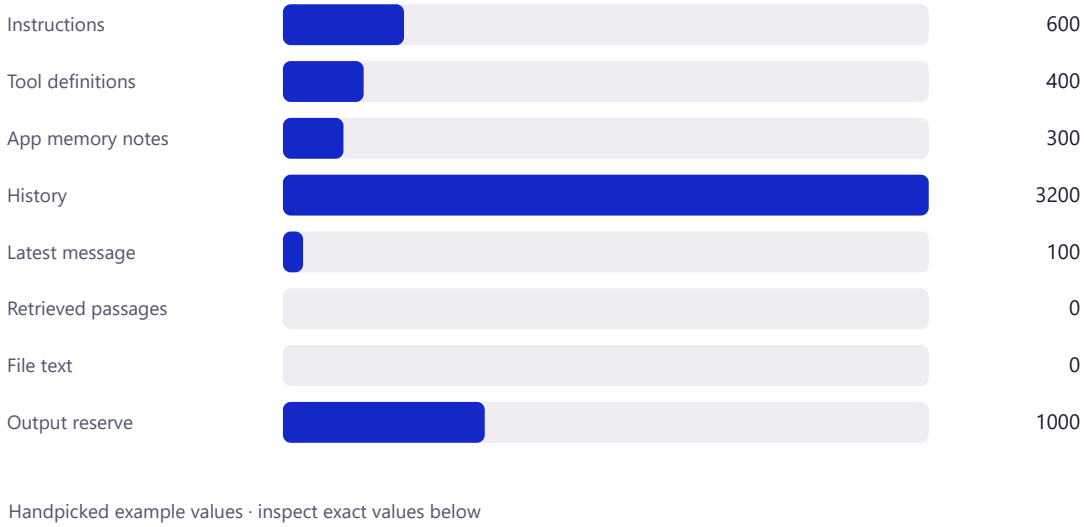
CHAPTER 12 / CONTEXT

Build the request.

Memory belongs to the application, not a weight update.

Memory belongs to the application, not a weight update.



Illustrative token counts, not provider limits. Fitting context does not guarantee equal use; summaries can omit details.

Assembly

The application chooses what information is included in each request.

Budget

Count input components and reserve room for the answer.

Persistence

An app can store and resend notes; request context does not update model weights.

↗ **The application chooses what information is included in each request.**

PUT IT TO WORK

The default uses 5,600 of an invented 8,000-token budget. Everything included uses 9,100; summarized history reduces that to 6,500.

CLOSE THE PAGE. EXPLAIN IT.

What must an app send again to make an assistant appear to remember last week?