How does the pre-training objective affect what large language models learn about linguistic properties?

Anonymous ACL submission

Abstract

Several pre-training objectives, such as masked language modeling (MLM), have been proposed to pre-train language models (e.g. BERT) with the aim of learning better language representations. However, to the best of our knowledge, no previous work so far has investigated how different pre-training objectives affect what BERT learns about linguistics properties. We hypothesize that linguistically motivated objectives (e.g. MLM) should help BERT to acquire better linguistic knowledge compared to using non-linguistically motivated objectives, i.e. hard for humans to guess the association between the input and the label to be predicted. To this end, we pre-train BERT with two linguistically motivated objectives and three non-linguistically motivated ones. We then probe for linguistic characteristics encoded in the representation of the resulting models. We find strong evidence that there is no actual differences in probing performance between the representations learned by the two different types of objectives. These surprising results question the dominant narrative of linguistically informed pre-training.[1]

1 Introduction

The most popular way to pre-train a transformer-based (Vaswani et al., 2017) language model (LM), e.g. BERT (Devlin et al., 2019), is by optimizing a masked language modeling (MLM) objective. The MLM task was inspired by the Cloze Task (Taylor, 1953), where humans were asked to guess omitted words in a sentence using its context, knowledge of syntax and other skills. The premise is that such an objective will guide a LM to encode linguistic information.

Apart from MLM, different types of objectives have been recently proposed. Yang et al. (2019) introduced a pre-training objective based on token order permutations. Clark et al. (2020) proposed a Replaced Token Detection pre-training task, that uses the output of a small MLM to corrupt the input by replacing some of the tokens. It then trains a discriminative model to predict if a token has been replaced or not. Aroca-Ouellette and Rudzicz (2020) explored various sentence and token-level auxiliary pre-training tasks (e.g. sentence ordering, term-frequency prediction), as better alternatives to the next sentence prediction (NSP) auxiliary task originally used to train BERT. Lan et al. (2020) introduced the sentence-order prediction task that focuses on the inter-sentence coherence, by predicting if two contiguous sentences have been swapped or not. Iter et al. (2020) proposed another inter-sentence pre-training task, that helps LMs to encode discourse relationships between sentences using contrastive learning. Yamaguchi et al. (2021) showed that a non-linguistically intuitive task (i.e. masked first character prediction) can effectively be used for pre-training.

Meanwhile, several studies have explored how well and to what extent LMs learn linguistic information. This is usually examined using probing tasks, i.e. simple classification tasks that test the LM's encodings for a single linguistic feature such as grammatical information. It has been found through probing that BERT encodes syntactic (Tenney et al., 2019; Liu et al., 2019; Miaschi and Dell'Orletta, 2020; Hewitt and Manning, 2019; Jawahar et al., 2019) and semantic information (Ettinger, 2020; Jawahar et al., 2019; Tenney et al., 2019). However, Hall Maudslay and Cotterell (2021) argue that BERT's syntactic abilities may have been overestimated.

In this paper, we hypothesize that linguistically motivated objectives (e.g. MLM) should help BERT to acquire better linguistic knowledge compared to using dummy or non-linguistically motivated objectives, i.e. tasks that are hard for humans to guess the association between the input and the label to be predicted. To this end, we seek to an-

[1] Code will be made publicly available.

swer the following research question: *How does the pre-training objective affect what LMs learn about the English language?*

Our findings challenge the MLM status quo, showing that pre-training with dummy, non-linguistically informative objectives (§2) results in models with similar linguistic capabilities, as measured by standard probing benchmarks (§3). These surprising results (§4) suggest that careful analysis of how LMs learn is critical to further improve language modeling (§5).

2 Pre-training Objectives

We experiment with five different pre-training objectives. Two of them are considered linguistically motivated while the rest are not.

2.1 Linguistically Motivated Objectives

Masked Language Modeling (MLM): We use MLM as our first linguistically motivated pre-training objective. First introduced by Devlin et al. (2019), MLM randomly chooses 15% of the tokens from the input sentence and replaces 80% of them with a [MASK] token, 10% with a random token, and 10% remain unchanged.

Manipulated Word Detection (S+R): We also experiment with a simpler linguistically motivated objective, where the model selects and replaces 10% of input tokens with shuffled tokens from the same input sequence. Concurrently, it selects and replaces another 10% of input tokens with random tokens from the vocabulary (Yamaguchi et al., 2021).

2.2 Non-Linguistically Motivated Objectives

We assume that tasks that are hard for humans (such as a completely random prediction task) will make less likely the deeper layers of BERT (i.e. closer to the output layer) to acquire meaningful information about language. We also hypothesize that layers closer to the input might learn word co-occurrence information (Sinha et al., 2021).

Masked First Character Prediction (First Char): For our first non-linguistically motivated pre-training objective, we use the masked first character prediction introduced by Yamaguchi et al. (2021). In this task, the model predicts only the first character of the masked token (e.g. '[c]at' and '[c]omputer' belong to the same class). The model predicts the first character as one of 29 classes, including the English alphabet and digit, punctuation mark, and other character indicators.

Masked ASCII Codes Summation Prediction (ASCII): We also propose a new non-linguistically motivated pre-training objective, where the model has to predict the summation of the ASCII code values of the characters in a masked token. To make this harder and keep the number of classes relatively small, we define a 5-way classification task by taking the modulo 5 of the ASCII summation: $V = [\sum_i ascii(char_i)] \,\%5$. Guessing the association between the input and such label, is an almost impossible task for a human.

Masked Random Token Classification (Random): Finally, we propose a completely random objective where we mask 15% of the input tokens and we assign each masked token a class from 0 to 4 *randomly* for a 5-way classification similar to the ASCII task. We assume that a model pre-trained with a random objective should not be able to learn anything meaningful about linguistic information.

3 Probing Tasks

Probing tasks (Adi et al., 2016; Conneau et al., 2018; Hupkes et al., 2018) are used to explore in what extent linguistic properties are captured by LMs. A model is normally trained, using the representations of a language model, to predict a specific linguistic property. If it achieves high accuracy, it implies that the LM encodes that linguistic property. In this work, we use six standard probing tasks introduced by Conneau et al. (2018) to examine the representation output for each layer of the different LMs we pre-train. These tasks probe for surface, syntactic and semantic information (i.e. two tasks per linguistic category). The dataset for each probing task contains 100k sentences for training, 10k sentences for validation and another 10k sentences for testing.[2] We train a logistic regression (LR) classifier for each probing task by only tuning the L2 regularization strength (Conneau et al., 2018).

Surface information tasks: **SentLen** aims for correctly predicting the number of words in a sentence. **WC** tests if the representations preserve information about the original words in a sentence by predicting which word the sentence contains out of 1000 words.

[2]The datasets are all publicly available by Conneau and Kiela (2018).

	SentLen					WC				
	1	3	6	9	12	1	3	6	9	12
Major.			20.0					0.5		
	BASE - Devlin et al. (2019) (Upper Bound)									
MLM+NSP	93.9	96.2	88.4	80.1	69.5	24.9	66.5	63.5	47.9	49.1
	BASE									
MLM	87.6	**88.0**	87.6	88.1	88.0	47.0	**71.1**	74.7	74.2	**73.1**
S+R	87.5	87.8	88.0	87.9	88.0	47.0	69.7	74.4	73.9	72.5
First Char	87.7	87.3	87.3	87.6	87.7	47.7	70.9	74.1	73.2	71.5
ASCII	**87.8**	**88.0**	**88.1**	**88.2**	**88.1**	**49.0**	70.8	**75.0**	**75.4**	72.9
Random	87.6	87.7	87.7	88.0	87.8	45.9	70.5	74.7	74.5	72.0

Table 1: Accuracy on surface information probing tasks for layers 1, 3, 6, 9, and 12 of each model.

Syntactic information tasks: **TreeDepth** tests if the representations preserve information about the hierarchical structure of a sentence, by predicting the depth of its parse tree. **TopConst** predicts the top constituents of the parse tree of a sentence.

Semantic information tasks: **SubjNum** predicts if the subject of the main clause is singular or plural. **ObjNum** tests if the direct object of the main clause is singular or plural.

