Analyzing Pre-trained and Fine-tuned Language Models

Marius Mosbach
Department of Language Science and Technology
Saarland University
mmosbach@lsv.uni-saarland.de

Abstract

Since the introduction of transformer-based language models in 2018, the current generation of natural language processing (NLP) models continues to demonstrate impressive capabilities on a variety of academic benchmarks and real-world applications. This progress is based on a simple but general pipeline which consists of pre-training neural language models on large quantities of text, followed by an adaptation step that fine-tunes the pre-trained model to perform a specific NLP task of interest. However, despite the impressive progress on academic benchmarks and the widespread deployment of pre-trained and fine-tuned language models in industry we still lack a fundamental understanding of how and why pre-trained and fine-tuned language models work, as well as they do. We make several contributions towards improving our understanding of pre-trained and fine-tuned language models, ranging from analyzing the linguistic knowledge of pre-trained language models and how it is affected by fine-tuning, to a rigorous analysis of the fine-tuning process itself and how the choice of adaptation technique affects the generalization of models. We thereby provide new insights about previously unexplained phenomena and the capabilities of pre-trained and fine-tuned language models.

1 Introduction

Since the introduction of transformer-based pre-trained neural language models in 2018 (Devlin et al., 2019; Liu et al., 2019b), the field of natural language processing (NLP) has witnessed a paradigm shift. Instead of designing and training highly task-specific models from scratch, the current default approach for most NLP tasks consists of adapting general-purpose pre-trained language models, a process which typically requires only very few task-specific changes to the model architecture, and therefore allows us to easily apply the same pre-trained model to different tasks. Over the last five years (2019 – 2023), this paradigm shift has led to impressive progress on a large variety of downstream NLP tasks, ranging from traditional computational linguistics tasks such as part-of-speech tagging and more challenging tasks like natural language inference, to text-based dialogue and assistant systems (Wang et al., 2018, 2019; OpenAI, 2023, *inter alia*).

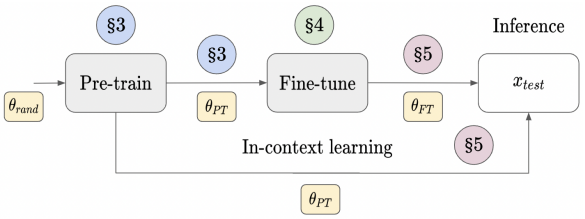


Figure 1: Our contributions positioned along the *pre-train then adapt pipeline* which is prevalent in modern-day NLP. §3 is concerned with how fine-tuning affects the linguistic knowledge of a model, §4 focuses on a better understanding of the fine-tuning process, and §5 is concerned with the generalization of models adapted via fine-tuning and in-context learning during inference.

At the core of this impressive progress lies a very simple but general pipeline which is illustrated in Figure 1 together with our contributions. The first step of this pipeline, which we will refer to as the *pre-train then adapt pipeline*, consists of pre-training a (large) neural language model on large quantities of text using self-supervised training. Due to the discrepancy between the pre-training objective (e.g., masked language modeling) and the downstream task (e.g., classification), the pre-training step is followed by an adaptation step which fine-tunes the pre-trained model to perform a specific task of interest. During fine-tuning, we either update all of the pre-trained parameters or update only a small fraction of them by leveraging parameter-efficient fine-tuning techniques. In both cases, however, fine-tuning results in a task-specific model which can be used for a single task. An alternative task-adaptation technique which was popularized by the most recent advances in training

pre-trained language models (Brown et al., 2020; OpenAI, 2023), allows us to bypass the fine-tuning step by treating the downstream task as a language modeling problem. This process, known as in-context learning, enables adapting a pre-trained model without updating any parameters and allows even non-expert users to easily leverage pre-trained language models.

Recent advancements in in-context learning have led to impressive progress on challenging reasoning benchmarks, surpassing the capabilities of fine-tuned language models by large margins (Wei et al., 2022a), a development which has resulted in unprecedented interest from the general public in the promises and potential risks associated with the use of large language models.

2 Research objectives

The previously described pipeline is ubiquitous in modern-day NLP and pre-trained and fine-tuned language models are now dominating research in academia as well as in industry. However, regardless of their impressive capabilities, pre-trained and fine-tuned language models are not without shortcomings. Our contributions center around three major shortcomings of pre-trained and fine-tuned language models. Each of the shortcomings concerns a specific component (or the interaction between two components) of the pre-train then align pipeline (see Figure 1).

