

## Lecture 24 — Large Language Models

Jeff Zarnett

2023-09-10

## Large Language Models and You

In November of 2022, OpenAI introduced ChatGPT to the world and suddenly everyone and their best friend was doing some combination of (1) writing news articles about how the advent of AI means we will all be unemployed and lead to the rise of Skynet; (2) founding a startup that used ChatGPT to do something something B2B/B2C SaaS something something profit; (3) rebranding themselves on LinkedIn as “prompt engineering” experts; or (4) generating lots social media cringe content about how to use it to get rich easily<sup>1</sup>.

Such large language models have existed before, but ChatGPT ended up a hit because it’s pretty good at being “conversational”, which is to say that it does a good job of responding to input in a way that seems like how a human would respond. For this reason, it’s referred to as NLP – Natural Language Processing. In the words of some experts [KSK<sup>+</sup>23]: “These models are trained on massive amounts of text data and are able to generate human-like text, answer questions, and complete other language-related tasks with high accuracy.”

Of course, just because such a tool can produce an answer to your question, doesn’t mean it is necessarily true or correct. You may enjoy this Legal Eagle video about a lawyer who used ChatGPT for “research” and found that the system returned an answer with completely made-up references. These are sometimes called “hallucinations”. We don’t have time to watch the video in lecture, but I encourage you to watch it to get an understanding of what went wrong: <https://www.youtube.com/watch?v=oqSYljRYDEM>.

Part of what makes the GPT-3 and GPT-4 models better at producing output that matches our expectations is that it relies on pre-training (it’s the PT part of GPT) on a very large data set, specifically a lot of stuff just out there on the internet [KSK<sup>+</sup>23]. This course is not one on neural networks, large language models, AI, or similar – there are other such technical electives which you may be taking. This does connect to the course content, however, because generating or updating the pre-trained model is computationally challenging... so performance increases matter here.

**Parameters.** One factor, but certainly not the only one, in how good the model is at responding to requests is the number of parameters. To explain a bit about why it matters, consider this quote from [MB23]: “LLM AI models are generally compared by the number of parameters — where bigger is usually better. The number of parameters is a measure of the size and the complexity of the model. The more parameters a model has, the more data it can process, learn from, and generate. However, having more parameters also means having more computational and memory resources, and more potential for overfitting or underfitting the data. Parameters are learned or updated during the training process, by using an optimization algorithm that tries to minimize the error or the loss between the predicted and the actual outputs. By adjusting the parameters, the model can improve its performance and accuracy on the given task or domain.”

## Generating or Training a Model

Before we get into the subject of how, we should address the question of why you would wish to generate or customize a LLM rather than use an existing one. To start with, you might not want to send your (sensitive) data to a third party for analysis. Still, you can download and use some existing models. So generating a model or refining an existing one may make sense in a situation where you will get better results by creating a more specialized model than the generic one. To illustrate what I mean, ChatGPT will gladly make you a Dungeons & Dragons campaign setting, but you don’t need it to have that capability if you want it to analyze your customer behaviours

<sup>1</sup> Watch this video on the subject of “Get Rich Easy” schemes: <https://www.youtube.com/watch?v=2bq3SdfzcA4>

to find the ones who are most likely to be open to upgrading their plan. That extra capability (parameters) takes up space and computational time and a smaller model that gives better answers is more efficient.

You may have guessed by the placement of this topic in the course that the GPU is an excellent choice for how to generate a large language model.

## References

- [KSK<sup>+</sup>23] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyiniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. URL: <https://www.sciencedirect.com/science/article/pii/S1041608023000195>, doi:10.1016/j.lindif.2023.102274.
- [MB23] John Maeda and Matthew Bolaños. What are Models?, May 2023. Online; accessed 2023-09-10. URL: <https://learn.microsoft.com/en-us/semantic-kernel/prompt-engineering/llm-models>.