Tokenization:

I request you to visit videos 13 - 18 from this playlist: Hugging face course
From the above material, you get an idea of the tokenizers used…

Now we explore a very specific kind of tokenizer used in gpt2 models, byte-pair tokenizers.
Byte Pair Encoding (BPE) is a subword tokenization technique that splits text into smaller units (subwords or characters) based on their frequency in the training data. It's widely used in models like GPT-2, where handling unseen words and maintaining efficiency are essential.
Here is a colab notebook link to explore it further:
Tokenizers.ipynb

Also, please go through these to get an idea of transformers
Transformers, explained: Understand the model behind GPT, BERT, and T5
Video 5-8 of Hugging face course
Illustrated Guide to Transformers Neural Network: A step by step explanation

Next steps are Hugging face ‘datasets’ library, dataloading, and pipeline functions…
Still a long way to go for a fine tuned model :)