9 LOCAL LLMS TESTED

Does Qwen 3.5 Live Up to the Hype?

I tested them on a Claude Code skill I actually use every day.

Not a coding benchmark.

A real multi-step agentic task described in natural language as a markdown file.

Khur Boon Kgim

AI-Powered Dad and Technopreneur

The Test

My **/linkedin-cover-image** skill works like this:

STEP 1

Read a LinkedIn post file and analyze the content

STEP 2

Decide if it is a post or article and choose the correct dimension

STEP 3

Pick the right layout template (quote, listicle, contrast)

STEP 4

Write a complete HTML cover image - typography, colors, spacing, decorative elements

STEP 5

Screenshot it to PNG

One command. No hand-holding. The model handles the entire workflow.

REFERENCE

Opus 4.6

Network graph decorations, gradient effects, clean text hierarchy.

Setting up an AI agent is a demo. Maintaining it is DevOps.

6 things nobody tells you about running AI agents in production

Budget 20x more time.

#1 USABLE - BEST

GLM 4.7 Flash

Highlighted both contrast keywords ("demo" and "DevOps"). Clean output.

Setting up an AI agent is a *demo*. Maintaining it in production is *DevOps*.

From personal assistant to business-critical service — 6 production realities you need to budget for.

Budget 20x more time for maintenance than setup

#2 USABLE

Qwen 3 VL (32B)

Nice layout, clean. Only highlighted "DevOps", missed the contrast on "demo".

Setting up an AI agent is a demo. Maintaining it in production is DevOps.

6 things that hit when you go from personal use to production

Budget 20x more time for maintenance than the setup.

#3 USABLE

Qwen 3 VL (30B)

Similar to 32B. Clean but missed the "demo" highlight.

Setting up an AI agent is a demo. Maintaining it is DevOps.

The moment your agent serves other people or runs a business-critical workflow, it is a production service.

Budget 20x more time for maintenance than setup.

#4 USABLE

GPT-OSS (20B)

Usable but not as polished.

Setting up an AI agent is a demo. Maintaining it in production is DevOps.

OpenClaw is great, but if you build for clients, budget 20x more time for maintenance.

#5 USABLE

Qwen 3 (30B)

Clean output though arguably highlighted the wrong keyword.

Setting up an AI agent is a demo. Maintaining it is DevOps.

6 production things to budget for

Budget 20x more time

#6 ISSUES

Qwen 3.5 (35B)

Got the content right but the layout of the last sentence is off.

Setting up an AI agent is a demo. Maintaining it in production is DevOps.

6 things that hit when you go from personal use to production — nobody posts about this.

Budget 20x more time for maintenance than setup.

#7 ISSUES

Qwen 3 Coder (30B)

Long winded, highlighted the same wrong keyword as Qwen 3, and created an article cover instead of a post cover. Needed a follow-up prompt.

Setting up an AI agent is a demo. *Maintaining* it in production is DevOps.

Budget 20x more time for the maintenance than the setup

OpenClaw is a great personal tool. But if you are building on it for clients or plugging it into workflows your business depends on, budget 20x more time for the maintenance than the setup.

#8-9 FAILED

Devstral Small 2 & Magistral 24B

Kept saying "I will create the HTML cover" then produced nothing. Repeated the loop and never wrote a file.

```
❯ /linkedin-cover-image  "/Users/boonkgim/codes/boon

● I'll create a LinkedIn cover image for your AI age

● Read 1 file (ctrl+o to expand)

● Now I'll create a cover image for this post. The p
  with 6 specific challenges. This calls for Layout

  Let me create the HTML cover:

* Crunched for 36s

❯ continue

● I'll create a LinkedIn cover image for your AI age
  message.

  Let me write the HTML file:

─────────────────────────────────────────────────────

❯
```

The Verdict

Same post. Same skill. One shot each.

THE REAL GAP

Multi-step agentic workflows where the model reads context, makes real world decisions, and writes a complete file - that is where you see the real gap.

WINNER

GLM 4.7 Flash got the closest I have seen from any local model. But "closest" is not "there yet."

THE TAKEAWAY

Everyone benchmarks local LLMs on some standard benchmark. I do it on a real daily use case.

Let's Connect

Which local LLM works best for your agentic tasks?

> Follow for more AI insights
> Repost to help others
> Comment with your experience

#AI #ClaudeCode #LocalLLM

Khur Boon Kgim

AI-Powered Dad and Technopreneur