

The same 55K MCP definitions occupy 27.5% at 200K, but 5.5% at 1M

Anthropic example: ≈55K tokens of loaded tool definitions before the first tool call

5× larger context share at 200K

200K context window



55K loaded definitions

145K remaining

27.5% occupied

Before task input, files, history, tool results, and output

1M context window



55K loaded definitions

945K remaining

5.5% occupied

Before task input, files, history, tool results, and output

No tool has been called yet.

But the definitions already occupy model context.

How much context are you allocating to tools the agent may never call?