Section Recap: Introduction to AI Models

- AI engineering is becoming a critical skill for modern developers, with real applications across industries.
- We explored how large language models (LLMs) are used in apps for summarization, content generation, translation, classification, and more.
- A large language model is a type of AI system trained on massive amounts of text to predict the next word in a sentence.
- LLMs are made up of billions of parameters that encode statistical patterns in human language.
- These models don't "understand" like humans do — they generate responses based on probabilities and training data.
- Tokens are the basic units of input/output in LLMs, representing chunks of text such as words or punctuation.
- The number of tokens used in a request directly impacts both cost and model limits.
- Every LLM has a context window, which defines how many tokens it can handle at once.
- We learned how to count tokens programmatically using a tokenizer, to help us estimate cost and stay within model limits.
- Choosing the right model depends on factors like reasoning ability, speed, cost, context window size, and support for different modalities.
- Some models are better suited for lightweight tasks, while others are designed for more complex reasoning.
- We explored how model settings like temperature, max tokens, and top_p affect the style, length, and variability of responses.
- Understanding these concepts helps us use LLMs more effectively and build more reliable AI-powered applications.