

GENESIS ITU

# LLM Engineering Fundamentals

Working with large language models responsibly in products

---

College-style student textbook & workbook · Clean · Practical · Professional

Zarco AI Systems LLC · Genesis ITU

## How to use this book

Read each chapter in order. Learn formal definitions the way employers and exams use them. Complete pause-and-review prompts. Use only authorized lab systems. This course teaches lawful, professional practice.

## Table of Contents

- Chapter 1 — LLM stack overview
- Chapter 2 — Tokens, cost, latency
- Chapter 3 — Prompt patterns
- Chapter 4 — RAG concept
- Chapter 5 — Safety & abuse resistance
- Chapter 6 — Evaluation
- Chapter 7 — Multi-model routing
- Chapter 8 — Build responsibly

## Chapter 1 — LLM stack overview

Applications call models through APIs, add prompts/system instructions, retrieve context, and filter outputs.

### System prompt

*Instructions that steer model behavior for an application.*

## Chapter 2 — Tokens, cost, latency

Billing and speed often track tokens. Design concise prompts and cache where appropriate.

### Token

*Unit of text processing used by model APIs for limits and pricing.*

**Pause and review.** *Why does a long chat history cost more?*

## Chapter 3 — Prompt patterns

Role, constraints, examples (few-shot), and structured outputs improve reliability.

- Separate system vs user messages
- Validate outputs
- Fail closed on safety

## Chapter 4 — RAG concept

Retrieval-augmented generation pulls trusted documents into context so answers can cite real sources.

### **RAG**

*Retrieve relevant docs, then generate with that context.*

## Chapter 5 — Safety & abuse resistance

Filter inputs/outputs, rate-limit, and never expose secret keys in browsers — server-side only.

- Store keys in env secrets
- Log for abuse
- Apply family-safe policies in product prompts

## Chapter 6 — Evaluation

Test with golden questions. Measure refusal quality, factuality on your domain, and tone.

***Pause and review.*** Name two tests you would run before shipping an education chatbot.

## Chapter 7 — Multi-model routing

Route teaching tasks to careful models; general chat to fast models; always apply product safety prompts.

## Chapter 8 — Build responsibly

Ship least privilege, clear user notices, and human escalation paths. Excellence includes restraint.

## Closing

Review every formal definition before practice exams. Continue with labs, instructor video when available, and job-path practice on NodePath or REACH. Genesis ITU wishes you steady progress.

---

Genesis ITU · Educational use · Authorized systems only