

13. Compacted Retrieval

A summary instead of raw chunks

An AI agent retrieves a 1,200-token document to answer one question. Instead of adding the raw chunks, the agent compacts them into a summary that keeps 25% of the document. Mark the resulting window, S: system, U: question, R: retrieved chunk, C: summary, O: answer.

Without compacting, the raw chunks go straight in

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

Step 1: Take the share the summary keeps

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

Step 2: Mark the window with the summary in place of the chunks

1 - 1,000	s	s	(	U	C	C	C	O	O	)
1,001 - 2,000										