

11. Compaction

Summarizing history to make room

An AI agent has a 2,000-token context window. Its system prompt is 300 tokens, and three turns of conversation follow, each a 100-token user message and a 300-token answer. The agent compacts, and the summary keeps 25% of the history. Mark the window as it stands after the compaction.

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

Step 1: Measure the history, which is everything but the system prompt

$$3 \times (100 + 300) = 1,200 \text{ tokens}$$

Step 2: Take the share the summary keeps

$$0.25 \times 1,200 = 300 \text{ tokens, or 3 cells}$$

Step 3: Mark the window after compacting, S: system, C: summary

1 - 1,000	S	S	S	C	C	C				
1,001 - 2,000										

11. Compaction (continued)

Practice 1

An AI agent has a 4,000-token context window. Its system prompt is 400 tokens, and three turns of conversation follow, each a 200-token user message and a 600-token answer. The agent compacts, and the summary keeps 25% of the history. Mark the window as it stands after the compaction.

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

Step 1: Measure the history, which is everything but the system prompt

Step 2: Take the share the summary keeps

Step 3: Mark the window after compacting, S: system, C: summary

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

11. Compaction (continued)

Practice 2

An AI agent has a 4,000-token context window. Its system prompt is 400 tokens, and three turns of conversation follow, each a 300-token user message and a 500-token answer. The agent compacts, and the summary keeps 25% of the history. Mark the window as it stands after the compaction.

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

Step 1: Measure the history, which is everything but the system prompt

Step 2: Take the share the summary keeps

Step 3: Mark the window after compacting, S: system, C: summary

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