Blog_Notes

In this document I will try to write my understandings or opinions or learnings from the content that I read on internet.

Random Learnings

- Semantic Search: Instead of just trying to compare words or tokens of an input question from user, this search will try to include context into searching to get better understanding of user intention.

Modern Bert : https://huggingface.co/blog/modernbert

This talks about the new encoder only model in machine learning space.

- This is the first time I am hearing about encoder-only and decoder-only models available.
    - GPT, LLama, Claude all are decoder only models which generally means generative models. These models are slow, big, and expensive. For many of practical situations we may not need these.
- Encoder only models produce output that is in vector format.
- Decoder only models can not look at future tokens because they are mathematically not allowed to do this. Encoder only models can look backwards and forward as well. Decoders can only look backwards.
- I really liked the way they distinguished between encoder only and decoder only models in this blog. Decoder only models are like Ferrari race cars and Encoder only models are like honda civic.
- In RAG we use document library which are relevant to user query. Here also we use representational models (encoder only) to select documents which are related to query.
- Recommendation systems, targeted ad systems use representational models.
- Cost of using encoder models is way cheaper than using decoder models.
- ModernBert uses 8192 tokens length of context, it is 16x more than any existing encoders.