Author Name

Self-Hosted AI Inference

A Systems Engineer's Guide

December 13, 2025

Springer Nature

Contents

1 Introduction to Self-Hosted Inference . . . 1
- 1.1 The Rise of Self-Hosted AI . . . 3
- 1.2 Training vs. Inference: Understanding the Difference . . . 3
  - 1.2.1 What Happens During Training . . . 3
  - 1.2.2 What Happens During Inference . . . 3
  - 1.2.3 Resource Comparison . . . 3
- 1.3 Why Self-Host? . . . 3
  - 1.3.1 Cost Control . . . 3
  - 1.3.2 Privacy and Data Sovereignty . . . 3
  - 1.3.3 Latency and Performance . . . 3
  - 1.3.4 Customization and Control . . . 3
  - 1.3.5 When NOT to Self-Host . . . 3
- 1.4 What You'll Build in This Book . . . 3
  - 1.4.1 The Progressive Journey . . . 3
  - 1.4.2 Control Plane Evolution . . . 3
- 1.5 Hands-On: Your First Local Inference . . . 3
  - 1.5.1 Installing Ollama . . . 3
  - 1.5.2 Running Your First Model . . . 3
  - 1.5.3 Making API Requests . . . 3
- 1.6 Understanding the Response . . . 4
- 1.7 Summary . . . 4
- Problems . . . 4
- References . . . 5

Chapter 1
Introduction to Self-Hosted Inference

Abstract This chapter introduces the concept of self-hosted AI inference, explaining why organizations and individuals choose to run their own models rather than relying on cloud APIs. We cover the fundamental differences between training and inference, the economic and strategic considerations for self-hosting, and provide a roadmap for what you'll build throughout this book. By the end of this chapter, you'll run your first local model and make your first inference request.

1.1 The Rise of Self-Hosted AI

1.2 Training vs. Inference: Understanding the Difference

1.2.1 What Happens During Training

1.2.2 What Happens During Inference

1.2.3 Resource Comparison

1.3 Why Self-Host?

1.3.1 Cost Control

1.3.2 Privacy and Data Sovereignty

1.3.3 Latency and Performance

1.3.4 Customization and Control

1.3.5 When NOT to Self-Host

1.4 What You'll Build in This Book

1.4.1 The Progressive Journey

1.4.2 Control Plane Evolution

1.5 Hands-On: Your First Local Inference

1.5.1 Installing Ollama

1.5.2 Running Your First Model

1.5.3 Making API Requests

First Inference Request

```
# Pull a 7B model
ollama pull llama3.2:7b

# Make an inference request
curl http://localhost:11434/api/generate \
  -d '{
    "model": "llama3.2:7b",
    "prompt": "Explain what AI inference is in one paragraph.",
    "stream": false
  }'
```

1.6 Understanding the Response

1.7 Summary

> Key Takeaways

- Inference is fundamentally different from training and accessible with consumer hardware
- Self-hosting offers cost control, privacy, and customization benefits
- Open models like Llama, Mistral, and Qwen make self-hosting viable
- You've run your first local inference with Ollama

Problems

1.1 Cost Comparison

Calculate the monthly cost of running 100,000 inference requests through OpenAI's GPT-4 API versus self-hosting a 7B model on an RTX 4060. Assume average request length of 500 tokens input and 200 tokens output.

1.2 Latency Measurement

Using Ollama, measure the time-to-first-token (TTFT) and tokens-per-second for three different prompts: (a) a simple question, (b) a code generation request, and (c) a creative writing prompt. What patterns do you observe?

1.3 Streaming vs Non-Streaming

Implement a simple Python script that makes both streaming and non-streaming requests to Ollama. Measure the perceived latency (time until user sees first output) for each approach.

References