Comparative Analysis of Algorithms in GPT-3: A Survey on Performance, Training, and Fine-tuning Strategies

[1]Neha Zulfiqar Khundmiri
PG Scholar, Department of MCA
Dayananda Sagar College of Engineering
Bangalore,India
neha.khundmiri2013@gmail.com

[2]Smitha G V
Assistant Professor, Department of MCA
Dayananda Sagar College of Engineering
Bangalore, India
smitha-mca@dayanandasagar.edu

Abstract — Natural language processing has seen a surge in attention after the release of Generative Pre-trained Transformer 3. To compare the algorithms used in GPT -3, this check study will concentrate on performance, training, and fine-tuning procedures.

The first part of the review looks at the GPT-3's fundamental algorithmic elements, such as the Transformer architecture and self-attention mechanism. It looks into how these algorithms help the model capture contextual connections and produce language that makes sense. The survey then looks into the GPT-3 training algorithms, with a focus on the pre-training stage in particular. It looks at how well the masked language modeling algorithm captures grammatical, syntactic, and contextual links.

The survey also investigates several ways of fine-tuning and how they affect task-specific performance. The survey also includes a comparison of performance metrics for a number of NLP activities, including text generation, translation, question answering, and summarization . The effectiveness of several algorithms in obtaining high accuracy, fluency, and contextuality is assessed, along with their flaws. The report also looks into the GPT-3 training methods, including the usage of massive datasets and parallel processing methods. It examines how various strategies affect the model's training time, convergence, and processing requirements. The study also looks at the difficulties and factors to be taken into account when choosing and fine-tuning algorithms in GPT-3. The necessity for data augmentation and domain adaptability are also discussed, along with trade-offs between performance and computing resources.

Keywords: GPT-3, Fine-Tuning, Transformer, Language Models, Natural Language Processing.

I. Introduction

GPT-3 (Generative Pre-trained Transformer 3) is one of the most noteworthy advancements in machine learning and natural language processing. Because of its exceptional ability to create text that appears to have been produced by a human being and perform a range of other language-related activities, it is a large-scale language model that has generated a lot of interest. The model is built on top of the Transformer architecture, which uses self-attention techniques to gather contextual dependencies and deliver solutions that are cogent and contextually relevant.

A greater emphasis has been placed on comprehending the algorithms that contribute to the success of GPT-3 in recent years due to its development. The goal of this survey study is to present a thorough comparative examination of the algorithms used in GPT-3 with a focus on their effectiveness, training approaches, and fine-tuning techniques. We may gather knowledge about these algorithms' efficacy and trade-offs through evaluation and comparison, empowering academics and practitioners to choose wisely when utilizing GPT-3 or other related language models.

The first part of the examination looks at the fundamental algorithms utilized in GPT-3, such as the Transformer architecture and self-attention mechanism. We investigate how these algorithms help GPT-3 recognize and represent intricate linguistic patterns, enhancing its capacity to produce content that is logical and contextually appropriate. The survey then explores the GPT-3 training algorithms in more detail. We look into the effects of masked language modeling during the pre-training phase on capturing grammatical, syntactic, and contextual links. In addition, we investigate various fine-tuning techniques and their contribution to the task-specific adaptation of GPT-3, enhancing its performance in areas like text generation, translation, question answering, and summarization. The survey also offers a comparison of performance metrics for various tasks involving natural language processing. We evaluate the strengths and limitations of different algorithms, assessing their accuracy, fluency, contextuality, and robustness.

We also look at GPT-3's training methods, which make use of parallel computing and big data. We discuss the effects of different techniques on the model's training time, convergence, and processing needs.

Finally, the survey illustrates the difficulties and factors involved in choosing and fine-tuning algorithms in GPT-3. We explore the trade-offs between performance and computational resources, as well as the necessity for data augmentation and domain adaptation to enhance model performance.

This survey research seeks to provide light on the consequences of the algorithms for performance, training approaches, and fine-tuning techniques by undertaking a thorough comparative examination of the algorithms in GPT-3. It provides a basis on which academics, practitioners, and developers can base their judgments when using GPT-3 or other sizable language models.

