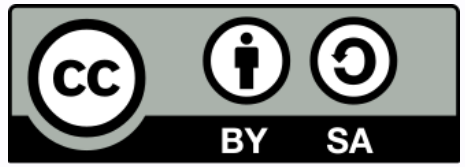
Things I did without (almost) a GPU

seLIA: 1st Workshop on Free Software and Open Artificial Intelligence

Fuenlabrada, Spain, July 6, 2026

Jesus M. Gonzalez-Barahona

https://jgbarah.github.io/presentations/



Usual equipment requirements

- Depends on architecture, and size of the model
- CPU can be enough, GPU can accelerate a lot
  - Cloud options for GPUs
- RAM enough so that model fits
- Code is usually FOSS

How far can we reach with easy-to-install software, and just a CPU?

Text generation

Ollama

```
curl -fsSL https://ollama.com/install.sh | sh
ollama serve
```

```
ollama run gemma3:1b
```

```
curl http://localhost:11434/api/generate -d '{
"model": "gemma3:1b",
"prompt":"Why is the sky blue?"
}'
```

Using Ollama to host an LLM on CPU-only equipment to enable a local chatbot and LLM API

I prefer:

```
wget https://ollama.com/download/ollama-linux-amd64.tar.zst
tar xvf ollama-linux-amd64.tar.zst
bin/ollama serve
```

```
bin/ollama run gemma3:1b
```

Si tienes suficiente RAM:

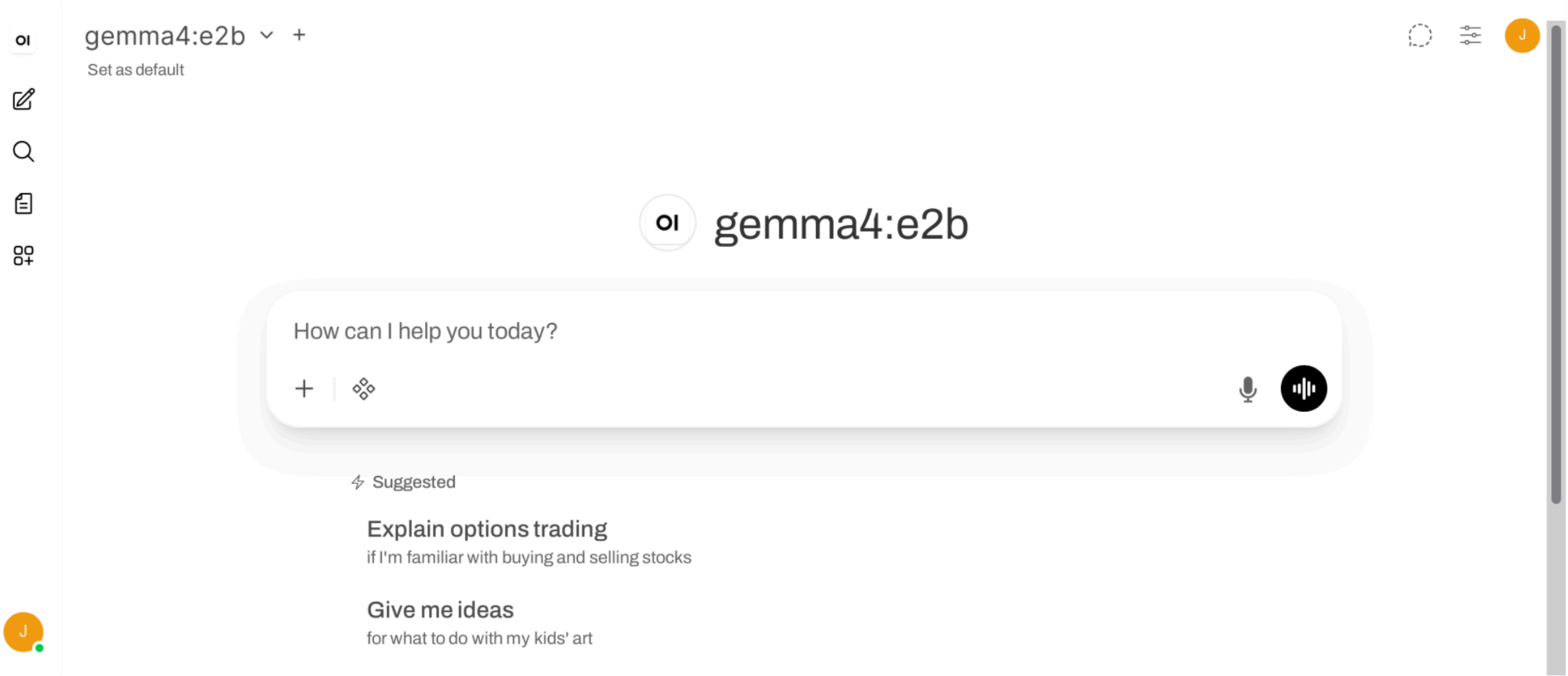
```
ollama run gemma4:e2b
ollama run gemma4:e4b
ollama run gemma4:12b
```

