Language Models are Few-Shot Learners

- SHORT SUMMARY

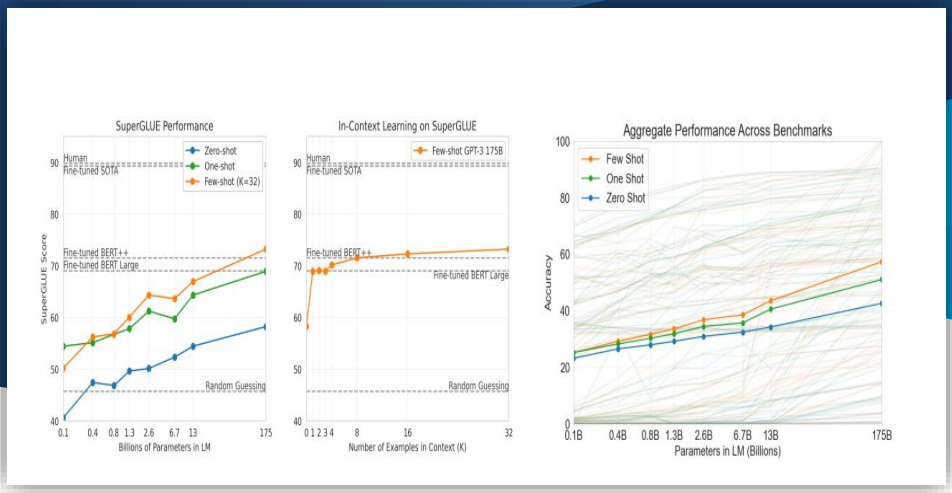
The research indicates a significant correlation between increased model size and improved performance on few-shot learning tasks across various NLP domains. Notably, GPT-3 demonstrates competitive results without requiring traditional fine-tuning techniques. By leveraging text-based interaction to specify tasks and provide minimal demonstrations, GPT-3 achieves strong performance on a variety of NLP benchmarks, including translation, question answering, and cloze tasks.

Introduction:

The field of Natural Language Processing has undergone a paradigm shift, transitioning from task-specific representations and architectures to a focus on task-agnostic pre-training and architectures. This shift has yielded significant advancements in various challenging NLP tasks.

While task-agnostic pre-training dominates the initial stages of model development, a final step of fine-tuning on task-specific datasets remains necessary to adapt the model for a specific NLP objective. However, recent research ("Language Models are Unsupervised Multitask Learners") suggests this fine-tuning step might be dispensable.

Furthermore, the study demonstrates the possibility of zero-shot transfer for standard NLP tasks using a single pre-trained language model, eliminating the need for dedicated training data. This work builds upon these findings by conducting an empirical evaluation to determine if scaling the model size further improves performance. By training a 175-billion parameter autoregressive language model, dubbed GPT-3, the authors investigate its transfer learning capabilities.



An analysis of performance on the SuperGLUE benchmark reveals a positive correlation with model size. BERT++, a fine-tuned variant of BERT-Large, achieves superior results despite utilizing a smaller dataset (630 examples) for fine-tuning compared to the original SuperGLUE training set (125k examples). These findings suggest that both increasing model capacity and leveraging larger datasets for fine-tuning contribute to enhanced performance on SuperGLUE tasks. Notably, zero-shot performance exhibits a consistent rise with model size, while few-shot performance demonstrates a more pronounced improvement. This observation suggests that larger models possess a greater aptitude for learning within context.

While the work presented in "Language Models are Unsupervised Multitask Learners" is framed as "zero-shot task transfer," their approach occasionally provides task examples within context. Due to the inclusion of these examples, which functionally serve as training data, these instances are more accurately characterized as "one-shot" or "few-shot" transfer learning. This observation suggests that one- and few-shot performance often surpasses true zero-shot performance. This aligns with the hypothesis that language models can be viewed as meta-learners, combining slow outer-loop learning via gradient descent with rapid "in-context" learning within the model's activation states.

Across various NLP tasks, GPT-3 demonstrates promising results in zero-shot, one-shot, and few-shot settings. In the few-shot setting, it even achieves performance competitive with, or occasionally exceeding, state-of-the-art models (despite them being fine-tuned). Generally, scaling the model capacity leads to relatively smooth performance improvement across all three settings. Notably, the gap between zero-shot, one-shot, and few-shot performance often widens with increasing model capacity, potentially suggesting that larger models possess a stronger meta-learning capability.

Fine-tuning (FT): This approach involves updating the weights of a pre-trained model using thousands of supervised labels specific to the desired NLP task. While a significant advantage of fine-tuning is its strong performance on various benchmarks, there are drawbacks. These include the requirement for a new large dataset for each task, potential

for poor generalization to unseen data (out-of-distribution), and susceptibility to exploiting irrelevant features in the training data (spurious correlations).

Few-shot (FS): This method leverages a pre-trained model at inference time. During this stage, the model receives a few task demonstrations as conditioning information (similar to [Language Models are Unsupervised Multitask Learners]), without any weight updates. The primary benefit of few-shot learning is the significantly reduced need for task-specific data. However, compared to state-of-the-art fine-tuned models, few-shot learning has typically yielded lower performance. Additionally, a small amount of task-specific data is still often required. It's important to note that few-shot learning in NLP aligns with the broader concept of few-shot learning in machine learning: both involve learning from a vast pool of tasks and rapidly adapting to new ones.