2.1 Interplay between fine-tuning and probing

It is well established that fine-tuned language models are often right for the wrong reasons and their good performance on downstream tasks can at least in part be explained by the tendency to pick up spurious correlations during the adaptation process (Jia and Liang, 2017; McCoy et al., 2019; Niven and Kao, 2019; Warstadt et al., 2020, *inter alia*). These results stand in contrast to a large body of evidence that pre-trained language models encode various forms of linguistic and factual knowledge (Liu et al., 2019a; Tenney et al., 2019a; Petroni et al., 2019; Goldberg, 2019; Hewitt and Manning, 2019, *inter alia*).

When combined, these findings require taking a nuanced perspective on the connection between the strong capabilities of language models, as shown by their impressive results on common NLP tasks, and their encoding of linguistic and factual knowledge. These findings also demonstrate the need for investigating the interplay between the linguistic capabilities of pre-trained language models and their downstream performance.

2.2 Investigating fine-tuning stability

Fine-tuned language models often exhibit striking variation in downstream task performance when performing small changes to the adaptation process such as changing the random seed used for initializing model weights, the order of training examples, or the format of a task instruction (Dodge et al., 2020; Webson and Pavlick, 2022; Lu et al., 2022). Large variations in fine-tuning performance are undesirable for several reasons such as hindering reproducible research and complicating the distinction between actual improvements due to modeling or algorithmic advances and comparisons against weak baselines.

Given the ubiquity of fine-tuned language models, it is therefore critical to gain a better understanding of the fine-tuning algorithms that are commonly applied to adapt language models to downstream tasks.

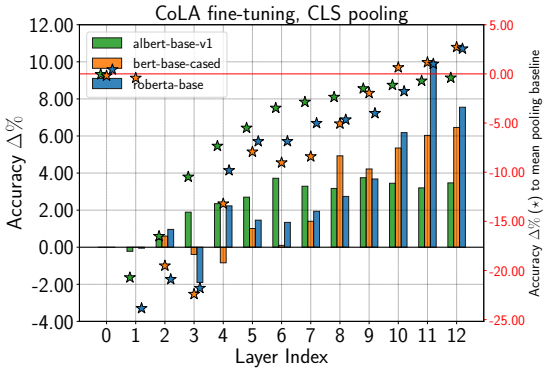
2.3 Generalization of task-adapted models

As mentioned in the previous section, the rapid progress in training ever larger language models has resulted in novel ways to adapt pre-trained language models to downstream tasks by simply instructing them to perform a task of interest via in-context learning. Instead of adapting a model via gradient based fine-tuning, in-context learning allows task adaptation via mere textual interaction and has lead to impressive progress on challenging reasoning benchmarks (Wei et al., 2022b,a). At the same time, there is growing evidence that in-context learning suffers from similar shortcomings to fine-tuning such as their sensitivity to changes in the data order (Min et al., 2022; Lu et al., 2022) and difficulties with generalizing to out-of-distribution inputs (Si et al., 2023).

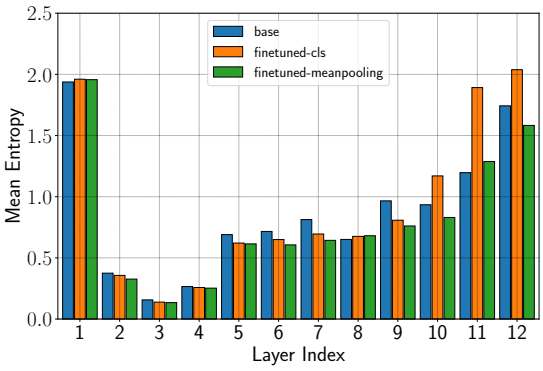
Given the prevalence of task adaptation via fine-tuning and in-context learning in modern NLP, it is necessary to investigate their respective benefits and downsides and provide a fair comparison of task adaptation approaches.

3 Interplay between fine-tuning and probing (Mosbach et al., 2020)

Our first contribution focuses on the connection between high performance on downstream tasks and



(a) Difference in probing accuracy before and after fine-tuning on CoLA using different models and pooling strategies.



