

2. Subword Tokens

When one word breaks into pieces

A tokenizer splits “preprocessing” into three pieces: “pre”, “process”, and “ing”. The agent receives the pipeline command “run preprocessing now”. How many tokens does it count?

Step 1: Write each token in a box

run	pre	process	ing	now			
-----	-----	---------	-----	-----	--	--	--

Step 2: Count the filled boxes

5 tokens

Practice 1

A tokenizer splits “embeddings” into “embed” and “dings”, and “retrieval” into “retriev” and “al”. The agent receives the query “run embeddings retrieval”. How many tokens does it count?

Step 1: Write each token in a box

--	--	--	--	--	--	--	--

Step 2: Count the filled boxes

--

Practice 2

A tokenizer splits “debugging” into “debug” and “ging”, and “codebase” into “code” and “base”. The agent receives the task “start debugging codebase issues”. How many tokens does it count?

Step 1: Write each token in a box

--	--	--	--	--	--	--	--

Step 2: Count the filled boxes

--