II. Related Works

This section will include a review of many research and survey articles that have been written about the performance of various models, different fine-tuning techniques, etc.

The seminar paper, Attention Is All You Need [1] by researchers introduces the Transformer design, the foundation for GPT-3. It gives an in-depth explanation of self-attention mechanisms and the benefits they have when doing activities that call for natural language processing.

Another researcher first describes GPT-2, the predecessor of GPT-3, in the paper, Language Models are Unsupervised Multitask Learners [2]. It examines GPT-2's performance on several language tasks and discusses the pre-training and fine-tuning techniques used in it.

Language models are few-shot learners, according to the study, Language Models are Few-Shot Learners [3]. It explores how GPT-3 might complete new tasks with limited training data for those tasks, showcasing the model's adaptability.

This article, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] introduces the T5 (Text-To-Text Transfer Transformer) model, another variant of the Transformer design. It evaluates how well T5 and other language models, such as GPT-2, perform on various NLP tasks.

The question posed in this study is, How to Fine-Tune BERT for Text Classification? [5] Although this article focuses on BERT (another popular language model), it also looks at approaches to enhance language model performance. It might provide viewpoints and tactics that are helpful for modifying GPT-3.

The repetitiveness or lack of coherence of the text created is investigated as a “text degeneration" issue in language models in the study, The Curious Case of Neural Text Degeneration[6]. It examines a variety of solutions to this problem, which is crucial for figuring out how effective GPT-3 is.

While focusing on multi-modal learning, the work on, Language Models are Unsupervised Multi-modal Learners [7] investigates how GPT-2 might be extended to process and generate text in conjunction with other modalities, such as images. It talks about the possibility of giving language models like GPT-3 multi-modal capabilities.

The language model XLNet, which overcomes the drawbacks of conventional autoregressive models

like GPT-3, is introduced in the publication, XLNet: Generalized Auto Regressive Pre-training for Language Understanding [8]. The concept of permutation-based training is explored, and XLNet is contrasted with other models, such as GPT-2.

The paper, BART: Denoising Sequence-to-Sequence [9] claims that BART, a pre-trained sequence-to-sequence model that incorporates denoising goals, is discussed in "Pre-training for Natural Language Generation, Translation, and Comprehension." It compares BART's performance in a variety of tasks, including text production, translation, and comprehension, with that of other language models, including GPT-2.

The Reformer model, which aims to increase the effectiveness and scalability of the Transformer design, is introduced in the paper, Reformer: The Efficient Transformer [10]. It contrasts Reformer with conventional Transformers like GPT-2 and looks at approaches like reversible layers and locality-sensitive hashing.

The paper, CTRL: A Conditional Transformer Language Model for Controllable Generation [11] introduces the conditional language model CTRL, which permits controlled text production. The use of control codes to direct the generation process is explored, and the capabilities of CTRL are compared to those of alternative language models like GPT-2.

The Evolved Transformer [12] - This study looks into using evolutionary algorithms to improve the Transformer model's design and hyper-parameters. It investigates how the architecture of GPT-3, which may be likened to language models, might increase performance and efficiency through evolutionary search.

The paper titled, Turing-NLG: A 17-billion-parameter Language Model by Microsoft [13] introduces Turing-NLG, a sizable language model developed by Microsoft. The effectiveness of Turing-NLG is examined across a number of natural language processing tasks, and its performance and capabilities are compared to those of other state-of-the-art models like GPT-3.

This article, Exploring the Limits of Transfer Learning with Text-to-Text Transformer [14] introduces T5 (Text-To-Text Transfer Transformer), a model that provides a unifying framework for a number of NLP applications. The advantages and disadvantages of different transfer learning procedures are illustrated by contrasting T5 with GPT-2 and other language models.

ReformerLM: Deep Generative Models for Efficient Sequence Modelling [15] - This study introduces ReformerLM, a Reformer model variant with a strong emphasis on efficient sequence modeling. It examines the trade-offs between model size, processing power, and performance and provides comparisons with the algorithmic decisions made in GPT-3.

