When Do You Need Billions of Words of Pretraining Data?

Yian Zhang,[*,1] **Alex Warstadt,**[*,2] **Haau-Sing Li,**[3] **and Samuel R. Bowman**[1,2,3]
[1]Dept. of Computer Science, [2]Dept. of Linguistics, [3]Center for Data Science
New York University
{yian.zhang, warstadt, xl3119, bowman}@nyu.edu

Abstract

NLP is currently dominated by language models like RoBERTa which are pretrained on billions of words. But what exact knowledge or skills do Transformer LMs learn from large-scale pretraining that they cannot learn from less data? To explore this question, we adopt five styles of evaluation: classifier probing, information-theoretic probing, unsupervised relative acceptability judgments, unsupervised language model knowledge probing, and fine-tuning on NLU tasks. We then draw learning curves that track the growth of these different measures of model ability with respect to pretraining data volume using the MiniBERTas, a group of RoBERTa models pretrained on 1M, 10M, 100M and 1B words. We find that these LMs require only about 10M to 100M words to learn to reliably encode most syntactic and semantic features we test. They need a much larger quantity of data in order to acquire enough commonsense knowledge and other skills required to master typical downstream NLU tasks. The results suggest that, while the ability to encode linguistic features is almost certainly necessary for language understanding, it is likely that other, unidentified, forms of knowledge are the major drivers of recent improvements in language understanding among large pretrained models.

1 Introduction

Pretrained language models (LMs) like BERT and RoBERTa have become ubiquitous in NLP. New models require massive datasets of tens or even hundreds of billions of words (Brown et al., 2020) to improve on existing models on language understanding benchmarks like GLUE (Wang et al., 2018). Much recent work has used probing methods to evaluate what these models do and do not learn (Belinkov and Glass, 2019; Tenney et al., 2019b; Rogers et al., 2020; Ettinger, 2020). Since most of these works only focus on models pretrained on a fixed data volume (usually billions of words), many interesting questions regarding the effect of the amount of pretraining data remain unanswered: What have data-rich models learned that makes them so effective on downstream tasks? How much pretraining data is required for LMs to learn different grammatical features and linguistic phenomena? Which of these skills do we expect to improve when we scale pretraining past 30 billion words? Which aspects of grammar can be learned from data volumes on par with the input to human learners, around 10M to 100M words (Hart and Risley)?

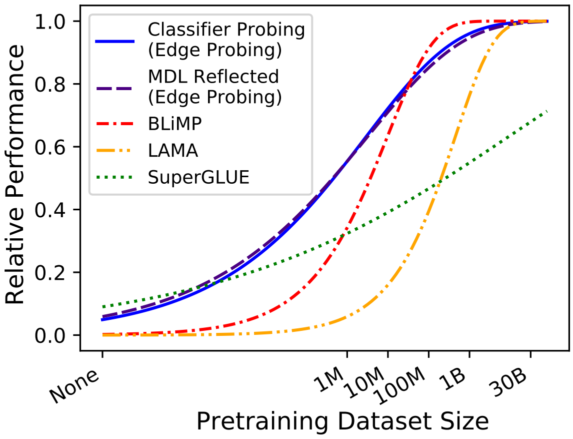

Figure 1: Overall learning curves for the five evaluation methods. For each method, we compute overall performance for each RoBERTa model tested as the macro average over sub-task's performance after normalization. We fit an exponential curve which we scale to have an initial value of 0 and an asymptote at 1. Classifier and MDL probing mainly test models' encoding of linguistic features; BLiMP tests model's understanding of linguistic phenomena; LAMA tests factual knowledge; SuperGLUE is a suite of conventional NLU tasks.

*Equal Contribution

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1112–1125
August 1–6, 2021. ©2021 Association for Computational Linguistics

With these questions in mind, we evaluate and probe the MiniBERTas (Warstadt et al., 2020b), a group of RoBERTa models pretrained on 1M, 10M, 100M, and 1B words, and RoBERTa$_{\text{BASE}}$ (Liu et al., 2019) pretrained on about 30B words, using five methods: First we use standard *classifier probing* on the edge probing suite of NLP tasks (Tenney et al., 2019b) to measure the quality of the syntactic and semantic features that can be extracted by a downstream classifier with each level of pretraining. Second, we apply *minimum description length (MDL) probing* (Voita and Titov, 2020) to the edge probing suite, with the goal of quantifying the accessibility of these features. Third, we test the models' knowledge of various syntactic phenomena using unsupervised acceptability judgments on the BLiMP suite (Warstadt et al., 2020a). Fourth, we probe the models' world knowledge and commonsense knowledge using unsupervised language model knowledge probing with the LAMA suite (Petroni et al., 2019). Finally, we fine-tune the models on five tasks from SuperGLUE (Wang et al., 2019) to measure their ability to solve conventional NLU tasks.

For each evaluation method, we fit an exponential learning curve to the results as a function of the amount of pretraining data, shown in Figure 1. We have two main findings: First, the results of classifier probing, MDL probing, and unsupervised relative acceptability judgement (BLiMP) show that the linguistic knowledge of models pretrained on 100M words and 30B words is similar, as is the description length of linguistic features. Second, RoBERTa requires billions of words of pretraining data to effectively acquire factual knowledge and to make substantial improvements in performance on dowstream NLU tasks. From these results, we conclude that there are skills critical to solving downstream NLU tasks that LMs can only acquire with billions of words of pretraining data. Future work will likely need to look beyond core linguistic knowledge if we are to better understand and advance the abilities of large language models.

2 Methods

We probe the MiniBERTas, a set of 12 RoBERTa models pretrained from scratch by Warstadt et al. (2020b) on 1M, 10M, 100M, and 1B words, the publicly available RoBERTa$_{\text{BASE}}$ (Liu et al., 2019), which is pretrained on about 30B words,[1] and 3 RoBERTa$_{\text{BASE}}$ models with randomly initialized parameters.

Descriptions of the five evaluation methods appear in the subsequent sections.[2] In each experiment, we test all 16 models on each task involved. To show the overall trend of improvement, we use non-linear least squares to fit an exponential learning curve to the results.[3] We upsample RoBERTa$_{\text{BASE}}$ results in regression in order to have an equal number of results for each data quantity. We use a four-parameter exponential learning curve used to capture diminishing improvement in performance as a function of the number of practice trials (Heathcote et al., 2000; Leibowitz et al., 2010):

$$E(P_n) = P_\infty - (P_\infty - P_0) \cdot e^{-\alpha \cdot n^\beta}$$

where $E(P_n)$ is the expected performance after n trials,[4] P_0 and P_∞ and are the initial and asymptotic performance, and α and β are coefficients to translate and dilate the curve in the log domain.

We plot the results in a figure for each task, where the y-axis is the score and the x-axis is the amount of pretraining data.[5] For some plots, we use min-max normalization to adjust the results into the range of [0, 1], where 0 and 1 are the inferred values of P_0 and P_∞, respectively.[6]

3 Classifier Probing

We use the widely-adopted probing approach of Ettinger et al. (2016), Adi et al. (2017), and others—which we call *classifier probing*—to test the extent to which linguistic features like part-of-speech and coreference are encoded in the frozen model representations. We adopt the ten probing tasks in the

[1] The miniBERTas' training data is randomly sampled from Wikipedia and Smashwords in a ratio of 3:1. These two datasets are what Devlin et al. (2019) use to pretrain BERT and represent a subset of the data used to pretrain RoBERTa. RoBERTa$_{\text{BASE}}$'s training data also includes of news and web data in addition to Wikipedia and Smashwords. Warstadt et al. ran pretraining 25 times with varying hyperparameter values and model sizes for the 1M-, 10M-, and 100M-word settings, and 10 times for the 1B-word setting. All the models were pretrained with early stopping on validation set perplexity. For each dataset size, they released the three models with the lowest validation set perplexity, yielding 12 models in total.

[2] Code: https://github.com/nyu-mll/pretraining-learning-curves

[3] We use SciPy's `curve_fit` implementation.

[4] In our case, a *trial* is one word of pretraining.

[5] We plot the no-pretraining random baseline with an x-value of 1.

[6] The unnormalized results are included in the appendix.

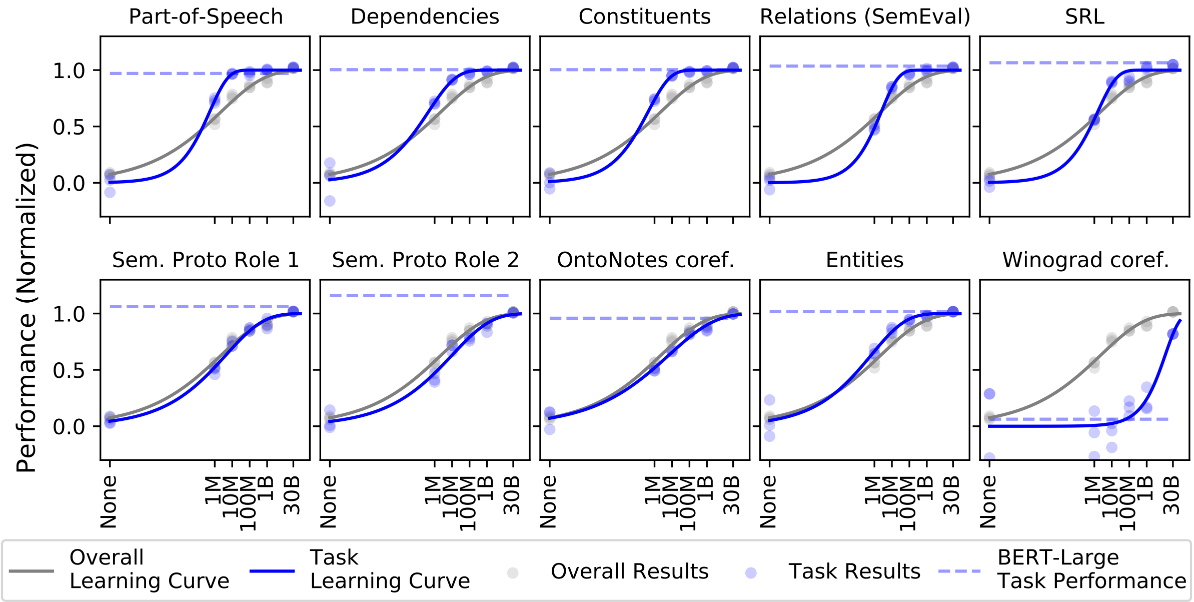

Figure 2: Classifier probing results for each task in the edge probing suite. Results are adjusted with min-max normalization for readability (see the Appendix for a non-normalized version). In each subplot we also plot the overall edge-probing performance, which we calculate for each MiniBERTa as its average F1 score on the 10 edge-probing tasks (after normalization). For context, we also plot BERT$_{\text{LARGE}}$ performance for each task as reported by Tenney et al. (2019a).

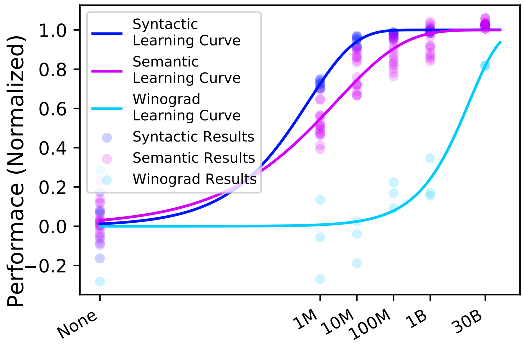

Figure 3: Edge Probing results for each group of tasks adjusted using min-max normalization. Syntactic tasks are Part-of-Speech, Dependencies, and Constituents. The commonsense task is Winograd coref. Semantic tasks are all remaining tasks.

edge probing suite (Tenney et al., 2019b).[7]

Classifier probing has recently come under scrutiny. Hewitt and Liang (2019) and Voita and Titov (2020) caution that the results depend on the complexity of the probe, and so do not precisely reveal the quality of the representations. However, we see two advantages to this method: First, the downstream classifier setting and F1 evaluation metric make these experiments easier to interpret in the context of earlier results than results from relatively novel probing metrics like minimum description length. Second, we focus on relative differences between models rather than absolute performance, and include a randomly initialized baseline model in the comparison. When the model representations are random, the probe's performance reflects the probe's own ability to solve the target task. Therefore, any improvements over this baseline value are due to the representation rather than the probe itself.

