

AROUND

RETRIEVAL AUGMENTED GENERATION



QUESTIONSONLY

“—

SOMETIMES IT'S THE
VERY PEOPLE WHO
NO ONE IMAGINES
ANYTHING OF WHO
DO THE THINGS NO
ONE CAN IMAGINE.

**THE Imitation Game
(2014)**

INTRODUCTION

Retrieval Augmented Generation, or RAG, stands as a pivotal technique shaping the landscape of applied generative AI. A novel concept introduced by Lewis et. al., in their seminal paper Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, RAG has swiftly emerged as a cornerstone, enhancing reliability and trustworthiness in the outputs from Large Language Models (LLMs).

A few months ago, Ali Ghodsi, co-founder and CEO of Databricks, revealed that their customers are actively embracing RAG, with 60% of their use cases involving LLMs being built upon this architecture. For developers, managers and business leaders, RAG has become as essential a concept to understand as Generative AI and Large Language Models.

One of the quickest ways to learn a concept is to study the common questions and their answers around the subject. This book contains a list of 80 questions divided in 8 groups for a quick, in-depth understanding of the technique.

THESE QUESTIONS ARE BASED ON

A SIMPLE GUIDE TO Retrieval Augmented Generation

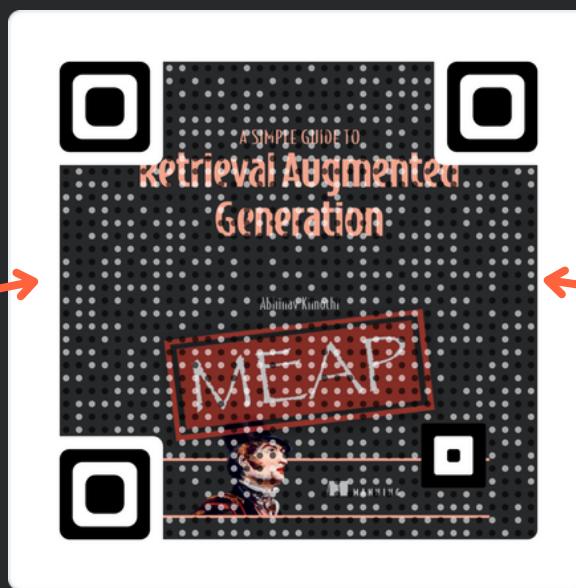
Abhinav Kimothi

MEAP



MANNING

IF YOU ARE LOOKING FOR AN IN-DEPTH INTRODUCTION TO RAG, CONSIDER PURCHASING THE BOOK



SCAN CODE OR CLICK HERE
TO GET AN EARLY ACCESS COPY

TABLE OF CONTENTS

RAG BASICS-----Q1 TO Q10

Questions on limitations of LLMs and the introduction to RAG

RAG SYSTEM DESIGN-----Q11 TO Q20

Questions on the high level design of RAG systems

INDEXING PIPELINE-----Q21 TO Q30

Questions on creating a knowledge brain for RAG systems

GENERATION PIPELINE-----Q31 TO Q40

Questions on real-time user interaction in RAG systems

EVALUATION-----Q41 TO Q50

Questions on the evaluation of RAG systems

BEYOND NAIVE RAG-----Q51 TO Q60

Questions on advanced RAG techniques

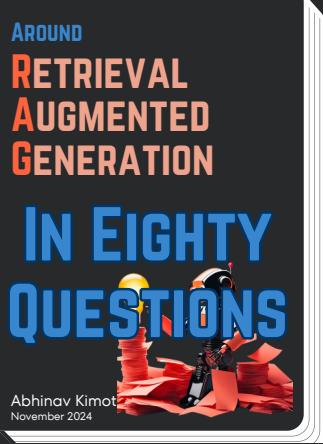
EVOLVING RAGOPS STACK-Q61 TO Q70

Questions on the technology stack for RAG

RAG VARIANTS-----Q71 TO Q80

Questions on RAG patterns like multimodal, graph, agents, etc.

FOR ANSWERS, THE ENTIRE E-BOOK IS AVAILABLE ON GUMROAD



SCAN CODE OR CLICK HERE
TO DOWNLOAD

“ —

THE CURSE OF LARGE LANGUAGE MODELS & THE NEED FOR RAG

PART I

RAG BASICS

Q1 TO Q10 OF 80

QUESTIONS

Q1 — Why are hallucinations a critical challenge in LLMs?

Q2 — What limitations of LLMs does RAG aim to overcome?

Q3 — What is Retrieval Augmented Generation (RAG)?

Q4 — Define parametric and non-parametric memory in the context of RAG.

Q5— What are the three core steps of RAG?

Q6 — How does RAG address hallucination issues in LLMs?

Q7 — How does RAG improve response reliability in LLMs?

Q8 — How does RAG support real-time updates in LLM responses?

Q9— What are some common applications of RAG systems?

Q10 — What is the difference between a naïve LLM and a RAG-powered LLM?

“ —

THE COMPONENTS OF RAG SYSTEMS IN PRODUCTION

PART II

RAG SYSTEM DESIGN

Q11 TO Q20 OF 80

QUESTIONS

Q11—What are the two main pipelines of a core RAG system?

Q12—Explain the purpose of the indexing pipeline.

Q13—What are the main components of the indexing pipeline?

Q14—What are the essential characteristics of information stored in the knowledge base?

Q15—What are the core responsibilities of the generation pipeline?

Q16—What are the main components of the generation pipeline?

Q17—What are the other components in the RAG system apart from the core pipelines?

Q18—How does the indexing pipeline change for third party knowledge sources?

Q19—What is the triad of RAG evaluation?

Q20—What are guardrails in RAG systems?

“—

BUILDING THE KNOWLEDGE BRAIN FOR RAG SYSTEMS

PART III

THE INDEXING PIPELINE

Q21 TO Q30 OF 80

QUESTIONS

Q21—What is data loading in the indexing pipeline?

Q22—Why is metadata important during data loading?

Q23—What are tokens and what is tokenization?

Q24—Why is chunking necessary in RAG systems?

Q25—Describe the different chunking methods.

Q26—What are embeddings, and why are they important?

Q27—What factors influence the choice of embeddings?

Q28—Name some popular embeddings models.

Q29—How do vector databases enhance the indexing pipeline?

Q30—What factors should be considered when choosing a vector database?

“—

GENERATING CONTEXTUAL LLM RESPONSES

PART IV

THE GENERATION PIPELINE

Q31 TO Q40 OF 80

QUESTIONS

Q31 - What are the three key steps in the generation pipeline of RAG systems?

Q32 - What are some popular retrieval methods used in RAG?

Q33 - What is the advantage of contextual embeddings over static embeddings?

Q34 - What are some key prompt engineering techniques used in RAG systems?

Q35 - What is Few Shot Prompting, and why is it useful?

Q36 - What are the criteria for choosing between fine-tuned and foundation LLMs

Q37 - What are the trade-offs between open-source and proprietary LLMs?

Q38 - How does the size of an LLM affect its suitability for RAG?

Q39 - How can chain-of-thought prompting improve complex task handling in RAG?

Q40 - What are the key challenges in optimizing retrieval for a RAG system?

“—
ACCURACY,
RELEVANCE,
FAITHFULNESS & MORE...

PART V
**EVALUATING
RAG**

Q41 TO Q50 OF 80

QUESTIONS

Q41 - What are the three key quality scores in RAG evaluation?

Q42 - Why is it important to evaluate the retrieval and generation components separately?

Q43 - What are some retrieval metrics commonly used in RAG evaluation?

Q44 - What is Ground Truth data and why is it important?

Q45 - What is Context Relevance?

Q46 - What is Faithfulness?

Q47 - What is Answer Relevance?

Q48 - What role do frameworks like RAGAs and ARES play in RAG evaluation?

Q49 - How do benchmarks like RGB, Multihop RAG, and CRAG contribute to RAG evaluation?

Q50 - What are some challenges and best practices in RAG evaluation methods?

“—

BEYOND NAÏVE RAG: BUILDING SMARTER AND RELIABLE AI SYSTEMS

PART VI

ADVANCED & MODULAR RAG

Q51 TO Q60 OF 80

QUESTIONS

Q51 - What are the limitations of naïve RAG systems?

Q52 - How does Advanced RAG pattern differ from Naïve RAG?

Q53 - What are the key strategies for query optimisation?

Q54 - Describe hybrid retrieval and its use cases in RAG systems.

Q55 - What is adaptive retrieval, and how does it differ from standard retrieval?

Q56 - Explain context compression and its role in RAG.

Q57 - How do reranking methods improve post-retrieval processes?

Q58 - Why is modularity important for RAG system design?

Q59 - What is the role of the memory module in Modular RAG?

Q60 - What are the trade-offs involved in implementing advanced RAG techniques?

“—

TECHNOLOGIES
THAT MAKE RAG
POSSIBLE

PART VII

EVOLVING
RAGOPS STACK

Q61 TO Q70 OF 80

QUESTIONS

Q61. - What are the different layers of the RAGOps stack?

Q62 - How does the data layer contribute to RAG systems?

Q63 - What are the components of the Model Layer?

Q64 - What are some examples of model deployment options for RAG systems?

Q65 - Why is application orchestration important in RAG?

Q66 - What are some essential layers in the RAGOps Stack?

Q67 - What are common security challenges in RAG systems, and how are they addressed?

Q68 - What are some enhancement layers in the RAGOps stack?

Q69 - What factors affect the choice of tools in the RAGOps stack?

Q70 - What are some production best practices for deploying RAG systems?

“—

MULTIMODAL, GRAPH, AGENTIC & OTHER RAG PATTERNS

PART VIII

RAG VARIANTS

Q71 TO Q80 OF 80

QUESTIONS

Q71 - What are RAG variants, and why do we need them?

Q72 - What is the primary purpose of Multimodal RAG?

Q73 - How do the RAG pipelines adapt for Multimodal RAG?

Q74 - What is a knowledge graph and how does it help in RAG?

Q75 - What enhancements are made to the RAG pipelines in Knowledge Graph RAG?

Q76 - What is Agentic RAG, and how does it differ from Standard RAG?

Q77 - How does Agentic RAG improve the RAG pipeline?

Q78 - What are the challenges associated with multimodal, graph and agentic RAG?

Q79 - What is Corrective RAG, and how does it improve accuracy?

Q80 - How does Speculative RAG reduce latency?

THESE QUESTIONS ARE BASED ON

A SIMPLE GUIDE TO
**Retrieval Augmented
Generation**

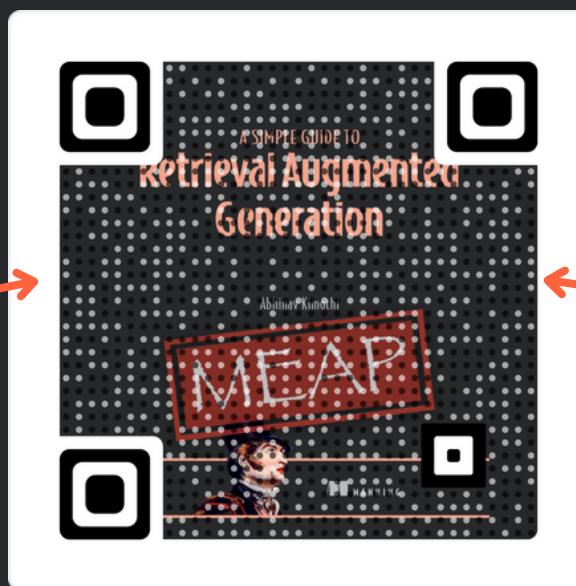
Abhinav Kimothi

MEAP



MANNING

**IF YOU ARE LOOKING FOR AN IN-
DEPTH INTRODUCTION TO RAG,
CONSIDER PURCHASING THE BOOK**



**SCAN CODE OR CLICK HERE
TO GET AN EARLY ACCESS COPY**

INTERESTED IN CODING RAG PIPELINES?

THE SOURCE CODE OF

A SIMPLE GUIDE TO

RETRIEVAL AUGMENTED GENERATION

ARE NOW AVAILABLE FOR FREE PUBLIC ACCESS

A SIMPLE GUIDE TO Retrieval Augmented Generation

Abhinav Kimothi

MEAP



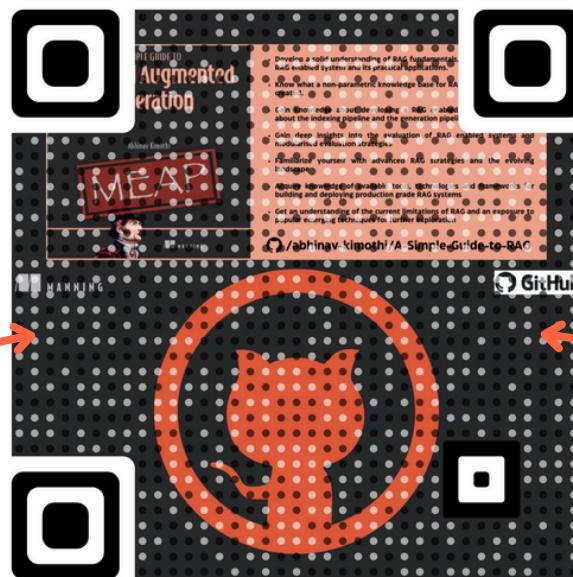
MANNING

- Develop a solid understanding of RAG fundamentals, the components of a RAG enabled system and its practical applications.
- Know what a non-parametric knowledge base for RAG means and how is it created.
- Gain knowledge about developing a RAG enabled system with details about the indexing pipeline and the generation pipeline.
- Gain deep insights into the evaluation of RAG enabled systems and modularised evaluation strategies
- Familiarize yourself with advanced RAG strategies and the evolving landscape
- Acquire knowledge of available tools, technologies and frameworks for building and deploying production grade RAG systems
- Get an understanding of the current limitations of RAG and an exposure to popular emerging techniques for further exploration

[/abhinav-kimothi/A-Simple-Guide-to-RAG](https://github.com/abhinav-kimothi/A-Simple-Guide-to-RAG)

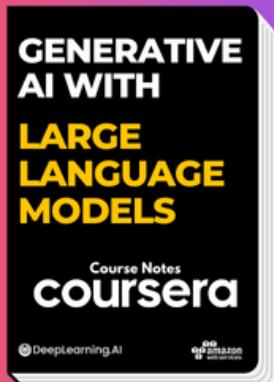
MANNING

GitHub



SCAN CODE OR CLICK HERE
TO VIEW SOURCE CODE

GENERATIVE AI EBOOKS



Generative AI
with Large
Language
Models (Course
Notes)



Generative AI
Terminology: An
evolving
taxonomy start
with Gen AI



A Taxonomy_
of
Retrieval
Augmented
Generation



**SCAN CODE OR CLICK HERE
TO CHECK THESE OUT**

A DOWNLOADABLE PDF VERSION OF THE ENTIRE EBOOK IS AVAILABLE ON GUMROAD

AROUND
RETRIEVAL
AUGMENTED
GENERATION

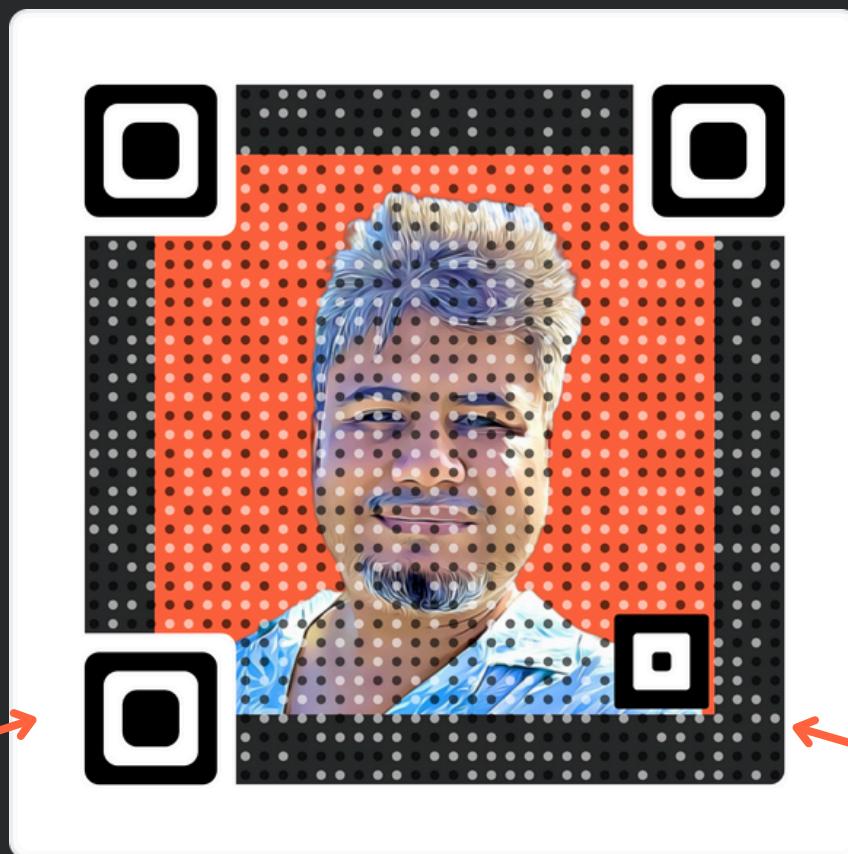
IN EIGHTY
QUESTIONS

Abhinav Kimot
November 2024



SCAN CODE OR CLICK HERE
TO DOWNLOAD

Hi! My name is Abhinav, and I talk about ML, RAG, LLMs & AI Agents. If our interests align, I'd love to stay connected



**SCAN CODE OR CLICK HERE
TO CONNECT**



linktr.ee/abhinavkimothi



[/in/Abhinav-Kimothi](https://www.linkedin.com/in/Abhinav-Kimothi)



[@akaiworks](https://www.instagram.com/@akaiworks)

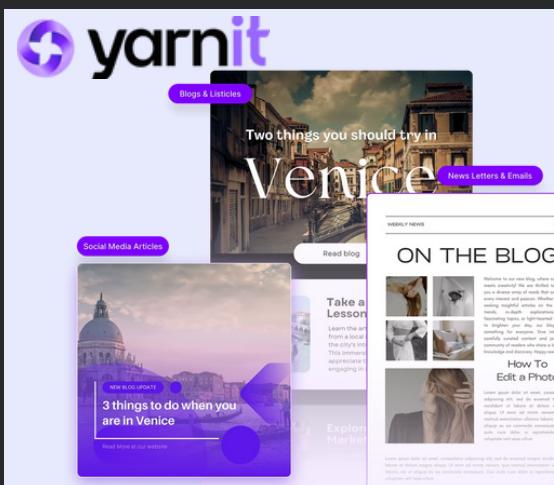


[@abhinav_kimothi](https://twitter.com/abhinav_kimothi)



[@abhinavkimothi](https://twitter.com/abhinavkimothi)

Check these out before you go



YARNIT - CONTENT MARKETING SaaS



RAG PUBLICATION

A Taxonomy of Retrieval Augmented Generation
Powering the rise of Contextual AI —Over 200 terms including Components, Pipelines, Ops Stack, Technologies & more
towardsai.net

Beyond Naive RAG: Advanced Techniques for Building Smarter and Reliable AI Systems
A deep dive into advanced indexing, pre-retrieval, retrieval, and post-retrieval techniques to enhance RAG performance
towardsdatascience.com

Stop Guessing and Measure Your RAG System to Drive Real Improvements
Key metrics and techniques to elevate your retrieval-augmented generation performance
towardsdatascience.com

MORE BLOGS ON RAG

Let's have a 1:1 call

● Career Chat & Learning Paths : AI, ML, DS
● AI & ML : Discovery Consultation
● Virtual Coffee
● Ask me anything

topmate.io/abhinav_kimothi

TALK TO ME

IF YOU LIKE MY CONTENT
PLEASE CONSIDER BUYING ME A COFFEE :)



Buy me a Coffee