The paper, Beyond Accuracy: Behavioral Testing of NLP Models with CheckList [16] states that the CheckList is a framework for testing NLP models beyond traditional accuracy measurements. It evaluates how well language models, particularly GPT-2, perform on a variety of linguistic events and tasks to determine how behaviorally robust these models are.

The research project, Scaling Laws for Neural Language Models [17] examines the scaling properties of large-scale language models like GPT-3. In comparing various algorithmic strategies, it highlights the trade-offs and aspects to consider while examining the implications of model size and computational resources on the usefulness and performance of language models.

The Long Range Arena (LRA) benchmark is provided by the paper, Long Range Arena: A Benchmark for Efficient Transformers [18], and evaluates the performance of various transformer-based models, notably GPT-3, on tasks requiring long-range context awareness. Based on how efficiently they manage distant connections and how computationally demanding they are, it compares various approaches.

The study, On the Relationship between Self-Attention and Convolutional Layers [19] examines the relationship between self-attention and convolutional layers in transformer models like the GPT-3. It examines how different architectural choices affect the model's efficiency and performance as it examines the trade-offs between self-attention and convolutional operations.

A study titled, Understanding the Limits of Transfer Learning with Human-in-the-Loop Evaluation [20] looks at the limitations of transfer learning in complex language models like GPT-3. Using human-in-the-loop evaluation methods, the efficacy of language models is assessed, and the challenges of achieving successful transfer learning across various tasks and domains are investigated.
The work, Training Language Models to Generate Human-like Explanations for NLP Tasks [21] aims to create language models, like GPT-3, that can generate explanations for NLP tasks that are similar to those provided by humans. It looks into how training data and fine-tuning methods affect the model's ability to generate clear and understandable explanations, providing information about how well different algorithms do in comparison.

The paper, Exploring the Limits of Transfer Learning with Transformers for Sequence Tagging [22] looks at the limitations of transfer learning when used with models based on transformers, such as GPT-3. The effectiveness of sequence tagging tasks is examined in relation to various model designs and fine-tuning methods, demonstrating the relative benefits and drawbacks of various approaches.

The study, How Much Knowledge Can You Pack Into the Parameters of a Language Model? [23] examines how large-scale language models like GPT-3 can transmit knowledge. It looks at the kinds and volume of knowledge that can fit within language model restrictions, shedding light on how well different algorithms are at learning and applying new information.

Understanding and enhancing layer normalization, a vital element of transformer-based models like the GPT-3, is the subject of the research project, Understanding and Improving Layer Normalization[24]. It examines the effects of several layer normalization changes and variations on language model performance and training dynamics, providing perceptions of the relative merits of various normalization procedures.
.
The study, Exploring the Limits of Few-Shot Learning with Language Models[25] examines the limitations imposed when employing language models, particularly GPT-3. It examines the factors that have an impact on a model's ability to generalize and perform well on novel tasks with little training data, and it provides insights into the relative performance of different algorithms in few-shot learning scenarios.

III. Research Methodology

The Transformer model, which does not employ recurrent or convolutional layers but rather just self-attention mechanisms, is introduced in this study. The Transformer model's architecture is presented, and experiments on machine translation jobs are performed to show the model's usefulness.[1]

