Building AI, Using AI

Workshop @ KCDC 2025

~ Shashank Daté ~

GitHub: /shanko

stdate@gmail.com

https://www.linkedin.com/in/shashankdate/

Attention is all you need *

* https://arxiv.org/abs/1706.03762 (Jun 12, 2017)

Agenda:

There will be four sub-sessions, about 1 hour each:

1. Intro and Set up: Intro to various AI/ML terms and prepare your device with necessary tools needed for the workshop.

2. AI Coding exercises: Using AI coding tools to develop scripts consuming various local and cloud based APIs

3. Building a RAG application: A fast paced tutorial on how to build a Retrieval-Augmented Generation model, using AI generated code

4. ~~Getting started with Fine-tuning: learn how to configure the model and the training loop on a resource rich cloud platform~~

4. Intro to Agentic AI: Build a workflow based on a couple of AI Agents

Set Up

Prerequisites: Laptop with at least **16 GB of RAM** and **40 GB of free disk space**

1. Signup for a Github Account: https://github.com and clone the repo:
2. If you are on Windows:
   a. install WSL2: https://learn.microsoft.com/en-us/windows/wsl/install
   b. then follow the remaining instructions on the installed Linux distribution
   c. OR install Ollama for Windows (new)
3. Install Python: https://python.org version 3.10 or above
4. Create and activate virtual env for python3 that you installed, in the cloned repo

```
$ python3 -m venv venv

$ source venv/bin/activate
```

5. Install the following Python packages in that environment using pip:

```
    $ pip install langchain langchain-chroma langchain-community langchain-core
langchain-ollama langgraph ollama pandas torch transformers
```

6. Install Ollama: https://ollama.com

Optional Set Up

Prerequisites: Please sign the Liability Waiver document

- Install Windsurf: https://windsurf.com
- Install ngrok: https://ngrok.com/downloads/mac-os
- Signup for Lovable: https://lovable.dev
- Signup for Google Colab: https://colab.research.google.com
- Signup for a Huggingface: https://huggingface.co

Start with ChatGPT

Prerequisites: Open ChatGPT: https://chatgpt.com/

Prompts:

- What is ChatGPT?
- What is LLM?
- What makes an LLM qualify as “Large”?
- Are there medium, small, tiny LMs?
- What is the LLM on which you are based?
- What is your knowledge cutoff date?
- What is the size of your context window in terms of English words?
- What is the "temperature" setting for GPT-4o on which you are based?

Basic Concepts

What Is:

- Language Model
- Generative AI
- GPT
- Prompt Engineering
- Vibe Coding
- RAG
- Agentic AI

Intro to AI coding tools: Lovable.dev

Competition: famous.ai, bolt.new, ...

Prompt: Create a Progressive Web App that fetches a list of LLMs from a REST API GET request. The display should have a 3-pane layout with the left pane for Navigation, the central wider pane for Chatbot interface, the right pane for Setting the parameters of the LLM. The API end-point URL will be entered via text box in the Settings pane . When the green “Get” button which is below the text-box is clicked, parse the JSON response that is generated by the GET call to extract the names of LLMs. Show the names of the LLMs in a dropdown. When a user selects a name, display a chatbot interface in the central pane, that interacts with the selected LLM. The UI/UX should be minimalistic, clean and have a button to toggle between light and dark mode.

Intro to AI coding tools: Windsurf.com

Competition: Github Copilot, Cursor, …

Prompt: Review the entire code in the /src folder and identify functions which can be unit tested. For these functions, write unit-tests in a separate test files, one test file per code file. All these test files must be in /test folder. Modify the README.md file to include instructions on how to run the tests. Do not change the original code in any other way.

Intro to Ollama

What is it: Ollama is a lightweight, extensible, open-source framework for building and running AI Language Models on the local machine.

