Towards the Future of GPT-3: A Examination of Limitations and Potential Research Directions

Qumeng SUN

August 1, 2023

1 Introduction

Language modeling is an unsupervised task that involves generating a probabilistic model for natural language. This is typically done by factorizing the joint-probability into conditional probabilities of the succeeding token given its preceding context or next-token predictions [30, 8, 41].

Neural networks, first applied to language modeling by Bengio et al. [6], have notably improved performance, especially when paired with recurrent strategy [25, 33, 36]. The emergence of **Transformer** models [46] using **attention mechanisms** [4] marks a significant step forward, solving the gradient problem faced by **Recurrent Neural Networks** (RNNs). Performance has further been boosted in **Large Language Models** (LLMs) like GPT-3 by scaling up data, compute resources, and model parameters [9]. GPT-3 remarkably demonstrated comprehension and multi-tasking abilities without needing fine-tuning [20].

GPT-3, a series of large language models released by OpenAI in 2017, is based on the Transformer architecture. The largest model in the series boasts a staggering 175 billion parameters, requiring at least $4,600,000 to train [35]. In alignment with other models in the GPT series, it exclusively employs the Decoder portion of the Transformer, operating as an autoregressive model that uses word embeddings as inputs and generates the probability of the next word as output. The model architecture is mainly consistent with that of GPT-2. As proposed by Brown et al. [9] in their paper, the model's performance on a variety of downstream tasks can be enhanced without updating the model parameters by using a few-shot learning strategy.

This paper thoroughly examines the GPT-3 technical report, with the main contributions as follows:

- A comprehensive **architecture diagram** of GPT-3 has been drawn by reviewing and summarizing relevant literature, providing readers with a more intuitive understanding of the model's structure.
- The analysis of GPT-3 from both the model and paper perspectives reveals several **shortcomings**.

- The paper identifies potential **future research directions** for GPT-3 and provides specific research examples.

The focus of this paper is to evaluate the paper GPT-3 in section Section 4. Before that, the paper will describe the architecture of GPT-3 in detail and briefly show which problems it performs better in sectionSection 2 and Section 3. Finally, the paper will look at the weaknesses of GPT-3 and point out potential future research directions in sectionSection 5.

2 Architecture

This section details the GPT-3 model architecture, covering inputs, special transformers, and outputs.

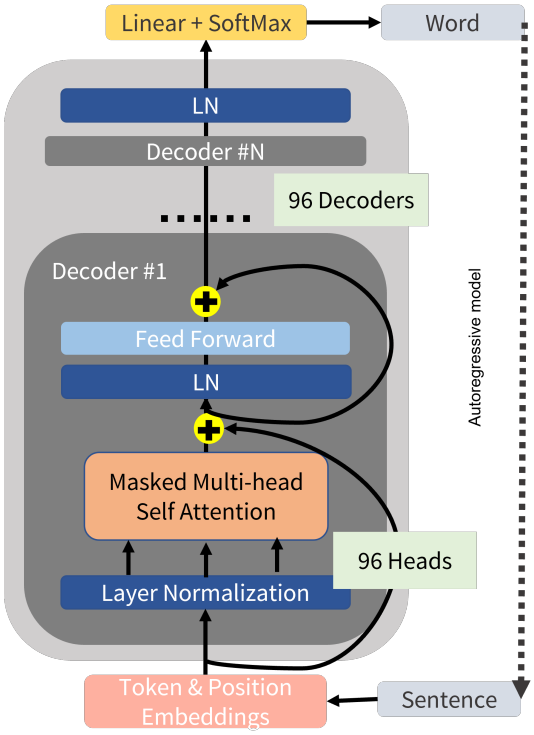

Figure 1: **GPT-3**: In the largest model in GPT-3, the number of Decoders, N, is 96, and the number of heads for the multi-attention mechanism is 96 with a dimension of 128. Multiplying and summing the two yields the model parameter $d_{model} = 12288$.

