

The
Pragmatic
Programmers

A Common-Sense Guide to AI Engineering

Build Production-Ready LLM Applications



Jay Wengrow
edited by Katharine Dvorak

This extract shows the online version of this title, and may contain features (such as hyperlinks and colors) that are not available in the print version.

For more information, or to purchase a paperback or ebook copy, please visit <https://www.pragprog.com>.

Copyright © The Pragmatic Programmers, LLC.

Contents

Change History	?
Preface	?

Part I — Foundations

1.	HeLLMo, World!	?
	Signing Up for an LLM-as-a-Service	?
	Creating Our First App	?
	Tweaking the Model and Temperature	?
	Checking API Usage	?
	Wrapping Up	?
2.	Understanding How LLMs Work	?
	What Is a Large Language Model (LLM)?	?
	Realizing LLMs Are Nondeterministic Creatures	?
	Gauging the Temperature	?
	Understanding the Challenges of Nondeterminism	?
	Wrapping Up	?
3.	Diving Deeper into LLMs	?
	Diving into Tokens	?
	Diving into Embeddings	?
	Diving into Fine-Tuning Behavior	?
	Wrapping Up	?

4.	Selecting an LLM	?
	Getting Your Hands on an LLM	?
	Comparing Different LLMs	?
	Deciding on an LLM	?
	Wrapping Up	?

Part II — Chatbots

5.	Building a Chatbot	?
	Getting User Input	?
	Augmenting the Prompt	?
	Adding Multi-Turn Dialogue	?
	Managing State with Memory Systems	?
	Adding a Developer Message	?
	Treating the Prompt as an Array	?
	Wrapping Up	?
6.	Augmenting a Prompt with Knowledge	?
	Building a Chatbot	?
	Augmenting with Knowledge	?
	Avoiding Context Window Limitations	?
	Preparing the Data	?
	Implementing the Knowledge Chatbot	?
	Running into PACKing Problems	?
	Wrapping Up	?
7.	Efficiently Adding Knowledge with RAG	?
	Augmenting with Documentation Chunks	?
	Getting into Search Engines, Retrieval, and RAG	?
	Searching with Meaning: Keywords Versus Semantics	?
	Using Embedding-Similarity Search	?
	Building a Starter Search Engine	?
	Implementing a RAG Chatbot	?
	Choosing the Right K	?
	Wrapping Up	?

8.	Measuring Quality with Evals	?
	Introducing Evals	?
	Setting Up Our App	?
	Conducting Error Analysis	?
	Open Coding	?
	Axial Coding	?
	Creating an Eval Test Framework	?
	Running Human Evals	?
	Wrapping Up	?
9.	Prompt Engineering	?
	Eliminating Ambiguity	?
	Utilizing the System Prompt	?
	Rewriting History	?
	Using Delimiters and Bullet Points	?
	Reordering Prompt Components	?
	Wrapping Up	?
10.	Reducing Hallucinations	?
	Understanding Why Our App Hallucinates	?
	Instructing the LLM to be Faithful	?
	Pleading and Threatening	?
	Upgrading the Model	?
	Citing Sources and Few-Shot Prompting	?
	Iterate, Iterate, Iterate	?
	Reviewing Our Current Chatbot Implementation	?
	Final Prompt Engineering Thoughts	?
	Checking On Our Evals	?
	Wrapping Up	?
11.	Evaluating and Optimizing RAG	?
	Discovering a RAG Failure	?
	Evaluating RAG	?
	Expanding the Query	?
	Metadata-Based Filtering	?
	Evaluating RAG Subcomponents	?

Dreaming Up an Agentic RAG Wish List	?
Wrapping Up	?

Part III — Agents

12. Equipping an LLM with Tools	?
Understanding an LLM's Limitations	?
Triggering a Function	?
Defining “Agents”	?
Feeding Tool Results Back to the LLM	?
Building a Website Reader Tool	?
Deciding to Use a Tool	?
Using the Tools API	?
Wrapping Up	?
13. Running the Agent Loop	?
Solving a Complex Problem	?
Constructing an Agent Loop	?
Building a News Podcast Agent	?
Exploring Agent Failure Modes and Evals	?
Giving the Agent a Plan	?
Asking the Agent to Create a Plan	?
Wrapping Up	?
14. Architecting Agentic Workflows	?
Designing an LLM Assembly Line	?
Implementing an LLM Assembly Line	?
Weighing Agentic Workflows Against Classic Agent Loops	?
Workflow Routing	?
Performing Tasks in Parallel	?
Wrapping Up	?
15. Enhancing Retrieval with Agentic RAG	?
Architecting an Agentic RAG Plan	?
Implementing a RAG Agent	?
Avoiding Unnecessary RAG	?

Generating Structured Outputs	?
Researching as an Agent	?
Conducting Multi-Hop Research	?
Wrapping Up	?
16. Building System-Integrated Agents	?

Part IV — Production

17. Observing AI Systems	?
18. Adding Guardrails	?
19. Exception Handling	?
20. Using an LLM-as-Judge	?