

18. Context Tax

Paying for the Past

An AI agent has a 200-token system prompt, and each turn is a 100-token question and a 200-token answer. Input costs 2 cents per 100 tokens, output 8 cents per 100. The window already holds three turns, S: system, U: question, O: answer.

Three turns, already in the window

1 - 1,000	S	S	(U1	O1	O1)	(U2	O2	O2)	(U3	O3
1,001 - 2,000	O3)									

Step 1: Count what each turn sends and gets back

Turn	Input	Input token count	Output	Output token count
1	S S U1	300	O1 O1	200
2	S S U1 O1 O1 U2	600	O2 O2	200
3	S S U1 O1 O1 U2 O2 O2 U3	900	O3 O3	200

Step 2: Price each turn, the two halves and their total

Turn	Input cost	Output cost	Total cost
1	$2 \times 3 = 6$	$8 \times 2 = 16$	$6 + 16 = 22 \text{ cents}$
2	$2 \times 6 = 12$	$8 \times 2 = 16$	$12 + 16 = 28 \text{ cents}$
3	$2 \times 9 = 18$	$8 \times 2 = 16$	$18 + 16 = 34 \text{ cents}$

18. Context Tax (continued)

Practice 1

An AI agent has a 100-token system prompt, and each turn is a 100-token question and a 100-token answer. Input costs 3 cents per 100 tokens, output 9 cents per 100. The window already holds three turns, S: system, U: question, O: answer.

Three turns, already in the window

1 - 1,000	S	(U1	O1)	(U2	O2)	(U3	O3)			
1,001 - 2,000										

Step 1: Count what each turn sends and gets back

Turn	Input	Input token count	Output	Output token count
1	S U1		O1	
2	S U1 O1 U2		O2	
3	S U1 O1 U2 O2 U3		O3	

Step 2: Price each turn, the two halves and their total

Turn	Input cost	Output cost	Total cost
1			
2			
3			

18. Context Tax (continued)

Practice 2

An AI agent has a 300-token system prompt, and each turn is a 200-token question and a 200-token answer. Input costs 2 cents per 100 tokens, output 5 cents per 100. The window already holds three turns, S: system, U: question, O: answer.

Three turns, already in the window

1 - 1,000	S	S	S	(U1	U1	O1	O1)	(U2	U2	O2
1,001 - 2,000	O2)	(U3	U3	O3	O3)					

Step 1: Count what each turn sends and gets back

Turn	Input	Input token count	Output	Output token count
1	S S S U1 U1		O1 O1	
2	S S S U1 U1 O1 O1 U2 U2		O2 O2	
3	S S S U1 U1 O1 O1 U2 U2 O2 O2 U3 U3		O3 O3	

Step 2: Price each turn, the two halves and their total

Turn	Input cost	Output cost	Total cost
1			
2			
3			