The Cost Reality of AI Apps

How much an AI application will cost to run, test, and maintain

- AI API usage is the basis for how much you will be charged. Usage is measured in tokens (MToks). Different models and APIs will charge you different amounts. Take into consideration that different models also have different strengths that are often not reflected in industry benchmarks

- Input and Output costs are often different prices, with output tokens typically costing more, meaning that an application that wants to output large AI responses consistently will pay more than applications with larger context windows

- The size of your context window will not just run up prices but could effect performance according to many studies on the matter, this is why managing the amount of tokens we're sending to the API matters and why standardized methods of controlling and organizing tokens into easily readable inputs is preferable

Apps that Evolve

Building for the unpredictability of AI

- AIs hallucinate. You likely already know this if you're considering building an AI platform, but consider that the scale of the hallucination and the downstream effects of them can expand exponentially with architecture and design that rests on the accuracy of AI responses. A lot of this course will cover this reality and demonstrate ways to hedge against these unfortunate issues or embrace them in interesting ways

- Luckily, many hallucination can be stamped out, accounted for, or predicted in the course of designing your app. This will require persistence, trial and error, and a level of systems-thinking

- All in all, this is a feature, not a bug. The reason AIs hallucinate is the same reason that we will receive different responses to the same input and the AI can mimic human creative capacity