(b) Entropy of the attention distribution for the cls-token of the RoBERTa model on the bigram-shift dataset.

Figure 2: A selection of our findings. (a) shows that when comparing to a stronger pooling baseline, fine-tuning has a negligible impact on probing performance. (b) shows that fine-tuning results in a more uniform attention which offers an alternative explanation for improved sentence-level probing performance.

the linguistic information encoded by a pre-trained model. Specifically, we investigate the hypothesis that the strong capabilities of fine-tuned language models can at least implicitly be attributed to the vast amount of linguistic knowledge which they encode (Pruksachatkun et al., 2020).

3.1 Previous work

A large body of previous work focused on analyzing the internal representations of neural models and the linguistic knowledge they encode via probing (Shi et al., 2016; Ettinger et al., 2016; Adi et al., 2016; Belinkov et al., 2017; Hupkes et al., 2018; Conneau et al., 2018; Krasnowska-Kieraś and Wróblewska, 2019). In a similar spirit to these first works on probing, Conneau et al. (2018) were the first to compare different sentence embedding methods based on the linguistic knowledge they encode. Krasnowska-Kieraś and Wróblewska (2019) extended this approach to study sentence-level probing tasks on English and Polish sentences.

Alongside sentence-level probing, a lot of recent work (Peters et al., 2018; Liu et al., 2019a; Tenney et al., 2019b; Lin et al., 2019; Hewitt and Manning, 2019) has focused on token-level probing tasks investigating more recent contextualized embedding models such as ELMo (Peters et al., 2018), GPT (rad), and BERT (Devlin et al., 2019). Two of the most prominent works following this methodology are Liu et al. (2019a) and Tenney et al. (2019b).

Limitations In contrast to our work, most studies that investigate pre-trained contextualized embedding models focus on pre-trained models and not fine-tuned ones. Therefore, little is known about the interaction between fine-tuning and probing. In our work, we aim to assess how probing performance changes with fine-tuning and how these changes differ based on the model architecture, as well as probing and fine-tuning task combination.

3.2 Our contributions

Setup We study three different pre-trained language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and ALBERT (Lan et al., 2020), and investigate via sentence-level probing (Conneau et al., 2018) how fine-tuning them on downstream tasks affects the linguistic information encoded in their representations.

We fine-tune on four datasets: CoLA (Warstadt et al., 2018), SST-2 (Socher et al., 2013), RTE (Dagan et al., 2005), SQuAD (Rajpurkar et al., 2016), and perform sentence-level probing experiments on three tasks from the SentEval probing suite (Conneau et al., 2018), each of which targets a different level of linguistic competence: bigram-shift, semantic-odd-man-out, and coordination inversion.

To evaluate the impact of fine-tuning on the linguistic information encoded by a model, we compare probing results before and after fine-tuning.

Fine-tuning mostly affects upper layers Comparing differences in probing performance before and after fine-tuning, we observe that fine-tuning mostly interacts with the upper layers of the pre-trained model. Changes in probing performance

are typically larger for higher layers and this finding is consistent across all models and tasks we experiment with.

Positive effect on probing performance is marginal When following the default strategy for sentence-level probing, i.e., constructing sentence representations based on the cls-token of the last hidden layer, we indeed observe large positive changes in probing performance due to fine-tuning, suggesting the encoding of new linguistic information during fine-tuning. However, when we change the pooling approach during probing to mean-pooling, the positive impact of fine-tuning on probing becomes negligible. This effect is illustrated in Figure 2a. For all models, we observe a large increase in probing performance when using cls-pooling to construct sentence representations. However, with mean-pooling, the difference in probing accuracy between the pre-trained and fine-tuned models becomes marginal and fine-tuning even hurts probing performance in lower layers.

Fine-tuning affects attention distribution To better understand the origin of the positive improvements in probing accuracy for cls-pooling, we investigate the attention distribution of the cls-token at every layer. We observe a large increase in entropy in the last three layers when fine-tuning on the cls-token (orange bars in Figure 2b). This is consistent with our hypothesis that during fine-tuning, the cls-token learns to take more sentence-level information into account, thus spreading its attention over more tokens, which offers an alternative explanation to why fine-tuning has a positive impact on probing performance.

