RAG 2.0 - The Next Generation Of RAG Systems

Wayfare



HI!
I'M DIANA.

Junior Developer @Wayfare

Passionate about AI and modern software development

Exploring how intelligent systems improve real-world applications

Always eager to learn and grow through technology

Wayfare

ABOUT WAYFARE

We've supported businesses of different sizes and across multiple industries to achieve their goals.

TOP TALENT, SHARED KNOWLEDGE, CLOSE COLLABORATION

Our cost-effective solutions and active community engagement foster innovation and efficiency.

We offer tailored services including Modernization, New Development, Maintenance & Support, Quality Assurance, UI/UX Design, Consultancy and Project Delivery.



OUR EXPERTISE

PROGRESS
OpenEdge, CorticonJS, Sitefinity

MOBILE React Native

DEVOPS Docker, Kubernetes, Jenkins

MANUAL AND AUTOMATION TESTING

WEB TECHNOLOGIES
Angular, React, Vue.JS, NodeJS, JS, TypeScript, Java, .Net

CLOUD AWS, Azure, Google Cloud Platform

UI/UX Figma, XD, Zeplin, Illustrator

LOW CODE/ NO-CODE FRAMEWORKS

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

INTRODUCTION



- How did we build AI systems before Generative AI?
- What does building an AI application mean today?
- Where does RAG fit into this picture?

Wayfare

INTRODUCTION





How did we build AI systems before Generative AI?



What does building an AI application mean today?



Where does RAG fit into this picture?

Wayfare

INTRODUCTION





How did we build AI systems before Generative AI?



What does building an AI application mean today?



Where does RAG fit into this picture?

Wayfare

AI SYSTEM

Building the Brain

TRAINING

Feed the model millions of examples

Adjust weights with backpropagation whenever it’s wrong

Needs massive compute (GPUs, cloud clusters)

Expensive and time-intensive, but this is where intelligence is created

INFERENCE

Putting the Brain to Work

Takes a small input (image, text, or voice)

Runs a quick forward pass to predict

Deployed everywhere - apps, cloud, edge devices

Optimized for speed so users get answers instantly

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

RAG 1.0 - The Foundations

Retrievel-Augmented Generation (RAG) is a method that improves the output of a Large Language Model (LLM) by referencing an external, authoritative knowledge base *at the time of the query*, rather than relying only on the model's pre-trained data.

RAG extends these capabilities to specific domains or an organization's internal knowledge base – *without needing to retrain the model*.

Large Language Models (LLMs) are trained on vast volumes of data and use billions of parameters to generate original content – for tasks like:

- Answering question
- Translating text
- Completing sentences

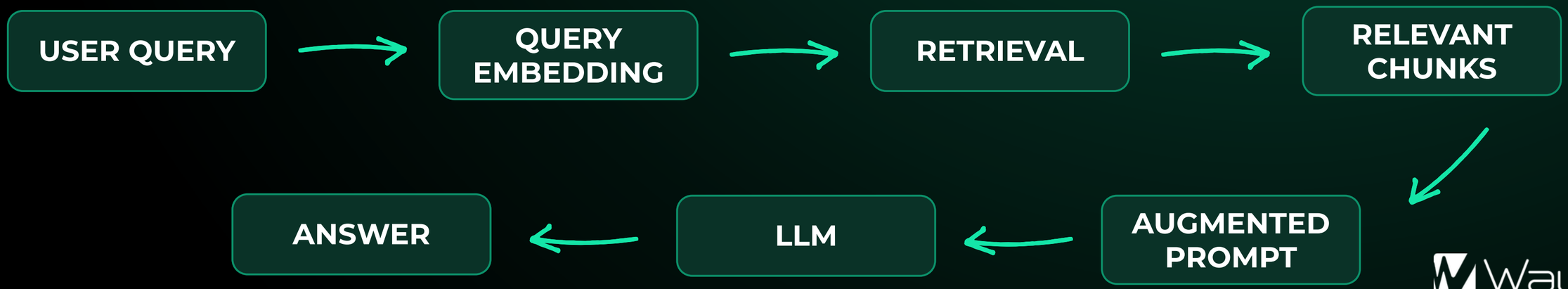
Wayfare

RAG KEY STAGES

KNOWLEDGE PREPARATION (Offline Process)



INFERENCE (Online Process)



Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

PREPARING THE KNOWLEDGE BASE

DATA



CHUNKING

Split documents into smaller, meaningful pieces

- ***Fixed-Size Chuncking -*** Splits text into fixed token windows
- ***Sentence-Based Chuncking -*** Splits at sentence boundaries
- ***Semantic Chuncking -*** Groups semantically similar content