Task formulation and training Following Tenney et al., we use attention pooling to generate representation(s) of the token span(s) involved in the task and train an MLP that predicts whether a given label correctly describes the input span(s). We adopt the "mix" representation approach described in the paper. To train the probes, we use the same hyperparameters used in Tenney et al. and tune the batch size and learning rate.[8]

Results We plot results in Figure 2. From the single-task curves we conclude that most of the

[7] Task data sources: Part-of-Speech, Constituents, Entities, SRL, and OntoNotes coref. from Weischedel et al. (2013), Dependencies from Silveira et al. (2014), Sem. Proto Role 1 from Teichert et al. (2017), Sem. Proto Role 2 from Rudinger et al. (2018), Relations (SemEval) from Hendrickx et al. (2010), and Winograd coref. from Rahman and Ng (2012); White et al. (2017).

[8] We randomly sample 5 pairs from the range $\{8, 16, 32, 64\} \times \{5\text{e}-5, 1\text{e}-4, 5\text{e}-4\}$.

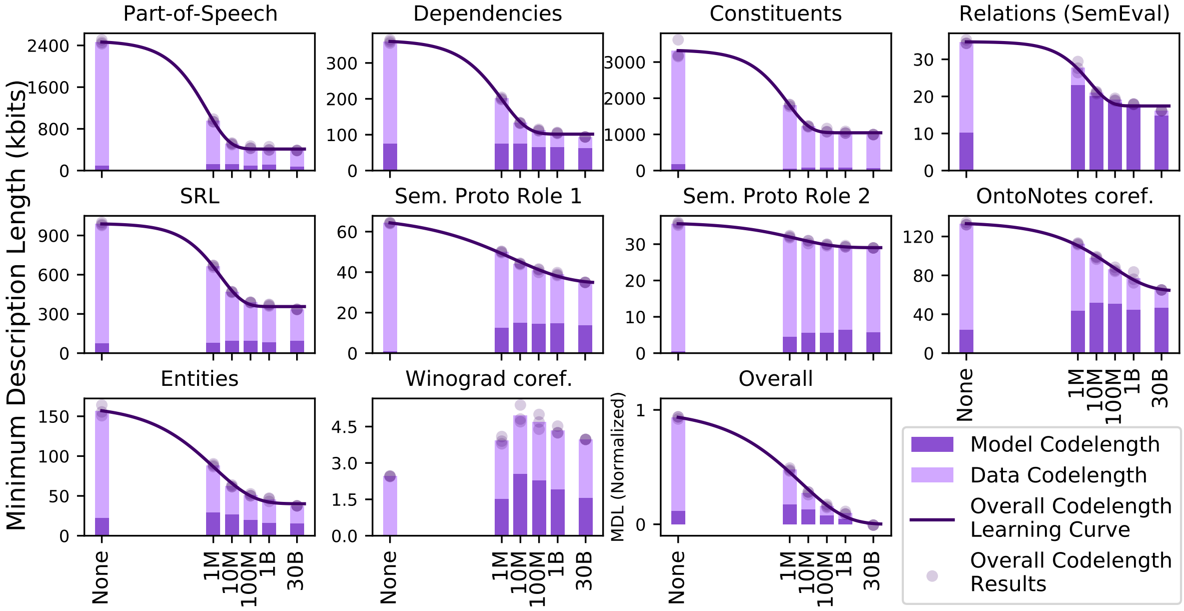

Figure 4: MDL results for each edge probing task. We do not plot a exponential curve for the Winograd coref. results because we could not find an adequate fit.

feature learning occurs with <100M words of pretraining data. Based on the best-fit curve, we can estimate that 90% of the attainable improvements in overall performance are achieved with <20M words. Most plots show broadly similar learning curves, which rise sharply with less than 1M words of pretraining data, reach the point of fastest growth (in the log domain) around 1M words, and are nearly saturated with 100M words. The most notable exception to this pattern is the Winograd task, which only rises significantly between 1B and 30B words of pretraining data.[9] As the Winograd task is designed to test commonsense knowledge and reasoning, the results suggest that these features require more data to encode than syntactic and semantic ones, with the caveat that the dataset is smaller than the other edge probing tasks, and results on Winograd tasks are highly sensitive to factors such as task formulation (Liu et al., 2020).

We observe some general differences between different types of tasks. Figure 3 shows the aggregated learning curves of syntactic, semantic, and commonsense tasks. The syntactic learning curve rises slightly earlier than the semantic one and 90% of the improvements in syntactic learning can be made with about 10M words, while the semantic curve still rises slightly after 100M. This is not surprising, as semantic computation is generally thought to depend on syntactic representations (Heim and Kratzer, 1998). The commonsense learning curve (for Winograd coref. only) rises far later, and is projected to continue to rise long after syntactic and semantic features stop improving.

4 Minimum Description Length Probing

In this experiment, we study the MiniBERTas with MDL probing (Voita and Titov, 2020), with the goal of revealing not only the total amount of feature information extracted by the probe, but also the effort taken by the probe to extract the features. MDL measures the minimum number of bits needed to transmit the labels for a given task given that both the sender and the receiver have access to the pretrained model's encoding of the data.

A well-trained *decoder* model can help extract labels from the representations and thus reduce the number of bits needed to transmit the labels. Since the model itself will also need to be transmitted, the total description length is a sum of two terms: The data codelength is the number of bits needed to transmit the labels assuming the receiver has the trained decoder model, i.e. the cross-entropy loss of the decoder. The model codelength is the number of bits needed to transmit the decoder parameters.

