

3. Growth per Turn

How a window fills, one turn at a time

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 context window, S: system, U: user, O: output, and () for turn boundaries.

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 context window, S: system, U: user, O: output, and () for turn boundaries.

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

Step 2: Total what the conversation has used

3. Growth per 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 context window, S: system, U: user, O: output, and () for turn boundaries.

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

Step 2: Total what the conversation has used