4 Experiments & Results

4.1 Experimental Setup

Models We pre-train BERT-BASE (Devlin et al., 2019) models by replacing MLM and the next sentence prediction (NSP) objectives, with one of the linguistically or non-linguistically motivated pre-training tasks (§2).[3]

Pre-training Data All models are pre-trained on the BookCorpus (Zhu et al., 2015) and English Wikipedia from Hugging Face.[4] The text is tokenized using Byte-Pair-Encoding (Sennrich et al., 2016), resulting to a total of 2.7 billion tokens.

Pre-training Details Due to limited computational resources, each model is pre-trained for 5 days using two NVIDIA Tesla V100 (SXM2 - 32GB). We use a batch size of 32 for BASE, and 64 for MEDIUM and SMALL. We optimize the models using Adam (Kingma and Ba, 2014).[5]

4.2 Probing Results

Surface information Table 1 shows results for the two surface information probing tasks (SentLen and WC), using the representations from the five BERT-BASE models as inputs to the LR model. We first observe that *the predictive performance of models trained on the representations learned using non-linguistically motivated objectives (e.g. First Char, ASCII, Random), are comparable to those trained on representations learned with linguistically motivated objectives (e.g. MLM)*. For example using representations from layer 12 on the SentLen probing task, representations learned with the non-linguistically motivated ASCII pre-training objective achieve the best performance with 88.1%, while the model pre-trained with the First Char learned representations achieves the lowest performance with 87.7%.

Syntactic information Similar to the surface information probing tasks, the results of the syntactic probing tasks in Table 2 show that the performance of models trained on representations learned with linguistically motivated objectives is very similar to ones trained on non-linguistically learned representations. For instance, in the TreeDepth probing task using representation from layer 1, the difference between the highest accuracy and the lowest accuracy is just 0.7%. In the TopConst probing task, both the model pre-trained using S+R and the model pre-trained using ASCII achieve the best

[3]For completeness, we also pre-train two smaller model architectures, MEDIUM and SMALL, from (Turc et al., 2019) as in Yamaguchi et al. (2021). The MEDIUM model has 8 hidden layers and 8 attention heads. The SMALL model has 4 hidden layers and 8 attention heads. Both, MEDIUM and SMALL, models have feed-forward layers of size 2048 and hidden layers of size 512. More hyperprameter details can be found in Appendix A

[4]`https://github.com/huggingface/datasets`

[5]We also include results from fine-tuning models on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) in Appendix B, examining their performance on downstream tasks.

	TreeDepth					TopConst				
	1	3	6	9	12	1	3	6	9	12
Major.			17.9					5.0		
	BASE - Devlin et al. (2019) (Upper Bound)									
MLM+NSP	35.9	39.7	41.3	38.5	34.7	63.6	71.5	83.3	83.1	76.5
	BASE - Five Days Pre-training									
MLM	31.6	31.2	31.7	30.9	**31.6**	67.8	68.1	68.5	**69.1**	**68.7**
S+R	31.4	**32.0**	31.6	**31.6**	31.4	67.8	**69.3**	**69.4**	68.9	68.6
First Char	31.1	31.5	**32.2**	**31.6**	**31.6**	67.6	68.3	68.7	68.5	68.2
ASCII	31.3	31.5	31.2	31.4	31.4	**67.9**	**69.3**	69.0	68.8	68.4
Random	**31.8**	31.3	31.7	**31.6**	**31.6**	67.8	68.8	68.5	68.6	68.4

Table 2: Accuracy on syntactic information probing tasks for layers 1, 3, 6, 9, and 12 of each model.

	SubjNum					ObjNum				
	1	3	6	9	12	1	3	6	9	12
Major.			50.0					50.0		
	BASE - Devlin et al. (2019) (Upper Bound)									
MLM+NSP	77.6	82.0	88.1	87.6	84.0	76.7	80.3	82.0	81.8	78.7
	BASE - Five Days Pre-training									
MLM	68.0	67.5	67.7	67.7	68.4	64.9	**65.6**	**64.7**	**64.9**	**65.0**
S+R	68.0	67.8	68.1	68.0	67.9	**65.3**	64.4	63.7	64.5	64.6
First Char	**68.9**	68.4	**68.4**	**68.9**	**68.9**	63.5	64.2	64.3	64.0	63.7
ASCII	68.6	**68.7**	67.8	67.6	67.9	63.5	62.8	63.3	62.5	62.9
Random	68.1	67.8	67.8	67.9	68.3	64.2	63.6	63.4	63.9	63.3

Table 3: Accuracy on semantic information probing tasks for layers 1, 3, 6, 9, and 12 of each model.

performance in layer 3 with 69.3%.

