

6. The Retrieved Document

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 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. The Retrieved Document (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									

7. Turns After Retrieval

An AI agent has a 4,000-token context window and a 300-token system prompt. On every turn it retrieves a 300-word document from its database at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 200 user tokens and 300 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

$$(300 \times 2)/100 = 6 \text{ chunks}$$

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

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

Step 3: Count the turns

$$(4,000 - 300)/(600 + 200 + 300) = 3.36, \text{ so } 3 \text{ turns}$$

7. Turns After Retrieval (continued)

Practice 1

An AI agent has a 3,000-token context window and a 200-token system prompt. On every turn it retrieves a 200-word document at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 100 user tokens and 200 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

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

Step 3: Count the turns

7. Turns After Retrieval (continued)

Practice 2

An AI agent has a 5,000-token context window and a 300-token system prompt. On every turn it retrieves a 900-word document at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 200 user tokens and 300 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

1 - 1,000									
1,001 - 2,000									
2,001 - 3,000									
3,001 - 4,000									
4,001 - 5,000									

Step 3: Count the turns

8. Three Rounds of Search

An AI agent has a 4,000-token context window and a 300-token system prompt. It searches its database three times. Each round costs a 200-token query and a 300-token answer. The first two searches find nothing, and an empty result still returns one 100-token chunk. The third finds a document of ten chunks. Mark the three rounds, then count the free space left.

Step 1: Mark the context window, S: system, U: query, R: retrieved, O: answer, and () for round boundaries.

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

Step 2: Count the free space

1,000 tokens free

Practice 1

An AI agent has a 3,000-token context window and a 200-token system prompt. It searches its database three times. Each round costs a 100-token query and a 200-token answer. The first two searches find nothing, and an empty result still returns one 100-token chunk. The third finds a document of eight chunks. Mark the three rounds, then count the free space left.

Step 1: Mark the context window, S: system, U: query, R: retrieved, O: answer, and () for round boundaries.

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

Step 2: Count the free space

8. Three Rounds of Search (continued)

Practice 2

An AI agent has a 5,000-token context window and a 300-token system prompt. It searches its database three times. Each round costs a 200-token query and a 300-token answer. The first search finds nothing and returns one 100-token chunk; the second finds six chunks and the third finds nine. Mark the three rounds, then count the free space left.

Step 1: Mark the context window, S: system, U: query, R: retrieved, O: answer, and () for round boundaries.

1 - 1,000									
1,001 - 2,000									
2,001 - 3,000									
3,001 - 4,000									
4,001 - 5,000									

Step 2: Count the free space

9. Sizing the Window

An AI agent must support a 30-turn conversation before its window fills. Each turn adds 250 tokens, and the system prompt is 500 tokens. What window does the agent need?

Step 1: Tokens needed for all the turns

$$30 \times 250 = 7,500 \text{ tokens}$$

Step 2: Add the system prompt

$$7,500 + 500 = 8,000 \text{ tokens}$$

Step 3: Take the smallest window that fits

☐ 4k
 ☒ 8k
 ☐ 16k
 ☐ 32k

Practice 1

An AI agent must support a 40-turn conversation before its window fills. Each turn adds 300 tokens, and the system prompt is 800 tokens. What window does the agent need?

Step 1: Tokens needed for all the turns

Step 2: Add the system prompt

Step 3: Take the smallest window that fits

☐ 4k
 ☐ 8k
 ☐ 16k
 ☐ 32k

9. Sizing the Window (continued)

Practice 2

An AI agent must support a 10-turn conversation before its window fills. Each turn adds 300 tokens, and the system prompt is 500 tokens. What window does the agent need?

Step 1: Tokens needed for all the turns

Step 2: Add the system prompt

Step 3: Take the smallest window that fits

- ☐ 4k
 ☐ 8k
 ☐ 16k
 ☐ 32k

10. Dropping the Oldest

An AI agent has a 4,000-token context window holding a 300-token system prompt and 7 turns of 500 tokens each. A new 1,500-token message arrives. How many of the earliest turns must be crossed out to make room?

Step 1: How much room is short

$1,500 - 200 = 1,300 \text{ tokens short}$

Step 2: Turns to drop

$1,300/500 = 2.6, \text{ so } 3 \text{ turns}$

Step 3: Cross out the earliest turns with a line

1 - 1,000	S	S	S	(	U	U	O	O	O	)	U	U
1,001 - 2,000	O	O	O	)	U	U	O	O	O	)	U	U
2,001 - 3,000	O	O	O	)	U	U	O	O	O	)	U	U
3,001 - 4,000	O	O	O	)	U	U	O	O	O	)		

10. Dropping the Oldest (continued)

Practice 1

An AI agent has a 4,000-token context window holding a 200-token system prompt and 9 turns of 400 tokens each. A new 1,000-token message arrives. How many of the earliest turns must be crossed out to make room?

Step 1: How much room is short

Step 2: Turns to drop

Step 3: Cross out the earliest turns with a line

1 - 1,000	S	S	(	U	O	O	O	)	U	O	O	O	
1,001 - 2,000	(	U	O	O	O	)	U	O	O	O	)	U	O
2,001 - 3,000	O	O	)	U	O	O	O	)	U	O	O	O	
3,001 - 4,000	(	U	O	O	O	)	U	O	O	O	)		

Practice 2

An AI agent has a 3,000-token context window holding a 300-token system prompt and 8 turns of 300 tokens each. A new 1,400-token message arrives. How many of the earliest turns must be crossed out to make room?

Step 1: How much room is short

Step 2: Turns to drop

Step 3: Cross out the earliest turns with a line

1 - 1,000	S	S	S	(	U	O	O	)	U	O	O	)	U
1,001 - 2,000	O	O	)	U	O	O	)	U	O	O	)	U	O
2,001 - 3,000	O	)	U	O	O	)	U	O	O	)			