2.1 Input Embedding

GPT-3's input comprises fixed-length embeddings of shape (bs, t, d_{model}), where t represents the input sequence length (number of tokens) and must satisfy $t \leq n_{ctx}$. This refers to a collection of embeddings with a batch size dimension, a width of 't', and length defined by the d_{model} hyperparameter indicating the length of an individual embedding. These embeddings contain the semantic and positional data of previous words (see Figure 2).

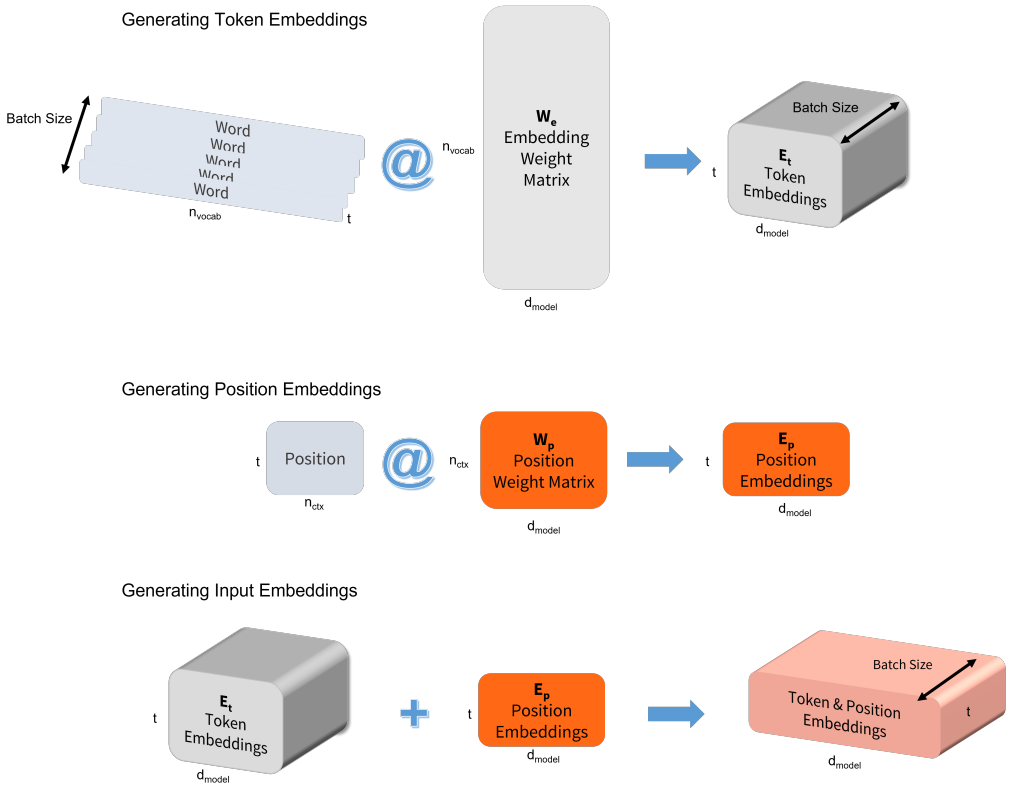

Figure 2: **Input**: The figure illustrates the process of generating the model input embeddings. The final embedding is obtained by summing the word embedding and position embedding tensor. Where t is the length of the input sequence (the number of tokens), the model hyperparameters are: Batch size = 3.2M, $n_{ctx} = 2048$, $d_{model} = 12288$, $n_{vocab} = 50257$ [9, 41]

GPT-3 tokenizes using Byte Pair Encoding (BPE)[21]. For example, *There is no life outside Goettingen* is tokenized into [There, is, no, life, outside, Go, etting, en], with IDs [1858, 318, 645, 1204, 2354, 1514, 35463, 268].

2.2 Transformer

Since their inception, Transformers have dominated multiple NLP tasks due to their enhanced parallelism over widely-used RNNs [52, 7].

