

7. Retrieval per Turn

Paying for the search every turn

An AI agent has a 4,000-token context window and a 300-token system prompt. On every turn it retrieves a 300-word document from its database at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 200 user tokens and 300 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

$$(300 \times 2)/100 = 6 \text{ chunks}$$

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

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

Step 3: Count the turns

$$(4,000 - 300)/(600 + 200 + 300) = 3.36, \text{ so } 3 \text{ turns}$$

7. Retrieval per Turn (continued)

Practice 1

An AI agent has a 3,000-token context window and a 200-token system prompt. On every turn it retrieves a 200-word document at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 100 user tokens and 200 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

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

Step 3: Count the turns

7. Retrieval per Turn (continued)

Practice 2

An AI agent has a 5,000-token context window and a 300-token system prompt. On every turn it retrieves a 900-word document at 2 tokens per word, split into chunks of 100 tokens, and the turn adds 200 user tokens and 300 output tokens of its own. How many turns can fit the window?

Step 1: The document in chunks

Step 2: Mark the context window, S: system, U: user, R: retrieved, O: output, and () for turn boundaries.

1 - 1,000									
1,001 - 2,000									
2,001 - 3,000									
3,001 - 4,000									
4,001 - 5,000									

Step 3: Count the turns