ANTHROP\C

Claude Builder Club LLM Interpretability Masterclass

@Penn

October 29, 2025

Check-in for attendance!

Introductions

Today’s Presenter

Albert Opher
Penn Claude Builder Club President
M&T | CIS + FNCE ‘25

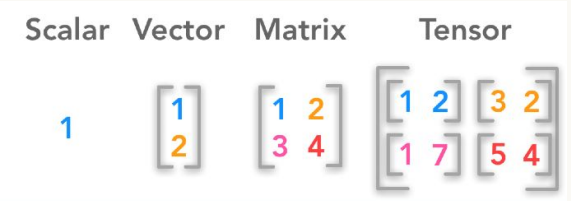
Today’s Agenda: AI Interpretability Research Demo

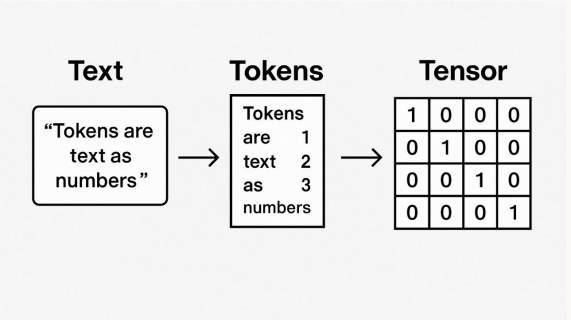
- What is Interpretability Research?
  - Basic Concepts Too
- Our Tools:
  - Ollama
  - Hugging Face
  - PyTorch
  - LangChain
  - StreamLit
- Demo:
  - Let’s Build An Interpretability Dashboard

Basics

What Is AI Interpretability Research?

- Trying to understand why and how models make decisions
- Backtracking from outputs to see what features were most important in a model's decision
- Open-Source Model weights are imperative to determining how a model determines it's weights
- Data is stored in Tensors: Multidimensional matrix stores used to house all our numerical data
  - Why? We have a lot of variables at play, each with numerous features
- Tokens: The basic unit of Data of an LLM
- Tokens are the fundamental units of text that models process, while tensors are the multi-dimensional numerical arrays used to represent these tokens and all other data within the model's computations





Ollama

Our Local LLM Inference Server

What is Ollama?

- Run powerful LLMs locally on your machine
- No API keys, no rate limits, no privacy concerns
- Simple HTTP API for integration

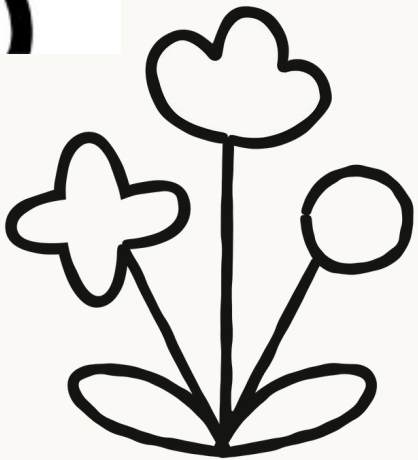
Why Ollama?

✓ Privacy - Your data never leaves your machine

✓ Speed - Local inference is fast

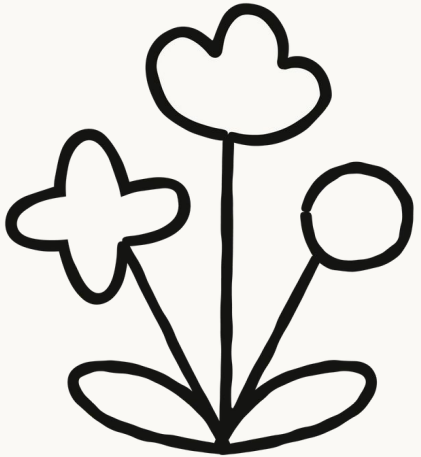
✓ Cost - Zero API costs

✓ Offline - Works without internet



Langchain

Agent Orchestration Framework



What is LangChain?

- Framework for building LLM applications
- Agent patterns and tool integration
- Production-ready abstractions

Agent Components:

1. **Tools** - Functions the agent can call
2. **Prompt** - Guides agent behavior
3. **LLM** - Powers decision making
4. **Callbacks** - Track agent reasoning

Why LangChain?

✓ Production agent patterns

✓ Tool integration

✓ Observability built-in

Hugging Face

Access to 100,000+ Pre-trained Models

What is HuggingFace?

- The GitHub of machine learning models
- Open-source library for loading and using models
- Access to internal model representations

What We'll Extract:

- Hidden states at each layer
- Attention weights across heads
- Token probabilities (logits)
- Model architecture details

Why HuggingFace?

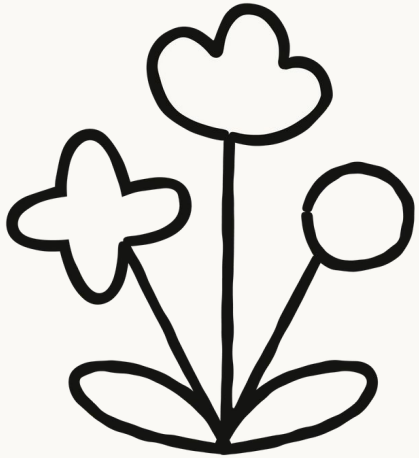
✓ Research-grade model access

✓ Standardized interfaces

✓ Active community

✓ State-of-the-art models

Hugging Face



Pytorch

Deep Learning Framework for Model Internals

What is PyTorch?

- Meta's deep learning library
- Powers most modern AI research
- Gives us direct access to model computations

Key Operations:

1. **Logits** → Raw model outputs
2. **Softmax** → Convert to probabilities
3. **Argmax** → Find most likely token
4. **Topk** → Get top-k predictions

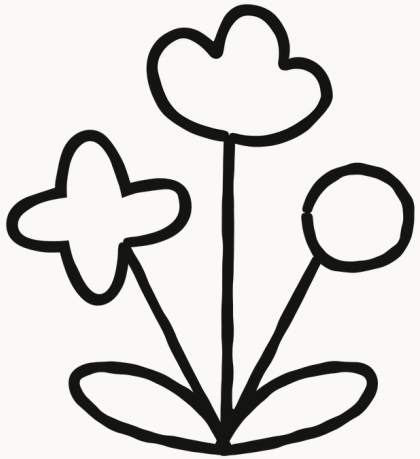
Why PyTorch?

✓ Full control over model internals

✓ Research standard

✓ Flexible and pythonic

✓ GPU acceleration



PyTorch

Demo

Interpretability Dashboard: Let's use our backend tools and a React Front End To Visualize how our LLM is making decisions.

https://github.com/Albinator3000/LLM-Interpretability-Dashboard-w-Ollama-HF-PyTorch-LangChain-CBC-PENN-

https://github.com/Albinator3000/LLM-Interpretability-Dashboard-w-Ollama-HF-PyTorch-LangChain-CBC-PENN-

YOU WIN FREE CLAUDE PRO

It's TRUE, for coming to this meeting, you have earned a free Claude Pro account.

To redeem:

- Fill out the form attached to the QR code
- Make sure you have a Claude.ai log-in that is connected to your **STUDENT EMAIL**

Note to students:
We will send out an email with your FREE Claude Pro login, this will take up to two business days!

Come to a meeting to get the QR code for Free Claude Pro

Q&A

ANTHROP\C

Thank you for coming!

ANTHROP\C