations and inductive biases, enabling their impressive performance. However, their entry into critical fields such as medicine, the justice system, and social media moderation often makes this trade-off a costly one. Consequently, there has been surging interest in the development of tools and suites for diagnosing and better understanding model behaviour, and gaining insight into what patterns and phenomena they have learned (§4.1).

Ideally, these diagnostics would not only help practitioners understand the failure modes and capabilities of large contemporary models, but also enable them to improve their models based on the diagnostics. To this end, we believe that model diagnostics are essential for making meaningful progress in natural language processing.

[*]Google AI Resident

Model diagnostics generally probe a model for specific learned qualities (§4.1). These may be a positive qualities (e.g., whether a model has acquired syntactic knowledge) or potentially problematic qualities (e.g., biases and stereotypes. These probes can be used to identify certain phenomena that can be used to further improve models.

Given the potential impact that model diagnostics can have for practitioners and the research community's fundamental understanding of contemporary models, this paper asks the important and inevitable question of whether these probes are actually reliable and robust, and to what extent they are. These diagnostics' explicit nature as a tool for understanding also imposes a greater bar for robustness, as inconsistencies may mislead and result in compounding errors.

Our findings demonstrate that model diagnostics can be unreliable on multiple fronts. To illustrate our point, we select three diagnostics tasks — StereoSet (Nadeem et al., 2020), CrowS-Pairs (Nangia et al., 2020), and SEATs (May et al., 2019) to base our empirical evaluation on. Overall, we find that **likelihood-based and representation-based diagnostics measured multiple times on the same training setup can result in wildly different findings.** Specifically, a substantial variance is observed when performing the same model diagnostics on identical BERT (Devlin et al., 2019) pre-training setups while varying minute details such as the initial random seed or choice of representation.

These findings are meant to caution researchers and practitioners that rely on such diagnostics so that they can be more mindful of these phenomena when analyzing their models in the future. We discuss the implications of our findings and propose recommendations for practitioners and researchers in §5.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1778–1785
August 1–6, 2021.

2 Methodology

2.1 Training setup

We pre-train 5 BERT BASE and LARGE uncased English models, each with the same configurations as in Devlin et al. (2019) using Tensorflow[1]. However, each model differs in its random seed, resulting in different parameter initializations and training data permutations. Hence, it is expected that the checkpoints will each end up at a different local minima. It should be noted that BERT uses static masking instead of dynamic masking, so the set of pre-training examples remains the same.

To decouple our findings from phenomena that occur as a result of using different training setups, we restrict our experiments to only those that require pre-trained BERT models, eliminating many probes mentioned in §4.3. Webster et al. (2020) report that patterns learned during pre-training are often resilient to fine-tuning, further supporting our reasoning.

2.2 Likelihood-base diagnostics

One approach to examining the behaviour of language models like BERT is to examine how they rank certain representative examples above others. We use two contemporary datasets that measure how often stereotypes are ranked above anti-stereotypes — StereoSet (Nadeem et al., 2020) and CrowS-Pairs (Nangia et al., 2020). Both datasets measure $ss = 100 * \sum_{n=1}^{|X|} 1_{[ll(x_n^{ster}) > ll(x_n^{anti})]} / |X|$.

StereoSet Nadeem et al. (2020) propose a benchmark that contains intra-sentence and inter-sentence examples of stereotypes and anti-stereotypes. Here, likelihoods are calculated as $ll(x) = p(x_\tau | x_{\setminus\tau})$ (where τ is the set of target demographic word(s) in x) and $ll(x) = p(\texttt{isNext} | x_1, x_2)$ for intra-sentence and inter-sentence examples respectively. They also propose and combine a language modeling score (lms) with ss into a hybrid metric ($icat$), but we only report ss to focus on StereoSet's primary purpose — measuring stereotypical preference in language models. We report results on the development set.

CrowS-Pairs Nangia et al., 2020 propose a test that contains intra-sentence examples, where likelihoods are calculated by conditioning on the target demographic word(s) in the sentence ($ll(x) = p(x_{\setminus\tau} | x_\tau)$) rather than vice-versa as in StereoSet.

