Max Tokens Planner - Summary
Generated: 2026-05-17 03:52:38

Model: Example
Context limit: 8192 tokens
Prompt tokens (est.): 2200
Expected output: 900
Total tokens: 3100
Remaining context: 5092
Recommended max output: 4500
Status: OK

Example scenarios:

- Short chat / quick answer | ctx 4096 | prompt 900 | out 300 | total 1200 | OK
- Medium chat / tool calls | ctx 8192 | prompt 2600 | out 800 | total 3400 | OK
- Long history / long output | ctx 16384 | prompt 8200 | out 3200 | total 11400 | OK
- Tight budget / risk of overflow | ctx 4096 | prompt 3700 | out 700 | total 4400 | OVER