

1. Mark the Context

An AI agent runs on a model with a 4,000-token context window, drawn below as cells of 100 tokens. The conversation so far uses 3,200 tokens. Mark each used cell with an X.

Step 1: Mark the used cells with an X

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

Practice 1

An AI agent runs on a model with a 2,000-token context window. The conversation so far uses 1,600 tokens. Mark each used cell with an X.

Step 1: Mark the used cells with an X

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

Practice 2

An AI agent runs on a model with a 3,000-token context window. The conversation so far uses 1,200 tokens. Mark each used cell with an X.

Step 1: Mark the used cells with an X

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

2. What Fills the Window

An AI agent has a 4,000-token context window. It holds a 400-token system prompt, 1,400 user tokens, and 1,000 output tokens. Mark the system prompt S, the user tokens U, and the output tokens O, then work out the free space.

Step 1: Mark what each part holds

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

Step 2: Subtract the used tokens from the window

$$4,000 - 400 - 1,400 - 1,000 = 1,200 \text{ tokens free}$$

Practice 1

An AI agent has a 4,000-token context window. It holds a 300-token system prompt, 1,100 user tokens, and 800 output tokens. Mark the system prompt S, the user tokens U, and the output tokens O, then work out the free space.

Step 1: Mark what each part holds

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

Step 2: Subtract the used tokens from the window

2. What Fills the Window (continued)

Practice 2

An AI agent has a 4,000-token context window. It holds a 200-token system prompt, 1,000 user tokens, and 800 output tokens. Mark the system prompt S, the user tokens U, and the output tokens O, then work out the free space.

Step 1: Mark what each part holds

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

Step 2: Subtract the used tokens from the window

3. Turn by Turn

An AI agent runs on a 4,000-token context window and starts with a 400-token system prompt. Each turn adds 100 user tokens and 300 output tokens. Mark the window after three turns.

Step 1: Mark the system prompt S, then each turn's user tokens U and output tokens O

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

Step 2: Total what the conversation has used

$400 + 3 \times 400 = 1,600 \text{ tokens}$

Practice 1

An AI agent runs on a 4,000-token context window and starts with a 300-token system prompt. Each turn adds 100 user tokens and 200 output tokens. Mark the window after four turns.

Step 1: Mark the system prompt S, then each turn's user tokens U and output tokens O

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

Step 2: Total what the conversation has used

3. Turn by Turn (continued)

Practice 2

An AI agent runs on a 4,000-token context window and starts with a 200-token system prompt. Each turn adds 200 user tokens and 200 output tokens. Mark the window after three turns.

Step 1: Mark the system prompt S, then each turn's user tokens U and output tokens O

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

Step 2: Total what the conversation has used

4. Will It Fit?

An AI agent's 3,000-token context window already holds 1,800 tokens, marked X below. A new 1,400-token user message arrives. Number the chunks into the empty cells, one chunk per cell, and see whether the whole message fits.

Step 1: The message in chunks

1,400/100 = 14 chunks

Step 2: Number the chunks into the empty cells

1 - 1,000	X	X	X	X	X	X	X	X	X	X
1,001 - 2,000	X	X	X	X	X	X	X	X	1	2
2,001 - 3,000	3	4	5	6	7	8	9	10	11	12

Step 3: Does the whole message fit?

☐ fits ☒ does not fit

Practice 1

An AI agent's 3,000-token context window already holds 1,200 tokens, marked X below. A new 900-token user message arrives. Number the chunks into the empty cells, one chunk per cell, and see whether the whole message fits.

Step 1: The message in chunks

Step 2: Number the chunks into the empty cells

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

Step 3: Does the whole message fit?

☐ fits ☐ does not fit

4. Will It Fit? (continued)

Practice 2

An AI agent's 3,000-token context window already holds 2,300 tokens, marked X below. A new 1,100-token user message arrives. Number the chunks into the empty cells, one chunk per cell, and see whether the whole message fits.

Step 1: The message in chunks

Step 2: Number the chunks into the empty cells

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

Step 3: Does the whole message fit?

☐ fits ☐ does not fit

5. Turns Until Full

An AI agent's 4,000-token context window already holds 800 tokens, marked X below. Each turn adds 200 user tokens and 300 output tokens. Mark each turn as U U O O O until the window is full, then count the turns.

Step 1: Mark each turn as U U O O O

1 - 1,000	X	X	X	X	X	X	X	X	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		

Step 2: Count the turns

$$(4,000 - 800)/(200 + 300) = 6.4, \text{ so } 6 \text{ turns}$$

Practice 1

An AI agent's 4,000-token context window already holds 1,200 tokens, marked X below. Each turn adds 100 user tokens and 300 output tokens. Mark each turn as U O O O until the window is full, then count the turns.

Step 1: Mark each turn as U O O O

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

Step 2: Count the turns

5. Turns Until Full (continued)

Practice 2

An AI agent's 4,000-token context window already holds 1,000 tokens, marked X below. Each turn adds 300 user tokens and 300 output tokens. Mark each turn as U U U O O O until the window is full, then count the turns.

Step 1: Mark each turn as U U U O O O

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

Step 2: Count the turns