LLM Tokenization Overview and Examples

Action Items

- Try the OpenAI tokenizer at platform.openai.com
- Test token counts for your own text, code snippets, and math expressions
- Estimate token usage using the 4-character ≈ 1 token rule of thumb
- Add the discussed code example to your cursor project (as Charles requested)

Tokenization Basics

- Each common word usually maps to a single token
- Tokens include a “beginning-of-word” marker (the preceding space)
- Rare or invented words split into multiple sub-tokens (e.g., “exquisitely” → “ex”, “qui”, “si”, “tely”)
- Tokens can represent word fragments, whole words, or punctuation

Model Choice & Tokenization

- Different models may tokenize slightly differently, but impact is minor
- Pick a model only if you need to study its specific tokenization behavior

Examples & Ratios

- Simple sentence: 50 characters → 9 tokens
- Complex sentence: 66 characters → 18 tokens
- Phone-number style: three-digit groups each become a token
- “LLM” originally split into two tokens; now often a single token due to popularity

Token-to-Word Estimates

- Rough rule: 1 token ≈ 4 characters or ≈ 0.75 words
- 1,000 tokens ≈ 750 words
- Complete Works of Shakespeare ≈ 900k words ≈ 1.2 M tokens

Special Cases

- Code, scientific terms, and variable names use more tokens (often near 1 token/character)
- Experiment by feeding code into the tokenizer to gauge token cost

Cost Context

- Pricing often quoted per million tokens; think of it as the cost of the entire Shakespeare corpus

Closing Note

- After the overview, the session will move to live coding with the cursor tool.