We follow Voita and Titov's *online code* estimation of MDL, where the decoder is implicitly transmitted. As in Section 3, we train decoders using the same hyperparameter settings and task

[9] These results are also noisier, similar to what Tenney et al. (2019b) find.

definitions as Tenney et al. (2019b).[10]

Results We plot the online code results in Figure 4. The overall codelength shows a similar trend to edge probing: Most of the reduction in feature codelength is achieved with fewer than 100M words. MDL for syntactic features decreases even sooner. Results for Winograd are idiosyncratic, probably due to the failure of the probes to learn the task.

The changes in model codelength and data codelength are shown on the bar plots in Figure 4. We compute the data codelength following Voita and Titov (2020) using the training set loss of a classifier trained on the entire training set, and the model codelength is the total codelength minus the data codelength. The monotonically decreasing data codelength simply reflects the fact that the more data rich RoBERTa models have smaller loss. When it comes to the model codelength, however, we generally observe the global minimum for the randomly initialized models (i.e., at "None"). This is expected, and intuitively reflects the fact that a decoder trained on random representations would provide little information about the labels, and so it would be optimal to transmit a very simple decoder. On many tasks, the model codelength starts to decrease when the pretraining data volume exceeds a certain amount. However, this trend is not consistent across tasks and the effect is relatively small.

5 Unsupervised Grammaticality Judgement

We use the BLiMP benchmark (Warstadt et al., 2020a) to test models' knowledge of individual grammatical phenomena in English. BLiMP is a challenge set of 67 tasks, each containing 1000 minimal pairs of sentences that highlight a particular morphological, syntactic, or semantic phenomena. Minimal pairs in BLiMP consist of two sentences that differ only by a single edit, but contrast in grammatical acceptability. A language model classifies a minimal pair correctly if it assigns a higher probability to the acceptable sentence. Since RoBERTa is a masked language model (MLM), we measure pseudo log-likelihood (Wang and Cho, 2019) to score sentences (Salazar et al., 2020).

Results We plot learning curves for BLiMP in Figure 5. Warstadt et al. organize the 67 tasks in BLiMP into 12 categories based on the phenomena tested and for each category we plot the average accuracy for the tasks in the category. We do not normalize results in this plot. For the no-data baseline, we plot chance accuracy of 50% rather than making empirical measurements from random RoBERTa models.

We find the greatest improvement in overall BLiMP performance between 1M and 100M words of pretraining data. With 100M words, sensitivity to contrasts in acceptability overall is within 9 accuracy points of humans, and improves only 6 points with additional data. This shows that substantial knowledge of many grammatical phenomena can be acquired from 100M words of raw text.

We also observe significant variation in how much data is needed to learn different phenomena. We see the steepest learning curves on agreement phenomena, with nearly all improvements occurring between 1M and 10M words. For phenomena involving *wh*-dependencies, i.e. filler-gap dependencies and island effects, we observe shallow and delayed learning curves with 90% of possible improvements occurring between 1M and 100M words. The relative difficulty of *wh*-dependencies can probably be ascribed to the long-distance nature and lower frequency of those phenomena. We also observe that the phenomena tested in the quantifiers category are never effectively learned, even by RoBERTa$_{\text{BASE}}$. These phenomena include subtle semantic contrasts—for example *Nobody ate* {*more than, *at least*} *two cookies*—which may involve difficult-to-learn pragmatic knowledge (Cohen and Krifka, 2014).

6 Unsupervised Language Model Knowledge Probe

LAMA is a test suite introduced by Petroni et al. to test LMs' factual knowledge. It contains over 50,000 cloze statements converted from subject-relation-object triples or question-answer pairs extracted from four datasets: GoogleRE,[11] TRE-x (Elsahar et al., 2018), ConceptNet (Speer and Havasi, 2012), and SQUAD (Rajpurkar et al., 2016). The Google-RE and T-REx tasks are each divided into three sub-tasks.

Results We plot the results on LAMA in Figure 6. The fastest growing point of most curves appears after 100M words. This relatively large quantity of

[10] Unlike us, Voita and Titov redefine the edge probing tasks as standard multi-class classification tasks.

[11] source: `https://code.google.com/archive/p/relation-extraction-corpus/`.

