

How to Serve LLMs

Building a Local LLM Server

Ken Hubbell

How to Serve LLMs

Building a Local LLM Server

By

Ken Hubbell

© Copyright 2023 by Printed Page Publishers - All rights reserved.

It is not legal to reproduce, duplicate, or transmit any part of this document in either electronic means or printed format. Recording of this publication is strictly prohibited.

This book is dedicated to

my son Isaac

who firmly believes

we are building relationships

with these amazing AI engines.

Table of Contents

[Introduction 1](#_Toc121308533)

[PART I: Mysterious Music 2](#_Toc121308534)

[Chapter One: An Ancient Song 3](#_Toc121308535)

[Chapter Two: Song of the Humpback Whale 5](#_Toc121308536)

[PART II: Blue Whale Ballads 7](#_Toc121308537)

[Chapter Three: Finding a Mate 8](#_Toc121308538)

[Chapter Four: Bridging the Gap 10](#_Toc121308539)

[Chapter Five: Saving the Singers 11](#_Toc121308540)

[Epilogue/Conclusion 12](#_Toc121308541)

[Bibliography 13](#_Toc121308542)

[Acknowledgments 14](#_Toc121308543)

[About the Author 15](#_Toc121308544)

# Introduction

This is a complete guide for beginners to dive into this new world of AI Chatbots and the Large Language Models that power them. In this guide I will walk you through creating a Language Learning Model (LLM) server from scratch. The LLM server allows you to run AI models locally. Once you have an LLM server, you can create the front-end (application interface) user experience to fit the requirements of your specific project.

We'll use Python and several supporting libraries to build this server.

Before starting, you should have:

* A computer running Ubuntu Linux (or similar distribution) – at this time Raspberry Pi is not recommended due processor speed and other form limitations
* Basic knowledge of Python programming
* Basic understanding of command line operations
* An NVIDIA GPU (recommended for better performance)

For those of you who are new to Python and the supporting libraries and AI/ML technologies we will be using to build the server, I will either include explanation as needed or provide links to other resources to help you get the most of this guide.

This guide is the first of three:

1. How to Serve LLMs
2. A Chat Webserver to Connect Them All
3. Make Your LLM Useful with a Custom Chat UX

My hope is this guide helps fast track you to build amazing products with generative AI.

# PART I: Introduction to LLM Servers

- Introduction and Overview (30 minutes)

- Core concepts

- System architecture

- Key components

- Learning objectives

## Chapter One: Environment Setup

Environment setup (2 hours)

### CUDA installation

(Capitalize First Letter of Each Word in Subheadings.)

jakljdklasda

### Python environment configuration

Insert content here…

An example for Subheading 2 is “Love Songs.” This chapter would talk about songs whales use during mating season.

### Package installation

Insert content here…

An example for Subheading 2 is “Love Songs.” This chapter would talk about songs whales use during mating season.

### Environment validation

Insert content here…

An example for Subheading 2 is “Love Songs.” This chapter would talk about songs whales use during mating season.

### Chapter Summary/Key Takeaways

### - Importance of proper environment setup

### - Each component's role in the system

### - Common installation issues

## Chapter Two: Core Server Components

Flask Server Implementation (2 hours)

### Basic server setup

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “An Oceanic Chorus.” You can talk about how different species of whale songs differ from each other.

### Route configuration

Insert content here…

For Subheading 2, use a subtitle such as “Phrases and Themes.” This subsection explains the differences between the elements of whale songs, the role those elements play in mating, their familial groups, etc.

### Error handling

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “An Oceanic Chorus.” You can talk about how different species of whale songs differ from each other.

### Testing procedures Model Management (2 hours)

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “An Oceanic Chorus.” You can talk about how different species of whale songs differ from each other.

### Model Management (2 hours)

### Model loading

Insert content here…

Sdfsdfs

### Configuration options

Insert content here…

Sdfsdfs

### Memory management

Insert content here…

Sdfsdfs

### Optimization techniques

Insert content here…

Sdfsdfs

### Practical Exercises:

1. Create basic Flask server

2. Implement model loading

3. Add error handling

4. Test server functionality

In the next chapter, you will learn…

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter, link the next chapter’s information with what has already been learned.

# PART II: Blue Whale Ballads

Use successive parts to cover the more detailed or complex areas of the book’s topic. Since Part I defined the topic/problem, consider using this section to provide solutions. In this case, a suitable example title is “Blue Whale Ballads” or “The Singing Blue Whale.” Don’t forget that the chapters in this part of the book must align with the Part title you have chosen.

## Chapter Three: Advanced Features

Document Processing (1.5 hours)

Chat Session Management (1.5 hours)

For the purposes of this example, this chapter’s title is “Finding a Mate.” This means that this chapter will be dealing with strategies for finding a sexual partner. Start off by providing a brief overview of the information contained in the chapter and then transition smoothly into your supporting points. Try to keep the language simple and understandable to generate a rapport with the reader and keep them engaged.

