

What Local AI can do?

First of all, we need to understand the overall of all tools, models, and services:

Programming:

- Claude is the king -> Critical Planner (Expensive)
- Codex is Queen -> Broad Planner (Expensive)
- GLM / Kimi K3 -> Worker
- Qwen, Gemma -> Free

Image Generation:

- Google Gemini
- GPT

Video Generation:

- Seedance
- Veo

CLI with good system prompts:

- Claude
- Pi
- Gemini

Top Agentic Apps:

- Hermes Agent

Local LLM

- Fable gemma
- Fable qwen
- Bonsai convert from big model to small

What can local llm can do right now:

1. does it smart enough to follow exactly the instruction of my language?
2. does it able to do long tasks?
3. how fast token generation it can do on our a 24 GB VRAM GPU 3090?
4. what dataset do we have and public have? old vs new?

Revision #3

Created 2026-07-18 03:24:24 UTC by Son.KyLuat

Updated 2026-07-18 03:52:14 UTC by Son.KyLuat