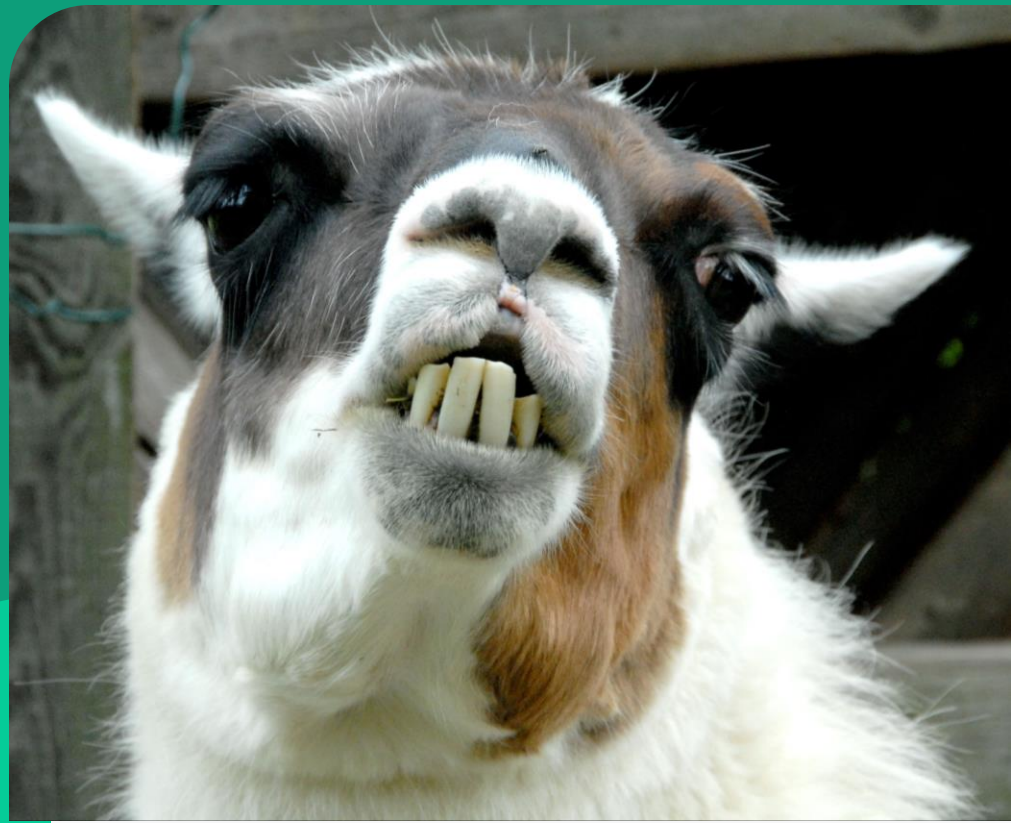
Tech Hive: LLMs for Developer

Introduction to Ollama

Presented by Paco Sarin Suriyakoon
Tech Lead
June 2024



What is LLM/SLM?

Background Check

PALO IT

Why

Why it matters do run LLM/SLM Locally

PALO IT

LLM for Developers

Why running LLMs locally matters

- **Encourage experiment and Learning** - Experimenting and Learning without paying money is awesome!
- **Help us improve writing prompt better because it less smart!**