3.3 Discussion

Our work provides novel insight into how to perform a fine-grained evaluation of the linguistic knowledge of pre-trained language models and on the interaction between probing performance and fine-tuning. Our findings demonstrate that there is no straightforward causal relationship between the linguistic information encoded by a model and its performance on NLP downstream tasks, which calls for a careful interpretation of changes in probing performance as a result of fine-tuning.

4 Investigating fine-tuning stability (Mosbach et al., 2021)

Our next contribution focuses on the second step of the pre-train then adapt pipeline. We analyze the fine-tuning process itself and study the intriguing finding that fine-tuned models tend to exhibit a large variance in performance, a phenomenon commonly referred to as fine-tuning instability.

4.1 Previous work

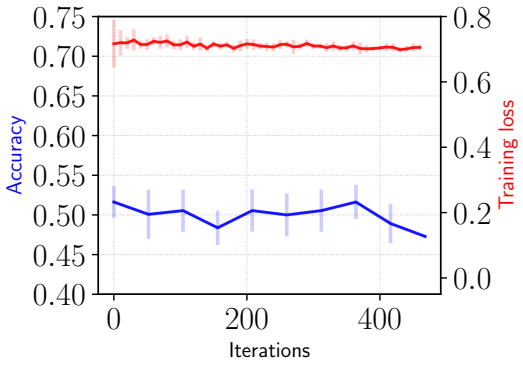
Previous work (Devlin et al., 2019; Lee et al., 2020; Dodge et al., 2020) has observed large differences in downstream task performance simply when fine-tuning models with different random seeds. Devlin et al. (2019) report instabilities when fine-tuning BERT-large on small datasets and resort to performing multiple restarts of fine-tuning and selecting the model that performs best on the development set. Dodge et al. (2020) performed a large-scale empirical investigation of the fine-tuning instability of BERT and found dramatic variations in fine-tuning accuracy across multiple restarts and argue how it might be related to the choice of random seed and the dataset size. Few approaches have been proposed to address the observed fine-tuning instability. Phang et al. (2018) study intermediate task training before fine-tuning with the goal of improving performance on the GLUE benchmark and find that their proposed method leads to improved fine-tuning stability. Lee et al. (2020) propose a new regularization technique termed Mixout which improves stability during fine-tuning.

Limitations While previous work on fine-tuning instability commonly states two hypotheses for the observed instability: catastrophic forgetting (Lee et al., 2020) and the small size of the training data (Dodge et al., 2020), there is no previous work that provides a sufficient understanding of why fine-tuning is prone to instability in the first place.

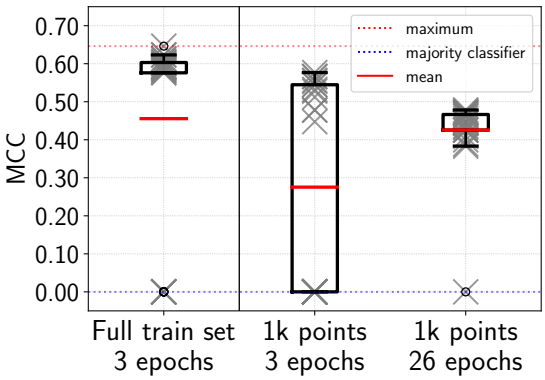
4.2 Our contributions

Motivated by the anecdotal observations stated in previous work, we perform a rigorous investigation of fine-tuning instability in order to determine its root cause.

Setup We analyze three different pre-trained language models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and ALBERT (Lan et al., 2020) and fine-tune them on widely used



(a) Training of failed models



(b) Results on down-sampled CoLA

Figure 3: Previous hypotheses fail to explain fine-tuning stability. (a) shows average training loss and validation accuracy across 3 failed fine-tuning runs on RTE. (b) shows validation performance of models fine-tuned on down-sampled CoLA.

datasets from the GLUE benchmark (Wang et al., 2018). We summarize our contributions below.

Previous hypotheses fail to explain instability First, we show that both catastrophic forgetting and the small size of the training data fail to explain the observed instability phenomenon. As shown in Figure 3a, failed fine-tuning runs in fact do not learn at all, violating the core assumption of catastrophic forgetting that the model performs well on the new task.