Semantic information We also note similar patterns in the results of the semantic information probing tasks in Table 3. Both types of pre-trained models achieve similar performance when probing for semantic information. For example, the model pre-trained using First Char achieves an accuracy of 68.9% while the model pre-trained with MLM achieves 68.4% in the SubjNum probing task using the representation of the last layer.

In general, similar behavior can also be observed for all layers and the two smaller model architectures, MEDIUM and SMALL. The full results of the probing tasks can be found in appendix C.

5 Discussion

Theoretically, LMs with dummy or non-linguistically motivated objectives would be expected to perform drastically worse than LMs pre-trained using MLM in both downstream tasks and linguistic capabilities. However, our results show that both types of LMs have surprisingly comparable performance (after fine-tuning in downstream tasks) and linguistic capabilities (after probing them) using the same training data, architecture and training scheme. We speculate that the pre-training data, and the size of the models have more impact on the effectiveness of LMs than the pre-training objectives. Furthermore, the comparable performance of the objectives in probing suggests that these models learn word co-occurrence information from pre-training (Sinha et al., 2021; Yamaguchi et al., 2021) and that the objectives may have a little effect.

6 Conclusions

In this work, we compared the linguistic capabilities of LMs. Surprisingly, our results show that pre-training with linguistically motivated objectives obtain similar performance to dummy objectives. This suggests that the data and the size of the model could be more influential than the objectives themselves in language modeling. In future work, we plan to extend our experiments into other languages and probing tasks.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Stéphane Aroca-Ouellette and Frank Rudzicz. 2020. On Losses for Modern Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4970–4981, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

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.

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.

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.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online. 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.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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 *International Conference on Learning Representations*.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

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.

2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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. *arXiv preprint arXiv:1804.07461*.

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.

Atsuki Yamaguchi, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras. 2021. Frustratingly simple pretraining alternatives to masked language modeling. *arXiv preprint arXiv:2109.01819*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

Appendices

A Hyperparameter Details

We implement the models using PyTorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). We use maximum 10 epochs for BASE and MEDIUM, and 15 epochs for SMALL. We also use a learning rate of 1e-4 for MLM. 5e-5 for BASE First Char, S + R, and ASCII. 5e-6 for BASE Random. 1e-4 for SMALL and MEDIUM First Char, ASCII and Random. We also use weight decay of 0.01, attention dropout of 0.1, 10000 warmup steps. We also use 1e-8 Adam ϵ, 0.9 Adam β_1 and 0.999 Adam β_2.

B Results on GLUE

We use the GLUE benchmark (Wang et al., 2018) to fine-tune each model using up to 20 epochs with early stopping. For each fine-tuning task, we test using five different seeds and report the average. Table 4 shows the performance of each model on 8 different fine-tuning tasks. We report matched accuracy for MNLI task, Matthews correlation for CoLA task, Spearman correlation for STS-B task, accuracy for MRPC task, F1 scores for QQP task, and accuracy for all other tasks. The WNLI task is disregarded following Aroca-Ouellette and Rudzicz (2020). The results on GLUE for the re-implemented models with MLM, Shuffle + Random and First Char pre-training tasks are in line with the results reported by Yamaguchi et al. (2021).

C Results of each Probing Task

Figures 1 to 6 show the performance of all model architectures for each of the 6 probing tasks.

