Transformer Architecture

- **Attention is All You Need** - This paper introduced the Transformer architecture, with the core "self-attention" mechanism. This article was the foundation for LLMs.
- **BLOOM: BigScience 176B Model** - 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.
- **Vector Space Models** - 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** - 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?** - The paper examines modeling choices in large pre-trained language models and identifies the optimal approach for zero-shot generalization.
- **HuggingFace Tasks** and **Model Hub** - Collection of resources to tackle varying machine learning tasks using the HuggingFace library.
- **LLaMA: Open and Efficient Foundation Language Models** - 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** - This paper investigates the potential of few-shot learning in Large Language Models.
- **Training Compute-Optimal Large Language Models** - 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** - LLM trained specifically for the finance domain, a good example that tried to follow chinchilla laws.

Multi-task, instruction fine-tuning

- **Scaling Instruction-Finetuned Language Models** - Scaling fine-tuning with a focus on task, model size and chain-of-thought data.

- **Introducing FLAN: More generalizable Language Models with Instruction Fine-Tuning** - This blog (and article) explores instruction fine-tuning, which aims to make language models better at performing NLP tasks with zero-shot inference.

Model Evaluation Metrics

- **HELM - Holistic Evaluation of Language Models** - HELM is a living benchmark to evaluate Language Models more transparently.
- **General Language Understanding Evaluation (GLUE) benchmark** - This paper introduces GLUE, a benchmark for evaluating models on diverse natural language understanding (NLU) tasks and emphasizing the importance of improved general NLU systems.
- **SuperGLUE** - This paper introduces SuperGLUE, a benchmark designed to evaluate the performance of various NLP models on a range of challenging language understanding tasks.
- **ROUGE: A Package for Automatic Evaluation of Summaries** - This paper introduces and evaluates four different measures (ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S) in the ROUGE summarization evaluation package, which assess the quality of summaries by comparing them to ideal human-generated summaries.
- **Measuring Massive Multitask Language Understanding (MMLU)** - This paper presents a new test to measure multitask accuracy in text models, highlighting the need for substantial improvements in achieving expert-level accuracy and addressing lopsided performance and low accuracy on socially important subjects.
- **BigBench-Hard - Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models** - The paper introduces BIG-bench, a benchmark for evaluating language models on challenging tasks, providing insights on scale, calibration, and social bias.

Parameter- efficient fine tuning (PEFT)

- **Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning** - This paper provides a systematic overview of Parameter-Efficient Fine-tuning (PEFT) Methods in all three categories discussed in the lecture videos.
- **On the Effectiveness of Parameter-Efficient Fine-Tuning** - The paper analyzes sparse fine-tuning methods for pre-trained models in NLP.

LoRA

- **LoRA Low-Rank Adaptation of Large Language Models** - This paper proposes a parameter-efficient fine-tuning method that makes use of low-rank decomposition matrices to reduce the number of trainable parameters needed for fine-tuning language models.
- **QLoRA: Efficient Finetuning of Quantized LLMs** - This paper introduces an efficient method for fine-tuning large language models on a single GPU, based on quantization, achieving impressive results on benchmark tests.

Prompt tuning with soft prompts

- **[The Power of Scale for Parameter-Efficient Prompt Tuning](#)** - The paper explores "prompt tuning," a method for conditioning language models with learned soft prompts, achieving competitive performance compared to full fine-tuning and enabling model reuse for many tasks.

Reinforcement Learning from Human-Feedback (RLHF)

- **[Training language models to follow instructions with human feedback](#)** - Paper by OpenAI introducing a human-in-the-loop process to create a model that is better at following instructions (InstructGPT).
- **[Learning to summarize from human feedback](#)** - This paper presents a method for improving language model-generated summaries using a reward-based approach, surpassing human reference summaries.

Proximal Policy Optimization (PPO)

- **[Proximal Policy Optimization Algorithms](#)** - The paper from researchers at OpenAI that first proposed the PPO algorithm. The paper discusses the performance of the algorithm on a number of benchmark tasks including robotic locomotion and game play.
- **[Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#)** - This paper presents a simpler and effective method for precise control of large-scale unsupervised language models by aligning them with human preferences.

Scaling human feedback

- **[Constitutional AI: Harmlessness from AI Feedback](#)** - This paper introduces a method for training a harmless AI assistant without human labels, allowing better control of AI behavior with minimal human input.

Advanced Prompting Techniques

- **[Chain-of-thought Prompting Elicits Reasoning in Large Language Models](#)** - Paper by researchers at Google exploring how chain-of-thought prompting improves the ability of LLMs to perform complex reasoning.
- **[PAL: Program-aided Language Models](#)** - This paper proposes an approach that uses the LLM to read natural language problems and generate programs as the intermediate reasoning steps.
- **[ReAct: Synergizing Reasoning and Acting in Language Models](#)** This paper presents an advanced prompting technique that allows an LLM to make decisions about how to interact with external applications.

LLM powered application architectures

- **LangChain Library (GitHub)** - This library is aimed at assisting in the development of those types of applications, such as Question Answering, Chatbots and other Agents. You can read the documentation here.
- **Who Owns the Generative AI Platform?** - The article examines the market dynamics and business models of generative AI.