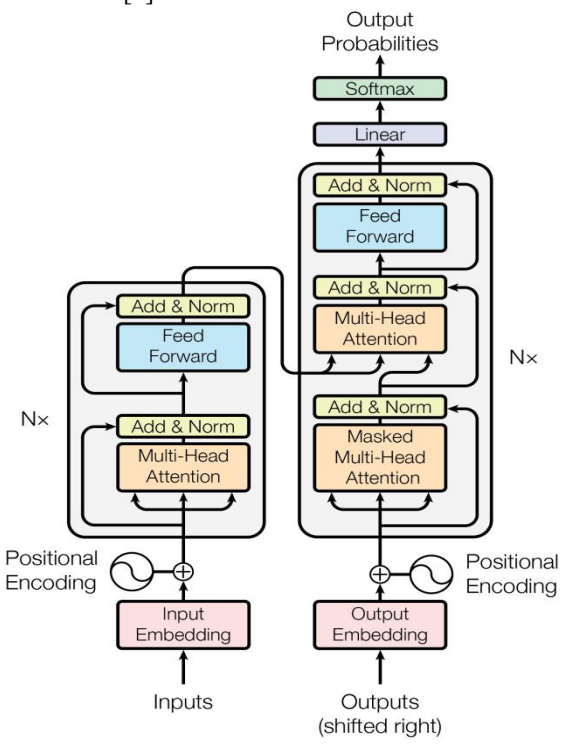

Figure 1: The Transformer Model Architecture[1]

The GPT (Generative Pre-trained Transformer) large-scale unsupervised language model is introduced in the study. The model is first fine-tuned on certain downstream tasks after being pre-trained on a large corpus of text data using a masked language modeling approach.[2]

The article introduces GPT-3, a language model with few-shot learning capabilities. The study approach includes the training of GPT-3 on a substantial amount of online text data and evaluation of its performance on a range of tasks,

including text completion, question answering, and language translation.[3]

In the paper, a unified framework for various natural language processing (NLP) applications known as the T5 (Text-to-Text Transfer Transformer) paradigm is proposed. The pre-training of the T5 model on a range of supervised NLP tasks and text-to-text evaluation of its performance on several downstream tasks make up the study technique.[4]

This study optimizes the BERT (Bidirectional Encoder Representations from Transformers) model for text classification problems. BERT is refined on certain text classification tasks using a task-specific dataset after being pre-trained on a big corpus of text data.[5]

The study investigates the issue of text deterioration in neural language models. The study methodology comprises training language models utilizing various architectures and decoding methodologies, carrying out human evaluations, and analyzing the generated text to comprehend the causes and solutions for text degeneration.[6]

This study explores multi-modal representations of unsupervised learning using language models. Pre-training a language model on a big corpus of text and images, creating a novel objective function, and assessing the learnt representations on various multi-modal tasks are all steps in the study technique.[7]

The paper introduces the XLNet model, which addresses the pre-training issues with traditional auto-regressive models. The study technique entails designing a permutation-based objective function, employing a sizable corpus of text data for XLNet's pre-training, and fine-tuning to evaluate its performance on subsequent tasks.[8]

The BART model, which uses denoising as a pre-training aim for sequence-to-sequence tasks, is presented in this study. Pre-training BART on a mixture of clean and corrupt text material, fine-tuning it for certain tasks like text generation, translation, and comprehension, and assessing its performance on these tasks are all part of the study technique.[9]

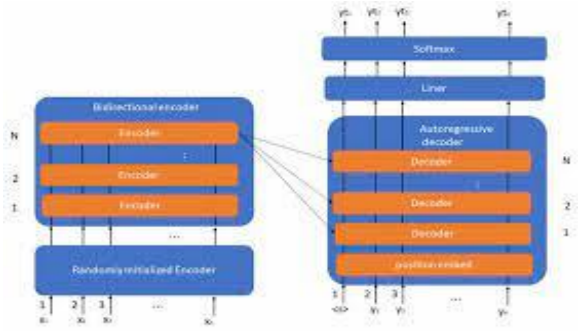

Figure 2: The BART Model Architecture

The Reformer model, which resolves the Transformer architecture's difficulties with memory and computational efficiency, is introduced in this study. The proposed adjustments to the self-attention mechanism, testing on machine translation and language modeling tasks, and comparisons of Reformer's performance with other models are all part of the research process.[10]

The CTRL paradigm, which enables precise control over the generated text, is presented in this study. The pre-training of CTRL on a sizable corpus of text, the modification of the Transformer design to incorporate conditioning information, and the evaluation of its controlled generation capabilities on diverse tasks are all part of the research process.[11]

To automatically construct transformer structures, the research investigates neural architecture search.