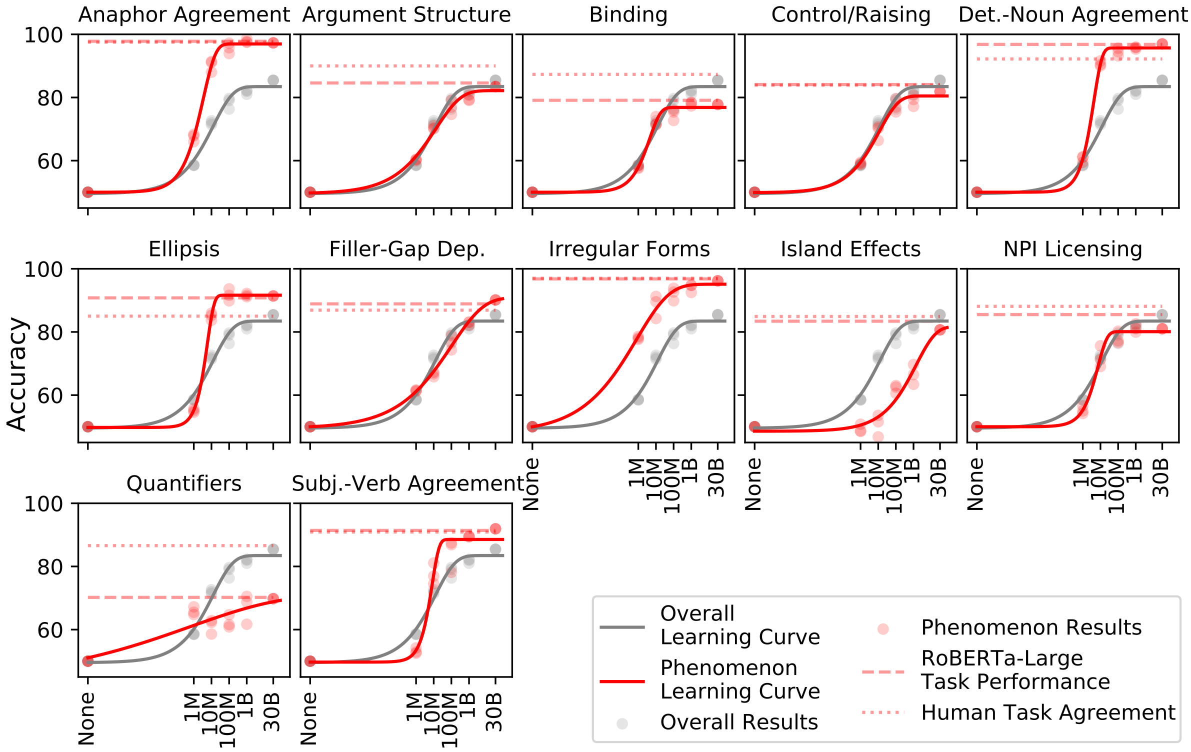

Figure 5: BLiMP results by category. BLiMP has 67 constituent datasets covering 12 linguistic phenomena. For each task the objective is to predict the more grammatically acceptable sentence of a minimal pair in an unsupervised setting. For context, we also plot human accuracy numbers from Warstadt et al. (2020a) and RoBERTa$_{\mathrm{LARGE}}$ performance from Salazar et al. (2020).

data may be needed for the model to be exposed to relevant factual knowledge. The learning curves for many LAMA tasks do not show clear signs of saturation in the range of 0 to 30B words, suggesting further improvements are likely with much larger data quantities. Among LAMA tasks, ConceptNet most directly tests commonsense knowledge. The steep slope of the ConceptNet curve between 100M and 30B words of pretraining data and the large precision jump (> 0.05) from 1B to 30B show that increasing the pretraining data to over 1B words significantly improve the LM's commonsense knowledge, which explains the shape of the Winograd coref. learning curve in Section 3.

7 Fine-tuning on NLU Tasks

SuperGLUE is a benchmark suite of eight classification-based language-understanding tasks (Wang et al., 2019). We test each MiniBERTa on five SuperGLUE tasks on which we expect to see significant variation at these scales.[12] The hyperparameter search range used for each task is described in the appendix.

Results We plot the results on the selected SuperGLUE tasks in Figure 7. Improvements in SuperGLUE performance require a relatively large volume of pretraining data. For most tasks, the point of fastest improvement in our interpolated curve occurs with more than 1B words. None of the tasks (with the possible exception of CommitmentBank) show any significant sign of saturation at 30B words. This suggests that some key NLU skills are not learnt with fewer than billions of words, and that models are likely to continue improving substantially on these tasks given 10 to 100 times more pretraining data.

8 Discussion

Figure 1 plots the overall learning curves for these five methods together. The most striking result is that good NLU task performance requires far more data than achieving good representations for linguistic features. Classifier probing, MDL

[12] Task data sources: CB from De Marneffe et al. (2019), BoolQ from Clark et al. (2019), COPA from Roemmele et al. (2011), WiC from Pilehvar and Camacho-Collados (2019); Miller (1995); Schuler (2005), and RTE from Dagan et al. (2006); Bar Haim et al. (2006); Giampiccolo et al. (2007); Bentivogli et al. (2009).

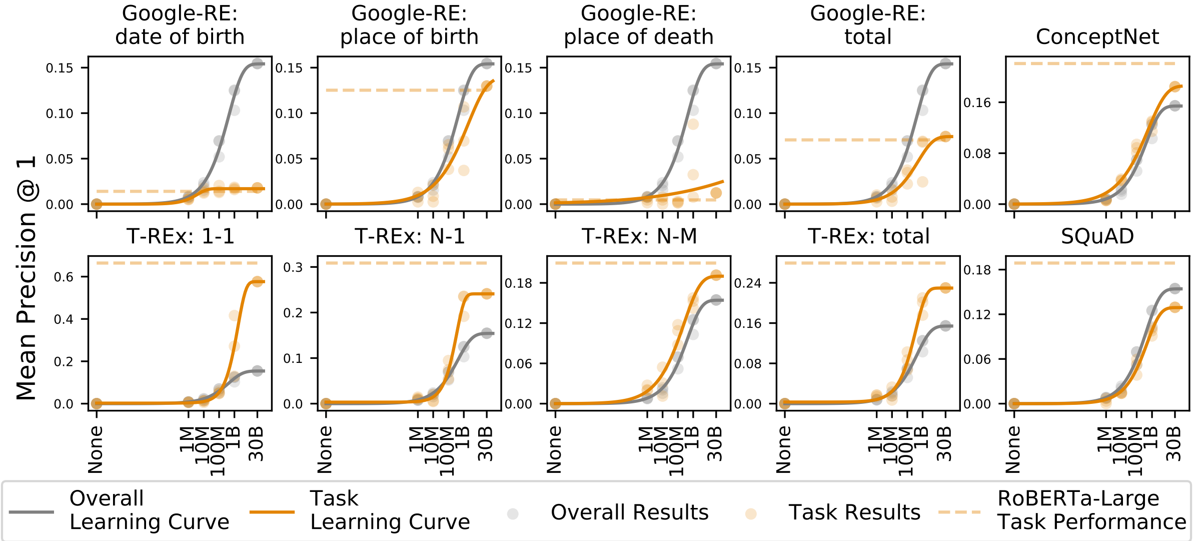

Figure 6: LAMA results. The metric for all tasks is mean precision at 1, i.e. the proportion of examples where the model assigns the highest probability to the ground truth token. For context, we also plot RoBERTa$_{\text{LARGE}}$ results.