The CrowS-Pairs diagnostic is expected to show higher variance than StereoSet for two reasons: (1) it is a smaller dataset ($\sim\frac{1}{3}$rd the size of StereoSet-dev) with more categories, so results are more sensitive to changes in individual predictions; and (2) the pseudo-likelihood it uses is more susceptible to the poor calibration (Jiang et al., 2020a; Desai and Durrett, 2020) of contemporary models, since the number of multiplied probabilities grows linearly with the number of words in a sentence.

2.3 Vector-space diagnostics

Directly examining representations learned by models is another way to understand their behavior. This is typically done by measuring relationships between different types of inputs, for example in terms of their relative orientations in a vector space.

SEATs We use Sentence Encoder Association Tests (SEATs; May et al., 2019), which extend the popular Word Embedding Association Tests (WEATs; Caliskan et al., 2017) by constructing "semantically bleached" sentences. A WEAT/SEAT measures the *effect size* $s(X, Y, A, B)$ of the association between two targets (e.g., X=`MentalDisease` and Y=`PhysicalDisease`) and two attributes (e.g., A=`Temporary` and B=`Permanent`), as well as the statistical significance of the association using a permutation test[2]. We conduct experiments using the same SEATs as in May et al. (2019). In addition to testing sentence (`[CLS]`) representations, we also test the contextualized word representations of the target/attribute words in the sentences. The reason we do this is that even for semantically bleached sentences, it is often non-trivial for models to encode information about an entire sentence in a single vector[3].

In addition to examining effect sizes, we also conduct an experiment to see how distinguishable representations of certain concepts are in vector space (e.g., do representations of `Pleasant` and `Unpleasant` sentences form their own clusters?). We do this by clustering (via k-means) sentence representations and subsequently examining how well the unsupervised clusters align with the actual categories. The aim of this experiment is to understand vector space diagnostics behave the way they do.

[1] https://github.com/tensorflow/models/tree/master/official/nlp/bert

[2] Please see Appendix A for how SEATs are computed.

[3] https://www.cs.utexas.edu/~mooney/cramming.html

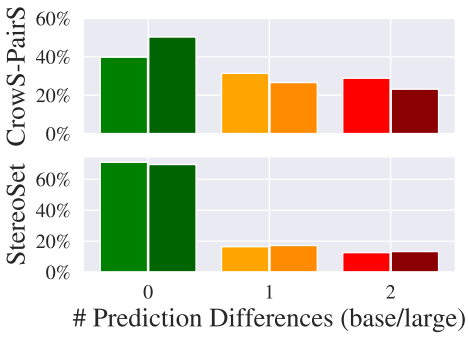

Figure 1: % of examples in likelihood-based tests that have d different predictions over 5 runs. Ideally, examples would always (100%) be predicted the same (d=0).

Test	Cat.	N	BERT results (%)	
			BASE	LARGE
CrowS-Pairs	Race	516	$54.4 \pm$ **4.7**	$55.9 \pm$ **2.7**
	Gen.	262	58.2 ± 2.5	61.1 ± 1.7
	S.O.	84	63.2 ± 3.4	$67.4 \pm$ **4.6**
	Rel.	105	$68.9 \pm$ **8.0**	72.2 ± 2.1
	Age	87	$55.4 \pm$ **4.2**	$60.9 \pm$ **5.6**
	Nat.	159	51.2 ± 1.2	$55.3 \pm$ **3.5**
	Dis.	60	$69.0 \pm$ **3.8**	79.0 ± 1.9
	P.A.	63	$59.1 \pm$ **4.9**	$64.4 \pm$ **4.3**
	Occ.	172	$54.9 \pm$ **4.5**	$58.0 \pm$ **4.2**
	all	1508	$57.1 \pm$ **2.8**	60.3 ± 1.7
StereoSet	Gen.	496	59.1 ± 0.7	62.4 ± 2.0
	Occ.	1636	60.5 ± 0.6	61.4 ± 0.8
	Race	1938	54.8 ± 1.1	56.4 ± 0.8
	Rel.	156	$51.8 \pm$ **2.8**	$54.4 \pm$ **3.3**
	all	4226	57.4 ± 0.7	59.0 ± 0.7

Table 1: Likelihood-based diagnostics over categories often have high standard deviation (bold) over pre-training runs, often varying from almost neutral ($\sim$50%) to a significant amount (highlighted).