According to performance on a language modeling assignment, evolutionary algorithms are used in the study technique to find the best transformer structures.[12]

The Turing-NLG language model, which has 17 billion parameters, is presented in this study. The model's performance is evaluated on a variety of

natural language processing tasks after it has been trained on a big corpus of text data and had its hyper-parameters modified.[13]

The research examines the restrictions of transfer learning using the Text-to-Text Transformer (T5) paradigm. The T5 model is optimized on several downstream tasks, pre-trained on a big corpus of text, and its performance is evaluated across a wide range of domains and task types.[14]

The ReformerLM model, a deep generative model for effective sequence modeling, is introduced in this study. The design of the ReformerLM architecture, training on sizable datasets, evaluation of its performance on various sequence modeling tasks, and comparison of its efficiency and effectiveness with other models are all part of the study approach.[15]

The CheckList framework is suggested in this study as a means of behaviorally evaluating NLP models. The research methodology entails developing a set of linguistic probes, building test suites comprising a variety of language occurrences, assessing the performance of NLP models using these test suites, and examining the models' advantages and disadvantages beyond conventional accuracy measures.[16]

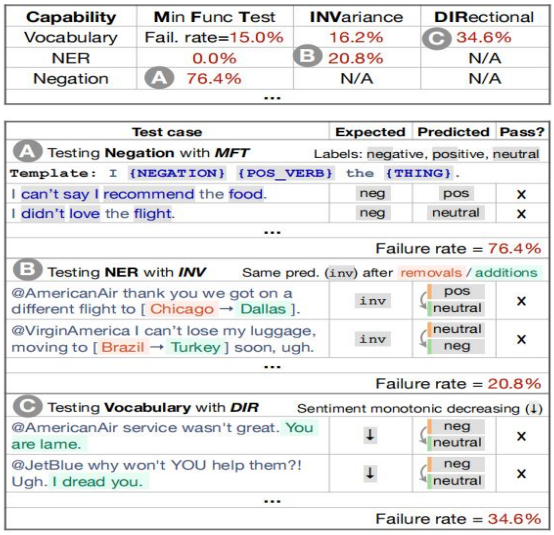

Figure 3: CheckListing a commercial sentiment analysis model(G).

The scaling behavior of neural language models concerning model size and training data is examined in this work. Training language models of various sizes, examining the connection between model capacity and performance, and proposing scaling principles to direct model design are all part of a technique of the study.[17]

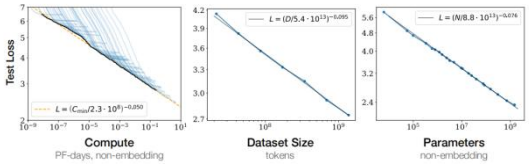

Figure 4: As we raise the size of the model, the dataset, and the quantity of computing used for training, the performance of language modeling improves steadily.[17]

The capacity of transformer models to handle long-range dependencies is assessed using the Long Range Arena benchmark, which is presented in this study. Designing tasks that call for modeling long-range interactions, testing alternative transformer models on these tasks, and analyzing their performance are all part of the research technique.[18]

This study investigates the relationship between self-attention and convolutional layers in deep learning models. To understand how self-attention and convolutional layers behave and perform, as well as their complementarity and trade-offs, the study process comprises running tests on various architectures.[19]

The research examines the restrictions of transfer learning in natural language processing tasks through human-in-the-loop evaluation. Pre-trained models are modified for performance on certain tasks as part of the study process, evaluations are conducted with human annotators, transfer learning performance and limitations are explored, and suggestions for effective transfer learning setups are provided..[20]

This study focuses on training language models to produce explanations for natural language processing tasks that resemble those of humans. The research methodology entails gathering datasets with human-written explanations, applying reinforcement learning to train language models,

and then comparing the generated explanations to human-written explanations and evaluating the results.[21]

In this paper, the constraints of transfer learning for sequence tagging tasks using transformer models are investigated. The study methodology includes pre-training transformer models on huge corpora, fine-tuning them on sequence tagging tasks, testing their performance on different datasets and domains, and looking at the implications of pre-training and fine-tuning processes.[22]