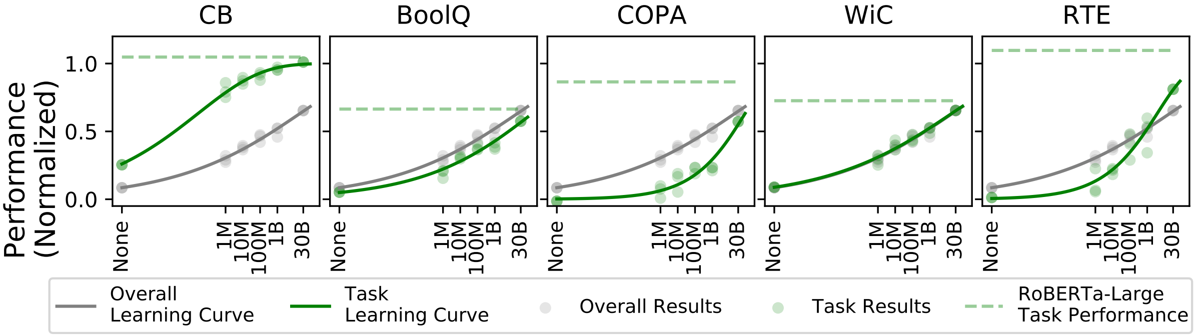

Figure 7: SuperGLUE results. The metric for BoolQ, COPA, WiC, RTE is accuracy, and for CB it is the average of accuracy and F1 score. Results are adjusted with min-max normalization for readability (see the Appendix for a non-normalized version). For context, we plot RoBERTa$_{\text{LARGE}}$ performance reported at https://github.com/pytorch/fairseq/tree/master/examples/roberta.

probing, and acceptability judgment performance all improve rapidly between 1M and 10M words and show little improvement beyond 100M words, while performance on the NLU tasks in SuperGLUE appears to improve most rapidly with over 1B words and will likely continue improving at larger data scales. While the linguistic features we test are undoubtedly needed to robustly solve most NLU tasks, a model that can extract and encode a large proportion of these features may still perform poorly on SuperGLUE. What drives improvements in NLU task performance at larger data scales remains an open question.

Factual knowledge may play a large role in explaining SuperGLUE performance. This hypothesis is backed up by results from the Winograd edge-probing task (Figure 2) and the LAMA tasks (Figure 6), which suggest that most of the improvements in the model's world and commonsense knowledge are made with over 100M words. However, the LAMA learning curve shows signs of slowing between 1B and 30B words, the SuperGLUE curve does not.

Another possible explanation is that linguistic features encoded by a model may not be easily accessible during fine-turning. Warstadt et al. (2020b) found that RoBERTa can learn to reliably extract many linguistic features with little pretraining data, but requires billions of words of pretraining data before it uses those features preferentially when generalizing.

In light of Warstadt et al.'s findings, we had initially hypothesized that feature accessibility as measured by MDL might show a shallower or later learning curve than standard classifier probing.[13]

[13] Warstadt et al.'s experiments are quite different to ours.

Our findings do not support this hypothesis: Figure 1 shows no substantial difference between the classifier probing MDL probing curves.

However, we do not totally rule out the possibility that linguistic feature accessibility continues to improve with massive pretraining sets. There are potential modifications to Voita and Titov's approach that could more faithfully estimate feature accessibility. First, although RoBERTa is actually fine-tuned in most applications, we and Voita and Titov measure MDL taking the outputs of the frozen RoBERTa model as input to a trainable MLP decoder. It may be more relevant to measure MDL by fine-tuning the entire model (Lovering et al., 2021). Second, MDL actually estimates the information content of a particular dataset, rather than the feature itself. Whitney et al. (2020) propose an alternative to MDL that measures feature complexity in a way that does not depend on the size of the dataset.

9 Related Work

Probing neural network representations has been an active area of research in recent years (Belinkov and Glass, 2019; Rogers et al., 2020). With the advent of large pretrained Transformers like BERT (Devlin et al., 2019), numerous papers have used classifier probing methods to attempt to locate linguistic features in learned representations with striking positive results (Tenney et al., 2019b; Hewitt and Manning, 2019). However, another thread has found problems with many probing methods: Classifier probes can learn too much from training data (Hewitt and Liang, 2019) and can fail to distinguish features that are extractable from features that are actually used when generalizing on downstream tasks (Voita and Titov, 2020; Pimentel et al., 2020; Elazar et al., 2020). Moreover, different probing methods often yield contradictory results (Warstadt et al., 2019).

There have also been a few earlier studies investigating the relationship between pretraining data volume and linguistic knowledge in language models. Studies of unsupervised acceptability judgments find fairly consistent evidence of rapid improvements in linguistic knowledge up to about 10M words of pretraining data, after which improvements slow down for most phenomena. van Schijndel et al. (2019) find large improvements in knowledge of subject-verb agreement and reflexive binding up to 10M words, and little improvement between 10M and 80M words. Hu et al. (2020) find that GPT-2 trained on 42M words performs roughly as well on a syntax benchmark as a similar model trained on 100 times that amount. Other studies have investigated how one model's linguistic knowledge changes during the training process, as a function of the number of updates (Saphra and Lopez, 2019; Chiang et al., 2020).

Raffel et al. (2020) also investigate how performance on SuperGLUE (and other downstream tasks) improves with pretraining dataset size between about 8M and 34B tokens. In contrast to our findings, they find that models with around 500M tokens of pretraining data can perform similarly on downstream tasks to models with 34B words. However, there are many differences in our settings that may lead to this divergence. For example, they pretrain for a fixed number of iterations (totaling 34B token updates), whereas the MiniBERTas we use were pretrained with early stopping. They also use prefix prompts in their task formulations, and adopt an encoder-decoder architecture and thus their model has roughly twice the number of parameters of the largest model we evaluate.