Regarding the small size of the training data, Figure 3b shows that fine-tuning on a down-sampled dataset for a small number of epochs does increase variance on the downstream task, however simply training for more iterations fully recovers the original variance in fine-tuning performance. This suggests that the observed instability on small datasets is connected to the number of training steps and not the size of the training set.

Optimization difficulties cause instability Next, we demonstrate that the observed instability is caused by optimization difficulties during fine-tuning that lead to vanishing gradients and models converging to sub-optimal local minima (illustrated in Figure 4). As we show in our work, this behavior is further amplified by choosing too large step sizes, fixing the number of epochs, and not warming up learning rates during the initial phase of fine-tuning.

A strong baseline for fine-tuning Based on our analysis, we present recommendations and a simple but strong baseline approach for fine-tuning. We recommend using small learning rates combined with warmup to avoid vanishing gradients during the initial fine-tuning phase. Additionally, when fine-tuning on small datasets, we suggest not fixing the number of epochs a priori (as was common practice) but rather fix the number of training steps.

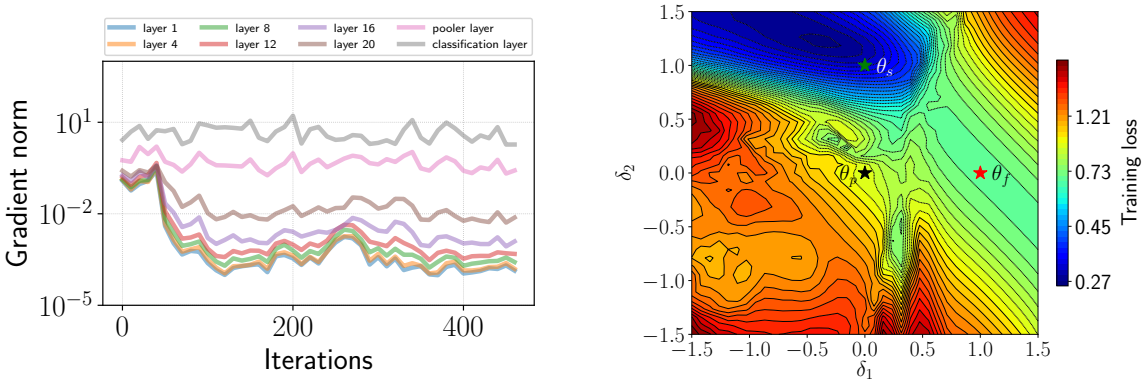
Approach	RTE			MRPC			CoLA		
	std	mean	max	std	mean	max	std	mean	max
Devlin	4.5	50.9	67.5	3.9	84.0	91.2	25.6	45.6	64.6
Lee	7.9	65.3	**74.4**	3.8	87.8	**91.8**	20.9	51.9	64.0
Ours	**2.7***	**67.3**	71.1	**0.8***	**90.3**	**91.7**	**1.8***	**62.1**	**65.3**

Table 1: Standard deviation, mean, and maximum performance on the development set of RTE, MRPC, and CoLA when fine-tuning BERT over 25 random seeds. Standard deviation: lower is better, i.e., fine-tuning is more stable. * denotes significant difference ($p < 0.001$) when compared to the second smallest standard deviation.

As can be seen in Table 1, our baseline makes fine-tuning pre-trained language models significantly more stable than previously proposed approaches while at the same time maintaining or even improving performance.

4.3 Discussion

Our work answers an open question about the instability of fine-tuning and shows that neither catastrophic forgetting nor small dataset sizes sufficiently explain fine-tuning instability. Instead, our analysis reveals that fine-tuning instability can be characterized by two distinct problems: (1) optimization difficulties early in training, characterized by vanishing gradients, and (2) differences in generalization, characterized by a large variance of de-



(a) Vanishing gradients during fine-tuning of BERT-large.

(b) 2D loss surface.

Figure 4: Fine-tuning instabilities are characterized by vanishing gradients (a) and convergence to sub-optimal local minima. The 2D loss surface in (b) is spanned by $\delta_1 = \boldsymbol{\theta}_f - \boldsymbol{\theta}_p$ and $\delta_2 = \boldsymbol{\theta}_s - \boldsymbol{\theta}_p$ on RTE.

velopment set accuracy for runs with almost equivalent training performance. Based on our analysis, we propose a simple but strong baseline strategy for fine-tuning BERT which outperforms previous works in terms of fine-tuning stability while maintaining or even increasing overall performance.