GPT-3 employs the decoder portion of the Transformer as shown in Figure 1. Using just half of a Transformer in an autoregressive mode makes it behave like an RNN, but without suffering from vanishing or exploding gradients. Each attention block and feed-forward neural network (FNN, or multilayer perceptrons) in the decoder has residual connections[23, 27], indicated by yellow plus symbols in the diagram. Residual connections, also known as skip connections, are where the output of one layer is added to the output of a later layer, effectively allowing the model to learn an identity function and mitigate the problem of vanishing gradients[51].

The masked multi-head attention block in the model uses two attention patterns: dense and locally banded sparse [9, 13]. This reduces computational complexity and memory requirements,

enabling the handling of longer sequences and becoming a canonical method for efficient attention computation [43].

In essence, GPT is an autoregressive probabilistic language model leveraging attention mechanisms, residual connections [27], and Layer Normalization [3]. During training, it uses teacher forcing, compelling the model to take the actual next token as the next input instead of the model's predicted token [48, 24, 15].

2.3 Model Output

GPT generates a token as output, which is then looped back into the model for autoregression until the model decides to stop generating, as shown in Figure 3.

For example, given the input "Even if it is life", GPT-3 will calculate the occurrence probability of the next token on the whole vocabulary list, and then select the desired token by sampling strategy. Assuming the model outputs "it", the phrase "Even if it is life, it" is then used as the new input to generate the next token. This process repeats until the output sentence "it is no life like here" is formed.

The sampling strategy employed in the GPT-2 paper is top-k random sampling, which limits the pool of potential words for sampling to the top k predicted words [41, 18]. For instance, when the top-k parameter is 1, we always select the most probable word.

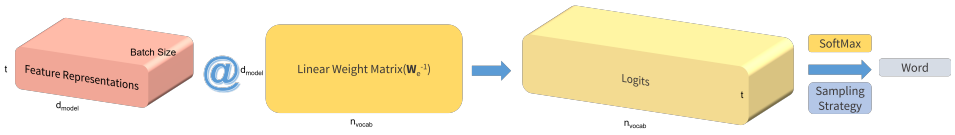

Figure 3: **Output**

3 Performance

This section describes the main advantages and disadvantages of GPT-3.

Meta-learning One of GPT-3's most intriguing capabilities is its in-context learning[1]. This allows it to rapidly adapt to new concepts from a few examples, without requiring explicit model gradient updates[11].

Such a learning method is manifested as Few-Shot learning in the paper. To employ this method, a handful of examples are directly integrated into the input. Consider a scenario where we want GPT-3 to execute a translation task - a completely new task for it, without prior specific training. Conventionally, we would employ a pre-trained model such as BERT-Large [16], fine-tune it on a specific translation dataset, update the base parameters optionally via gradient updates, train an additional output layer, and finally apply this fine-tuned model to the translation task[51].

In contrast, with GPT-3, we can straightforwardly input the following:

```
1) Translate the following Latin phrases to English:
2) Bonum mane -> Good morning
3) In publica commoda -> For the good of all
4) Extra Gottingam non est vita, si est vita non est ita. ->
```

Adding examples and the new task description into the conditions as shown by

$$p(\text{output} \mid \text{input, task, example}_1, \text{example}_2, \text{example}_3, \ldots)$$

The paper identifies the gradient updates during pre-training and in-context learning as two distinct yet cooperative processes. The interplay of these elements encapsulates the essence of meta-learning.

However, the root of this unique ability remains a subject of debate[22]. Some researchers propose its origin to be the distribution of the training data[11], while others attribute it to the design of the transformer models and the gradient descent optimization process[1, 14, 39].