Ollama

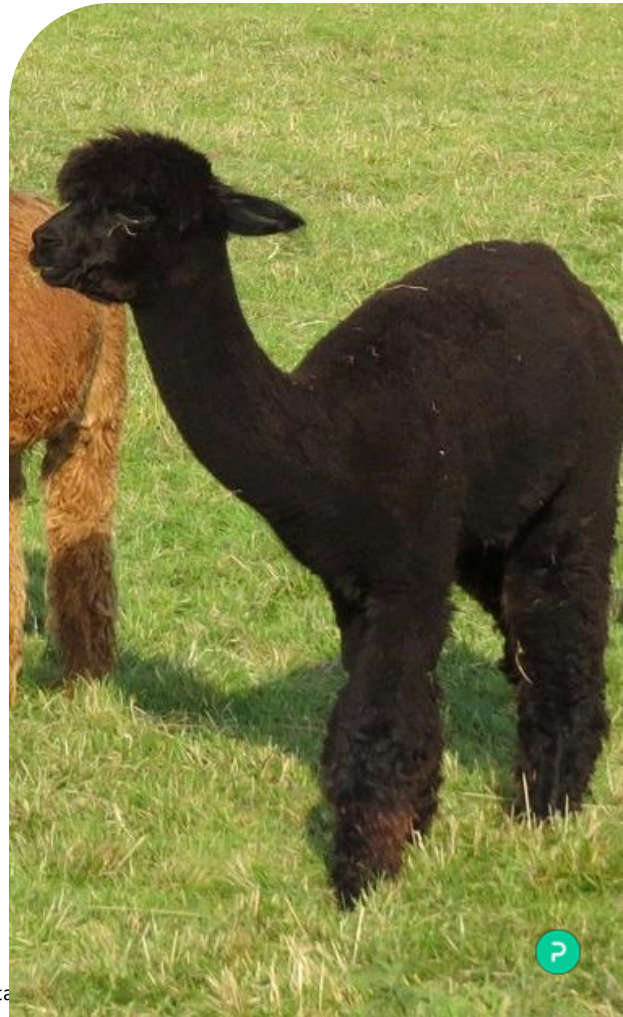
Applause for **Ollama**

PALO IT

LLM for Developers

What is Ollama

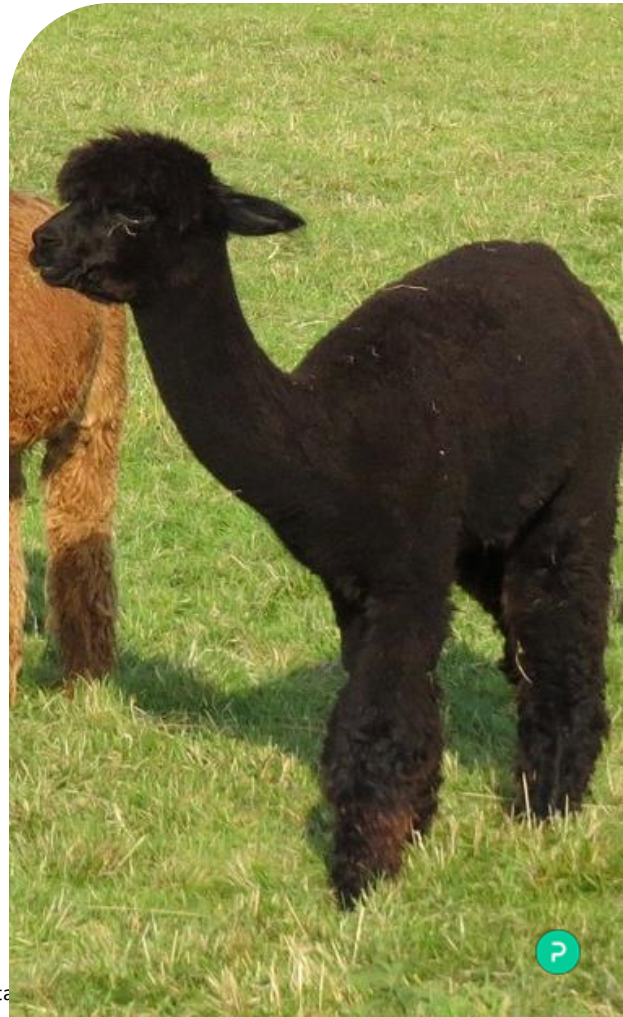
- Based on llama.cpp
- REST API
- Docker, Kubernetes
- NodeJS, Python
- Langchain, LLamaIndex
- A lot of integration
- Support Multimodal
- OpenAI compatible
- Many more!
- Open for run your own Fine-Tune model
- Open to experiment with prompt engineering
- Amazing community



