Local LLM (Offline AI) — Ek Standard Guide (Hinglish)

1. Introduction aur Definition

Local LLM = aisa Large Language Model jo internet pe depend na kare aur aapke **machine/server** par hi chale (offline mode).
Benefits: Data privacy, low latency, offline use, aur pura control aapke paas.

2. Model Versions / Types

1. **Chhote / Mid-size models (1B–13B parameters)** – Laptop/consumer GPU pe run karne layak. (e.g., GPT4All models)
2. **Large models (30B–70B)** – 48GB+ VRAM ya multi-GPU setup required.
3. **Very Large models (100B+)** – Datacenter level GPUs (A100/H100) chahiye.
4. **Optimized On-device builds (Apple M-series, AMD Ryzen AI)** – CPU / M-series chips pe bhi chalne wale optimized runtime available.

3. Hardware Suggestions (GPU / CPU)

A. Consumer-level (Personal Use / Prototype)

- Example: **NVIDIA RTX 4080 / 4090**
- Achhi VRAM aur compute power ke saath 7B–13B models easily run ho jate hain.

Cost Jayda mehaneg hai buy iski 1.25 lakh something but perofrmace is too good

B. Pro / Server-level (Bigger Models ke liye)

- Example: **NVIDIA A6000, L40, RTX 6000** (48GB–80GB VRAM)
- 30B–70B models ke liye suitable, fine-tuning bhi ho sakti hai.

C. Data-center GPUs (Production / Enterprise)

- Example: **NVIDIA A100 / H100**
- 100B+ models ya large-scale serving ke liye best option.

Note on CPU / Apple Silicon

- Apple M-series (M1/M2/M3) aur AMD Ryzen AI chips small/mid models ke liye kaafi optimized hai.

AMD Radeon RX 550

this cost **is** 6000 – 8000 smart but slow response gpu

---

4 Tools & Frameworks (Popular Choices)

- **llama.cpp** → Local inference ke liye lightweight C/C++ based runtime.
- **GPT4All (Nomic)** → Desktop-friendly local LLMs (3B–13B models).
- **Hugging Face (Transformers, Accelerate, bitsandbytes)** → Training, fine-tuning aur quantization ke liye.
- **vLLM / Ollama / FastAPI serving** → API aur production deployment ke liye.
- **AMD Gaia / ONNX Runtime** → AMD/Windows optimized solutions.

5. Manual Implementation — Step by Step

Step A. Environment Setup

```
python -m venv llm_env
source llm_env/bin/activate
pip install --upgrade pip
pip install transformers accelerate bitsandbytes
```

Step B. Model Download

- Hugging Face / GPT4All se required model (7B/13B quantized) download karo.

Step C. Quantization / Conversion

- Model ko **4-bit/8-bit (GGUF, GGML format)** me convert karo.

Step D. Local API / Server

```
from fastapi import FastAPI
app = FastAPI()

@app.post("/generate")
async def gen(prompt: str):
    # Call local model inference here
    return {"text": "Generated response"}
```

Step E. RAG (Retrieval-Augmented Generation) add karna

1. Documents → Embeddings (Sentence Transformers).
2. Store in FAISS/Milvus vector DB.
3. Query + Context → LLM → Better domain-answers.

Step F. Optimization

- Use batching, streaming, quantized kernels.
- GPU profiling tools (nvidia-smi, nsys) se monitor karo.

---

6. Costing (Approximate, 2025 ke hisaab se)

- **Laptop/CPU-only**: ₹0–₹25,000 extra (agar machine already hai).
- **Consumer GPUs (RTX 4070/4080/4090)**: ₹80,000 – ₹1.8 Lakh.
- **Pro GPUs (A6000 / L40)**: ₹3 Lakh – ₹10 Lakh.
- **Data-center GPUs (A100/H100)**: ₹10 Lakh – ₹50 Lakh+.
- **Cloud GPU hourly rental**: ₹150 – ₹2000/hour (model aur GPU type ke hisaab se).
- **Costing in sahi ki almost mahngi hi hai sir**
