Lecture 03 - Contents

An overview of the parts in the medical RAG systems lecture.

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 3:

RAG for Healthcare

Evidence-Based AI

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

RAG Architecture for Healthcare

Query Processing → Knowledge Retrieval → Evidence Generation

📚 Literature Search

Search 35M+ PubMed articles with semantic understanding

📋 Clinical Guidelines

Access WHO, CDC guidelines with real-time updates

💊 Drug Information

Query DrugBank, RxNorm for interactions and side effects

🔬 Diagnostic Support

Evidence-based differential diagnosis recommendations

🏥 Treatment Planning

Protocol recommendations based on latest research

⚠️ Safety Monitoring

Real-time adverse event detection and reporting

✅ Factual Accuracy

🔗 Source Citations

🔄 Always Up-to-date

🎯 Domain-Specific

Part 1:

Building Medical Knowledge Bases

Clinical Guidelines Ingestion

🌍 **WHO** Global health standards and protocols	🏛️ **CDC** Disease prevention and control guidelines	🏥 **Professional Societies** Specialty-specific recommendations
💊 **FDA** Drug approval and safety information	🔬 **NIH** Research-based clinical protocols	**EU EMA** European medical guidelines

📋 Metadata Schema

- **Source:** Organization name
- **Version:** Publication date
- **Topic:** Medical category
- **Evidence:** Quality level
- **Updates:** Revision history
- **Language:** Multi-lingual support

Reserved Slot (L03_05)

추후 내용이 추가될 자리입니다. 강의 흐름의 연속성을 위해 번호를 보존합니다.

Drug Database Integration

💊 DrugBank

13,000+

Comprehensive drug data with molecular structures

🏥 RxNorm

150,000+

Standardized medication nomenclature

📋 FDA Labels

100,000+

Official prescribing information

🕸️ Drug Knowledge Graph

Drug Entity ↔ Interactions ↔ Side Effects ↔ Indications

⚠️ Interaction Matrix

Drug-drug, drug-food, drug-disease interaction checking

🔬 Pharmacokinetics

ADME properties, half-life, metabolism pathways

📊 Adverse Events

FDA FAERS database with 10M+ reports

💰 Pricing & Access

Cost information and formulary status

Vector Embedding Strategies

📷 Dense Embeddings

BERT, BioBERT, Sentence-BERT

High-dimensional
continuous space

✅ Semantic similarity
✅ Context understanding
❌ Computational cost

Best for: "chest pain" ≈ "cardiac discomfort"

📊 Sparse Embeddings

BM25, TF-IDF

Most values = 0
Only keywords

✅ Fast retrieval
✅ Interpretable
❌ No semantics

Best for: Exact term "ICD-10 I21.0"

🔀 Hybrid Approach

Dense + Sparse fusion

✅ Best of both
✅ High accuracy (95%+)
⚠️ More complex

Recommended for medical applications

Retrieval Accuracy Comparison

📏 Dimension Selection

384d - Fast, general purpose

768d - BERT standard

1024d - High precision

Dense vs Sparse Retrieval

Aspect	Dense Retrieval	Sparse Retrieval
Similarity Type	Semantic meaning	Keyword matching
Speed	Medium (ANN search)	Fast (inverted index)
Accuracy	High for concepts	High for exact terms
Medical Terms	Understands synonyms	Exact match required

Dense best for:

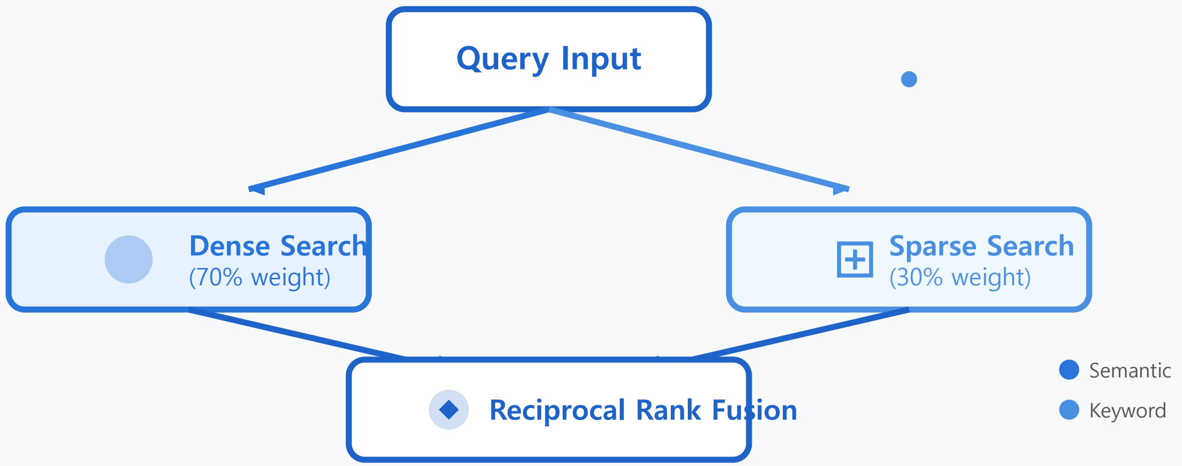
"Patient with chest pain and shortness of breath"

