
Running Local LLMs
on Your Laptop
A practical starter cheatsheet for Apple Silicon and small open-source models
Top Local Models
Context Lengths
Performance Tips
Your guide to experimenting with LLMs on your own machine.

## Mental model

TTFT: how fast the model starts responding
Throughput (tokens per second): how fast it continues generating
Context length: how much fits into a single prompt
Quantization (Q4 vs Q8): speed and memory versus stability and accuracy

## Recommended starting profiles

**Fast interactive coding**
Model: Qwen3-VL-4B
Context: 2048 tokens
Best for quick edits, short prompts, and fast iteration

**Long-context coding and refactors**
Model: DeepSeek-Coder-V2-Lite (Q8_0)
Context: 4096 tokens
Best for multi-file refactors and stable code generation

**Code review and explanation**
Model: Qwen3-4B Kimi-K2 Thinking
Context: 4096 tokens
Best for reasoning-heavy explanations and structured reviews

## LM Studio settings that matter

Context length: use 2048 for speed, 4096 for larger refactors
Threads: Auto works well on Apple Silicon
Batch size: Auto is recommended; higher values may improve throughput at the cost of latency

## Sampling defaults for coding

Temperature: 0.2
Top-p: 0.9
Repeat penalty: 1.1

## Q4 vs Q8 rule of thumb

Use Q8 for correctness, refactors, and long prompts
Use Q4 for speed, lower memory usage, and quick experimentation

## Simple daily workflow

Coding: small model with 2048-token context
Refactoring: coder-focused model with 4096-token context and Q8
Reviewing: reasoning-focused model with 4096-token context

## Troubleshooting

If everything feels slow, reduce context length or model size
If output feels unstable, lower temperature or switch to Q8
If the model repeats itself, slightly increase repeat penalty