Results In Table 1, GPT-3 excels in next-word prediction and English translation tasks, matching or exceeding state-of-the-art (SOTA) performance. For other tasks, while not reaching SOTA, GPT-3, without fine-tuning, shows impressive capabilities. Key observations include:

- **More data results in better performance**: Translation, logic reasoning, and arithmetic tasks show better results when the outputs resemble the training data. Conversely, GPT-3 is less sensitive to the nature of the input data. In translation tasks, for example, the quality of results is quite similar irrespective of the input language when the output text is in English. However, if the output language is less frequent in the dataset and the input is in English, there is a noticeable disparity in result quality.
- **GPT-3 struggles with generating long texts**: A series of tests on various QA datasets reveals that GPT-3 performs worse on the Natural QS dataset, which requires longer answers, compared to the Trivia QA dataset.
- **GPT-3 exhibits limited comprehension of context**: GPT-3's performance on the DROP and QuAC tasks is over 40% lower than SOTA models. In these particular tasks, the model needs to read lengthy texts, understand the context, and remember previous questions and answers. We attribute this limitation to the restricted context window size and propose an increased focus on key information on a larger scale.
- **GPT-3 fails to accomplish complex reasoning and arithmetic**: GPT-3's poor performance in reasoning tasks may be due to its inability to understand complex issues or execute tasks that require a deeper understanding. Additionally, GPT-3's lack of exposure to the physical world and even audio and video makes its limited understanding of the natural world reasonable. This could be a contributing factor to GPT-3's underperformance in tasks requiring graphical understanding and worldly knowledge, such as ARC and PIQA.

It's important to clarify that GPT-3 is mostly compared with supervised and fine-tuned models in most tasks. However, GPT-3 itself is trained unsupervised. While this comparison may not seem entirely fair, it is meaningful considering that the SOTA models it is compared with are at least an order of magnitude smaller than GPT-3.

Table 1: Comparison of results of different types of tasks with SOTA.

Task Types	Comparison to SOTA
Classic Language Modeling	↑ 27%
Translate to English	↑ 5%
QA Tasks	↓ 33%, ↑ 5%
Continuation of a story	↓ 6%
Logical Reasoning	↓ 20%
Translate to Other Languages	↓ 34%
Reading Comprehension	↓ 36%

4 Shortcomings

4.1 Paper Side

This section discusses some practical issues of GPT-3.

The Bug GPT-3 utilizes a large quantity of internet data, primarily the Common Crawl data from 2016 to 2019, along with several other databases totaling 570GB[9], constitutes the training data for GPT-3, and was trained unsupervised on only one task, language modeling. With such a massive database and vast parameter count, data leakage, where test data is prematurely exposed to the model during training, can easily occur. To mitigate this, the authors could have removed the test set from the training set, avoiding claims that the promising results were due to data leakage. However, due to a bug, they didn't manage to remove all overlapping data in time, resulting in over 90% of downstream tasks being contaminated. This led to a whole section discussing whether data contamination significantly impacted the model's performance.

LLMs as few-shot learners The original paper presents three strategies: Zero-Shot, where the model is given no examples and is directly asked to perform the task; One-Shot, where the model is given one example, then asked to perform the task; and Few-Shot, where the model is given more than one example, then asked to perform the task.

This paper argue that only large-scale language models can benefit from the Few-Shot learning strategy. For smaller models, Few-Shot learning may even impair their performance, making it the worst-performing strategy for smaller models and the best-performing one for larger models.

In the GPT-3 paper, eight figures depict the accuracy relationship between these three strategies, with five reflecting the phenomenon mentioned above.

Writing Interestingly, in a 41-page paper, the authors did not discuss the model's details or provide a structure diagram. They assume the audience has already read their old papers rather than providing a brief introduction to the method they are using. This gives the paper the appearance of a technical report showcasing model performance to an audience with a specialized knowledge background.

Additionally, there are some implicit logical issues. For instance, when discussing the data contamination of PIQA, they claim that this dataset was published after they trained the model, suggesting that the model's performance would not be affected, which is not the case.

