# Raspberry Pi 4 — Stanford Alpaca

*Offline AI Chatbot at the Edge*





Local LLM · No Cloud · No Internet · Runs Entirely on a Pi 4

## Project Overview

**One-line takeaway:** A fully offline, air-gapped AI chatbot powered by Stanford's Alpaca 7B model — fine-tuned from Meta's LLaMA — running locally on a Raspberry Pi 4 (8GB). No internet, no cloud, no subscription required.

**Purpose:** Demonstrate that a capable instruction-following AI language model can run entirely on a $35–$75 single-board computer using CPU-only inference. Stanford's Alpaca 7B is fine-tuned from Meta's LLaMA 7B on 52,000 instruction-following examples and served via the llama.cpp engine with 4-bit GGML quantization for Pi-compatible memory usage.

## About Stanford Alpaca


### Stanford CRFM

Alpaca was developed by the Center for Research on Foundation Models (CRFM) at Stanford University and released in March 2023 as an open, replicable instruction-following model for academic research.


### Alpaca 7B Model

Fine-tuned from Meta's LLaMA 7B on 52,000 instruction-following demonstrations generated using OpenAI's text-davinci-003 via the self-instruct method. Training cost under $600 total.


### llama.cpp Inference

Pure C++ inference engine by Georgi Gerganov. Enables LLaMA-family models to run on CPU-only hardware. 4-bit GGML quantization reduces the 7B model from ~13 GB to ~4 GB for Pi compatibility.

## Key Capabilities

### Fully Air-Gapped

No Wi-Fi, no Ethernet — the Pi runs completely offline. The model and runtime are entirely self-contained on the SD card.

### Instruction-Following Q&A

Handles writing tasks, explanations, summaries, creative prompts, and general Q&A similar in style to early GPT-3.5.

### Terminal Chat Interface

Interactive terminal session via llama.cpp's built-in chat mode. Type a prompt, receive a response — no browser or GUI needed.

### CPU-Only Inference

No GPU required. The quantized 7B model runs on the Pi 4's quad-core Broadcom BCM2711 ARM Cortex-A72 CPU.

### 4-Bit GGML Quantization

Model weights compressed from FP16 (~13 GB) to 4-bit (~4 GB), fitting within the Pi 4's 8 GB RAM with room to spare.

### Academic Demonstration

Intended for education and research. Showcases the frontier of edge AI — what's possible on minimal, affordable hardware.

## How It Works

### ① Model Preparation (done on a PC)

The LLaMA 7B base model is downloaded and converted to GGML FP16 format on a Linux PC (requires 16 GB RAM — too much for the Pi alone). It is then quantized to 4-bit using llama.cpp's quantize tool, reducing the file to ~4 GB, and transferred to the Pi via USB drive.

### ② llama.cpp Setup on the Pi 4

llama.cpp is cloned and compiled from source on the Raspberry Pi 4 (Raspberry Pi OS). The quantized model file is placed in the models/ directory. Python dependencies (NumPy, SentencePiece) are installed for support utilities.

### ③ Running the Chatbot

The chat session is launched from the terminal using llama.cpp's interactive mode. The user types prompts directly into the terminal and the model generates responses in real time — entirely on the Pi's CPU with zero network calls.

## Technologies Used

**HARDWARE**

- Raspberry Pi 4 (8GB RAM)
- MicroSD card (32GB+)
- HDMI display
- USB keyboard & mouse

**SOFTWARE**

- Raspberry Pi OS
- llama.cpp (C++ engine)
- Python 3, NumPy
- SentencePiece tokenizer

**AI MODEL**

- Stanford Alpaca 7B
- Meta LLaMA 7B base
- 52K instruction examples
- 4-bit GGML quantization

## Known Limitations & Notes

🐢 **Response Speed:** CPU-only inference is slow — expect 1–5 tokens/sec on Pi 4. Responses may take 30–90 seconds for longer outputs.

🌀 **Hallucination:** Like all LLMs, Alpaca can confidently generate incorrect facts. Treat outputs as demonstrations, not authoritative answers.

🎓 **Academic Use Only:** Alpaca is released for academic research only. Commercial use is prohibited per Stanford, Meta LLaMA, and OpenAI license terms.

💾 **Memory Constraint:** 8GB Pi 4 is the recommended minimum. The 4-bit quantized model uses ~4 GB of RAM, leaving limited headroom for the OS.

🦙 **Fun Fact:** Stanford's Alpaca 7B performs comparably to OpenAI's text-davinci-003 on instruction-following benchmarks — winning 90 vs. 89 pairwise comparisons — yet runs on a $75 Raspberry Pi 4 with no internet connection. The entire training run cost under $600.