There is also some recent work that investigates the effect of pretraining data size of other languages. Micheli et al. (2020) pretrain BERT-based language models on 10MB, 100MB, 500MB, 1GB, 2GB, and 4GB of French text and test them on a question answering task. They find that the French MLM pretrained on 100MB of raw text has similar performance to the ones pretrained on larger datasets on the task, and that corpus-specific self-supervised learning does not make a significant difference. Martin et al. (2020) also show that French MLMs can already learn a lot from small-scale pretraining.

Concurrent work (Liu et al., 2021) probes RoBERTa models pretrained on different numbers of iterations using a set of probing tasks similar to ours. They find that linguistic abilities are acquired fastest, world and commonsense knowledge learning takes more iterations, and reasoning abilities are never stably acquired. Both studies show that linguistic knowledge is easier to learn than factual knowledge.

They measure RoBERTa's preference for linguistic features over surface features during fine-tuning on *ambiguous* classification tasks.

10 Conclusion

We track several aspects of RoBERTa's ability as pretraining data increases. We find that ability in syntax and semantics largely saturates after only 10M to 100M words of pretraining data—on par with the data available to human learners—while learning factual knowledge requires much more data. We also find that scaling pretraining data size past billions of words significantly improves the NLU performance, though we cannot fully explain what abilities drive this improvement. Answering this question could be a stepping stone to more data-efficient models.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant no. 1850208. 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. We would like to thank Udit Arora, Jason Phang, Clara Vania, and ML^2 for feedback on an earlier draft. Thanks also to Kyunghyun Cho, Tal Linzen, Grusha Prasad, and Emin Orhan for suggestions regarding the exponential learning curve, and to Elena Voita, Ian Tenney, and Haokun Liu for the discussion about the implementation of the probing methods.

Ethical Considerations

There are several ethical reasons to study LMs with limited pretraining data. Training massive LMs like RoBERTa from scratch comes with non-trivial environmental costs (Strubell et al., 2019), and they are expensive to train, limiting contributions to pretraining research from scientists in lower-resource contexts. By evaluating LMs with *limited* pretraining, we demonstrate that smaller LMs match massive ones in performance in many respects. We also identify a clear gap in our knowledge regarding *why* extensive pretraining is effective. Answering this question could lead to more efficient pretraining and ultimately reduce environmental costs and make NLP more accessible. On the other hand, there is a danger that our work, by projecting substantial gains in model performance by increasing pretraining size, could legitimize and encourage the trend of ever growing datasets.

Massive LMs also replicate social biases present in training data (Nangia et al., 2020). By establishing benchmarks for smaller LMs and highlighting their efficacy for certain purposes, we hope to spur future work that takes advantage of smaller pretraining datasets to carefully curate the data distribution, as advocated by Bender et al. (2021), in order to build LMs that do less to reproduce harmful biases and are more inclusive of minority dialects.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track. Toulon, France.*

Roy 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*.

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Textual Analysis Conference (TAC)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. 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.

Ariel Cohen and Manfred Krifka. 2014. Superlative quantifiers and meta-speech acts. *Linguistics and Philosophy*, 37(1):41–90.

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 Textual Entailment*. Springer.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. When BERT forgets how to POS: Amnesic probing of linguistic properties and MLM predictions. *arXiv preprint 2006.00995*.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

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*. Association for Computational Linguistics.

Betty Hart and Todd R. Risley. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6):1096.

Andrew Heathcote, Scott Brown, and Douglas JK Mewhort. 2000. The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, 7(2):185–207.

Irene Heim and Angelika Kratzer. 1998. *Semantics in generative grammar*. Blackwell Oxford.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*. 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.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Nathaniel Leibowitz, Barak Baum, Giora Enden, and Amir Karniel. 2010. The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, 54(3):338–340.

Haokun Liu, William Huang, Dhara Mungra, and Samuel R. Bowman. 2020. Precise task formalization matters in Winograd schema evaluations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8275–8280, Online. Association for Computational Linguistics.

Leo Z. Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? *CoRR*, abs/2104.07885.

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. *arXiv preprint arXiv:1907.11692*.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of fine-tuned models. In *International Conference on Learning Representations*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*.

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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. 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. 2020. 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.

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.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. 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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. In *Findings of EMNLP*.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-Davidsonian semantic proto-role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. 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 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew Gormley. 2017. Semantic proto-role labeling. In *AAAI Conference on Artificial Intelligence*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of ICLR*.

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*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *33rd Conference on Neural Information Processing Systems*.

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

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of EMNLP-IJCNLP*, pages 2870–2880.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Marcus Mitchell, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0 LDC2013T19. Linguistic Data Consortium.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

William F Whitney, Min Jae Song, David Brandfonbrener, Jaan Altosaar, and Kyunghyun Cho. 2020. Evaluating representations by the complexity of learning low-loss predictors. *arXiv preprint arXiv:2009.07368*.

A Appendices

Task	**Batch Size**	**Learning Rate**	**validation interval**	**Max Epochs**
BoolQ	{2,4,8}	{1e-6, 5e-6, 1e-5}	2400	10
CB	{2,4,8}	{1e-5, 5e-5, 1e-4}	60	40
COPA	{16,32,64}	{1e-6, 5e-6, 1e-5}	100	40
RTE	{2,4,8}	{5e-6, 1e-5, 5e-5}	1000	40
WiC	{16,32,64}	{1e-5, 5e-5, 1e-4}	1000	10

Table 1: Hyperparameter search ranges for the SuperGLUE tasks. Our search ranges are largely based on those used in Pruksachatkun et al. (2020).

