17. Retrieval Cost

Retrieved Chunks Are Input

An AI agent has a 200-token system prompt, and each turn is a 100-token question and a 200-token answer. On turn two it retrieves a 300-token document. Mark three turns, then write out the chunks each request carries, S: system, U: question, R: retrieved chunk, O: answer.

Step 1: Mark three turns, with the retrieval on turn two

1 - 1,000	S	S	(U1	O1	O1)(U2	R2	R2	R2	O2
1,001 - 2,000	O2)(U3	O3	O3)						

Step 2: Write the chunks each turn sends and gets back

Turn	Input	Output
1	S S U1	O1 O1
2	S S U1 O1 O1 U2 R2 R2 R2	O2 O2
3	S S U1 O1 O1 U2 R2 R2 R2 O2 O2 U3	O3 O3

17. Retrieval Cost (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. On turn two it retrieves a 200-token document. Mark three turns, then write out the chunks each request carries, S: system, U: question, R: retrieved chunk, O: answer.

Step 1: Mark three turns, with the retrieval on turn two

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

Step 2: Write the chunks each turn sends and gets back

Turn	Input	Output
1		
2		
3		

Practice 2

An AI agent has a 100-token system prompt, and each turn is a 100-token question and a 100-token answer. On turn one it retrieves a 300-token document. Mark three turns, then write out the chunks each request carries, S: system, U: question, R: retrieved chunk, O: answer.

Step 1: Mark three turns, with the retrieval on turn one

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

Step 2: Write the chunks each turn sends and gets back

Turn	Input	Output
1		
2		
3		