LLM for Developers

What is NOT Ollama

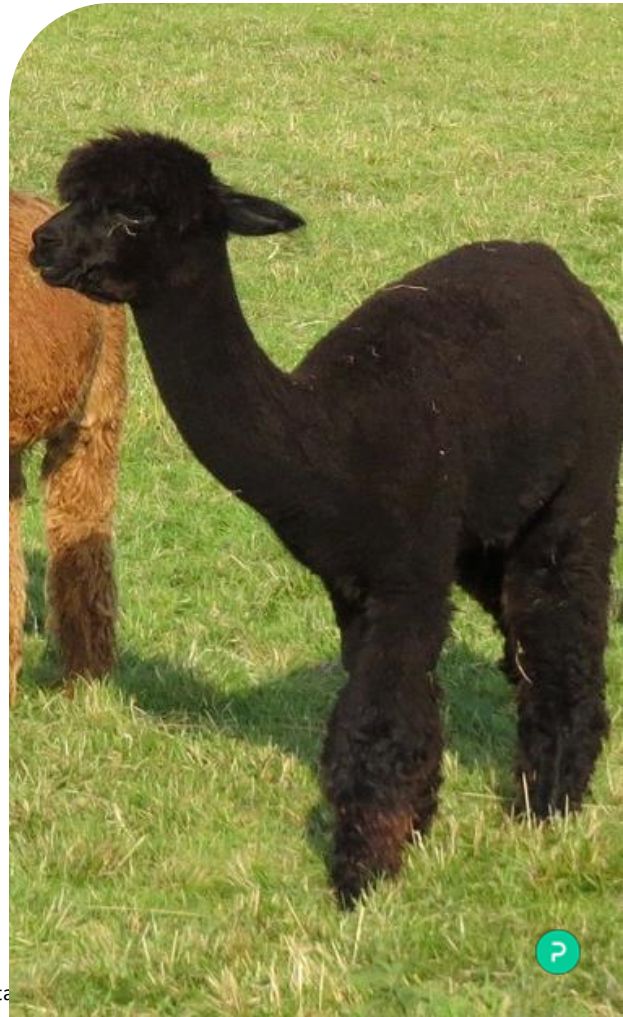
- Fastest tool
- Depends on your CPU or GPU
- NOT A MODEL



LLM for Developers

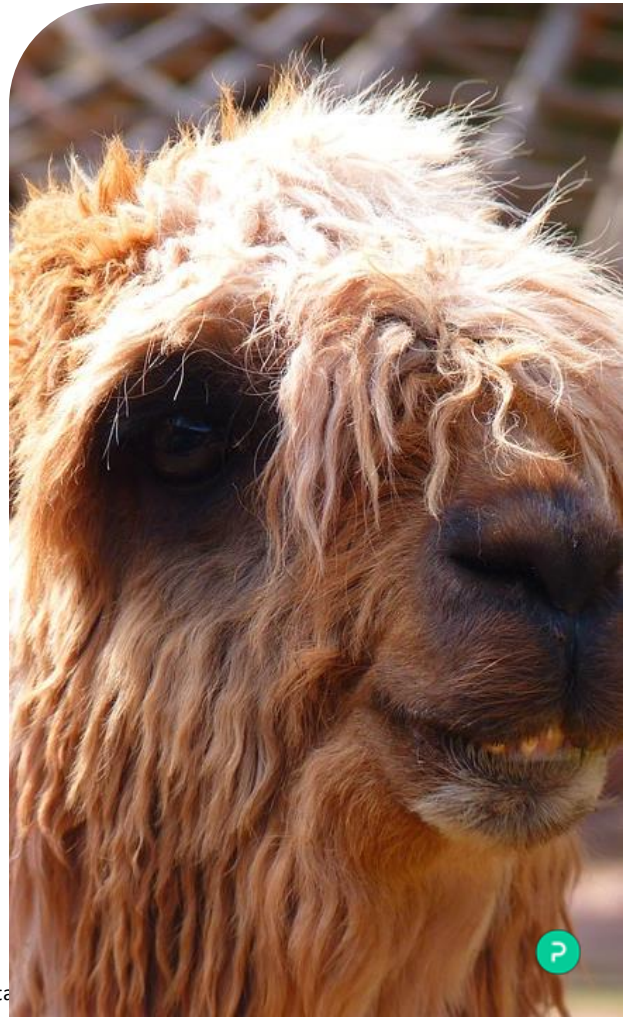
Get our hands dirty!

- Get to know it at Ollama.com
- Let's go to Github
- https://github.com/ollama
- Download and Run
- ollama run phi3
- ollama run llama3