5 Generalization of task-adapted models (Mosbach et al., 2023)

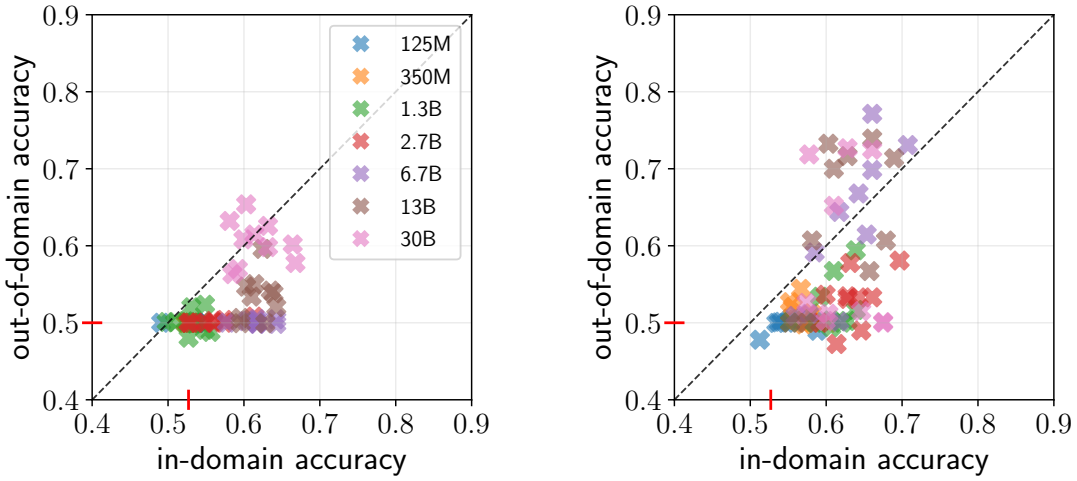
Our final contribution is concerned with the last step of the NLP pipeline, namely, inference. We compare the generalization behavior of task-adaptation via few-shot fine-tuning and in-context learning (ICL), which has recently gained popularity over fine-tuning due to its simplicity and strong performance on challenging reasoning tasks.

5.1 Previous work

Brown et al. (2020) compared GPT-3's few-shot in-context learning performance with fine-tuned language models trained in the fully supervised setting and found that both approaches lead to similar results in question answering. More recently, Liu et al. (2022) compared parameter-efficient few-shot FT of T0 (Sanh et al., 2022) to in-context learning with GPT-3, finding that their parameter-efficient fine-tuning approach outperforms in-context learning when evaluated on in-domain data. Focusing on out-of-domain (OOD) performance, Si et al. (2023) investigated the generalization of GPT-3 along various axes, including generalization under covariate shift. They observed much better OOD performance for in-context learning than fine-tuning, concluding that in-context learning with GPT-3 is more robust than fine-tuning using BERT or RoBERTa. Another work that compares the OOD generalization of different adaptation approaches is Awadalla et al. (2022). They investigate the robustness of question answering models under various types of distribution shifts and find that in-context learning is more robust to distribution shifts than fine-tuning. Moreover, they argue that for fine-tuning, increasing model size does not have a strong impact on generalization.

Utama et al. (2021) investigate the OOD generalization of encoder-only models adapted via pattern-based few-shot fine-tuning. For MNLI and HANS, they find that these models adopt similar inference heuristics to those trained with vanilla fine-tuning and hence perform poorly OOD. They observe that models rely even more on heuristics when fine-tuned on more data. Lastly, Bandel et al. (2022) show that masked language models can generalize well on HANS if fine-tuned for a sufficient number of steps.

Limitations A common limitation in the previous literature is the comparisons of generalization abilities under unequal conditions. Most studies either compare the in-context learning abilities of large models (e.g., GPT-3, 175B; Brown et al., 2020) to the fine-tuning abilities of much smaller models (e.g., RoBERTa-large, 350M; Liu et al., 2019b), or compare models fine-tuned on large datasets to few-shot in-context learning (Si et al., 2023). These comparisons raise the question of whether fine-tuning leads to weaker OOD generalization than in-context learning, or whether this is



(a) In-context learning

(b) Fine-tuning

Figure 5: In-domain (RTE) and out-of-domain performance (HANS) for in-context learning and fine-tuning with OPT models of various sizes. We fine-tune models using pattern-based fine-tuning. We report results using 10 different data seeds. When using 16 samples, in-context learning's performance with a 30B model is comparable to that of fine-tuning with smaller models (6.7B) and for most model sizes, fine-tuning outperforms in-context learning. — in the x- and y-axes indicates majority class accuracy.

just a byproduct of the experimental setup.