Closed AI GPT-3 has not been open-sourced by OpenAI, and it remains a closed-source model as of the time of writing. People can only access the API of its successor, GPT-4, without touching the model itself. Meanwhile, some open-source large models, such as the LLaMA series [45, 44], offer performance comparable to GPT.

4.2 Model Side

This subsection discusses several GPT-3 issues in the model itself.

Data and Model The evolution from GPT-2 to GPT-3 extends beyond the mere augmentation of model parameters. It also encompasses a significant boost in the volume of training data. This leap forward has substantially upgraded its capabilities, enhancing its performance across various tasks beyond the anticipated scaling laws [52]. Moreover, GPT-3 has showcased emergent competencies in understanding and multitasking without the necessity for fine-tuning.

Yet, the escalation in data volume and model parameters introduces substantial challenges. The massive resource consumption required for such scaling culminates in high operational expenses and potential environmental impacts. The excessive demand for computational resources could lead to a monopolistic AI landscape, where only a select few corporations possess the requisite resources for large-scale AI training[37].

Such circumstances underline the importance of adhering to ethical and responsible practices in data collection and usage, as well as the training and deployment of large language models (LLMs). The primary focus should be on minimizing costs and environmental impacts [5]. In weighing environmental considerations, it's critical to account for the model's social and environmental benefits, energy consumption, carbon emissions, and indirect effects such as future energy demand, desertification, ecological diversity reduction, and pollution [7].

The financial implications of model scaling are also evident in storage and computational costs. With a whopping 175 billion parameters, GPT-3 necessitates significant resources for operation. This paper proposes that GPT-3 is considerably redundant, suggesting that a large portion of its parameters are ineffective. Research by Hsieh et al. [28] indicates that a model one two-thousandth the size could potentially deliver the same performance. Furthermore, the method introduced by Frantar and Alistarh [19] enables the elimination of 100 billion parameters post-training without any functional loss.

Lastly, the risk of low-quality data is a significant concern for model performance. As Gunasekar et al. [26] points out, contemporary databases are rife with noise, ambiguity, and incompleteness, which can obstruct learning processes—even for humans. It's therefore crucial to uphold stringent data quality standards, given that a model's performance is closely linked to the quality of its input.

Hallucination Hallucination is an anomaly observed in AI models, specifically text-generation models like GPT-3, where the output does not align with the input data or lacks meaningful relevance. These generated texts can seem nonsensical or disconnected from the given source input. Hallucination is not an uncommon phenomenon, occurring frequently in GPT-3 and models of its kind [31].

Recent studies, such as those led by Jin and Rinard [32], argue that these language models demonstrate a foundational ability to understand language and context. They cite instances where AI models were able to produce shorter, yet precise and accurate answers beyond their input dataset during various tasks. However, the primary argument in this paper is that models like GPT-3, essentially designed as next-token predictors, do not truly form their own understanding or comprehension of the world.

What these models demonstrate, in delivering improved answers, is not understanding but a manifestation of the creativity inherent in them. This creative element is not unique to GPT-3; all preceding generative models exhibit some degree of creativity. They generate novel constructs, like denoised images, that do not exist per se. It is worth considering that this inbuilt creativity may inadvertently lead to the occurrence of hallucination [31].

One significant area where GPT-3 and similar models struggle is with long texts and comprehension of extended contexts. Given the maximum input limit of GPT-3 at 2048 tokens, understanding the comprehensive meaning of long-form text, even when segmented, proves to be a considerable challenge.

This paper posits that one possible cause for this difficulty lies in the unidirectional nature of these autoregressive models. This structure may inhibit their ability to retain contextual information throughout a conversation. Fundamentally, these models lack the essential capability to loop back to prior text segments to review and use relevant information.

Thus, it is clear that hallucination in AI models like GPT-3 is not solely attributable to poor-quality input data. In fact, certain inherent characteristics of the model itself contribute significantly to the emergence of hallucination. Understanding and addressing these intrinsic features could be a significant step towards mitigating this issue.

