Week 1 resources

Below you'll find links to the research papers discussed in this weeks videos. You don't need to understand all the technical details discussed in these papers - **you have already seen the most important points you'll need to answer the quizzes** in the lecture videos.

However, if you'd like to take a closer look at the original research, you can read the papers and articles via the links below.

Generative AI Lifecycle

- **Generative AI on AWS: Building Context-Aware, Multimodal Reasoning Applications** - This O'Reilly book dives deep into all phases of the generative AI lifecycle including model selection, fine-tuning, adapting, evaluation, deployment, and runtime optimizations.

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 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