https://ollama.com/library/gemma4

Open WebUI: how to run

```
uv venv --python 3.11
uv pip install open-webui
uv run open-webui serve
```

Now, open http://localhost:8080


gemma4:e2b
Set as default
J
gemma4:e2b
How can I help you today?
Suggested
Explain options trading
if I'm familiar with buying and selling stocks
Give me ideas
for what to do with my kids' art
J

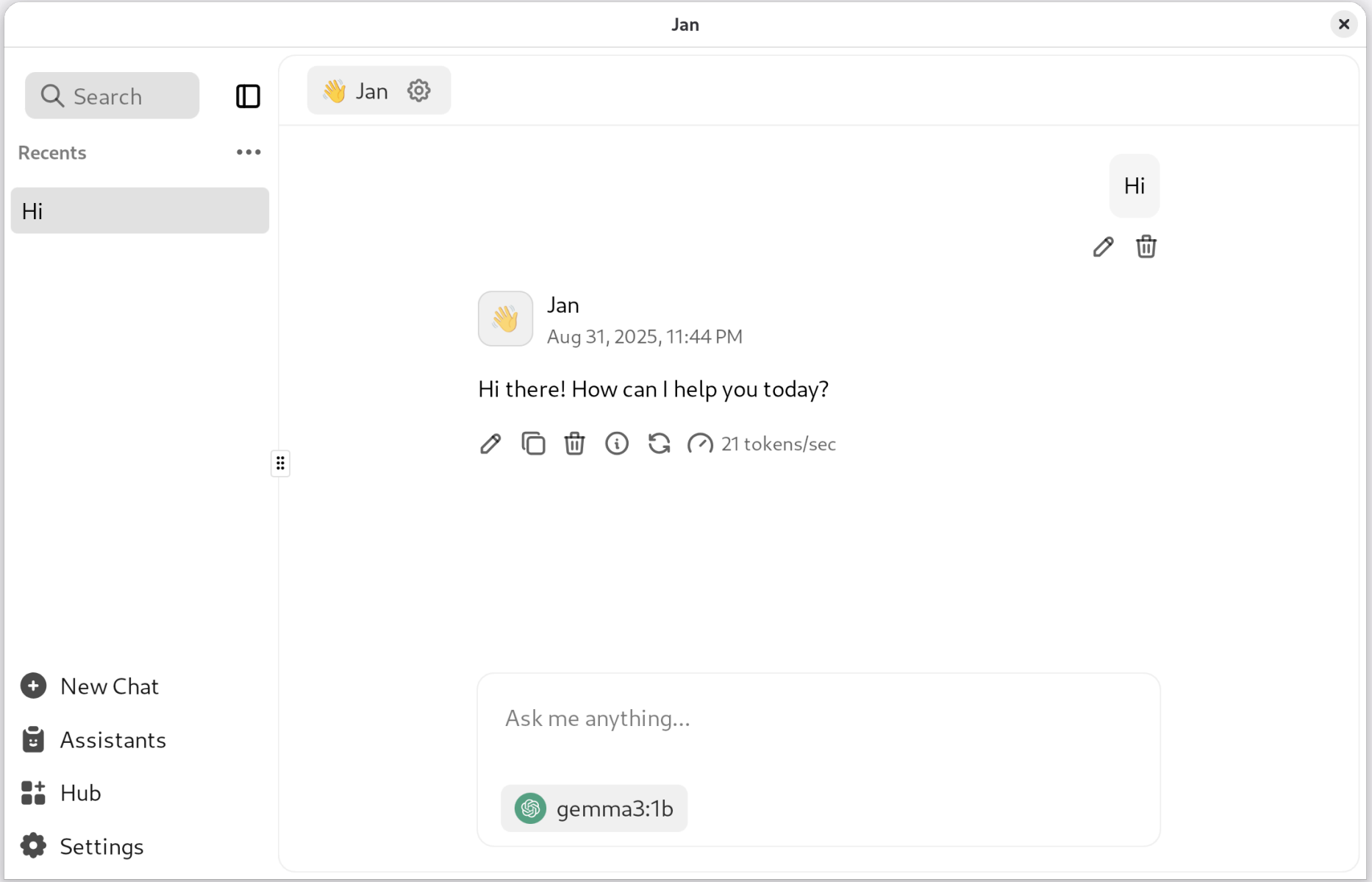
Jan: how to run

- Fetch the Debian package (or the one for your OS)

```
sudo pkg -i Jan_0.6.9_amd64.deb
Jan
```

- Settings > Model Providers > OpenAI
- API Key "ollama"
- Base URL: http://localhost:11434/v1
- Models: add a new one ("+"), "gemma3:1b"
- Select the model in the Chat

https://github.com/janhq/jan


Jan
Search
Recents
Hi
New Chat
Assistants
Hub
Settings
Jan
Hi
Jan
Aug 31, 2025, 11:44 PM
Hi there! How can I help you today?
21 tokens/sec
Ask me anything...
gemma3:1b

Speech to text

Whisper

Whisper, MIT License

```
uv venv
uv pip install openai-whisper
uv run whisper speech.wav --language Spanish
```

```
#!/usr/bin/python3
import whisper

model = whisper.load_model('tiny')
transcription = model.transcribe('recording.wav')
print(transcription['text'])
```

VocaLinux

Speech to text in any application, works with English, Spanish...

```
curl -fsSL raw.githubusercontent.com/jatinkrmalik/vocalinux/main/install.sh \
-o /tmp/vl.sh && bash /tmp/vl.sh
```

Backends:

- whisper.cpp (supports Vulkan)
- Whisper
- VOSK (for low-RAM devices)

https://github.com/jatinkrmalik/vocalinux/

Adding subtitles to a video

The OfiLibre way

- Speech to text: Whisper
- Diarization (who speaks): Pyannote