Pre-training task	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	GLUE Avg.
	BASE - 40 Epochs Pre-training (Upper Bound)								
MLM + NSP	83.8	90.8	87.8	69.9	91.9	85.0	58.9	89.3	82.1 (0.4)
	BASE - Five Days Pre-training								
MLM	**79.9**	88.3	86.0	60.3	**90.2**	81.9	52.9	85.0	78.1 (0.3)
Shuffle + Random	79.8	**88.7**	**86.4**	**65.8**	87.9	**86.7**	**58.2**	**86.6**	**80.0 (0.1)**
First Char	78.1	86.1	85.2	55.3	88.5	81.5	43.7	82.8	75.2 (0.7)
ASCII	75.4	83.6	83.5	57.7	87.6	81.5	37.5	79.4	73.3 (0.3)
Random	69.7	78.3	77.8	54.4	82.0	70.9	17.4	25.0	59.4 (0.4)
	MEDIUM - Five Days Pre-training								
MLM	**78.9**	86.2	**85.9**	59.8	**89.5**	82.7	44.8	84.3	76.5 (0.6)
Shuffle + Random	78.7	**87.4**	85.8	**64.3**	87.3	**85.3**	**52.7**	**85.6**	**78.4 (0.5)**
First Char	76.4	84.9	84.9	55.1	87.5	81.4	38.4	82.4	73.9 (0.4)
ASCII	75.1	83.9	83.9	59.1	87.5	81.5	39.1	79.4	73.7 (0.4)
Random	73.3	82.6	82.9	57.5	86.2	80.2	33.3	78.6	71.8 (0.5)
	SMALL - Five Days Pre-training								
MLM	76.2	85.1	84.9	**59.1**	**88.6**	**81.7**	36.3	**84.2**	74.5 (0.2)
Shuffle + Random	**76.5**	**85.7**	**85.3**	58.1	86.4	80.7	**46.6**	83.8	**75.4 (0.1)**
First Char	74.6	84.0	84.3	55.0	87.5	78.2	31.7	80.5	72.0 (0.4)
ASCII	73.0	81.7	83.2	57.4	85.0	77.1	32.9	77.8	71.0 (0.3)
Random	71.3	82.1	83.0	57.8	85.7	74.3	27.6	78.0	70.0 (0.2)

Table 4: Results on GLUE dev sets with standard deviations over five runs in parentheses. **Bold** values denote the best performance across each GLUE task and GLUE Avg. for each model setting.