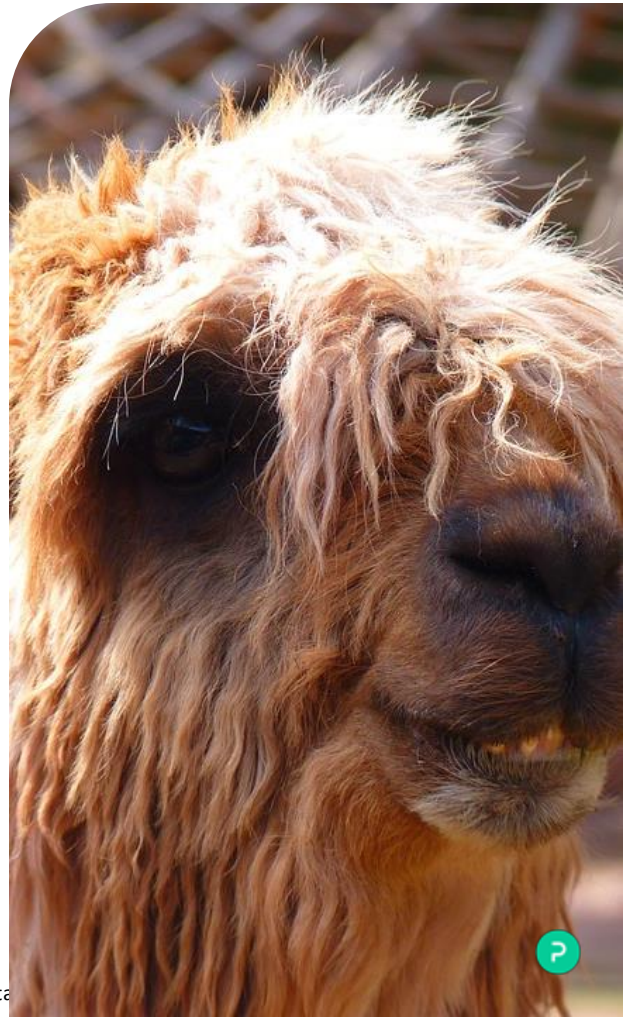
Usage

- Ollama pull
- Ollama run
- Ollama list
- http://localhost:11434/
- Modelfile
- https://github.com/ollama/ollama/blob/main/docs/modelfile.md
- https://github.com/ollama/ollama/blob/main/docs/api.md



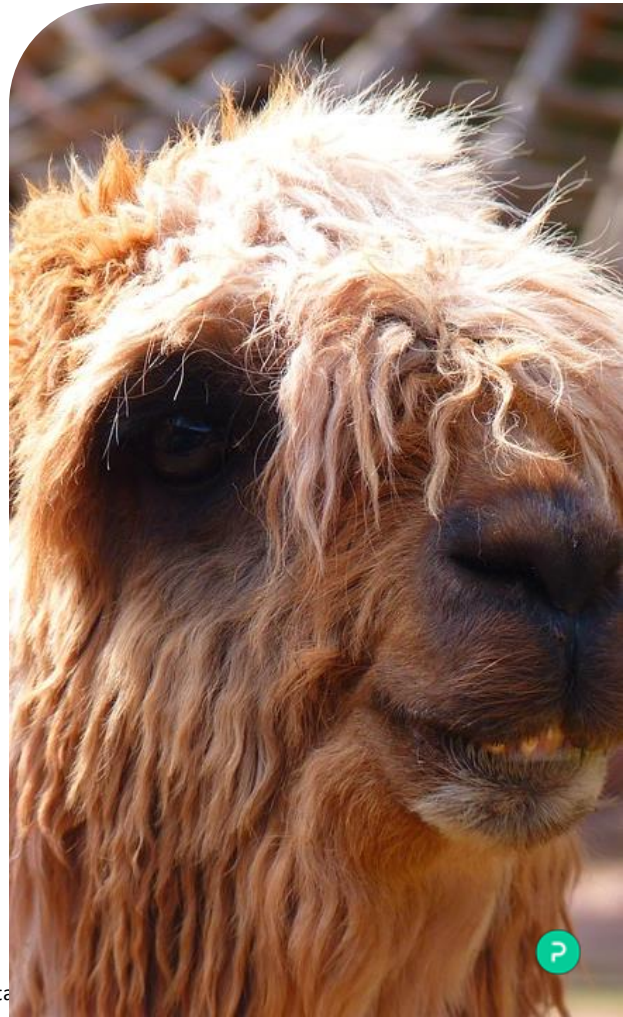
Alternative For Prompt Engineering

- https://huggingface.co/chat/



What next?

- Kubernetes with Ollama?
- Create model file?
- Use it with langchain
- Fine-Tune from Unsloth and use it in Ollama?



Some more ideas and question

- **How to solve less smart version on local?"You are a Human Wish to Applescript translator, return a Applescript file content only"**
- How to use model from Huggingface? GGUF
- How to customize template?LLama2, Claude template, Qwen
- How to train a model with Ollama?*
- Alternative tool is Huggingface? Spot the different? Huggingface Transformers js/python