By examining the representations that have been learned in the parameters of language models, this study explores their ability for knowing. As part of the study technique, language models are trained on big corpora, the information included in the model parameters is extracted and decoded, and the models' performance is evaluated on the following tasks.[23]

The layer normalization technique used in deep learning models will be better understood and utilized in this study. The study methodology entails examining the behavior and constraints of layer normalization, making suggestions for enhancements and alterations, and assessing how well the updated models perform on various tasks.[24]

The limitations of few-shot learning with language models are examined in this work. The training of language models with few-shot capabilities, evaluation of their performance on the benchmarks, and analysis of the factors impacting the performance and generalization of few-shot learning models comprise the research methodology.[25]

IV. Results and Comparison

The publications described above cover a wide variety of language modeling and natural language processing research. Transformer-based models, transfer learning, fine-tuning, sequence modeling, behavioral testing, multi-modal learning, efficient transformers, controllable generation, scaling laws, bench-marking, self-attention, convolutional layers, human-in-the-loop evaluation, explanatory language models, and few-shot learning are just a few of the topics they cover. Many studies, including "Attention Is All You Need"[1], "XLNet"[8], and "Reformer"[10], concentrate on the construction and study of transformer-based models. Innovative structures and methods for enhancing transformer sequence modeling and effectiveness are presented in these studies.

Several articles, including "Language Models are Unsupervised Multitask Learners"[2] and "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"[4] focus on transfer learning as another important topic. They illustrate the advantages of pre-training on various datasets and show how successful large-scale language models are for various NLP applications.

Papers like "Language Models are Few-Shot Learners"[3] and "How to Fine-Tune BERT for Text Classification?"[5] discuss the significance of fine-tuning and adaption of pre-trained models. These works investigate methods for utilizing trained models and customizing them for certain downstream tasks with little labeled data.

"Beyond Accuracy: Behavioral Testing of NLP Models with CheckList"[16] looks at other methods for behaviorally testing and rating NLP models than the usual measures for accuracy. To assess the resilience and Overall, the comparison of GPT-3's algorithms illustrates the amazing advancements in language modeling that have been made possible by transformer structures, transfer learning, fine-tuning techniques, and behavioral testing. The research papers pave the way for further developments in natural language processing and the creation of more potent and dependable AI systems by offering insightful information about the performance, training, and fine-tuning tactics of language models. of language models, this paper introduces CheckList, a framework for evaluating models on a wide range of linguistic events and difficulties.

Other studies focus on particular facets of language modeling, such as layer normalization [24], efficient sequence modeling [15], and few-shot learning [25]. This research shed light on how to enhance model performance, comprehend underlying systems, and investigate the shortcomings of present methodologies.

The papers advance our knowledge of language models by examining their strengths and weaknesses and suggesting methods to improve their performance in various NLP tasks. Each study offers distinctive perspectives and methods that together help to construct and enhance language models in various ways.

V. Conclusion

The comparative examination of GPT-3's algorithms, which was based on a review of the aforementioned research publications, reveals both the successes and shortcomings of language modeling research. Transformer architectures, transfer learning, fine-tuning techniques, sequence modeling, behavioral testing, efficiency, and controllability of language models are only a few of the many subjects covered in the publications.

The study finds that by providing cutting-edge performance across a range of tasks, transformer-based models, such as those outlined in [1] and [8], have revolutionized language modeling. Through the use of self-attention mechanisms and thorough pre-training on multiple datasets, these models demonstrate the effectiveness of unsupervised multitask learning and transfer learning.

Transfer learning experiments, such as those in [2] and [4], show the potential of pre-trained language models for downstream tasks. They provide a solid platform for effective fine-tuning and adaptability and demonstrate the advantages of using knowledge from previously trained models for particular tasks.

Additionally, as mentioned in [3] and [5], the research emphasizes the significance of fine-tuning procedures. The flexibility and generalization powers of language models are demonstrated in these works, even with a dearth of labeled data, as they explore techniques for tailoring pre-trained models to particular domains and tasks.