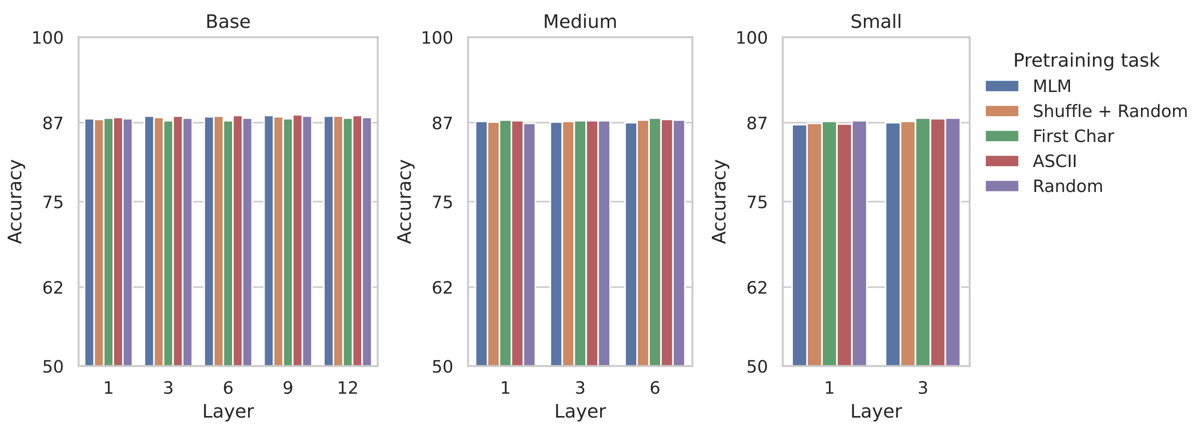

Figure 1: Results of the Sentence Length (**SentLen**) probing task

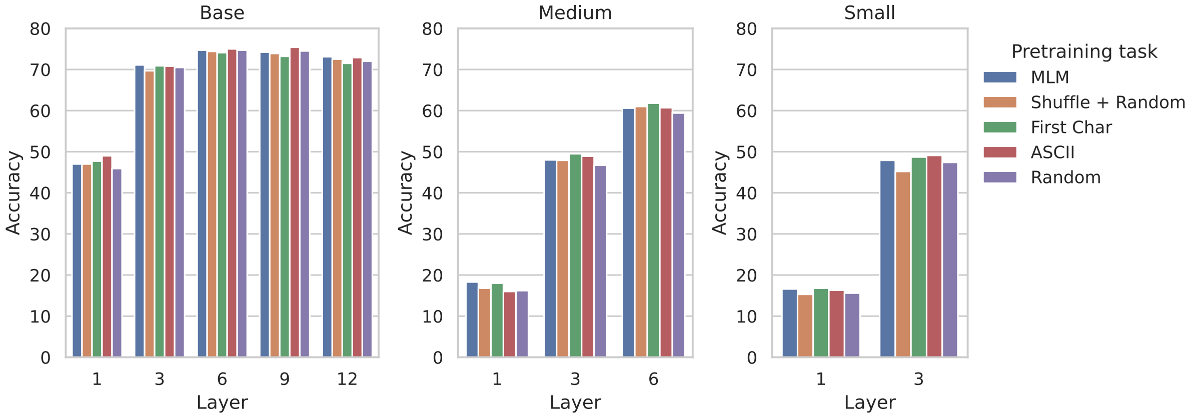

Figure 2: Results of the Word Content (**WC**) probing task

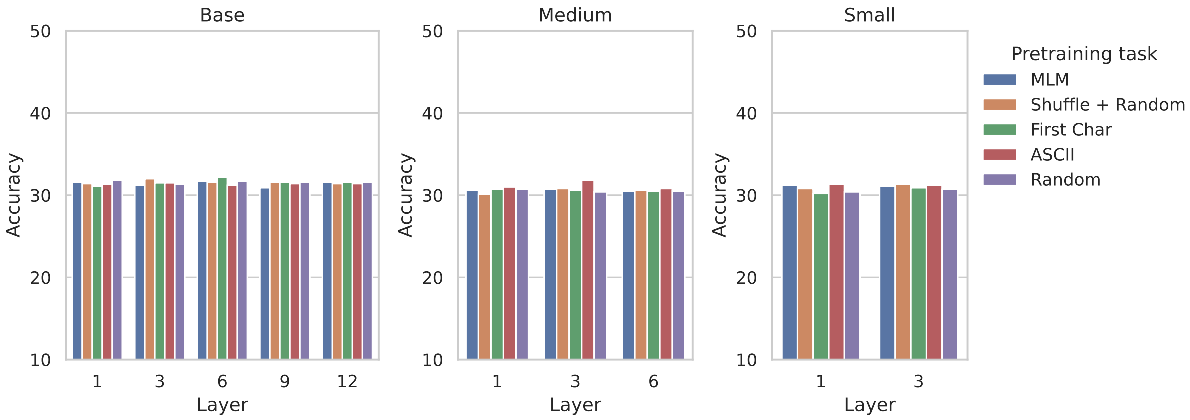

Figure 3: Results of the Tree Depth (**TreeDepth**) probing task

