# RAG IS OVERRATED

## Here's What You Actually Need

⚠️ **CONTRARIAN TAKE**

Get and at: community.nachiketh.in

## THE TEST

Can you solve your problem with:

Structured Few-Shot Examples

Chain-of-Thought Prompting

Function Calling
+
Clear Schemas

**IF YES → YOU DON'T NEED RAG**

Get **source code** and **production patterns** at:
community.nachiketh.in

# RAG ADDS...

🏗️

## INFRASTRUCTURE COMPLEXITY

Vector databases, embedding pipelines, deployment overhead

💰

## DATABASE COSTS

Pinecone, Weaviate, Qdrant - $50-500/month minimum

⚙️

## CHUNKING OVERHEAD

Splitting strategies, overlap logic, metadata management

⏱️

## RETRIEVAL LATENCY

Additional 200-500ms per request for embedding + search

Get **source code** and **production patterns** at:
community.nachiketh.in

# DECISION FRAMEWORK

## 1

**Can it fit in the prompt? (<8k tokens)**

YES → Use prompt engineering. Stop.

**SIMPLE**

## 2

**Is the knowledge static?**

YES → Consider fine-tuning

**SIMPLE**

## 3

**Does it change frequently?**

YES → NOW you need RAG

**RAG**

Get **source code** and **production patterns** at:
community.nachiketh.in

↓

# WHEN TO ACTUALLY USE RAG

📚

### Knowledge base is LARGE (>100k tokens)

Example: Product documentation, legal corpus, research papers

🔄

### Information changes FREQUENTLY

Example: News, pricing, inventory, policy updates

🔌

### Context must be EXTERNAL to model

Example: User-specific data, company-internal docs, compliance

Get **source code** and **production patterns** at: community.nachiketh.in

# START SIMPLE.

# ADD COMPLEXITY ONLY WHEN FORCED.

📥

**JOIN COMMUNITY**
community.nachiketh.in

🎓

**LEARN PRODUCTION SYSTEMS**
bootcamp.nachiketh.in

Get **source code** and **production patterns** at:
community.nachiketh.in

Speaker notes