# Claude Code Without the Cloud Bill

A Practical Guide to Running AI-Assisted Development Locally

# Why This Matters

Sonnet 4.5: $3/$15 per M tokens

Opus 4.5: $5/$25 per M tokens

Opus 4.6: $5/$25 per M tokens

**$600-1,200**

estimated monthly cost per developer using cloud APIs

# What You'll Need

## Hardware Requirements

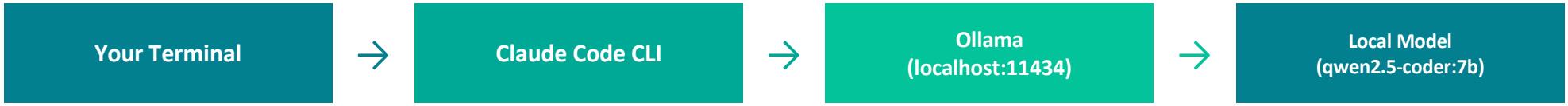
- macOS 14+ (Sonoma or later)
- 8GB+ RAM (16GB+ recommended)
- Basic terminal familiarity

## Software Requirements

- Ollama v0.15+ (free, open-source)
- Claude Code CLI (free to install)
- Node.js (for npm scripts)

**Tested: MacBook Pro M1 Max, 64GB RAM, macOS Tahoe**

# How It All Connects




**Everything runs on YOUR machine. Zero cloud calls. Zero billing.**

# Step 1: Install Ollama

```
# Check version
ollama --version
ollama --version
# Download from ollama.com/download (~104MB)


# Download from ollama.com/download (~104MB)
```

Ollama is your local AI runtime — think of it as Docker for AI models

# Model

## qwen2.5-coder:7b

- Size: 4.7GB
- RAM: 8GB+
- Speed: Fast
- Best for: Coding tasks
- MCP Tools: Limited

**RECOMMENDED FOR STARTERS**

## gpt-oss:20b

- Size: 11GB
- RAM: 16GB+
- Speed: Slower
- Best for: General + MCP tools
- MCP Tools: Full support

# Step 3: Launch

```
ollama launch claude --model qwen2.5-coder:7b
```

**NEVER run `ollama launch claude` without --model flag! It may default to cloud models and charge your account.**

## What the launch command does:

- Sets up environment variables automatically
- Connects Claude Code to your local Ollama
- Starts your AI coding session

# Verify Local Setup

```
Model: qwen2.5-coder:7b
Base URL: http://localhost:11434
Token usage: ↑ 0 tokens ← confirms no billing
```

**Red flags to watch for:**

- Model shows "claude-sonnet-4-5" = Cloud (BILLING!)
- Auth conflict warnings = Cloud API active

**Green flags:**

- Model shows your local model name
- 0 tokens used

# Using the Repo

| | |
|---|---|
| `npm start` | Launch with default model (qwen) |
| `npm run start:gpt-oss` | Launch with gpt-oss:20b |
| `npm run setup:web-search` | Add free web search |
| `npm run reset:mcp` | Reset MCP servers |
| `npm test` | Run workflow tests |

# Add Web Search — Still Free

## How:

- `npm run setup:web-search`

Requirements: Works best with gpt-oss:20b model

## What you get:

- DuckDuckGo search (no API key)
- URL content fetching
- Real-time data access

# Watch Out For These

**Don't type /model**

Switches to paid cloud models

**Always specify --model**

Default may use cloud

**Check token usage**

0 tokens = local, otherwise billing

**MCP tool limits**

qwen2.5-coder can't execute tools

# What You Save

**Before**

$600-1,200/month per developer

(cloud API)

**After**

$0/month

(local models)

**Save $7,200-14,400 per developer per year**

# Resources & Next Steps

- GitHub: github.com/TheRobBrennan/how-to-setup-local-ollama-with-claude-code
- Ollama: ollama.com
- Claude Code: docs.anthropic.com/claude/docs/claude-code

**Questions? rob@sploosh.ai**