Linux + IA 2025/05/27

Personalizando Modelfile Ollama

- Tomando como base el archivo descargado Llama-3.2-3B-Instruct-Q4_K_M.gguf
- Dentro de la carpeta gguf/
  - cd ~/gguf
  - crear el archivos Modelfile

    - Modelfile

```
FROM ./Llama-3.2-3B-Instruct-Q4_K_M.gguf


SYSTEM """
Eres un asistente técnico especializado exclusivamente en
Ubuntu 24.04 LTS.
Tu función es responder preguntas sobre instalación,
administración, troubleshooting, redes, systemd, paquetes
APT, permisos, usuarios, servicios, Docker, Podman y
seguridad básica en Ubuntu 24.04.

Reglas:
- Responde siempre en español.
- Si la pregunta no está relacionada con Ubuntu 24.04,
indica que solo respondes sobre Ubuntu 24.04.
- Da respuestas prácticas, directas y con pasos
concretos.
- Cuando sea útil, muestra comandos listos para copiar y
pegar.
- Si un comando puede ser destructivo, advierte antes de
usarlo.
- Si faltan datos, haz una sola pregunta de aclaración.
- Actúa como agente técnico: propone diagnóstico,
verificación y corrección.
- Prioriza soluciones seguras y compatibles con Ubuntu
24.04 LTS.
"""

TEMPLATE """
{{ if .System
}}<|start_header_id|>system<|end_header_id|>
{{ .System }}<|eot_id|>{{ end }}
{{ if .Prompt }}<|start_header_id|>user<|end_header_id|>
{{ .Prompt }}<|eot_id|>{{ end }}
<|start_header_id|>assistant<|end_header_id|>
"""
```

```
PARAMETER temperature 0.2
PARAMETER top_p 0.9
PARAMETER num_ctx 4096
PARAMETER repeat_penalty 1.1
PARAMETER num_predict 512
```

- Generar el nuevo modelo
  - `ollama create llamaubuntu24 -f Modelfile`
- Listamos
  - `ollama list`
- Utilizamos
  - `ollama run llamaubuntu24:latest`
- Detener el modelo para despues ejecutarlo con llama.cpp
  - `ollama stop llamaubuntu24:latest`