<!DOCTYPE html>

<html lang="en">

<head>

<meta charset="UTF-8" />

<meta name="viewport" content="width=device-width, initial-scale=1.0"/>

<title>Local LLM Deployment Guide – DeepSeek & MiniMax-M1</title>

<style>

body {

font-family: Arial, sans-serif;

margin: 40px;

line-height: 1.6;

background-color: #fdfdfd;

}

code {

background: #f4f4f4;

padding: 2px 5px;

border-radius: 4px;

font-family: monospace;

}

pre {

background: #f4f4f4;

padding: 12px;

border-radius: 6px;

overflow-x: auto;

}

h1, h2, h3 {

color: #333;

}

a {

color: #007acc;

text-decoration: none;

}

a:hover {

text-decoration: underline;

}

.section {

margin-bottom: 60px;

}

</style>

</head>

<body>

<h1>🚀 Local LLM Deployment Guide: DeepSeek & MiniMax-M1</h1>

<p><strong>For:</strong> Windows-based laptop deployment with Python scripting and agent integration.</p>

<div class="section">

<h2>1. Prerequisites</h2>

<ul>

<li><a href="https://www.python.org/downloads/windows/">Python 3.10+</a></li>

<li><a href="https://www.nvidia.com/Download/index.aspx">CUDA Drivers (NVIDIA GPU)</a></li>

<li><a href="https://git-lfs.github.com/">Git LFS</a></li>

<li><a href="https://visualstudio.microsoft.com/downloads/">Visual Studio C++ Build Tools</a></li>

<li><a href="https://ollama.com/windows">Ollama (optional deployment tool)</a></li>

<li><a href="https://www.docker.com/products/docker-desktop">Docker Desktop (optional)</a></li>

</ul>

</div>

<div class="section">

<h2>2. Python Environment Setup</h2>

<pre><code>python -m venv llm\_env

.\llm\_env\Scripts\activate

pip install --upgrade pip

pip install transformers accelerate torch

pip install vllm langchain</code></pre>

<p>Make sure to match your <code>torch</code> install with your CUDA version using <a href="https://pytorch.org/get-started/locally/">this selector</a>.</p>

</div>

<div class="section">

<h2>3. Model Downloads</h2>

<h3>DeepSeek (Distilled)</h3>

<pre><code>git lfs install

git clone https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B</code></pre>

<h3>MiniMax-M1</h3>

<pre><code>git clone https://huggingface.co/MiniMaxAI/MiniMax-M1-40k</code></pre>

<p>Use the <code>huggingface-cli</code> tool if you'd like to download files individually.</p>

</div>

<div class="section">

<h2>4. Basic Model Load (Python)</h2>

<pre><code>from transformers import AutoModelForCausalLM, AutoTokenizer

model\_path = "path/to/DeepSeek-R1-Distill-Qwen-7B"

model = AutoModelForCausalLM.from\_pretrained(model\_path, trust\_remote\_code=True, device\_map="auto")

tokenizer = AutoTokenizer.from\_pretrained(model\_path, trust\_remote\_code=True)

prompt = "User: What's the Pythagorean theorem?\nAssistant:"

inputs = tokenizer(prompt, return\_tensors="pt").to(model.device)

outputs = model.generate(\*\*inputs, max\_new\_tokens=200)

print(tokenizer.decode(outputs[0], skip\_special\_tokens=True))</code></pre>

</div>

<div class="section">

<h2>5. Ollama Deployment (Optional)</h2>

<pre><code>ollama pull deepseek-r1:7b

ollama run deepseek-r1:7b</code></pre>

<p>Use the Ollama UI or API for programmatic interaction.</p>

</div>

<div class="section">

<h2>6. LangChain Agent with Shell Tool</h2>

<pre><code>from langchain.agents import initialize\_agent, Tool

import subprocess

def run\_shell(command):

result = subprocess.run(command, shell=True, capture\_output=True, text=True)

return result.stdout.strip() if result.stdout else result.stderr.strip()

shell\_tool = Tool(

name="ShellExecutor",

func=run\_shell,

description="Runs a shell command on the local machine."

)

# Initialize LangChain agent (assuming Ollama is running)

from langchain.llms import Ollama

llm = Ollama(model="deepseek-r1:7b", base\_url="http://localhost:11434")

agent = initialize\_agent(tools=[shell\_tool], llm=llm, agent="zero-shot-react-description", verbose=True)

print(agent.run("Create a folder named testFolder on my desktop."))</code></pre>

</div>

<div class="section">

<h2>7. References & Resources</h2>

<ul>

<li><a href="https://huggingface.co/deepseek-ai">DeepSeek on Hugging Face</a></li>

<li><a href="https://huggingface.co/MiniMaxAI">MiniMaxAI on Hugging Face</a></li>

<li><a href="https://github.com/deepseek-ai/DeepSeek-V3">DeepSeek GitHub</a></li>

<li><a href="https://github.com/MiniMax-AI/MiniMax-M1">MiniMax GitHub</a></li>

<li><a href="https://python.langchain.com/">LangChain Docs</a></li>

<li><a href="https://ollama.com">Ollama Docs</a></li>

</ul>

</div>

<p><em>Created with ♥ by your AI companion. Happy deploying!</em></p>

</body>

</html>