EMBEDDINGS

Convert each chunk into a numerical representation

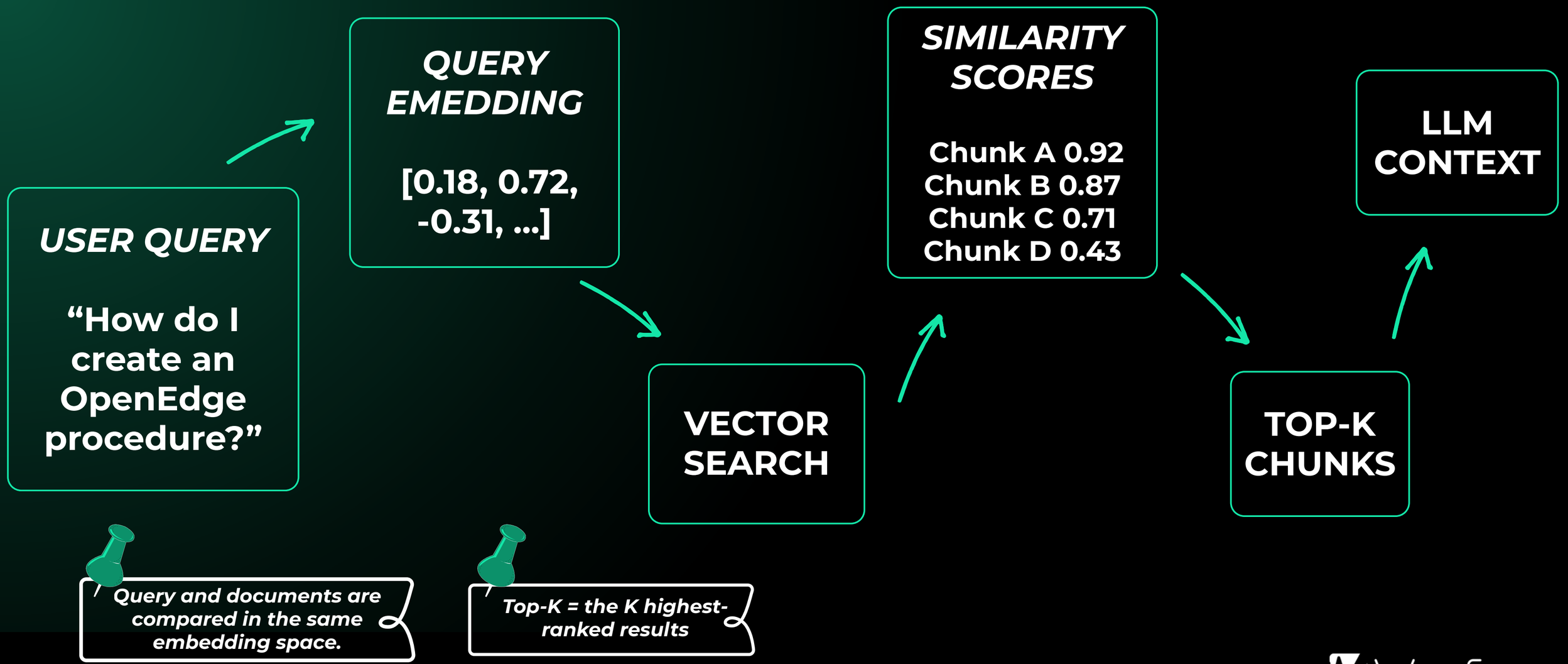
- [0.05,-0.6,...]
- [0.4,0.7,...]
- [0.2,0.01,,...]
- [-0.4,-0.6,...]
- [0.35,0.1,...]
- [0.9,0.3,...]

VECTOR DATABASE

Store and index the embeddings for efficient search



RETRIEVING THE RIGHT CONTEXT



Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

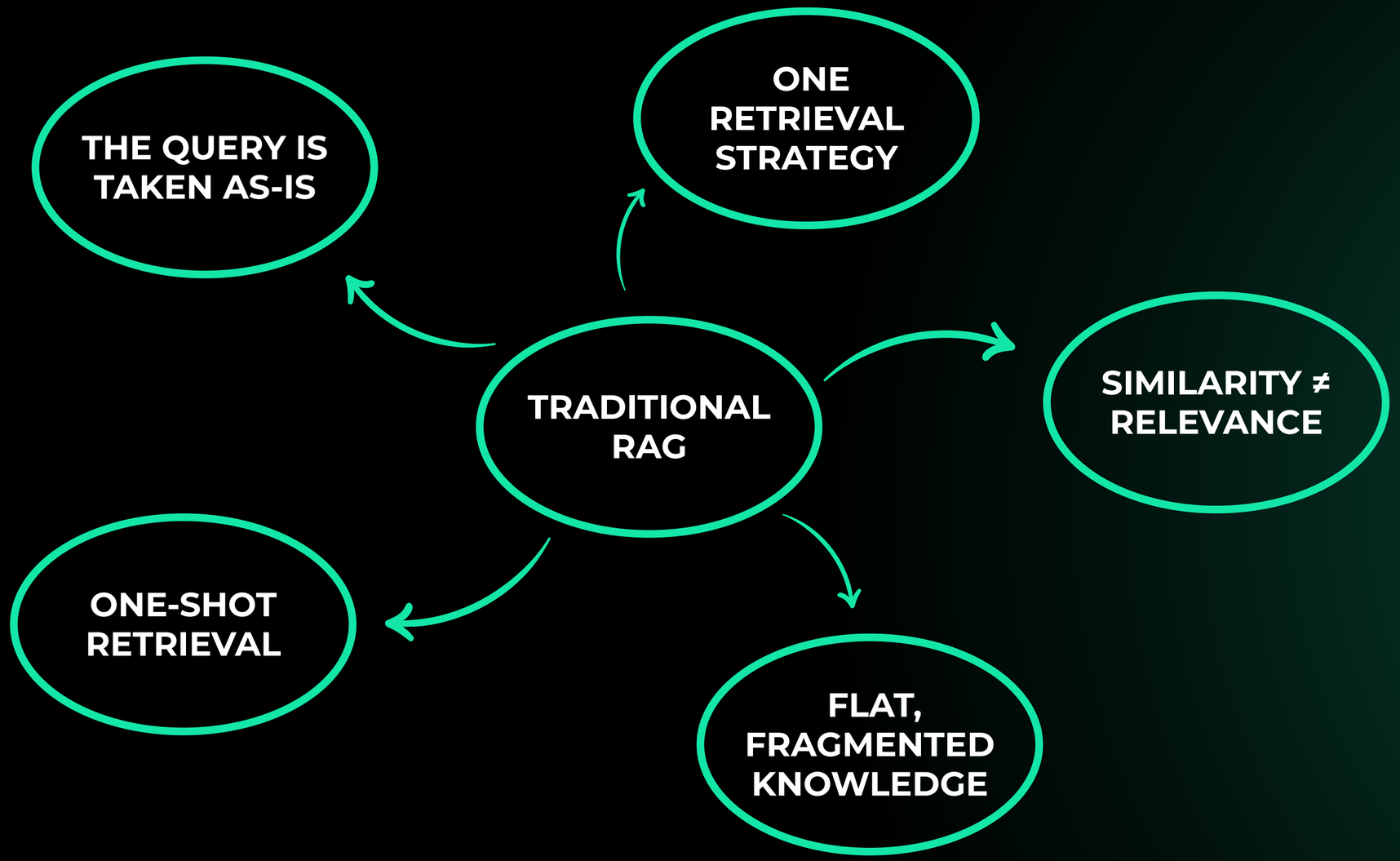
DEMO

Key Takeaways

Q&A

Wayfare

Where Traditional RAG Breaks



Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

Naive RAG: The Starting Point



One query. One retrieval step. One generated answer.

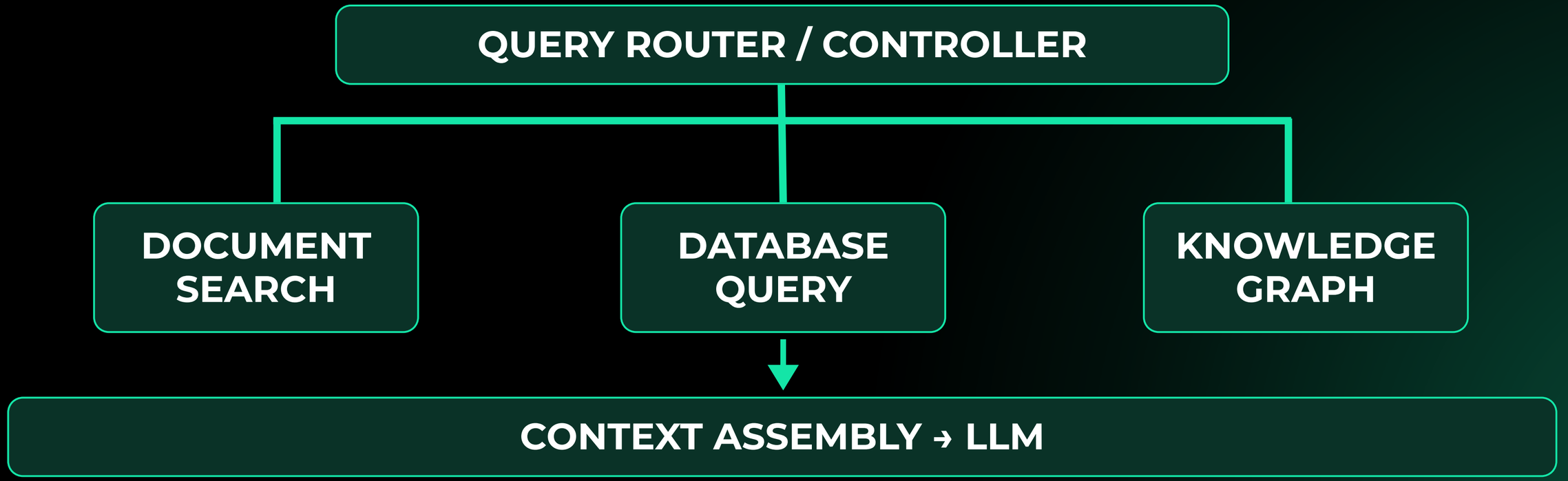
Wayfare

Advanced RAG: Optimizing the Pipeline

PRE-RETRIEVAL	RETRIEVAL	POST-RETRIEVAL
QUERY REWRITING	HYBRID SEARCH	RERANKING
QUERY EXPANSION	CANDIDATE SELECTION	CONTEXT FILTERING

Wayfare

Modular RAG: Beyond a Fixed Pipeline



Wayfare

Agentic RAG



AI AGENT

PLAN → CHOOSE TOOLS → EVALUATE

Select tools & retrieve evidence

Vector Database

Code Repository

External API

Evaluate → Retrieve again if needed

GENERATE THE GROUND ANSWER

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

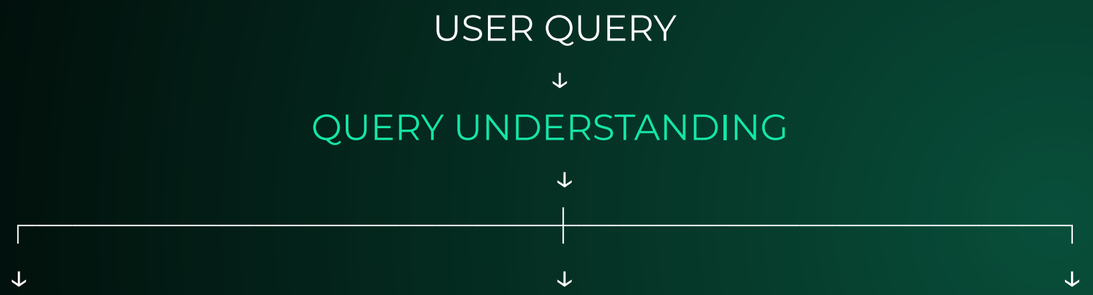
DEMO

Key Takeaways

Q&A

Wayfare

SMARTER QUERY UNDERSTANDING



How has authentication changed, and where are user permissions checked?

REWRITE

Clarify intent

"Changes to the authentication mechanism..."

EXPAND

Add related terms

authentication, authorization, permissions, access control

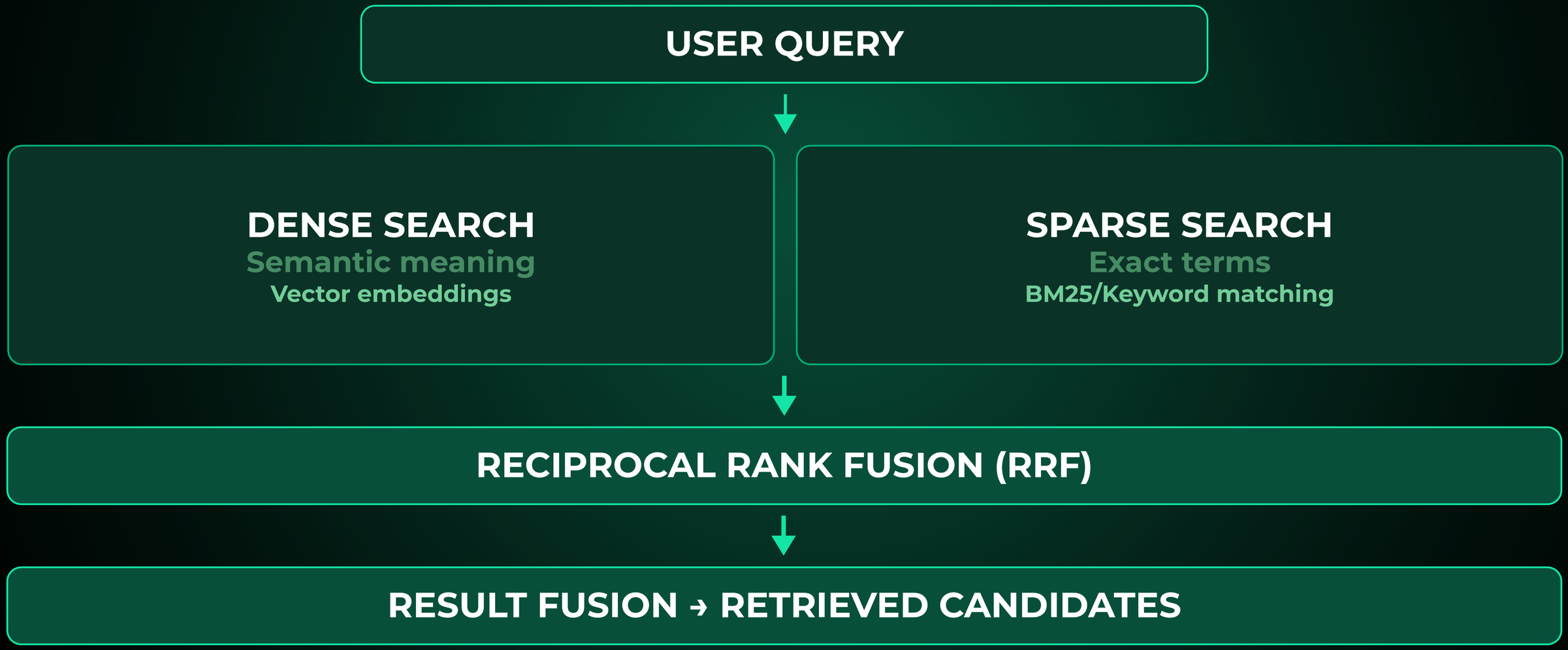
DECOMPOSE

Split into sub-questions

1. How has authentication changed?
2. Where are permissions checked?

Wayfare

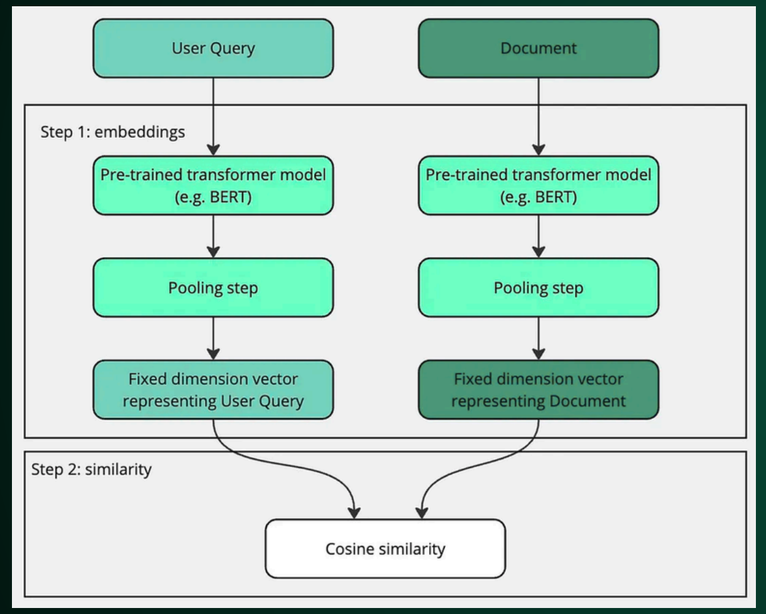
Hybrid Retrieval



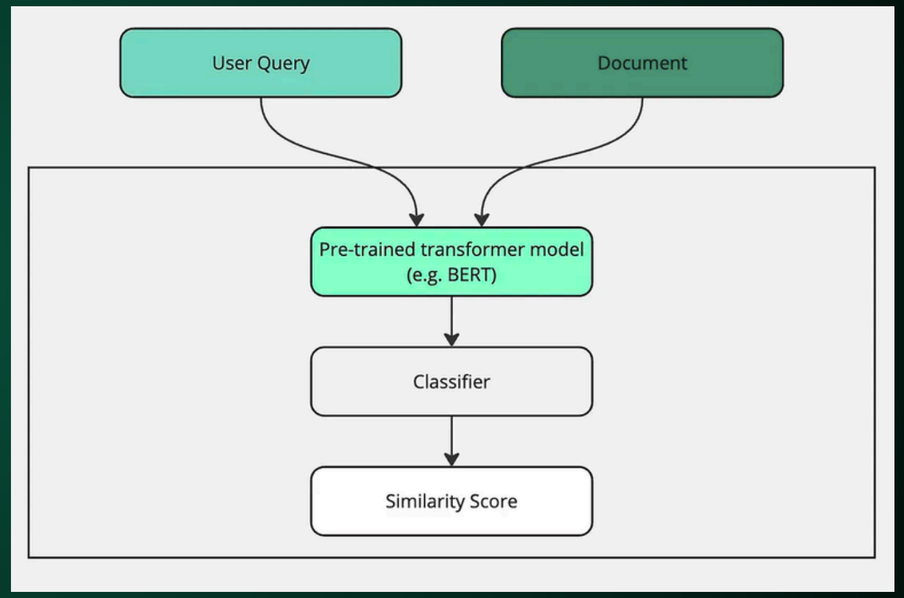
Wayfare

Smarter Ranking

BI-ENCODER



CROSS-ENCODER



USER QUERY

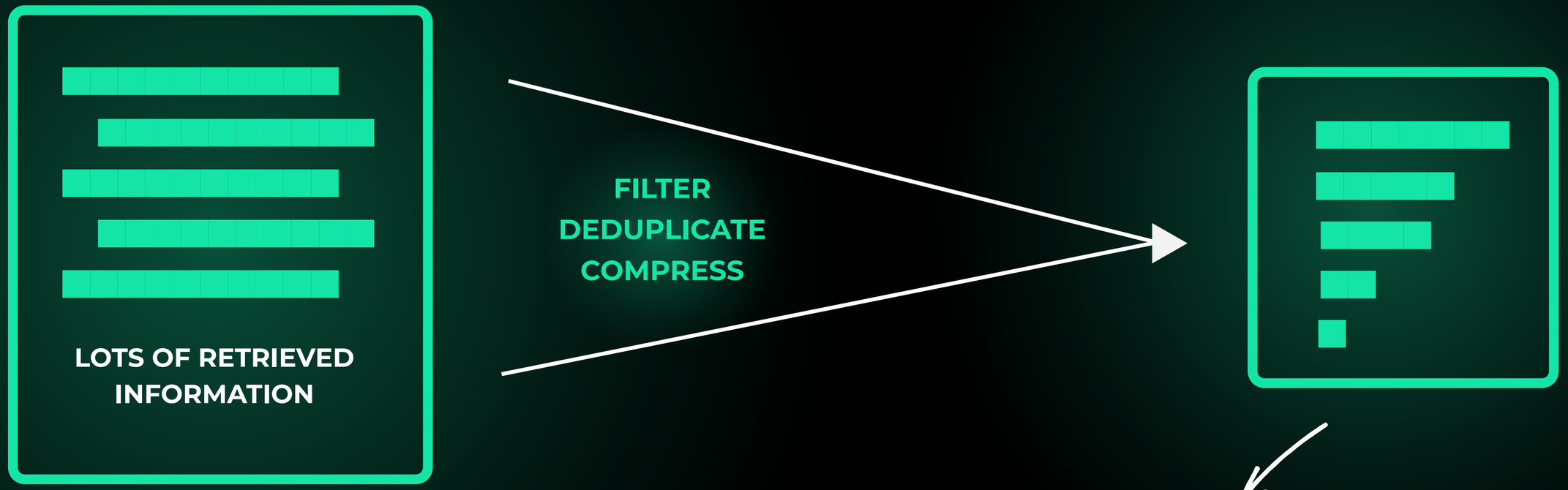
TOP 20-50 CANDIDATES

TOP 5 BEST CONTEXT

Similarity finds candidates.
Reranking finds relevance.

Wayfare

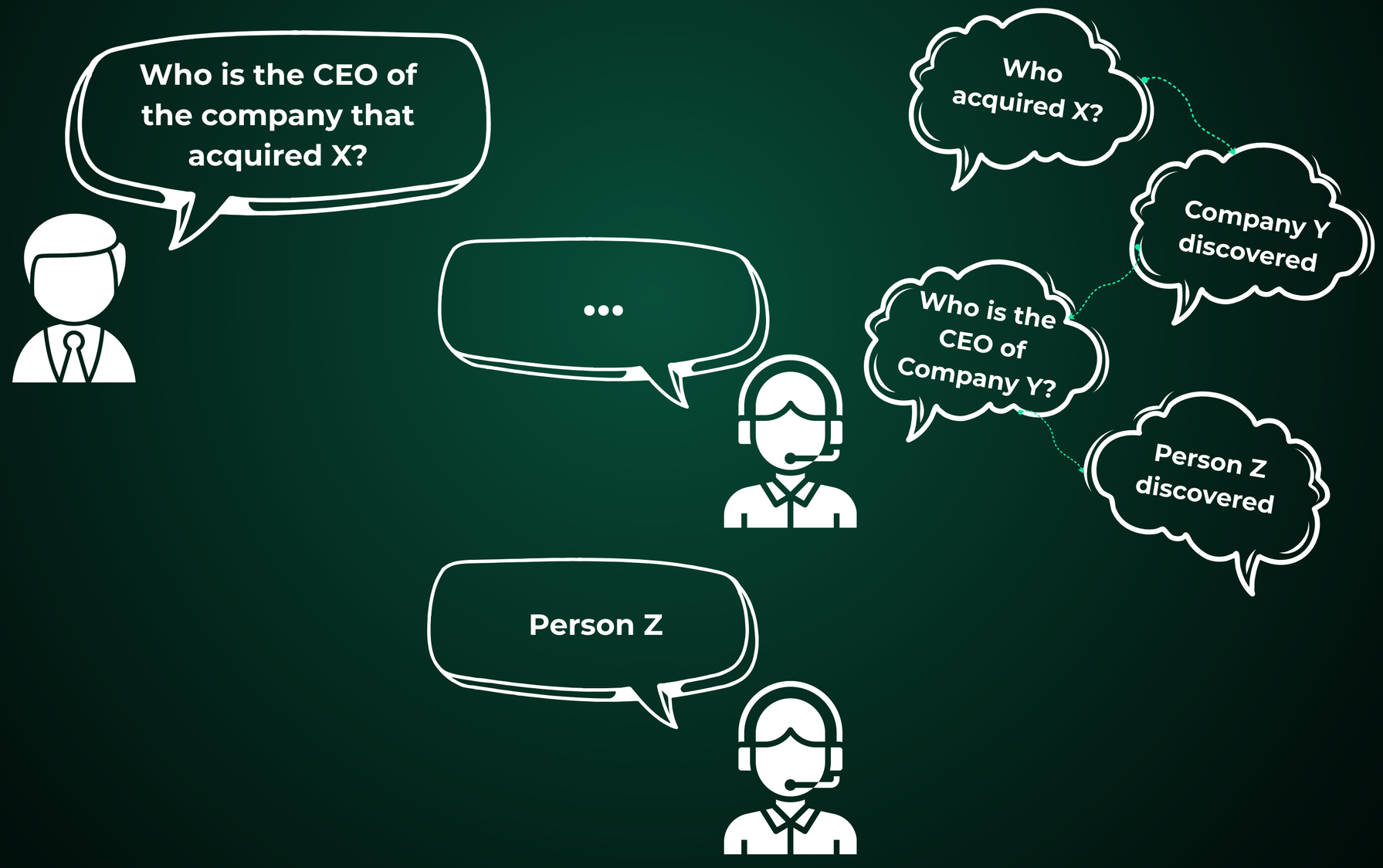
Context Optimization



More Context ≠ Better Context

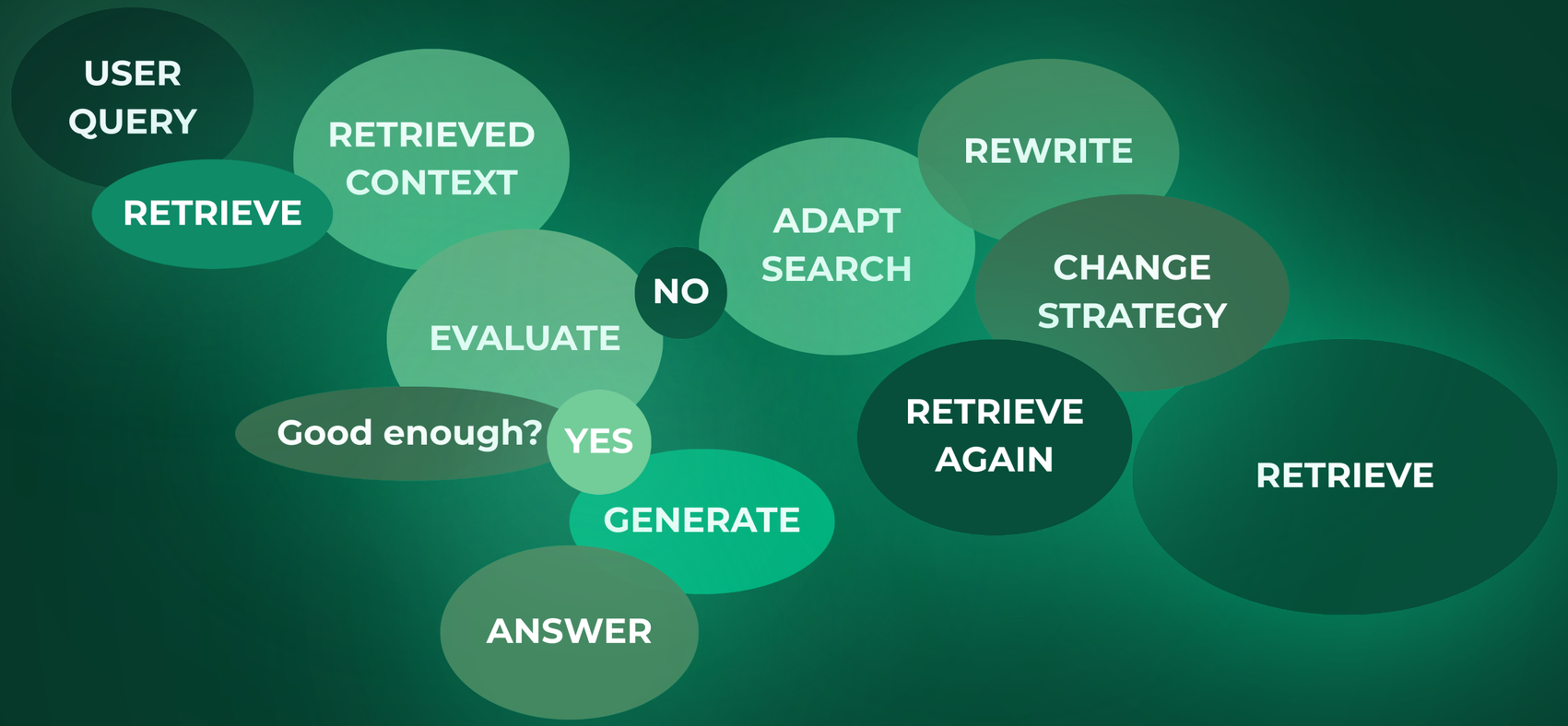
Wayfare

MULTI-HOP RETRIEVAL



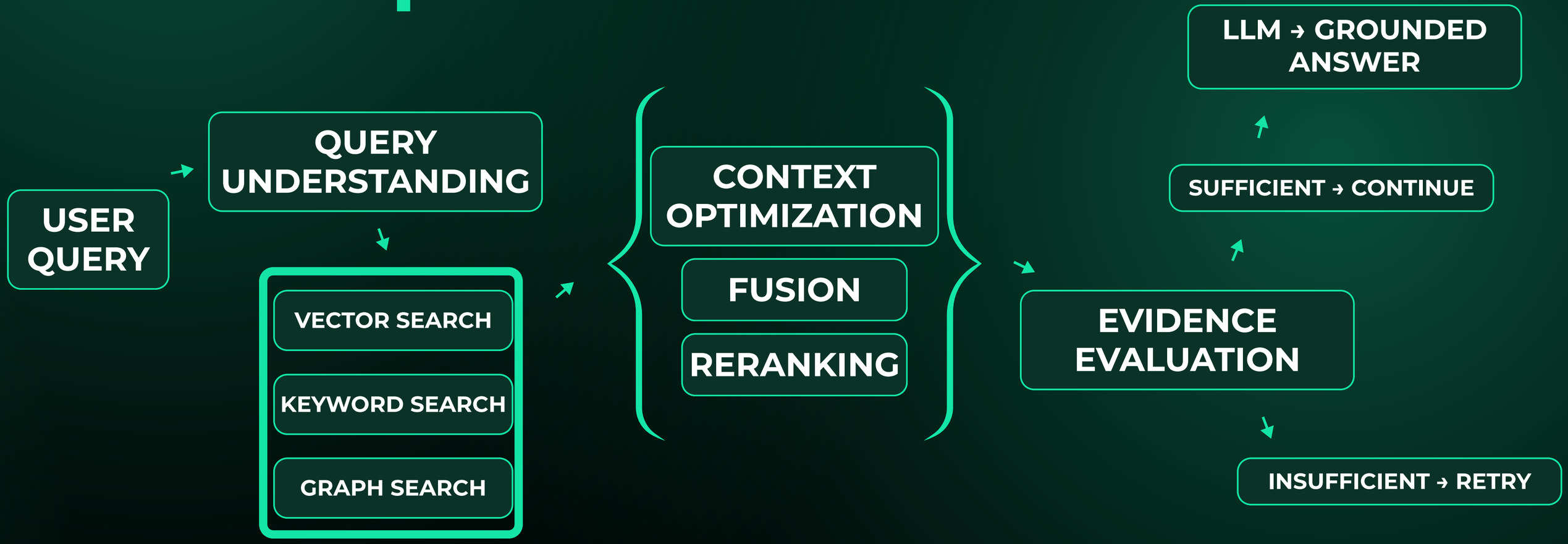
Wayfare

ADAPTIVE RETRIEVAL



Wayfare

The Modern RAG Pipeline



More Context ≠ **Better** Context

Wayfare

Evaluating Modern RAG Systems

RETRIEVAL	GENERATION	OPERATIONS
Recall@K	Faithfulness	Latency
Precision@K	Answer accuracy	Cost
Ranking quality	Citation quality	Reliability

Measure retrieval quality, answer quality, and operational performance.

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

Demo

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

KEY TAKEAWAYS



RAG Has Evolved

From simple retrieval pipelines to optimized, modular, and adaptive systems.



Better Retrieval, Better Context

The right evidence matters more than the amount of information retrieved.



No One-Size-Fits-All Architecture

The right approach depends on the task, data, and system requirements.



Measure What Matters

Balance answer quality and reliability with latency, cost, and complexity.

Wayfare

TABLE OF CONTENTS

Introduction

RAG 1.0 - The Foundations

Preparing the Knowledge Base

Where Traditional RAG Breaks

From RAG 1.0 to RAG 2.0

Inside RAG 2.0

DEMO

Key Takeaways

Q&A

Wayfare

Q&A

wayfare.ro

business@wayfare.ro

Wayfare

Our PUG Challenge sessions



DIANA

RAG 2.0 - The Next Generation of RAG Systems

Wednesday, November 4, 4:30 PM



CARLA

AI Security Auditor: Real-Time Threat Detection for OpenEdge Applications

Friday, November 6, 10:15 AM



CLAUDIU

Understanding database performance

Friday, November 6, 1:45 PM



DIANA

One BA, two teams: How I standardized workflows without slowing anyone down

Friday, November 6, 9:00 AM



ALEXANDRU

Exploring Local AI: Capabilities and Limits of Self-Hosted Models for Development and Documentation

Thursday, November 5, 3:00 PM



OANA

Guideline to an Accessible Web Application

Thursday, November 5, 9:00 AM

Wayfare

Wayfare



thank you

wayfare.ro

wayfare.community

wayfare.academy

business@wayfare.ro