One-shot (1S) and Zero-shot (0S): These methods are similar to few-shot learning. However, one-shot learning provides only one demonstration, while zero-shot learning relies solely on a natural language description of the task, forgoing any examples.

While the paper highlights few-shot results with the highest performance, one-shot or even zero-shot approaches might be more comparable to human performance, and they represent important areas for future research.

Model and Architectures:

This section of the paper utilizes the same model and architecture as GPT-2. This includes the modified initialization, pre-normalization, and reversible tokenization techniques described in the original GPT-2 paper ([gpt2,]). One exception is the use of alternating dense and locally banded sparse attention patterns in the transformer layers, similar to the Sparse Transformer approach.

Training Considerations:

Batch Size and Learning Rate: Consistent with prior research, the authors observe that larger models benefit from increased batch sizes, but necessitate proportionally lower learning rates. To guide the selection of optimal batch size, they leverage the concept of gradient noise scale measured during training. Scalable Training: To accommodate the immense memory requirements of these large models, the training process employs a combination of model parallelism techniques. This involves distributing the workload across multiple devices: within each matrix multiplication operation (intra-layer parallelism) and across different network layers (inter-layer parallelism).

Results:

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Table 3.2: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

Setting	ARC (Easy)	ARC (Challenge)	CoQA	DROP
Fine-tuned SOTA	**92.0**[a]	**78.5**[b]	**90.7**[c]	**89.1**[d]
GPT-3 Zero-Shot	68.8	51.4	81.5	23.6
GPT-3 One-Shot	71.2	53.2	84.0	34.3
GPT-3 Few-Shot	70.1	51.5	85.0	36.5

Table 3.3: GPT-3 results on a selection of QA / RC tasks. CoQA and DROP are F1 while ARC reports accuracy. See the appendix for additional experiments. [a][KKS+20] [b][KKS+20] [c][JZC+19] [d][JN20]

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Table 3.4: Few-shot GPT-3 outperforms previous unsupervised NMT work by 5 BLEU when translating into English reflecting its strength as an English LM. We report BLEU scores on the WMT'14 Fr↔En, WMT'16 De↔En, and WMT'16 Ro↔En datasets as measured by `multi-bleu.perl` with XLM's tokenization in order to compare most closely with prior unsupervised NMT work. SacreBLEU[f] [Pos18] results reported in the appendix. Underline indicates an unsupervised or few-shot SOTA, bold indicates supervised SOTA with relative confidence. [a][EOAG18] [b][DHKH14] [c][WXH+18] [d][oR16] [e][LGG+20] [f][SacreBLEU signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.intl+version.1.2.20]

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.5: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

This work presents a 175-billion parameter language model demonstrating strong performance on various NLP tasks and benchmarks across zero-shot, one-shot, and few-shot settings. In some cases, it even approaches the performance of leading fine-tuned systems. Additionally, the model generates high-quality samples and exhibits impressive qualitative performance on dynamically defined tasks. The study reveals a predictable pattern of performance improvement with model size scaling, without the need for fine-tuning. Furthermore, the authors discuss the potential social impacts associated with this class of models. While acknowledging limitations and weaknesses, these findings suggest that very large language models could be a crucial component in developing versatile and general-purpose language systems.

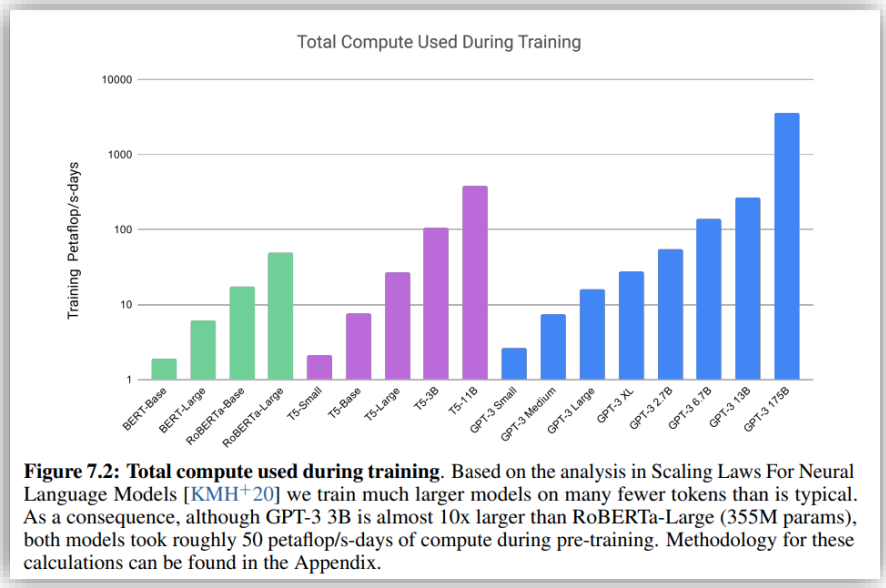


Figure 7.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in the Appendix.