- Embed subtitles: FFmpeg

https://ofilibre.gitlab.io/guias/automatizacion-contenido-audiovisual/

Text to speech

Kokoro-82M

```
uv venv
uv pip install kokoro
uv run kokoro -m ef_dora -i texto.txt -o texto.wav
```

https://github.com/hexgrad/kokoro

Koko Clone

Some voice-cloning capabilities on top of Kokoro

```
git clone https://github.com/Ashish-Patnaik/kokoclone
cd kokoclone
uv sync
uv run cli.py --text "Texto a leer" --lang es \
  --ref orginal.wav --out output.wav
uv cli.py --mode convert --source original.wav \
  --ref target.wav --out revoiced.wav
```

https://github.com/Ashish-Patnaik/kokoclone

Some other options

Text generation

- llama.cpp: designed for speed and efficiency
- LocalAI: Drop-in OpenAI API replacement
- Oobabooga TextGen: specialized in text generation
- KoboldCpp: text generation based on llama.cpp

Image generation

- ComfyUI, repo: front-end and UI for several self-hostable text-to-image and text-to-video models
- Wan2GP: front-ed and UI for several self-hostable text-to-video models

Agents and other applications

- OpenCode: assistant, agentic
  - CLI, web and desktop app
- Hive: orchestrator for coding agents
- AnythingLLM: assistant, agentic
- FastSDCPU: image generation optimized for CPU
- Lemonade: LLM, image and speech generation tool
  - Works well with Vulkan, apparently recognizing my Intel Iris GPU

Agents and other applications (2)

- OpenClaw: Agentic system
- Hermes agent: Agentic system
- Autoresearch: Self-improvign agent
- Good night, have fun: Autoresearch-style agent, for coding

If speed is not a (big) problem, you can reach really far