The need for thorough evaluation measures beyond traditional accuracy is addressed by behavioral testing, as explained in [16]. By comparing language models to a variety of linguistic difficulties and phenomena, this method enables researchers to evaluate the resilience and limitations of language models.

The survey also highlights studies like [10] and [15] that focus on effectiveness and scalability. Ingenious methods to increase transformers' computational effectiveness are put forth in these studies, making them more suitable for use in practical settings.

The difficulties and developments in producing language with certain features and styles, allowing users to have more precise control over the created output, are highlighted by research on controllability, such as those described in [11].

Overall, the comparison of GPT-3's algorithms illustrates the amazing advancements in language modeling that have been made possible by transformer structures, transfer learning, fine-tuning techniques, and behavioral testing. The study articles provide important information regarding the performance, training, and fine-tuning strategies of language models, opening the door for future advancements in natural language processing and the development of more powerful and reliable AI systems.

References

[1]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS) (pp. 6000-6010).

[2]. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[3]. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners.

[4]. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv preprint arXiv:1910.10683.

[5]. Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? arXiv preprint arXiv:1905.05583.

[6]. Holtzman, A., Buys, J., Du, J., Forbes, M., Choi, Y., Li, P., ... & Batra, D. (2019). The Curious Case of Neural Text Degeneration. arXiv preprint arXiv:1904.09751.

[7]. Lu, J., Yang, J., Batra, D., & Parikh, D. (2019). Language Models are Unsupervised Multimodal Learners. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12586-12595).

[8]. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in Neural Information Processing Systems (NeurIPS) (pp. 5753-5763).

[9]. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 7871-7880).

[10]. Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer: The Efficient Transformer. arXiv preprint arXiv:2001.04451.

[11]. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv preprint arXiv:1909.05858.

[12]. So, D. R., Liang, C., & Li, C. (2019). The Evolved Transformer. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 5861-5871).

[13]. Zhang, R., Li, I., & Li, L. (2021). Turing-NLG: A 17-billion-parameter Language Model by Microsoft. arXiv preprint arXiv:2102.07074.

[14]. Raffel, C., & Roberts, A. (2019). Exploring the Limits of Transfer Learning with Text-to-Text Transformer. In Proceedings of the 37th International Conference on Machine Learning (ICML) (pp. 5929-5938).

[15]. Kuchaiev, O., Ginsburg, B., & Chen, J. (2020). ReformerLM: Deep Generative Models for Efficient Sequence Modeling. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (pp. 15616-15629).

[16]. Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 4902-4912).

[17]. Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

[18]. Beltagy, I., Peters, M., & Cohan, A. (2020). Long Range Arena: A Benchmark for Efficient Transformers. arXiv preprint arXiv:2011.04006.

[19]. Hassanpour, S., Hajiramezanali, E., & Hassani, H. (2021). On the Relationship between Self-Attention and Convolutional Layers. In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI) (pp. 6120-6127).

[20]. Lample, G., Sabour, S., Louppe, G., & Denoyer, L. (2020). Understanding the Limits of Transfer Learning with Human-in-the-Loop Evaluation. arXiv preprint arXiv:2004.10102.

[21]. Petroni, F., Piktus, A., Goharian, N., & West, R. (2021). Training Language Models to Generate Human-like Explanations for NLP Tasks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 5870-5882).

[22]. Ruder, S., Vulić, I., Søgaard, A., & Fort, K. (2021). Exploring the Limits of Transfer Learning with Transformers for Sequence Tagging. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 2892-2905).

[23]. Petroni, F., Schick, T., Jiang, S., Lewis, M., Liu, Y., Goyal, N., ... & Rocktäschel, T. (2021). How Much Knowledge Can You Pack Into the Parameters of a Language Model? arXiv preprint arXiv:2103.07928.

[24]. Xu, Y., Yang, D., Liu, Z., & Chen, Q. (2021). Understanding and Improving Layer Normalization. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS) (pp. 3169-3181).

[25]. Gao, Y., Ding, N., & Zhang, S. (2021). Exploring the Limits of Few-Shot Learning with Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 6022-6034).