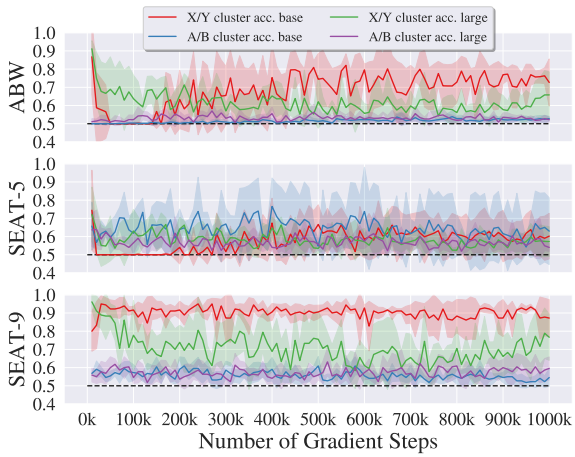

Figure 2: Representations of targets and/or attributes often don't cluster over pre-training. The dashed line is when representations are indistinguishable (acc. = 0.5).

3 Findings and Insights

3.1 Likelihood-based diagnostics are unstable

Experiments on StereoSet and CrowS-Pairs show that while likelihood-based ranking diagnostics may be stable across all categories, instability is evident in the results of individual categories (Table 1). Many categories have a standard deviation of over 2.5 percentage points. Some categories also vary from almost no stereotypical preference to a significant amount (highlighted in Table 1) — a result that could potentially cause practitioners to draw false conclusions.

Additionally, from Figure 1 it is evident that many examples are assigned different labels over the 5 pre-trained models, often having 3 models assign them one label and 2 models assigning them the opposite label — almost as random as a coin flip! The implies that the models are probably uncertain about their predictions for these datapoints, motivating the consideration of model uncertainty in diagnostic measures instead of simply making a binary decision by comparing likelihoods.

Worryingly, both tests report wildly differing results on religious stereotypes ("Rel."), with CrowS-Pairs detecting strong stereotypical preference and StereoSet detecting almost none. It is also worth noting that results on CrowS-Pairs exhibit far higher variance compared to StereoSet (Table 1, Figure 1), as hypothesized in §2.2.

3.2 Vector-space diagnostics are unstable

Representation-based experiments exhibit high variance across multiple pre-training runs, choices of representation, and model sizes (Figure 3). Notably, SEAT results are often on both sides of the "neutral" mark (0), and their statistical significance is often erratic. In other words, it is possible for two models to be pre-trained with the exact same configurations but different random seeds to yield completely opposite conclusions on some SEATs. Moreover, the same checkpoint often yields different results depending on whether sentence or pooled target-word representations are used. Ideally, a SEAT would *always* or *never* be statistically significant, and yield effect sizes with the same sign over multiple pre-training runs and (seemingly innocuous) choices of representation.

From Figure 2, the representational instability of semantically bleached SEAT sentences is further evident — how these representations cluster

