

6. Retrieval Footprint

What a document costs the window

An AI agent has a 4,000-token context window holding a 300-token system prompt and a 200-token user question. To answer it, the agent retrieves a 2,000-token document, split into chunks of 100 tokens. The model's answer will take 200 tokens. Mark the system prompt S, the question U, each retrieved chunk R, and the answer O.

Step 1: The document in chunks

$2,000 / 100 = 20 \text{ chunks}$

Step 2: Mark what the window holds

1 - 1,000	S	S	S	U	U	R	R	R	R	R
1,001 - 2,000	R	R	R	R	R	R	R	R	R	R
2,001 - 3,000	R	R	R	R	R	O	O			
3,001 - 4,000										

Practice 1

An AI agent has a 4,000-token context window holding a 200-token system prompt and a 300-token user question. It retrieves a 1,600-token document, split into chunks of 100 tokens. The answer will take 300 tokens. Mark the system prompt S, the question U, each retrieved chunk R, and the answer O.

Step 1: The document in chunks

Step 2: Mark what the window holds

1 - 1,000										
1,001 - 2,000										
2,001 - 3,000										
3,001 - 4,000										

6. Retrieval Footprint (continued)

Practice 2

An AI agent has a 4,000-token context window holding a 400-token system prompt and a 200-token user question. It retrieves a 2,400-token document, split into chunks of 100 tokens. The answer will take 400 tokens. Mark the system prompt S, the question U, each retrieved chunk R, and the answer O.

Step 1: The document in chunks

Step 2: Mark what the window holds

1 - 1,000									
1,001 - 2,000									
2,001 - 3,000									
3,001 - 4,000									