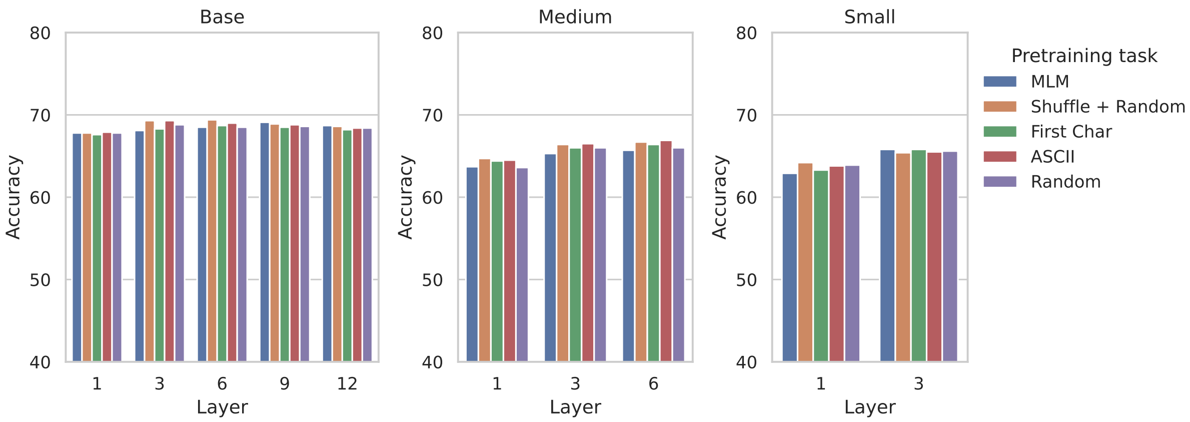

Figure 4: Results of the Top Constituent (**TopConst**) probing task

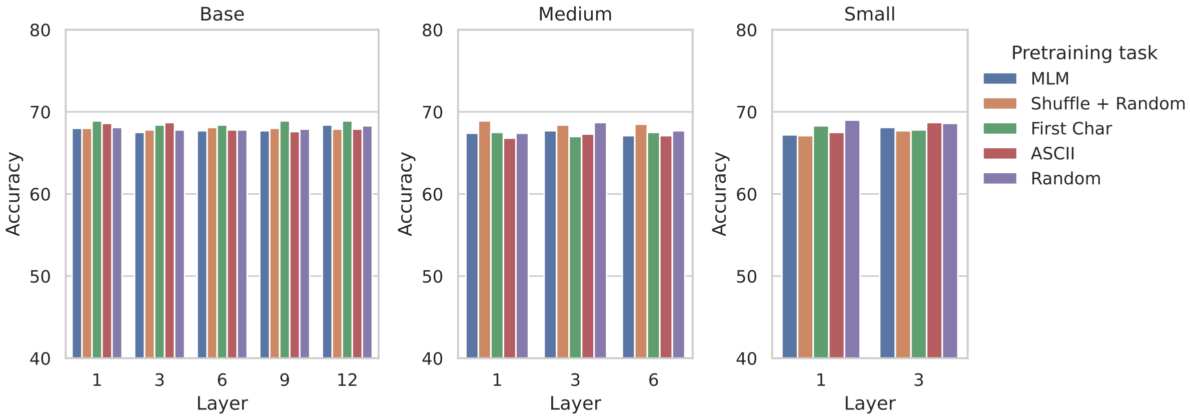

Figure 5: Results of the Subject Number (**SubjNum**) probing task

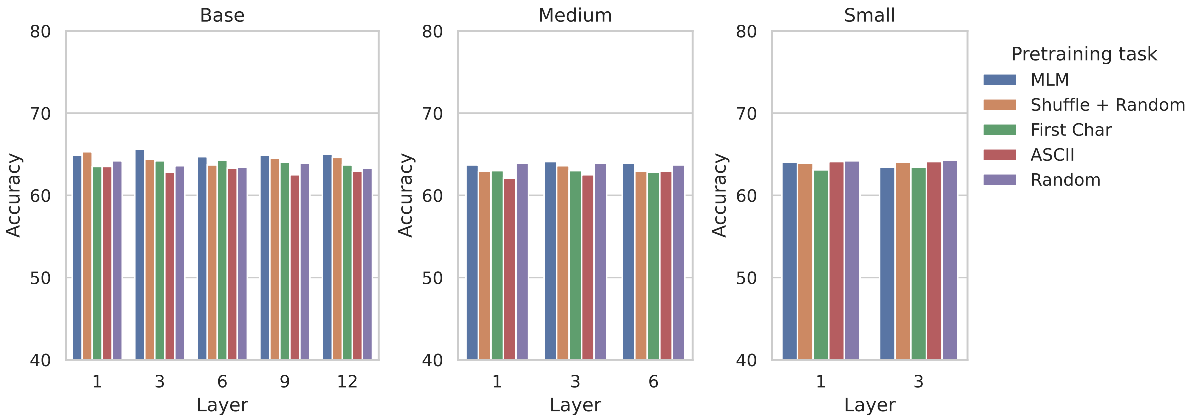

Figure 6: Results of the Object Number (**ObjNum**) probing task