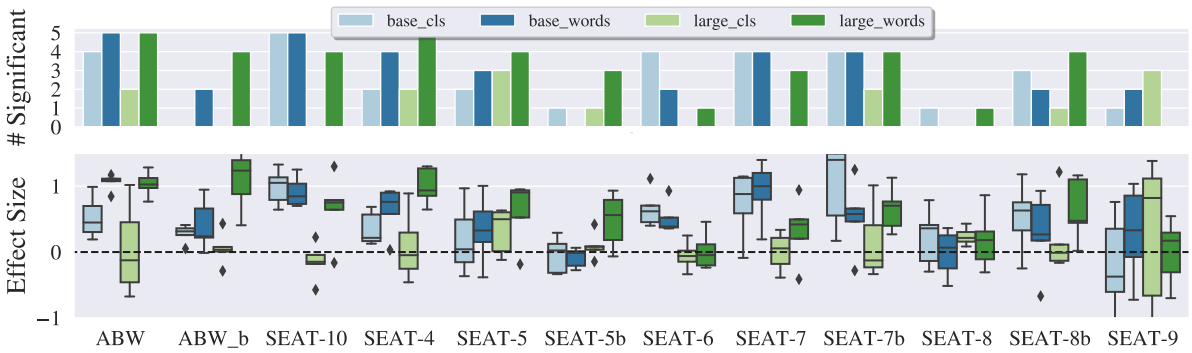

Figure 3: SEAT results exhibit high variance across pre-train runs, model sizes, and choice of representation. Moreover, effect sizes often vary around the "neutral" mark (0) and also have different statistical significances (at $p = 0.01$). Ideally, a test would always (5) or never (0) be significant, and yield effect sizes with the same sign.

together is erratic both across pre-training steps as well as across multiple pre-training runs. This result gives us further insight into why high variance is observed for vector-space diagnostics — representations often can't form their own clusters for certain concepts, so simply examining their relative orientations is insufficient. Our findings provide empirical arguments for what May et al. (2019) surmise — there is scope for sentence embedding-based tests that do more than naturally extend word embedding-based tests with semantically bleached sentences.

We surmise that representation-based diagnostics are less stable than likelihood-based diagnostics because large models like BERT are optimized to be good at modeling likelihoods via their pre-training objective. However, there is no constraint on how sentences must be represented other than it should be possible to "extract" correct likelihoods from them. In other words, there is no reason to expect the orientations of these representations to provide deep insight into what these models learn.

3.3 Diagnostic instability is despite equivalent downstream performance

We fine-tuned the 10 checkpoints on SST-2 (Socher et al., 2013), RTE (Dagan et al., 2006; Bar Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and QNLI (Rajpurkar et al., 2016) from the GLUE benchmark (Wang et al., 2019). Development-split results show that performance was largely the same across checkpoints (Table 2) despite diverging behaviour on the model diagnostics as shown in §3.1 and §3.2. **This shows that the different local optima still perform largely the same on downstream tasks despite behaving differently with respect to model diagnostics**.

Dataset	BERT fine-tuning results	
	BASE	LARGE
SST-2	91.2 ± 0.3	93.0 ± 0.3
RTE	71.3 ± 1.2	76.8 ± 1.8
QNLI	92.1 ± 0.2	92.1 ± 0.3

Table 2: The checkpoints generally exhibit equivalent performance on downstream tasks.

Dev-set performance is also largely consistent with what is expected of BERT BASE and LARGE models. It should be noted that we only used one set of hyperparameters and did not perform the hyperparameter sweep as in Devlin et al. (2019), so further tuning would likely improve results.

4 Related Work

4.1 Model Diagnostics

Models have been probed to understand what exactly they learn beyond traditional language tasks, ranging from their linguistic capabilities (Adi et al., 2017; Tenney et al., 2019; Conneau et al., 2018; Ribeiro et al., 2020; Belinkov et al., 2017; Hewitt and Manning, 2019; Marvin and Linzen, 2018), multilingual capabilities (Pires et al., 2019; Kudugunta et al., 2019), world knowledge (Jiang et al., 2020b; Petroni et al., 2019), and social bias (Nadeem et al., 2020; Nangia et al., 2020; May et al., 2019) among other phenomena.

Another axis to compare model diagnostics on is whether they are intrinsic or extrinsic, i.e., whether they directly analyze models for certain phenomena that aren't tied to any downstream task or do so keeping particular tasks in mind. This paper restricts itself to intrinsic tasks for reasons mentioned in §2.1. An example of an extrinsic task is

Rudinger et al. (2018), which probes models for gender bias through the lens of coreference resolution. We refer readers to Belinkov and Glass (2019) for a more comprehensive survey on model analysis for natural language processing.