### Call of the Wild

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are always related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “Call of the Wild.” Offer the reader insight into the necessity of whales to attract a mate.

### Complexity and Frequency

Insert content here…

For Subheading 2, use a subtitle such as “Complexity and Frequency.” This example chapter provides visual examples of how whale songs travel in the ocean and are received by potential mates.

### Chapter Summary/Key Takeaways

Insert content here…

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

Developing an attractive song is an important part of securing a potential mate

Point 2 from your text…

Point 3 from your text…

etc.

In the next chapter, you will learn…

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter, link the next chapter’s information with what has already been learned.

## Chapter Four: Performance Optimization

Begin a new chapter here…

For the purposes of this example, this chapter’s title is “Bridging the Gap.” It covers the ideas and theories around human to animal communication. Start off by providing a brief overview of the information contained in the chapter and then transition smoothly into your supporting points. Try to keep the language simple and understandable to generate rapport with the reader and keep them engaged.

### Communicating with Nature Through Music

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “Communicating with Nature Through Music.” Offer the reader information and insight into new technologies allowing us to be closer than ever to communicating directly with animals.

### Our Most Intelligent Species

Insert content here…

For Subheading 2, use a subtitle such as “Our Most Intelligent Species.” It provides fascinating facts on whale social life, emotions, etc.

### Chapter Summary/Key Takeaways

Insert content here…

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

The more we learn about mammals, the more we learn about ourselves

Point 2 from your text…

Point 3 from your text…

etc.

In the next chapter, you will learn…

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter with a paragraph, link the next chapter’s information with what has already been learned.

## Chapter Five: Common Issues and Solutions

Begin a new chapter here…

Assuming this is your last chapter in the book, create a title that is somewhat forward-looking, for example, “Saving the Singers.” Here, present an outlook of how the information in this book, as well as new scientific insights, will affect the future. Also discuss what the world will look like if we don’t do our part to keep the ocean clean.

### Ocean Conservation

Insert content here…

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is “Ocean Conservation” where you then discuss the importance of being eco-conscious and “green”, etc.

### How to Volunteer

Insert content here…

For Subheading 2, use a subtitle such as “How to Volunteer.” Here, wrap everything up by discussing how all the strategies provided will ensure a better tomorrow for all.

### Chapter Summary/Key Takeaways

Insert content here…

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

Taking care of our oceans is in the best interest of ourselves and the planet…

Point 2 from your text…

Point 3 from your text…

etc.

# Epilogue/Conclusion

Insert content here…

Keep it short and sweet. Mention the key highlights of the book and the action steps to solve the problems. Remind the reader of how taking the said action steps will benefit them.

# Bibliography

Below is an example of a list of works cited using APA style. Arrange your list of references alphabetically.

When citing books, use the format as shown in the following examples, applying Bibliography style to the format:

Author’s last name, first initial or initials. (Publication date). *Book title*. Additional information. City, State of publication: Publishing company.

King, S. (2000). *On writing: A memoir of the craft.* New York, NY: Pocket Books.

When citing online resources, use the format as shown in the following examples:

For internet documents

Author’s last name, first initial or initials. (Date of publication). Title of article. *Title of work*. Retrieved from full URL

Amir, N. (2018, October 17). 4 tips for staying on track with your writing. *Write Nonfiction now!* Retrieved from http://writenonfictionnow.com/tips-staying-track-writing/

For online periodicals

Author’s last name, first initial or initials. (Date of publication). Title of article. *Title of Periodical*, volume and page numbers. Retrieved from full URL

Brewer, R. L. (2018, October 4). How to write better titles: 7 effective title tips for books, articles, and conference sessions. *Writer’s Digest.* Retrieved from http://www.writersdigest.com/whats-new/how-to-write-better-titles

When citing magazines, use the format as shown in the following examples:

For magazines and periodicals

Author’s last name, first initial or initials. (Publication date). Article title. *Title of periodical*, *volume number (issue number if available*), inclusive pages.

McPhee, J. (2013, April 29). Draft No. 4. *New Yorker*, *89*, 20-25.

For more details and guidelines, consult the APA Publication Manual.

# Acknowledgments

Insert content here…

Thank the key people who inspired you and helped you throughout the process of writing and publishing your work. This is somewhat similar to the dedication page, except here you can elaborate and include more people.

# About the Author

Insert content here…

Write this page in the third person. (For example, use “the author” or your name, not “I.”) The information within establishes your credibility with readers. Avoid being too wordy. Simply provide your background and expertise on the topic of your book, as well as other information that will build trust. For example:

Professional and personal achievements related to the topic at hand

List of other published works and a link to your website

Educational background

Mention other notable experts in the field that you have worked with

Your area of residence, family status, hobbies, etc.