Don't believe the BS that you can use Claude Code for free.

I tested Claude Code + 7 local LLMs

Khur Boon Kgim

AI-Powered Dad and Technopreneur

Claude Code without Opus 4.5 is not Claude Code

But it is still great for cost optimization if we can delegate easier tasks to local LLMs.

In this post, I run an experiment to test if local LLMs can follow simple copy-paste and update instructions.

This is my 1st round interview for 7 local LLMs.

My Experiment

Testing local LLMs' ability to follow instructions and code based on templates

INITIAL FINDINGS

Local LLMs struggle with research and planning tasks.

Let's see if they can handle copy-paste and update well.

WHAT I DID

1. Used Opus to create an HTML template for a slide
2. Turned one slide into a template and deleted the rest
3. Asked each LLM to refer to the template to code its own slide
4. Added footnote to each slide

→

[MODEL_NAME]

[TAG_1] [TAG_2] [TAG_3]

Size	[SIZE]
Parameters	[PARAMETERS]
Context	[CONTEXT]

KEY FEATURES / INTENDED USE

[KEY_FEATURES]

devstral-small-2

Vision Tools Agentic

Size	15 GB
Parameters	24B
Context	384K

KEY FEATURES / INTENDED USE

Agentic LLM for codebases; includes vision encoder.

Result: Hallucinated "Agentic" as a capability.

GLM-4.7-Flash

Thinking Tools

Size	19 GB
Parameters	30B (MoE)
Context	198K

KEY FEATURES / INTENDED USE

High-efficiency model; production-ready tool calling.

Result: Perfect.

gpt-oss:20b

Thinking Tools Python execution

Size	13 GB
Parameters	20B
Context	128K

KEY FEATURES / INTENDED USE

Low-latency reasoning, Python execution, and agentic tasks.

Result: Hallucinated "Python execution" as a capability.

qwen3:30b

Thinking Tools MoE

Size	18 GB
Parameters	30B
Context	256K

KEY FEATURES / INTENDED USE

General-purpose MoE with enhanced reasoning and tool use.

Result: Hallucinated "MoE" as a capability.

Qwen3-Coder:30B

Tools | Software Engineering | Agentic

Size	18 GB
Parameters	30B (3.3B active)
Context	256K - 1M

KEY FEATURES / INTENDED USE

Specialized agentic model for software engineering. Designed for complex coding tasks, code generation, and development workflows. Excels in understanding and generating code across multiple programming languages with deep contextual understanding.

Result: Hallucinated "Software Engineering" and "Agentic" as capabilities. Also hallucinated the Key Features / Intended Use.

Qwen3-VL:30B

Model Name	Parameters	Context	Memory	Tools	Description
qwen3-vl:30b	30B (3.3B active)	256K - 1M	18 GB	Vision-Language	Specialized agentic model for software engineering.

Result: Failed!

qwen3-vl:32b

Vision Thinking Tools

Size	20 GB
Parameters	32B
Context	256K

KEY FEATURES / INTENDED USE

Thinking-enabled vision model; improved visual coding.

Result: Perfect. But it only has a 40k tokens context window with 32GB VRAM.

The Verdict

INTERVIEW SUMMARY

GLM-4.7-Flash and **qwen3-vl:32b** followed instructions perfectly with no hallucinations.
Qwen3-VL:30B failed completely in the task.
Other models had minor hallucinations but were still usable for basic tasks.

NEXT STEPS

This 1st round of interviews was probably too easy. I will conduct a more challenging second round to better evaluate the models.

Want to join as one of the interviewers?

GET INVOLVED

1. Drop a comment on what you want me to test
2. Connect with me
3. I will share the interview results afterward

Let's Connect

Which local LLM works for you?

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

#Ollama #AI #LLM #ModelSpecs

Khur Boon Kgim

AI-Powered Dad and Technopreneur