

12. The Compaction Bill

What a summary call costs

An AI agent compacts the window below, S: system, U: user, O: answer, C: summary. Each cell is 100 tokens. Input is priced at 2 cents per 100 tokens, output at 8 cents per 100 tokens. What does the compaction call cost?

Before, the window the model reads

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

After, the summary it writes in their place

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

Step 1: Count the input, every cell the model reads

$$300 + 3 \times (100 + 300) = 1,500 \text{ tokens}$$

Step 2: Count the output, the summary alone

$$3 \times 100 = 300 \text{ tokens}$$

Step 3: Price each at its own rate and add

$$1,500/100 \times 2 + 300/100 \times 8 = 30 + 24 = 54 \text{ cents}$$

12. The Compaction Bill (continued)

Practice 1

An AI agent compacts the window below, S: system, U: user, O: answer, C: summary. Each cell is 100 tokens. Input is priced at 2 cents per 100 tokens, output at 8 cents per 100 tokens. What does the compaction call cost?

Before, the window the model reads

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		

After, the summary it writes in their place

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

Step 1: Count the input, every cell the model reads

Step 2: Count the output, the summary alone

Step 3: Price each at its own rate and add

12. The Compaction Bill (continued)

Practice 2

An AI agent compacts the window below, S: system, U: user, O: answer, C: summary. Each cell is 100 tokens. Input is priced at 3 cents per 100 tokens, output at 9 cents per 100 tokens. What does the compaction call cost?

Before, the window the model reads

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

After, the summary it writes in their place

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

Step 1: Count the input, every cell the model reads

Step 2: Count the output, the summary alone

Step 3: Price each at its own rate and add