AJENTIC AI - Local LLaMA Setup Guide (Offline - No Ollama Needed)

==

1) Download llama.cpp binary (llama.exe for Windows)

CPU-only (x64):

https://github.com/ggml-org/llama.cpp/releases/download/b6962/llama-b6962-bin-win-cpu-x64.zip

CUDA 12.4 (x64, faster on NVIDIA):

https://github.com/ggml-org/llama.cpp/releases/download/b6962/llama-b6962-bin-win-cuda-12.4-x64.zip

Extract and place 'llama.exe' at:

C:\temp\ajentic-ai-local\bin\llama.exe

2) Download GGUF Model (Qwen2.5 Coder)

Recommended (Q5_K_M, ~5.44 GB):

https://huggingface.co/bartowski/Qwen2.5-Coder-7B-Instruct-GGUF/resolve/main/Qwen2.5-Coder-7B-Instruct-Q5_K_M.gguf

Alternative Smaller (~4.68 GB):

https://huggingface.co/bartowski/Qwen2.5-Coder-7B-Instruct-GGUF/resolve/main/Qwen2.5-Coder-7B-Instruct-Q4_K_M.gguf

Higher Quality (~6.52 GB):

https://huggingface.co/bartowski/Qwen2.5-Coder-7B-Instruct-GGUF/resolve/main/Qwen2.5-Coder-7B-Instruct-Q6_K_L.gguf

Place the model file at:

C:\temp\ajentic-ai-local\models\model.gguf

3) Folder Structure

```
C:\temp\ajentic-ai-local\
  bin\
    llama.exe
  models\
    model.gguf
```

4) Verify llama.cpp works

Open PowerShell:

```
cd C:\temp\ajentic-ai-local
. in\llama.exe -m .\models\model.gguf -p "Say HELLO WORLD once and stop." -n 64 --temp 0.1
```

5) Run Ajentic AI with Local LLaMA

```
py server.py
```

Open browser at: http://localhost:8099/

Click: "Generate testcase (Local LLM)" -> It will auto-use ./bin/llama + ./models/model.gguf

6) Notes

- CPU build works on all Windows 10/11 x64 systems.
- CUDA build is recommended if you have an NVIDIA GPU and CUDA 12.4 installed.
- Q4_K_M and Q5_K_M models are suitable for 16-32 GB RAM systems.
- Q6_K_L gives higher accuracy but needs more RAM.

Author: ChatGPT GPT-5

Date: November 2025