4.2 Diagnostic Fragility

It has been shown that classifier probes — which require an additional classifier (like an MLP) to be trained on top of frozen model representations — are unstable (Voita and Titov, 2020), and that it might not be clear from their results whether the probe *itself* learned a phenomena or whether the diagnosed representations learned it (Hewitt and Liang, 2019). Similarly, Wang et al. (2020) find that gradient-based analysis of language technologies based on neural networks can often be unreliable and manipulable. Attention-based interpretation can also be unreliable and manipulable to the point of deceiving practitioners, as Pruthi et al. (2020) and Jain and Wallace (2019) show. The works mentioned above all support our arguments, and some raise similar concerns to those expressed in this paper.

4.3 Inconsistencies between equivalent checkpoints

This paper's findings can be linked to the problems caused by underspecification in machine learning (D'Amour et al., 2020), i.e., when multiple unique predictors trained with the same configuration have the same performance but differ in subtle ways. In a setting where practitioners might train and thoroughly analyze one model but then retrain it and assume that the first checkpoint's model diagnostics hold for the second one, this issue is highly relevant. McCoy et al. (2020) also find that separately fine-tuned BERT models often vary significantly in generalizing to auxiliary tasks.

5 Discussion

Recommendations No probe is perfect, but it is clear that model diagnostics are not as reliable as previously assumed. Our empirical findings — coupled with the works mentioned in §4.2 and §4.3 — motivate careful scrutiny of model diagnostics. **We recommend that:**

- Practitioners not generalize a single diagnostic result to the entire training setup, and instead restrict conclusions to a specific checkpoint.
- Researchers proposing probes not only test on publicly available checkpoints, but rather examine a probe's performance and robustness across a range of model/probe configurations.

Future Work While this paper primarily aims to motivate further scrutiny of model diagnostics, we hope it motivates studies that ask *why* these diagnostics often behave unreliably. One future research direction we are excited about is analyzing correlations between the properties of the models' local minima in the loss landscape and behaviour on model diagnostics. This would not only be another step towards a better understanding of how contemporary deep language models work, but also enable researchers to use that information to design better, more robust model diagnostics. Such a study may even help inform the optimization process for future state-of-the-art language technologies.

It should also be noted that this paper is restricted to three diagnostics spanning likelihood-based and representation-based probes, and that future work is needed to determine the extent to which other diagnostic probes are reliable.

6 Conclusion

In this paper, we motivate further scrutiny of model diagnostics that aim to understand the behaviour of contemporary "black-box" language technologies. Our results show that model diagnostics are often fragile and can yield different conclusions as a result of seemingly innocuous configuration changes. We hope that our results over multiple pre-train runs will encourage researchers and practitioners to be mindful of the reliability of such model diagnostics when verifying hypotheses about their models and training setups.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks.

Roy Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second PASCAL recognising textual entailment challenge.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. How can we know when language models know?

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, Online. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pages 2383–2392. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pages 1631–1642.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In the Proceedings of ICLR.

Junlin Wang, Jens Tuyls, Eric Wallace, and Sameer Singh. 2020. Gradient-based analysis of NLP models is manipulable. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 247–258, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models.

A SEAT computation

The effect size of a SEAT — characterized by two target (X, Y) and two attribute (A, B) sets of sentences — is calculated as:

$$d = \frac{\text{mean}_{x \in X}\, s(x, A, B) - \text{mean}_{y \in Y}\, s(y, A, B)}{\text{stdev}_{z \in X \cup Y}\, s(z, A, B)}$$

where:

$$s(sent, A, B) = \text{mean}_{a \in A} \cos(\overrightarrow{sent}, \overrightarrow{a}) - \text{mean}_{b \in B} \cos(\overrightarrow{sent}, \overrightarrow{b}).$$

The p-value of the permutation test to determine the statistical significance of the effect size is calculated as::

$$p = \Pr[S(X_i, Y_i, A, B) > S(X, Y, A, B)]$$

over partitions (X_i, Y_i) of $(X \cup Y)$ such that $|X_i| = |Y_i|$, where:

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$