5 Future Progress

In this section, this paper will list some potential research directions for GPT-3 to address the various shortcomings mentioned in the previous section.

5.1 Data

As mentioned earlier, we can put some effort into finding better databases. Improving the quality of training data can significantly reduce the required data size and computational consumption. Further, it can reduce environmental damage. There are some research contributions in this direction. Gunasekar et al. [26] proposed a new large language model **phi-1**. It is characterized by pre-training using high-quality textbook-like databases, and then fine-tuning with textbook-like problem data. This approach is similar to human feedback proposed by [40], where a lot of manpower is invested to optimize the machine. **phi-1** outperformed almost all models in HumanEval and MBPP.

In addition to building new databases, filtering duplicate elements in the database can also improve model performance and even reduce hallucinations[34]. Similarly, Cao, Kang, and Sun [10] proposed a method to perform high-quality sampling of instruction sets, and tried to fine-tune the model with a small amount of high-quality instructions after screening[49], and also achieved performance exceeding regular sampling strategies.

5.2 Long-context information

GPT-3 has a maximum context window of only 2048, which makes it ineffective in dealing with long texts and information in historical conversations. This indirectly leads to hallucinations, severely affecting the user experience of LLMs. To address this issue, Wang et al. [47] proposed a new method **LongMem**, which allows the model to remember a longer context and can handle 31 times as many tokens as GPT-3 at once. This paper believes that similar methods have the potential to further expand the impact of LLMs like GPT-3, as they address one of the largest pain points of GPT-3.

5.3 Physical Interact

Since GPT-3 can only accept text input, its understanding of the world is partial. GPT-3's successor, GPT-4, is a multimodal large language model[38], and there are several other multimodal LLMs on the market[12]. This can help the model better understand world knowledge in a user-friendly way and complete a wider range of tasks more comprehensively [50].

In addition to multimodal models, embodied AI is another direction that combines physical interaction with LLMs[17]. The goal of embodied artificial intelligence is to create agents, such as robots, that can learn to solve challenging tasks requiring interaction with the environment in creative ways. These agents are capable of accomplishing a variety of tasks in the real world through seeing, conversing, listening, acting, and reasoning. **ViperGPT** uses LLMs and Python for compliant visual reasoning[42]. Further, Huang et al. [29]'s **VoxPoser** pushes the visual reasoning ability into the real world.**VoxPoser** can convert robot manipulation tasks from free-form language instructions into robot trajectories. This is achieved by observing how large language models (LLMs) infer the affordances and constraints of manipulation tasks. By utilizing these models' encoding writing abilities, they can interact with visual language models (VLMs) to ground knowledge into the agent's observation space. Then, these composite value maps are used in a model-based planning framework to synthesize closed-loop robot trajectories that are robust to dynamic disturbances in a Zero-Shot way.

5.4 Retrieval-based Model

Due to the significant improvement in model performance brought about by expanding the amount of data, people tend to train larger models, resulting in increasingly high training costs. Retrieval-based language models are a type of language model that expands size by connecting to an (extremely large) external database, rather than through more parameters or larger training databases. The **RETRO** model proposed by Borgeaud et al. [8] uses an external database of 1.75 trillion tokens, and finds similar text for reading in the external database by comparing similarities, which does not significantly increase computational load compared

to traditional methods. Since additional information outside the model can be accessed, this enables repeated and selective reading, which helps to resolve the hallucination phenomenon that GPT-3 exhibits when dealing with long texts [31].

According to Asai et al. [2], models similar to **RETRO** have the potential to be combined with multimodal models, and are particularly suitable for scenarios requiring knowledge updates and involving privacy security. Its greatest advantage lies in its relatively low computational cost, which is replaced by the storage cost of hardware storage devices. However, judging from the current market, the price of storage devices is relatively cheap.