Model	***Overall***	***ANA. AGR***	***ARG. STR***	***BINDING***	***CTRL. RAIS.***	***D-N AGR***	***ELLIPSIS***	***FILLER GAP***	***IRREGULAR***	***ISLAND***	***NPI***	***QUANTIFIERS***	***S-V AGR***
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	**71.3**	89.0
BERT$_{\text{BASE}}$	84.2	97.0	80.0	**82.3**	79.6	**97.6**	89.4	83.1	**96.5**	73.6	**84.7**	71.2	**92.4**
RoBERTa$_{\text{BASE}}$	**85.4**	97.3	**83.5**	77.8	**81.9**	97.0	91.4	**90.1**	96.2	**80.7**	81.0	69.8	91.9
1B-1	82.3	97.7	80.7	77.3	80.7	95.8	91.6	83.1	92.5	69.7	79.9	68.7	89.4
1B-2	81.0	97.5	79.1	78.3	79.4	96.0	**92.2**	82.1	94.8	63.4	81.2	61.7	89.6
1B-3	82.0	**98.6**	79.3	78.5	77.2	95.3	91.2	83.1	94.8	66.5	82.6	70.5	89.5
100M-1	76.3	93.9	74.6	72.7	77.0	93.2	89.9	74.3	89.9	60.6	76.6	61.6	78.1
100M-2	79.7	97.2	79.1	75.4	79.6	94.5	91.6	78.8	92.7	63.0	77.2	64.7	87.5
100M-3	79.1	95.8	76.9	76.0	75.4	95.6	93.7	76.8	93.9	62.5	80.2	60.9	86.9
10M-1	72.0	88.0	70.3	74.0	70.3	90.0	83.7	66.8	89.6	51.5	71.3	62.9	74.5
10M-2	72.6	91.1	70.1	71.6	70.7	91.6	86.0	67.3	84.3	53.6	75.6	58.6	77.0
10M-3	71.4	91.4	71.1	71.4	66.4	90.5	85.3	65.8	91.3	46.8	69.1	62.3	81.1
1M-1	58.5	67.9	60.4	58.5	59.4	59.5	54.6	61.6	78.1	50.8	54.2	64.8	52.5
1M-2	58.5	66.0	60.0	57.8	58.8	61.1	55.7	61.5	78.6	48.7	55.0	65.5	54.2
1M-3	58.7	68.4	60.3	57.5	59.1	61.3	55.1	61.2	77.7	48.5	56.6	67.2	52.9

Table 2: BLiMP results. 5-gram, LSTM, TXL, GPT-2 scores come from Warstadt et al. (2020a). BERT$_{\text{BASE}}$ scores come from Salazar et al. (2020).

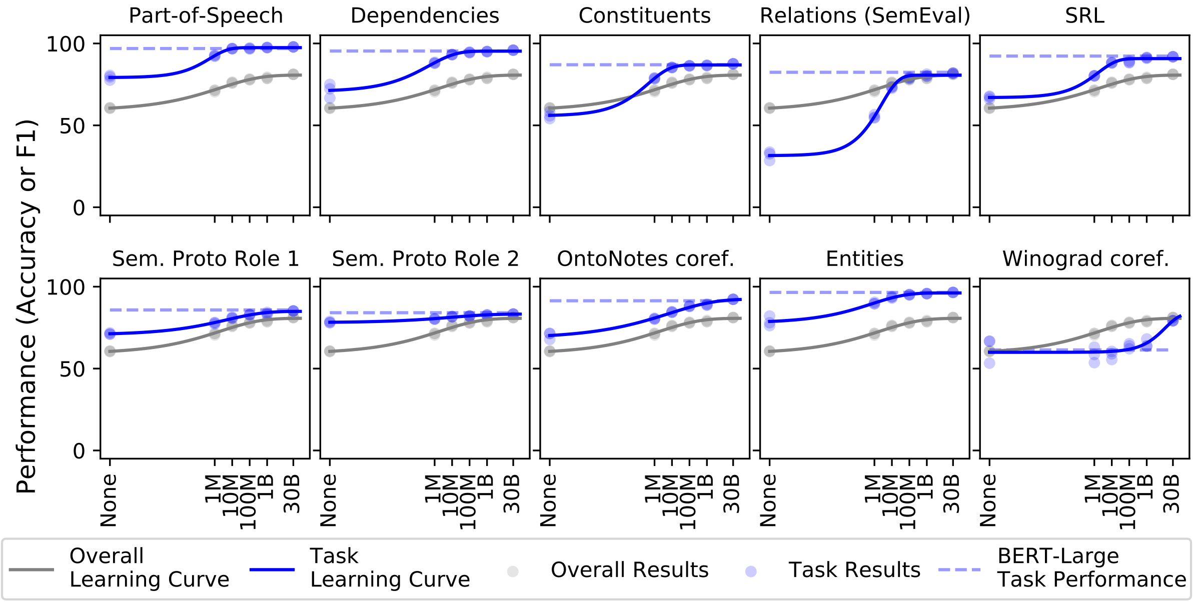

Figure 8: Our absolute edge probing dev set results (not normalized) compared to BERT$_{\text{LARGE}}$ test set results from Tenney et al. (2019b).

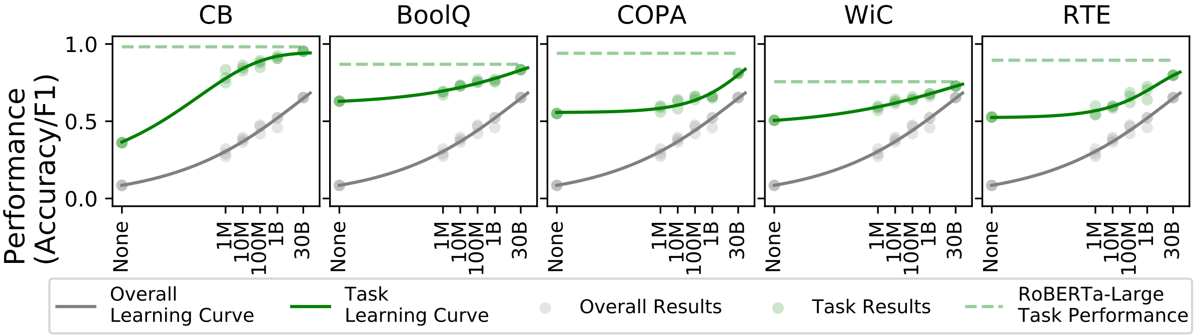

Figure 9: Our absolute SuperGLUE results (not normalized) compared to RoBERTa$_{\text{LARGE}}$ results from Liu et al. (2019).