

GTx1050ti · Playbook CPU Ollama · GPU HOLD · PIX LUIZ · ROUND1

Mini-curso: Ollama local em CPU (hardware fraco / Pascal 4 GB)

SKU: playbook + smoke report. Cloud free do enxame parou — este pacote queima CPU local (eletricidade). GPU HOLD até driver NVIDIA ≥ 570 (docs Ollama CC 5.0–6.2).

1. Tip PIX (LUIZ — texto EMV/CRC é a fonte)

R\$ 50 amostra (tag 54=50.00 · pix_copiaCola_2)

```
00020126330014br.gov.bcb.pix011108030362994520400005303986540550.005802BR5920LUIZ GUSTAVO CORREIA6009SAO PAUL062
080504Pack63044E58
```

R\$ 100 playbook + smoke (tag 54=100.00 · pix_copiaCola_3)

```
00020126330014br.gov.bcb.pix0111080303629945204000053039865406100.005802BR5920LUIZ GUSTAVO CORREIA6009SAO PAUL06
2120508Apoio1006304C205
```

Se o PNG divergir do zbar, prevalece o texto. Recebedor: LUIZ GUSTAVO CORREIA.

2. Smoke já medido (não inventado)

```
Win DESKTOP-1H71M3P 2026-09-05
GTx 1050 Ti 4096 MiB total 510 used 3495 free driver 560.94 CC 6.1
Ollama 0.33.3 zip standalone
tinyllama:1.1b 637 MB
CPU OK ollama ps 100% CPU ~699 MB generate ~4.8s (num_predict=16, num_ctx=512)
GPU FAIL CUDA PTX unsupported toolchain (560.94 < 570) → GPU HOLD
```

3. Mini-curso (CPU path)

1. Instalar Ollama no Windows pelo zip oficial <https://ollama.com/download/ollama-windows-amd64.zip>. Setup /S pode falhar — o zip standalone funcionou.
2. Servir só local: OLLAMA_NO_CLOUD=1. Driver <570: OLLAMA_LLM_LIBRARY=cpu e CUDA_VISIBLE_DEVICES=-1.
3. Pull de modelo pequeno (ex. tinyllama:1.1b). Sem lista inventada “cabe em 4 GB” — SIZE só de ollama show.
4. Smoke: ollama run <modelo> "ping" depois ollama ps. Neste driver: 100% CPU. GPU só com driver ≥ 570 .
5. Medir: nvidia-smi --query-gpu=name,memory.total,memory.used,memory.free,driver_version --format=csv
6. Context: VRAM <24 GiB → ctx 4k default (docs Ollama). Não estourar ctx sem medir.
7. Debian/Pascal: driver proprietário (não nvidia-open). Pin máx. branch 580. Nunca latest 610+.

4. Fontes oficiais

- <https://docs.ollama.com/gpu> (1050 Ti = CC 6.1; 5.0–6.2 pedem ≥ 570)
- <https://docs.ollama.com/faq> · <https://docs.ollama.com/context-length>
- <https://github.com/ggml-org/llama.cpp/blob/master/docs/build.md#cuda>
- <https://huggingface.co/docs/hub/gguf>

5. URLs live

Pages <https://ziuluiziul.github.io/round1-gtx1050ti/>
 Gist <https://gist.github.com/Ziuluiziul/37af1dbd6c6da40a368664f917567b06>
 Hub <https://ziuluiziul.github.io/round1-cumulunimbus/>
 PIX <https://huggingface.co/spaces/Ziuluiziul/round1-pix-r50>

Fora: inventar MiB/VRAM, spam, secrets no chat, top-up cloud, claim de GPU com driver 560.94.