**Transformer Architecture**

* [**Attention is All You Need**](https://arxiv.org/pdf/1706.03762) - This paper introduced the Transformer architecture, with the core “self-attention” mechanism. This article was the foundation for LLMs.
* [**BLOOM: BigScience 176B Model**](https://arxiv.org/abs/2211.05100) - BLOOM is a open-source LLM with 176B parameters (similar to GPT-4) trained in an open and transparent way. In this paper, the authors present a detailed discussion of the dataset and process used to train the model. You can also see a high-level overview of the model [here](https://bigscience.notion.site/BLOOM-BigScience-176B-Model-ad073ca07cdf479398d5f95d88e218c4).
* [**Vector Space Models**](https://www.coursera.org/learn/classification-vector-spaces-in-nlp/home/week/3) - Series of lessons from DeepLearning.AI's Natural Language Processing specialization discussing the basics of vector space models and their use in language modeling.

**Pre-training and scaling laws**

* [**Scaling Laws for Neural Language Models**](https://arxiv.org/abs/2001.08361)- empirical study by researchers at OpenAI exploring the scaling laws for large language models.

**Model architectures and pre-training objectives**

* [**What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?**](https://arxiv.org/pdf/2204.05832.pdf) - The paper examines modeling choices in large pre-trained language models and identifies the optimal approach for zero-shot generalization.
* [**HuggingFace Tasks**](https://huggingface.co/tasks) **and** [**Model Hub**](https://huggingface.co/models) - Collection of resources to tackle varying machine learning tasks using the HuggingFace library.
* [**LLaMA: Open and Efficient Foundation Language Models**](https://arxiv.org/pdf/2302.13971.pdf) - Article from Meta AI proposing Efficient LLMs (their model with 13B parameters outperform GPT3 with 175B parameters on most benchmarks)

**Scaling laws and compute-optimal models**

* [**Language Models are Few-Shot Learners**](https://arxiv.org/pdf/2005.14165.pdf)-This paper investigates the potential of few-shot learning in Large Language Models.
* [**Training Compute-Optimal Large Language Models**](https://arxiv.org/pdf/2203.15556.pdf)- Study from DeepMind to evaluate the optimal model size and number of tokens for training LLMs. Also known as “Chinchilla Paper”.
* [**BloombergGPT: A Large Language Model for Finance**](https://arxiv.org/pdf/2303.17564.pdf) - LLM trained specifically for the finance domain, a good example that tried to follow chinchilla laws.