LLM Apps

How to overcome the context window limits

Remember: What is the context window?

- The context window is the maximum size of the context that we can give to an LLM.

- For example, an LLM like chatGPT has the following context windows:
  - chatGPT 3.5 supports a context window of up to 4,096 tokens (approx. 3,000 words, 6 pages).
  - chatGPT 4 supports a context window of up to 8,192 tokens (approx. 6,100 words, 12 pages).

What limits does the context window impose on us?

- The context window prevents us from things like:
  - Asking chatGPT to summarize a 100-page report.
  - Asking chatGPT to use a database.
  - Etc.

How to overcome the context window limits?

- Training an LLM from scratch with our data.
  - Hugely expensive. Impractical for most in reality.
- Adding our data to an already trained LLM (fine-tuning).
  - Very expensive and technically very complex. Impractical for most in reality.
- With the RAG (Retrieval-Augmented Generation) technique, we divide our data into small segments, allowing the LLM to use them within the limits of its context window.
  - This is the technique used today by virtually all LLM applications.