5.2 Our contributions

In our work, we investigate whether the observed weaker out-of-domain generalization of fine-tuned models by previous work is an inherent property of fine-tuning or an artifact of their experimental setup and provide a fair comparison between the generalization of fine-tuning and in-context learning.

Setup For our experiments, we consider few-shot pattern-based fine-tuning (Schick and Schütze, 2021; Gao et al., 2021, *inter alia*) and in-context learning (Brown et al., 2020). We perform a fair comparison of task adaptation focusing on in-domain and OOD generalization under *covariate shift* (Hupkes et al., 2022). We run all experiments using 7 different OPT models (Zhang et al., 2022) ranging from 125 million to 30 billion parameters. During fine-tuning, we update all model parameters if not stated otherwise.

Fine-tuned models can generalize well OOD For our first experiment, we compare fine-tuning and in-context learning using 16 examples for each. We plot the results of this experiment in Figure 5. For in-context learning, we observe an increase in in-domain performance with model size and non-trivial OOD performance only for the largest model (30B). For fine-tuning, we similarly observe that in-domain performance increases with model size. However, as model size increases, OOD performance increases as well, demonstrating that even in the challenging few-shot setting, fine-tuned models can generalize OOD. In Table 2 we provide significance tests that further support our findings. In-context learning only outperforms fine-tuning when comparing large models adapted via in-context learning to small fine-tuned models, which is unfair. Comparing models of the same size however, reveals that fine-tuned models either perform significantly better or similarly to models adapted via in-context learning.

		PBFT						
		125M	**350M**	**1.3B**	**2.7B**	**6.7B**	**13B**	**30B**
ICL	**125M**	−0.00	0.01	0.02	0.03	0.12	0.14	0.09
	350M	−0.00	0.01	0.02	0.03	0.12	0.14	0.09
	1.3B	−0.00	0.01	0.02	0.03	0.12	0.14	0.09
	2.7B	−0.00	0.01	0.02	0.03	0.12	0.14	0.09
	6.7B	−0.00	0.01	0.02	0.03	0.12	0.14	0.09
	13B	−0.04	−0.02	−0.01	−0.00	0.09	0.11	0.05
	30B	−0.11	−0.09	−0.08	−0.08	0.02	0.03	−0.02

Table 2: Difference between average **out-of-domain performance** of ICL and FT on RTE across model sizes. We use 16 examples and 10 random seeds for both approaches. We perform a Welch's t-test and color cells according to whether: ICL performs significantly better than FT, FT performs significantly better than ICL. For cells without color, there is no significant difference.

Generalization improves with more data In contrast to in-context learning, where the maximum number of demonstrations is limited by the context size of a model, fine-tuning allows us to perform task adaptation using arbitrary amounts of training data. Therefore, we analyze how the relationship between in-domain and OOD performance is impacted by training on more data. For the smallest models, we find that while in-domain performance increases with more training data, OOD performance remains low, which is consistent with previous work (Utama et al., 2021). However, for larger models, OOD performance improves as the amount of training data increases.

Findings generalize beyond OPT To test the generality of our findings beyond the OPT models, we run the same experiments using Pythia models of different sizes (Biderman et al., 2023). Similarly to OPT, we observe a clear effect of model size on both in-domain and OOD performance. For most model sizes, fine-tuning leads to significantly better OOD performance than in-context learning. Additionally, both the in-domain and OOD performance of Pythia models improve drastically as we fine-tune on more data.

Findings generalize to parameter-efficient fine-tuning We additionally experiment with parameter-efficient fine-tuning via LoRA (Hu et al., 2022) to demonstrate the generality of our findings beyond full fine-tuning. Using LoRA makes adaptation via fine-tuning more similar to adaptation via in-context learning as it allows the re-use of a large fraction of the weights of a pre-trained language model across tasks. Figure 6 shows that fine-tuning via LoRA leads to similar performance as training all parameters (shown in Figure 5b) which demonstrates the generality of our findings beyond a specific fine-tuning method.

5.3 Discussion

Our findings are an important first step towards a better understanding of the fundamental differences in model behavior between different task adaptation approaches. We demonstrate that fine-tuned language models can generalize well both in and out-of-domain. In fact, we find that the generalization of fine-tuning and in-context learning is highly similar as both approaches exhibit large variation in performance and strongly depend on properties such as model size and the number of examples. Hence, our work provides evidence that the poor out-of-domain generalization of fine-tuned models observed in previous work is not a fundamental flaw of fine-tuning but rather a result of their experimental setup, highlighting that truly robust task adaptation remains a challenge.

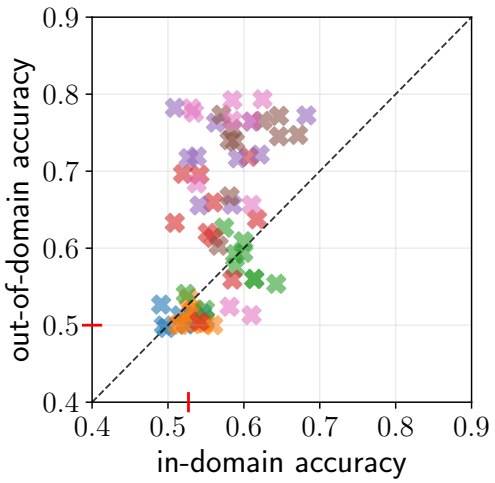


Figure 6: In-domain and OOD performance of parameter-efficient fine-tuning via LoRA on RTE. — in the x- and y-axes indicates the accuracy of the majority class label.

6 The bigger picture

Adapting pre-trained language models via fine-tuning or in-context learning is an integral part of modern-day NLP. While from late 2018 to mid-2020, fine-tuning was the dominating strategy for task adaptation, i.e., converting a pre-trained (masked) language model into a classifier, the introduction of GPT-3 (Brown et al., 2020) in 2020 and the demonstration of its in-context learning abilities resulted in an increasing interest in in-context learning as a new promising paradigm for task adaptation. Recently however, driven by work on instruction fine-tuning (Sanh et al., 2022; Wang et al., 2022, *inter alia*) and alignment to human preferences (Ouyang et al., 2022; Zhou et al., 2023, *inter alia*), fine-tuning[1] is again gaining significant interest from the NLP research community.

Given the ubiquity of language model adaptation in modern-day NLP and machine learning research, it is crucial to make progress towards a better understanding of the inner workings of commonly used

[1]Due to the dominance of decoder-only language models fine-tuning is however no longer used to explicitly adapt language models into classifiers but is instead used to adapt language models to assign higher probability to specific distributions, e.g., instructions and information seeking questions.

adaptation techniques as well as their limitations. The work presented in this paper demonstrates how empirical research can help to achieve this goal and hopefully serves as an inspiration for future research that critically investigates the rapid progress made along the pre-train then adapt pipeline.

7 Summary

Our work makes several contributions towards improving our understanding of pre-trained and fine-tuned language models by carrying out a detailed analysis of various parts of the pre-train then adapt pipeline. Our contributions range from analyzing the linguistic knowledge of pre-trained language models and how it is affected by fine-tuning, to a rigorous analysis of the fine-tuning process itself and how the choice of adaptation technique affects the generalization of models. We provide new insights about previously unexplained phenomena and the capabilities of pre-trained and fine-tuned language models and overall a better understanding of a crucial component of the modern NLP toolbox. Beyond our empirical contributions, we hope that our work demonstrates the importance of taking a critical perspective on previous work and shows that despite the rapid progress in our field, there is a need for work that critically analyzes this progress.

Acknowledgements

Marius Mosbach acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

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*.

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elron Bandel, Yoav Goldberg, and Yanai Elazar. 2022. Lexical generalization improves with larger models and longer training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4398–4410, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

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. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020.

Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

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, Berlin, Germany. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2022. State-of-the-art generalisation research in nlp: a taxonomy and review. In *arXiv*.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2020. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Muqeeth Mohammed, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064,

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516, Online. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

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

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. 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.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on STILTS: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

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.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

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 the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Prasetya Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9063–9074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. 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.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment.