Chapter 2

- Tokens and embeddings are two of the central concepts of using large language models (LLMs). As we've seen in the rst chapter, they're not only important to understanding the history of Language AI, but we cannot have a clear sense of how LLMs work, how they're built, and where they will go in the future without a good sense of tokens and embeddings
- The input goes too llm and generated token by token
- The First token </s> a special token put at the beginning of the first sentence .
- Some tokens are complete words (e.g., Write, an, email). Some tokens are parts of words (e.g., apolog, izing, trag, ic). Punctuation characters are their own token.
- e, the creator of the model chooses a tokenization method. Popular methods include byte pair encoding (BPE) (widely used by GPT models) and WordPiece (used by BERT).
- Second, after choosing the method, we need to make a number of tokenizer design choices like vocabulary size and what special tokens to use.
- Third, the tokenizer needs to be trained on a specic dataset to establish the best vocabulary it can use to represent that dataset
- Third, the tokenizer needs to be trained on a specic dataset to establish the best vocabulary it can use to represent that dataset.
- Even if we set the same methods and parameters, a tokenizer trained on an English text dataset will be different from another trained on a code dataset or a multilingual text dataset
- The tokenization scheme we just discussed is called subword tokenization. It's the most commonly used tokenization scheme but not the only one. (word tokenizer , character tokenizer - subword tokenizer - Byte Tokenizer)
- The problem with word tokenizer which it is used first in nlp that huge number of words as well there is a small difference between words which is not acceptable but it is used in recommendation system now
- , you may be able to t about three times as much text using subword tokenization than using character tokens
- The GPT-2 and RoBERTa tokenizers do bytes pair encoding tokenization (subword) and used in coding and multilingual ,Flan-T5 uses a tokenizer implementation called SentencePiece
- GPT4 use the the same tokenization like gpt2 but with a new enhancement of how to deal with capital words , spaces , some tricks of coding like elif and encode the list of white spaces as single token etc
- Tokenizers have been adapted to this direction by the addition of tokens that indicate the turns in a conversation and the roles of each speaker. These special tokens include: <|user|> <|assistant|> <|system|>
- There are three major groups of design choices that determine how the tokenizer will break down text: the tokenization method, the initialization parameters, and the domain of the data the tokenizer targets
- 30K and 50K are often used as vocabulary size values, but more and more we're seeing larger sizes like 100K
- d if we train a good-enough model on a large-enough set of tokens, it starts to capture the complex patterns that appear in its training dataset
- Achieving a good threshold of language coherence and better-than-average factual generation, however, starts to present a new problem. Some users start to trust the model's fact generation ability (e.g., at the beginning of 2023 some language models were being dubbed "Google killers").

It didn’t take long for advanced users to recognize that generation models alone aren’t reliable search engines. This led to the rise of retrieval-augmented generation (RAG), which combines search and LLMs.

- LLM doesn’t take a fixed embedding vector of llm instead it using dynamic depends on context and that’s called contextualized word embeddings
- The book started to show we used to use word2vec to generate and predict nex word
- 4. Skip-gram and negative sampling are two of the main ideas behind the word2vec algorithm and are useful in many other problems that can be formulated as token sequence problems.
- s. Imagine if we treated each song as we would a word or token, and we treated each playlist like a sentence. These embeddings can then be used to recommend similar songs that often appear together in playlists and it is based to build recommendation system
-