6 Conclusion

In summary, GPT-3 is a vast, attention mechanism-based, autoregressive probabilistic language model. There are numerous areas for improvement both in its construction and in its documentation. Regarding the paper itself, it falls short in providing an overarching view of the model's architecture, fails to delineate clearly the effects of the Few-shot strategy on the model's learning, and makes significant errors during model training. As for the model, the main issues emanate from its large size leading to memory constraints, unavoidable data overlap, and ethical concerns linked to the data utilized. This paper also outlines several key directions for the improvement of large language models, such as enhancing the quality of data, the use of external databases, and the incorporation of physical interactions.

Acknowledgements

This report was prepared as a course assignment for the seminar: Topics in Machine Learning and Computational Neuroscience. I extend my gratitude to everyone who participated in the seminar section for their patience and invaluable feedback. Special thanks go to Michaela Vystrcilová for her kind suggestions regarding the outline of this paper.

References

[1] Ekin Akyürek et al. "What learning algorithm is in-context learning? Investigations with linear models". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=0g0X4H8yN4I.

[2] Akari Asai et al. "ACL 2023 Tutorial: Retrieval-based LMs and Applications". In: *ACL 2023* (2023).

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization*. 2016. arXiv: 1607.06450 [stat.ML].

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473 (2014). URL: https://api.semanticscholar.org/CorpusID:11212020.

[5] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

[6] Yoshua Bengio et al. "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435.

[7] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". In: *CoRR* abs/2108.07258 (2021). arXiv: 2108.07258. URL: https://arxiv.org/abs/2108.07258.

[8] Sebastian Borgeaud et al. "Improving language models by retrieving from trillions of tokens". In: *CoRR* abs/2112.04426 (2021). arXiv: 2112.04426. URL: https://arxiv.org/abs/2112.04426.

[9] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165.

[10] Yihan Cao, Yanbin Kang, and Lichao Sun. *Instruction Mining: High-Quality Instruction Data Selection for Large Language Models*. 2023. arXiv: 2307.06290 [cs.CL].

[11] Stephanie C. Y. Chan et al. *Data Distributional Properties Drive Emergent In-Context Learning in Transformers*. 2022. arXiv: 2205.05055 [cs.LG].

[12] Feilong Chen et al. *X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages*. 2023. arXiv: 2305.04160 [cs.CL].

[13] Rewon Child et al. "Generating Long Sequences with Sparse Transformers". In: *CoRR* abs/1904.10509 (2019). arXiv: 1904.10509. URL: http://arxiv.org/abs/1904.10509.

[14] Damai Dai et al. *Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers*. 2023. arXiv: 2212.10559 [cs.CL].

[15] Jurafsky Daniel and James Martin. *Speech and Language Processing*. Jan. 2023. URL: https://web.stanford.edu/~jurafsky/slp3/10.pdf.

[16] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[17] Jiafei Duan et al. "A Survey of Embodied AI: From Simulators to Research Tasks". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 6.2 (2022), pp. 230–244. DOI: 10.1109/TETCI.2022.3141105.

[18] Angela Fan, Mike Lewis, and Yann N. Dauphin. "Hierarchical Neural Story Generation". In: *CoRR* abs/1805.04833 (2018). arXiv: 1805.04833. URL: http://arxiv.org/abs/1805.04833.

[19] Elias Frantar and Dan Alistarh. *SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot*. 2023. arXiv: 2301.00774 [cs.LG].

[20] Hao Fu Yao; Peng and Tushar Khot. "How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources". In: *Yao Fu's Notion* (Dec. 2022). URL: https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc

[21] Matthias Gallé. "Investigating the Effectiveness of BPE: The Power of Shorter Sequences". 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1375–1381. DOI: 10.18653/v1/D19-1141. URL: https://aclanthology.org/D19-1141.

[22] Shivam Garg et al. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 30583–30598. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c529dba08a146ea8d6cf715ae8930cbe-Paper-Conference.pdf.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[24] Anirudh Goyal et al. "Professor Forcing: A New Algorithm for Training Recurrent Networks". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 4608–4616. ISBN: 9781510838819.