Provides:

1. Command-Line Interface
2. REST like API
3. Quantization: reduce the computational load by optimizing model performance
4. Customization: via “model file” - a text file that defines how a model should be built, customized, and configured.
5. Lightweight Finetuning: using a method called LoRA (Low-Rank Adaptation)
6. Distribution: via Huggingface

Ollama Continued

- Accessing via ngrok:

  ```
  ngrok http 11434 --host-header="localhost:11434"
  ```

  - GET /api/version
  - GET /api/ps
  - GET /api/tags
- Importing a GGUF based model
- Quantizing a Model

Intro to Retrieval Augmented Generation (RAG)

What is it: a technique that enhances the capabilities of AI Language Models by combining their generative abilities with information retrieval from external sources such as databases.

How it works:

1. **Retrieval:** When a user asks a question, a RAG system first searches a knowledge base (which could be anything from a database of documents to a company's internal knowledge base) to find relevant information.
2. **Augmentation:** This retrieved information is then combined with the original user query to create an augmented prompt.
3. **Generation:** The LLM then uses this augmented prompt to generate a response, drawing on both its pre-trained knowledge and the retrieved information.

Intro to Agentic AI

What is it: An system of “AI agents” that acts autonomously to accomplish specific goals with minimal or no human supervision, using planning, reasoning, and contextual adaptation.

Key characteristics:

- **Autonomy:** Acts independently to achieve predetermined goals without constant human oversight.
- **Proactive reasoning and planning:** Analyzes context, reasons about options, and develops plans rather than following preset rules.
- **Adaptability:** Alters its approach in response to new information or shifting conditions.
- **Multi-agent orchestration:** Often combines multiple AI models or agents that collaborate to complete complex tasks.

Types of AI Agents

- **Simple Reflex Agent:** needs predefined condition-action rules to which it reacts to, e.g. Thermostat
- **Model Based Reflex Agent:** needs state, which it remembers, in addition to the condition-action rules, e.g. Robotic Vacuum Cleaner
- **Goal Based Agent:** no need to have predefined rules, but has goals and it predicts which actions are needed to achieve them, e.g. Self Driving Cars
- **Utility Based Agent:** needs goals prediction plus evaluates how desirable the outcome is based on the utility (scoring/ranking) e.g. Drones
- **Learning Agent:** has all of the above, but improves by learning about the environment from experience e.g. Chess Bots

Caveat: Human-In-The-Loop is still needed!

Resources

Attention is all you need: https://arxiv.org/abs/1706.03762

Windows WSL2: https://learn.microsoft.com/en-us/windows/wsl/install

Python: https://python.org

Ollama: https://ollama.com

Windsurf: https://windsurf.com

Google Colab: https://colab.research.google.com

Lovable: https://lovable.dev

Huggingface: https://huggingface.co

ChatGPT: https://chatgpt.com

LangGraph: https://www.langchain.com/langgraph

Lambda Labs: https://lambda.ai

Runpod: https://www.runpod.io

IBM Risk Atlas: https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas

Bad LLMs: https://blog.sshh.io/p/how-to-backdoor-large-language-models

Titanium Sponsors

Progress®

aws

Platinum Sponsors

Particular Software

outsystems

ORACLE Database

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

NIPR NATIONAL INSURANCE PRODUCER REGISTRY

Nx

LocalStack

DON'T PANIC LABS™

Couchbase

descope

Chainguard

ascend LEARNING

dynatrace

Gold Sponsors

BUILDERTREND

TEKsystems Own change

LT LEAN TECHNIQUES

TURNBERRY SOLUTIONS

sonatype

Red Hat

Veterans United Home Loans

QUEST ANALYTICS

Coder

H&R BLOCK

TEXT CONTROL

SHAMROCK™ TRADING CORPORATION

touchnet® A Global Payments Company

Duende.

Chocolatey

snyk

Aviron Labs

propio

reboot

VML

Other Awesome Organizations

Teleport

IowaComputer GURUS

neo4j

Gradle

Leading EDJE

FLOX

360PICKC.COM

KC DIGITALDRIVE

KVC

Thanks!

- Event Organizers: esp. Gabby Spurling
- Aalap Patil: https://www.linkedin.com/in/aalap-patil/
- Rothbright: https://rothbright.com
- GrowthNet: https://growthnet.biz
- 360Sequrity: https://360sequrity.com
- Apisdor: https://www.apisdor.com

Full Workshop: Discount

https://sysprotech.com/AIWorkshop.html

Discount Code: Text 913-322-0779

Feedback, please!