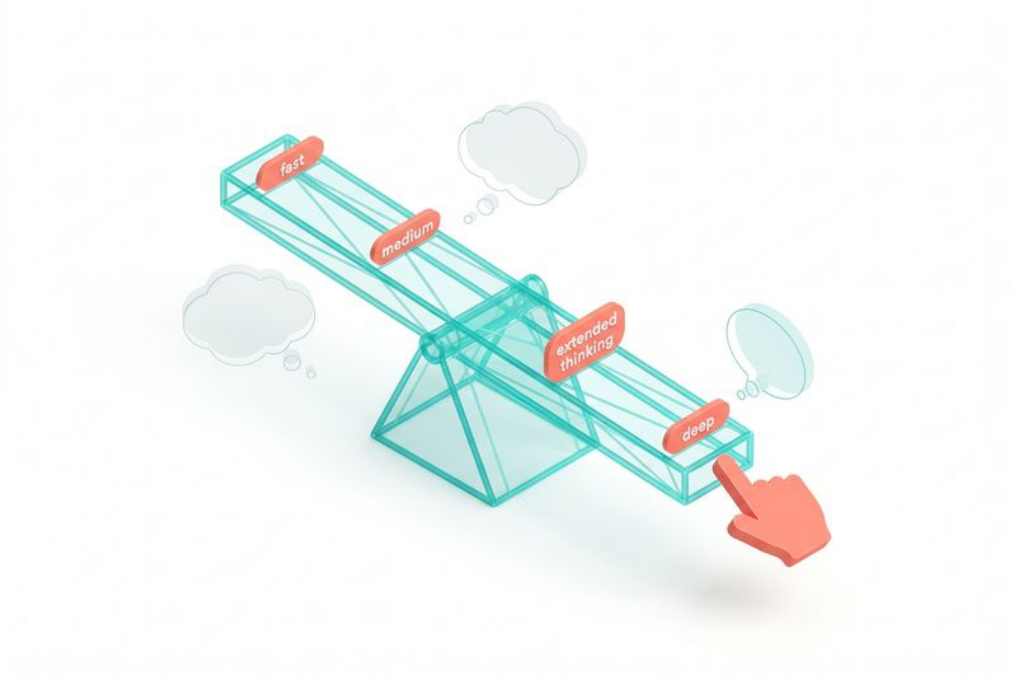
The Thinking Lever

When to pull extended thinking and how to tune the budget without burning context.



When to enable extended thinking

- The task requires planning across multiple steps
- The work touches more than three files and benefits from a written plan
- You would normally write a design doc before coding

When to skip it

- The task is a one-line config change
- The agent already has a clear path and stalling is the only risk
- Latency budget cannot tolerate the extra round trip

Tuning the budget

- Start at the default and only raise it if the model fails to plan
- A bigger budget rarely fixes a bad prompt
- For production traffic, measure the tradeoff before turning it on globally

From "Claude Code, Definitive Guide for 2026" — Chapter 4. Source: The thinking lever talk, Code with Claude London 2026.