[25] Alex Graves. "Generating Sequences With Recurrent Neural Networks". In: *CoRR* abs/1308.0850 (2013). arXiv: 1308.0850. URL: http://arxiv.org/abs/1308.0850.

[26] Suriya Gunasekar et al. *Textbooks Are All You Need*. 2023. arXiv: 2306.11644 [cs.CL].

[27] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[28] Cheng-Yu Hsieh et al. *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. 2023. arXiv: 2305.02301 [cs.CL].

[29] Wenlong Huang et al. *VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models*. 2023. arXiv: 2307.05973 [cs.RO].

[30] Fred Jelinek and Robert L. Mercer. "Interpolated estimation of Markov source parameters from sparse data". In: *Proceedings, Workshop on Pattern Recognition in Practice*. Ed. by Edzard S. Gelsema and Laveen N. Kanal. Amsterdam: North Holland, 1980, pp. 381–397.

[31] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *CoRR* abs/2202.03629 (2022). arXiv: 2202.03629. URL: https://arxiv.org/abs/2202.03629.

[32] Charles Jin and Martin Rinard. *Evidence of Meaning in Language Models Trained on Programs*. 2023. arXiv: 2305.11169 [cs.LG].

[33] Rafal Józefowicz et al. "Exploring the Limits of Language Modeling". In: *CoRR* abs/1602.02410 (2016). arXiv: 1602.02410. URL: http://arxiv.org/abs/1602.02410.

[34] Katherine Lee et al. "Deduplicating Training Data Makes Language Models Better". In: *CoRR* abs/2107.06499 (2021). arXiv: 2107.06499. URL: https://arxiv.org/abs/2107.06499.

[35] Chuan Li. *OpenAI's GPT-3 Language Model: A Technical Overview*. June 2020. URL: https://lambdalabs.com/blog/demystifying-gpt-3#1 (visited on 07/30/2023).

[36] Tomas Mikolov et al. "Recurrent neural network based language model." In: *Interspeech*. Vol. 2. 3. Makuhari. 2010, pp. 1045–1048.

[37] Gabriel Axel Montes and Ben Goertzel. "Distributed, decentralized, and democratized artificial intelligence". In: *Technological Forecasting and Social Change* 141 (2019), pp. 354–358. ISSN: 0040-1625. DOI: https://doi.org/10.1016/j.techfore.2018.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0040162518302920.

[38] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[39] Johannes von Oswald et al. "Transformers learn in-context by gradient descent". In: *ArXiv* abs/2212.07677 (2022). URL: https://api.semanticscholar.org/CorpusID:254685643.

[40] Long Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL].

[41] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[42] Dídac Surís, Sachit Menon, and Carl Vondrick. *ViperGPT: Visual Inference via Python Execution for Reasoning*. 2023. arXiv: 2303.08128 [cs.CV].

[43] Yi Tay et al. "Sparse Sinkhorn Attention". In: *CoRR* abs/2002.11296 (2020). arXiv: 2002.11296. URL: https://arxiv.org/abs/2002.11296.

[44] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[45] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[46] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[47] Weizhi Wang et al. *Augmenting Language Models with Long-Term Memory*. 2023. arXiv: 2306.07174 [cs.CL].

[48] Ronald J. Williams and David Zipser. "A Learning Algorithm for Continually Running Fully Recurrent Neural Networks". In: *Neural Computation* 1.2 (1989), pp. 270–280. DOI: 10.1162/neco.1989.1.2.270.

[49] Can Xu et al. *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. 2023. arXiv: 2304.12244 [cs.CL].

[50] Shukang Yin et al. *A Survey on Multimodal Large Language Models*. 2023. arXiv: 2306.13549 [cs.CV].

[51] Aston Zhang et al. "Dive into Deep Learning". In: *arXiv preprint arXiv:2106.11342* (2021).

[52] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL].