AI PRODUCT MANAGEMENT 2 · DETAILED COURSE

Retrieval-Augmented Generation (RAG)

Grounding model answers in your own data: architecture, retrieval quality, evaluation and production operation.

Chapter 14 of 26 · 29 topics · Edition 2026 · skillnavora.in

CHAPTER 14

Retrieval-Augmented Generation (RAG)

Grounding model answers in your own data: architecture, retrieval quality, evaluation and production operation.

RAG Fundamentals

Retrieve relevant documents at request time and give them to the model to ground its answer.

Retrieval-augmented generation finds relevant documents at request time and gives them to the model so its answer is grounded in them. It lets a product answer from current company data without retraining the model.

Why RAG Is Required

Models lack your private or recent data and can hallucinate; RAG supplies sources and enables citations.

Models do not know your private or recent data and can hallucinate. RAG supplies sources, which enables citations and makes answers checkable.

RAG vs Fine-Tuning

RAG adds knowledge at query time and is easy to update; fine-tuning changes behaviour or style. They are often combined.

RAG adds knowledge at query time and is easy to update. Fine-tuning changes behaviour, style or format. Many products start with RAG and fine-tune only if behaviour still falls short.

RAG Architecture

Ingest and index, retrieve, rerank, build context, generate, cite, and monitor.

The pipeline ingests and indexes documents, retrieves candidates for a query, reranks them, builds the context, generates the answer with citations and is monitored for quality.

Data Ingestion

Collecting documents from sources and keeping the index in sync with changes.

Ingestion collects documents from sources such as drives, wikis and databases, then keeps the index in sync when content changes. Stale content is a common cause of wrong answers.

Document Parsing

Extracting clean text, tables and structure from PDFs, web pages and office files.

Parsing extracts clean text, tables and headings from PDFs, web pages and office files. Poor parsing silently degrades every later step, so check parsed output by hand.

Data Cleaning

Removing noise, duplicates and boilerplate and fixing encoding so retrieval is accurate.

Cleaning removes boilerplate, duplicates and noise, and fixes encoding problems. Clean data makes retrieval more accurate and smaller.

Document Chunking

Splitting documents into passages sized for retrieval and context limits.

Chunking splits documents into passages sized for retrieval and the context window. Chunks that are too small lose meaning; chunks that are too large dilute relevance.

Chunking Strategies

Fixed size, recursive, by heading or semantic boundaries, with overlap to preserve context.

Options include fixed-size chunks, recursive splitting, splitting by heading, and semantic boundaries, often with some overlap to keep context. Evaluate each option on your own questions.

Embedding Models

Models that convert text to vectors; choose by quality, language coverage, dimensions and cost.

Embedding models turn text into vectors. Compare them on retrieval quality for your language and domain, plus dimensions, cost and speed.

Vector Databases

Stores optimised for nearest-neighbour search over embeddings.

Vector databases store embeddings and run fast nearest-neighbour search. They also support metadata filters. Choose based on scale, hosting needs and integration.

Pinecone

A managed vector database service with simple scaling.

Pinecone is a managed vector database service with simple scaling. Teams that do not want to run infrastructure often choose a managed option.

Weaviate

An open-source vector database with hybrid search and modules.

Weaviate is an open-source vector database with hybrid search and modules for embedding. It can be self-hosted or used as a managed service.

Chroma

A lightweight open-source vector store popular for prototypes.

Chroma is a lightweight open-source vector store, popular for prototypes and small applications. Check its production readiness for your scale before committing.

FAISS

Meta's library for efficient similarity search over vectors, run in your own process.

FAISS is a library from Meta for efficient similarity search, run inside your own process. It is fast but does not provide a full database service by itself.

Semantic Search

Finding passages by meaning using embeddings, even without shared keywords.

Semantic search finds passages by meaning, even when they do not share keywords. It handles paraphrases well but can miss exact codes and names.

Keyword Search

Matching exact terms (BM25), strong for names, codes and rare words.

Keyword search, often BM25, matches exact terms. It is strong for names, product codes and rare words that embeddings can blur.

Hybrid Search

Combining keyword and semantic scores for better recall.

Hybrid search combines keyword and semantic scores. It usually improves recall over either method alone, at the cost of more tuning.

Metadata Filtering

Restricting retrieval by attributes such as date, source, permission or product.

Filters restrict retrieval by attributes such as date, source, product or user permission. Permission filters are essential so users only retrieve documents they may read.

Retrieval Strategies

Top-k, multi-query, parent-document, query rewriting and filters tuned to the content.

Options include top-k retrieval, multi-query retrieval, parent-document retrieval and query rewriting. Choose based on the failure patterns you see in evaluation.

Reranking

A second model reorders retrieved candidates by relevance before generation.

A reranker, a second model, reorders retrieved candidates by relevance before generation. It improves precision at the cost of extra latency.

Context Construction

Selecting, ordering and formatting retrieved passages inside the prompt.

Context construction selects, orders and formats retrieved passages inside the prompt, with source labels. Include only what supports the answer, since extra text can distract the model.

Grounded Generation

Instructing the model to answer only from provided sources and say when they are insufficient.

The model is instructed to answer only from the provided sources and to say when they are insufficient. Test that it does this on questions the documents do not cover.

Citation and Source Attribution

Linking each claim to its source so users can verify.

Each claim links to its source so users can verify it. Citations build trust, but they must point to the passage that actually supports the claim.

RAG Evaluation

Measuring retrieval and answer quality separately on a labelled question set.

Evaluate retrieval and answer quality separately on a labelled set of questions. A wrong answer may come from retrieval or from generation, and you need to know which.

Retrieval Precision and Recall

Precision is the share of retrieved passages that are relevant; recall is the share of relevant passages retrieved.

Precision is the share of retrieved passages that are relevant. Recall is the share of relevant passages that were retrieved. Both must be measured to tune chunking and ranking.

Faithfulness and Context Relevance

Faithfulness checks the answer is supported by the context; context relevance checks the context fits the question.

Faithfulness checks the answer is supported by the context. Context relevance checks the retrieved context fits the question. Both can be scored with a model-based judge validated against human labels.

RAG Performance Optimisation

Tuning chunking, embeddings, hybrid search, reranking, caching and prompts for quality, speed and cost.

Tune chunking, embeddings, hybrid search, reranking, caching and prompts for quality, speed and cost. Change one factor at a time and re-run the evaluation.

Production RAG Monitoring

Tracking retrieval quality, unanswered questions, latency, cost and freshness in live use.

Monitor retrieval quality, unanswered questions, latency, cost and document freshness in live use. Questions with no good source are a useful signal for content gaps.

Summary of terms

Term	Meaning

RAG Fundamentals	Retrieve relevant documents at request time and give them to the model to ground its answer.
Why RAG Is Required	Models lack your private or recent data and can hallucinate; RAG supplies sources and enables citations.
RAG vs Fine-Tuning	RAG adds knowledge at query time and is easy to update; fine-tuning changes behaviour or style. They are often combined.
RAG Architecture	Ingest and index, retrieve, rerank, build context, generate, cite, and monitor.
Data Ingestion	Collecting documents from sources and keeping the index in sync with changes.
Document Parsing	Extracting clean text, tables and structure from PDFs, web pages and office files.
Data Cleaning	Removing noise, duplicates and boilerplate and fixing encoding so retrieval is accurate.
Document Chunking	Splitting documents into passages sized for retrieval and context limits.
Chunking Strategies	Fixed size, recursive, by heading or semantic boundaries, with overlap to preserve context.
Embedding Models	Models that convert text to vectors; choose by quality, language coverage, dimensions and cost.
Vector Databases	Stores optimised for nearest-neighbour search over embeddings.
Pinecone	A managed vector database service with simple scaling.
Weaviate	An open-source vector database with hybrid search and modules.
Chroma	A lightweight open-source vector store popular for prototypes.
FAISS	Meta's library for efficient similarity search over vectors, run in your own process.
Semantic Search	Finding passages by meaning using embeddings, even without shared keywords.
Keyword Search	Matching exact terms (BM25), strong for names, codes and rare words.
Hybrid Search	Combining keyword and semantic scores for better recall.
Metadata Filtering	Restricting retrieval by attributes such as date, source, permission or product.
Retrieval Strategies	Top-k, multi-query, parent-document, query rewriting and filters tuned to the content.
Reranking	A second model reorders retrieved candidates by relevance before generation.
Context Construction	Selecting, ordering and formatting retrieved passages inside the prompt.
Grounded Generation	Instructing the model to answer only from provided sources and say when they are insufficient.
Citation and Source Attribution	Linking each claim to its source so users can verify.
RAG Evaluation	Measuring retrieval and answer quality separately on a labelled question set.
Retrieval Precision and Recall	Precision is the share of retrieved passages that are relevant; recall is the share of relevant passages retrieved.
Faithfulness and Context Relevance	Faithfulness checks the answer is supported by the context; context relevance checks the context fits the question.

RAG Performance Optimisation	Tuning chunking, embeddings, hybrid search, reranking, caching and prompts for quality, speed and cost.
Production RAG Monitoring	Tracking retrieval quality, unanswered questions, latency, cost and freshness in live use.

Practice exercise: Build a small RAG assistant over ten documents and measure retrieval precision and recall on twenty questions.

Self-check

1. Explain RAG Fundamentals in two sentences to a non-technical stakeholder, then say what you would measure for it.
2. Explain FAISS in two sentences to a non-technical stakeholder, then say what you would measure for it.
3. Explain Production RAG Monitoring in two sentences to a non-technical stakeholder, then say what you would measure for it.

Related SkillNavora material

- AI Tutor : Ask questions on any topic.
- Study Library : Every lesson, class and PDF by programme.

© SkillNavora · skillnavora.in · Facts and tools change; check official sites before relying on a figure.

ORIGINAL CONTENT

Original AI Product Management content

The AI Product Management programme already on SkillNavora, included unchanged so the full course is in one place.

Quick links Study Library (all material) Learning sites for PM AI builder tools (how-to) Webhooks & Playwright MCP Webhook URLs and keys (start here) AI Skills Academy AI Skills Map Context engineering (PDF) AI skills & templates (PDF) Orchestration (PDF) PRD library (120 ideas) AI PRD Writer AI PM Toolkit AI Evals (PDF) Browser automation (PDF) AI second brain (PDF) PM handbook Top 100 PM questions AI product sense Q&A (86) PM practice test Practice Lab Tool tutorials AI Agents & MCP for PMs

PRD documents

PRD library: template, framing guide and 120 PRD ideas

Learn how top PMs frame a PRD, download the free template, and turn any of 120 ideas across 15 industries into a full PRD with AI.

Open the PRD library ✨ AI PRD Writer PRD template (Word) AI Evals for PMs (PDF)

Context window

Context window: what a model can see at once

The context window is the most tokens a model can consider in one request. It covers your prompt, any documents you paste, the conversation so far and the reply. Anything outside it is invisible to the model. For PMs: a longer window fits more material but costs more and can slow responses; put the most important facts near the instructions; use retrieval (RAG) rather than pasting whole libraries; and test answers on long inputs, where quality can drop.

Context engineering course Context windows in the AI PM 2 course

Tokens and context in the Grokbot course

AI-First Product Management

Resources for Product Managers

The PM Master Handbook, AI and product tools with their official websites, interview practice, a frameworks cheat sheet, a PRD template and free courses.

Open the PM handbook library: all 68 PDFs ↗ PM Glossary Handbook Tokens, Pricing and Tokenization Vibe Coding addendum

PM Master Handbook by SkillNavora
23 modules and 620+ pages: 12 Product Core modules, an 11-class AI Tools Lab, frameworks, case studies, projects and interview prep.

Open the interactive handbook ↗ Complete handbook PDF (620+ pages, 23 MB) ↗

Product Core modules

1Fundamentals of Product Management*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
2Product Strategy and Strategic Thinking*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
3Problem Discovery and Market Research*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
4Problem Framing and Prioritization*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
5UX and Design Collaboration*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
6Effective Communication and Stakeholder Management*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
7Product Analytics and Metrics*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
8Product Launch, Growth, and Monetization*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
9Tech 101 and System Design for PMs*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
10LLMs, RAGs, and AI Foundations*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
11Vibe Coding for Product Managers*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
12Agentic AI and AI Systems*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides

AI Tools Lab

13AI Tools Landscape and Your Starter StackPDF ↗ 14Claude for Daily PM WorkPDF ↗ 15Claude Code: Basic to AdvancedPDF ↗ 16Lovable: Basic to AdvancedPDF ↗ 17Replit and Other AI App BuildersPDF ↗ 18Figma and AI Design ToolsPDF ↗ 19Notion, PRD Tools and AI DocumentationPDF ↗ 20Automation with Zapier, Make and n8nPDF ↗ 21AI Chatbots for BusinessPDF ↗ 22Voice Agents for BusinessesPDF ↗ 23Learning Paths, Courses and CommunitiesPDF ↗

Business AI Stack (chapters 24 to 31)

24The business AI stack: tools and where to learn themPDF ↗ 25Next.js: customer-facing dashboardsPDF ↗ 26Zoho CRM: managing customers and leadsPDF ↗ 27WhatsApp Business Platform: customer messagingPDF ↗ 28Deployment with Vercel and AWSPDF ↗ 29Jira: tracking product and development workPDF ↗ 30Meta Business Suite, social media marketing and Google AdsPDF ↗ 31A one-page website that generates leadsPDF ↗

Complete AI tool guides (15)

1n8n: workflows, Gmail and an AI reply agentPDF ↗ 2Make: scenarios, routers and lead routingPDF ↗ 3Zapier: invoices from Gmail to Drive and a sheetPDF ↗ 4Webhooks: how apps talk in real timePDF ↗ 5Replit: one app for the web and a phonePDF ↗ 6Lovable: a job application trackerPDF ↗ 7Figma: components, auto layout and handoffPDF ↗ 8Claude Code: an AI coding agent in your projectPDF ↗ 9MCP: connecting AI assistants to your toolsPDF ↗ 10LLMs: how they work and how to call themPDF ↗ 11RAG: answers from your own documentsPDF ↗ 12AI agents: a job-search agent with approvalsPDF

↗ 13Agentic AI: planning, guardrails and teamsPDF ↗ 14Generative AI: uses, limits and a weekly reportPDF ↗ 15Other AI tools: choosing one per jobPDF ↗

Complete AI tool guides (one PDF, 30 pages) ↗

Libraries

★Framework Library*29 frameworks with worked examples and templates*PDF ↗ ★Case Study Library*12 product case studies*PDF ↗ ★Practical Projects and Portfolio*5 end-to-end portfolio projects*PDF ↗ ★Interview Handbook*100+ PM interview questions with model answers*PDF ↗ ★AI Tools Lab*Every AI tools class with step-by-step labs*PDF ↗

PM interview practice (free, on the original sites)

Hand-picked public question banks and guides from other PM communities. Each link opens the original page in a new tab.

PM frameworks cheat sheet

Framework	Use it to	How it works
RICE	Prioritise a backlog	Reach × Impact × Confidence ÷ Effort. Score each idea, then sanity-check the ranking.
Kano	Decide which features delight	Classify features as Must-be, Performance, Attractive, Indifferent or Reverse using a short survey.
MoSCoW	Scope a release	Must have, Should have, Could have, Won't have (this time).
Jobs-to-be-done	Understand motivation	When [situation], I want to [motivation], so I can [outcome].
AARRR (Pirate metrics)	Map the growth funnel	Acquisition, Activation, Retention, Referral, Revenue.
HEART	Measure user experience	Happiness, Engagement, Adoption, Retention, Task success, each with goals, signals and metrics.
North Star metric	Align the team	One metric that captures the value users get, supported by input metrics the team can move.
Five whys	Find the root cause	Ask 'why?' repeatedly until you reach a cause you can act on.
CIRCLES	Answer product design questions	Comprehend, Identify users, Report needs, Cut through prioritisation, List solutions, Evaluate trade-offs, Summarise.

PRD template

PRD: [Feature name] Owner: [Name] · Status: Draft · Last updated: [Date] 1. Problem Who has the problem, what it is, and the evidence (data, interviews, tickets). 2. Goal and success metrics Primary metric (with baseline and target), secondary metrics, guardrails. 3. Users and use cases Target segment, top 3 user stories: "As a [user], I want [action] so that [benefit]." 4. Solution What we'll build (link designs), what's out of scope, open questions. 5. AI behaviour (if applicable) Inputs, model, prompt/approach, what it must never do, human review, evaluation set and pass criteria, cost per request. 6. Launch plan Experiment design, rollout stages, support and comms, rollback criteria. 7. Risks and dependencies Technical, legal/privacy, operational; owners and mitigations.

Free courses