f0gkdwbnz

December 4, 2024

Understanding Tokenizer and Context Window

```
[ ]: # visit this link : https://platform.openai.com/tokenizer
     # to check tokenization in action
```

```
[ ]: # tokens - indiviual units passed to language model
     # in early days neural nets we trained on character level
     # meaning predict the next character in the seq

     # small vocab size, lesse possibilities at input level
     # Later, neural nets we started training on word level
     # predict next word in the sequence
     # it lead to enormous vocab
     # so many places for names, places, animals etc...

     # Rather than training on each word, or each character
     # a middle ground was achiueved to take the token and work with series of tokens
     # this had lot of benefits : handling became easier

     # it was good at handling word stems

     # you can check GPT tokenization
     # check how the text is turned into a series of tokens
     # tokens are highlighted in colours
     # When you check tokenization , you might also find that the whitespace␣
      ↪prefixing the token
     # is also highlighted as the break between two tokens is equally important

     # somtimes when you give some rare words, you can find tokens being broken down␣
      ↪differently

     # Refer GPT tokenizer 1 png file to check the break in tokens
```

As suggested on Open AI website .. A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

Example : The complete works of Shakespeare are ~9,00,000 words or 1.2 million tokens

[ ]:

What is Context Window ?

1. The main job an LLM is next token prediction.
2. Context windows tells you total no.of tokens that an LLM can examine at any point when generating the next token.
3. Usually higher the context window for an LM, better its reasoning and logical abilities.
4. User prompt and system prompt are inputs to LLMs, which predicts next token.
5. It is like a chained reaction to predict next token in case of context window.

For example, when you input a prompt to chatgpt, it generates some response. Say after that, you make a follow up question. Now while generating the next token , it will take in all of this as input to generate the next token .

[ ]:

[ ]:

[ ]: