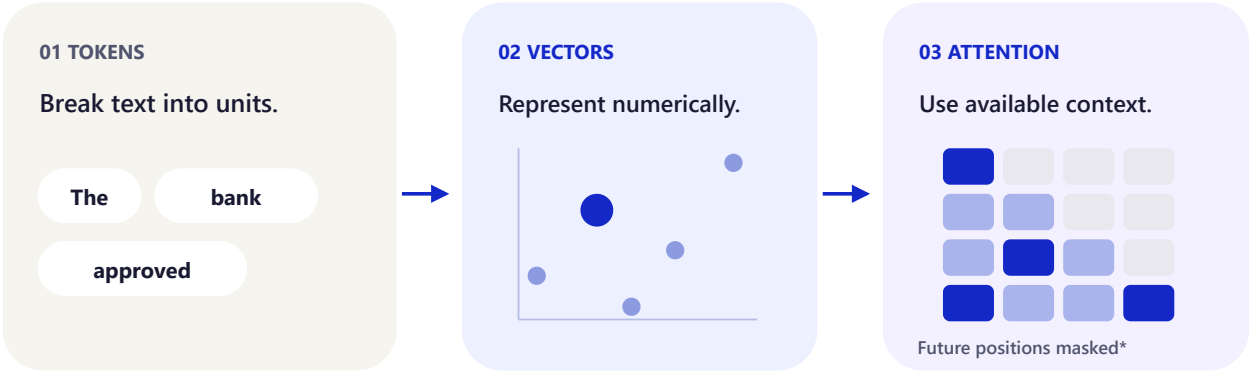
CHAPTER 05 / BUILD A MENTAL MODEL

Give context a shape.

Tokens are units. Embeddings are vectors. Context has limits.

From pieces of text to contextual representations.



A FINITE CONTEXT BUDGET



*Illustrative causal text attention. Budget proportions are examples, not provider limits.

Token splits, vector positions, and attention weights are illustrative. Actual representations vary by tokenizer and model.

Tokens

Chunks of text, not necessarily whole words. Count with the selected model's tokenizer rather than a fixed words-to-tokens ratio.

Embeddings

A token ID selects an initial vector. Later hidden representations incorporate permitted context. Similarity is not truth.

Context window

Finite request capacity. Check input/output accounting and overflow policy. Fitting does not guarantee reliable recall.

↗ **More context is not always better context. Include what the task needs.**

PUT IT TO WORK

For document Q&A, retrieve the relevant passages, remove duplicates, and reserve room for the answer. Check the provider's accounting rules.

CLOSE THE PAGE. EXPLAIN IT.

Which context could you remove without losing the evidence needed for the answer?