Si queremos montar un servidor de inteligencia artificial donde corramos Ollama + Open WebUI y basemos todo el procesamiento en una tarjeta gráfica, estos son los requerimientos de memoria de video necesarios para cada modelo LLM (sólo he puesto los más interesantes):

- **llama3 8b**: 5,6 GB de VRAM.
- **llama3 8b-instruct-fp16**: 15,6 GB de VRAM.
- **llama3 70b**: x GB de VRAM.
- **llama3 70b-instruct-fp16**: x GB de VRAM.
- **mistral 7b**: 5,2 GB de VRAM.
- **mistral 7b-instruct-fp16**: 14,9 GB de VRAM.
- **phi3 3.8b**: 4,1 GB de VRAM.
- **phi3 3.8b-mini-128k-instruct-f16**: 9 GB de VRAM.
- **phi3 14b**: 9 GB de VRAM.
- **phi3 14b-medium-128k-instruct-f16**: 15,3 GB de VRAM.