Sparse best for:

"ICD-10 code I21.0" or "Aspirin 81mg"

Hybrid Search Implementation

🔀 Hybrid Architecture



Weighted Sum

`score = α·dense + (1-α)·sparse`

RRF (Recommended)

`score = Σ 1/(k + rank)`

Ensemble

`Multiple models voting`

📈 Performance Improvement

Precision@10: **89% → 95%**

Recall@10: **76% → 92%**

Similarity Metrics for Medical Text

📐 Cosine Similarity

$\cos(\theta) = A \cdot B / (||A|| \, ||B||)$

Range: [-1, 1]

Best for: Dense embeddings

📏 Euclidean Distance

$d = \sqrt{\Sigma (ai - bi)^2}$

Range: [0, ∞]

Best for: Spatial similarity

🎯 Jaccard Index

$J = |A \cap B| / |A \cup B|$

Range: [0, 1]

Best for: Set overlap

🧠 Semantic Similarity

Based on medical ontology

Range: [0, 1]

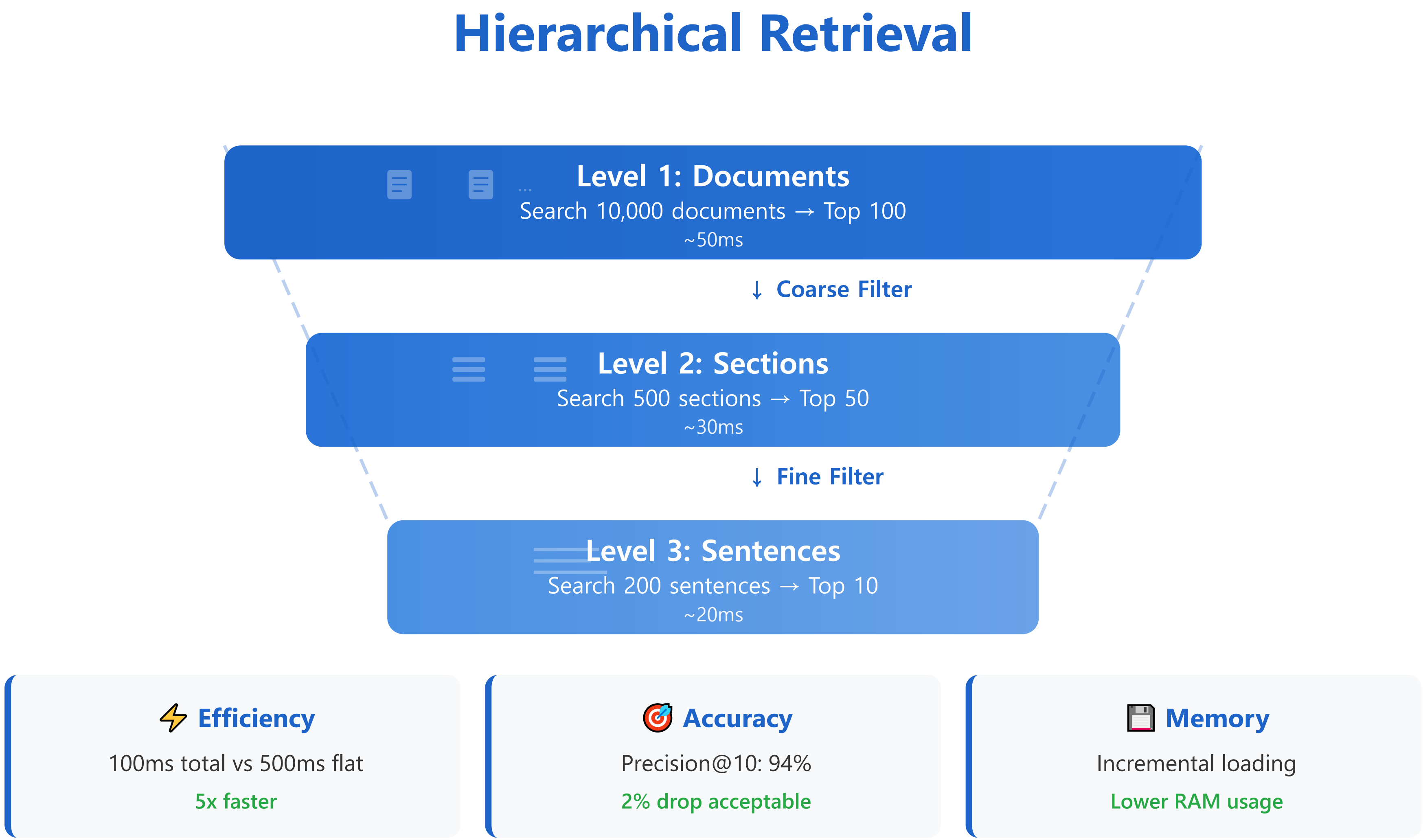
Best for: Medical concepts

💡 Medical Text Example

Text 1: "Patient has myocardial infarction"

Text 2: "Heart attack diagnosed"

Cosine: **0.89** ✅ Jaccard: **0.12** ❌ Semantic: **0.95** ✅

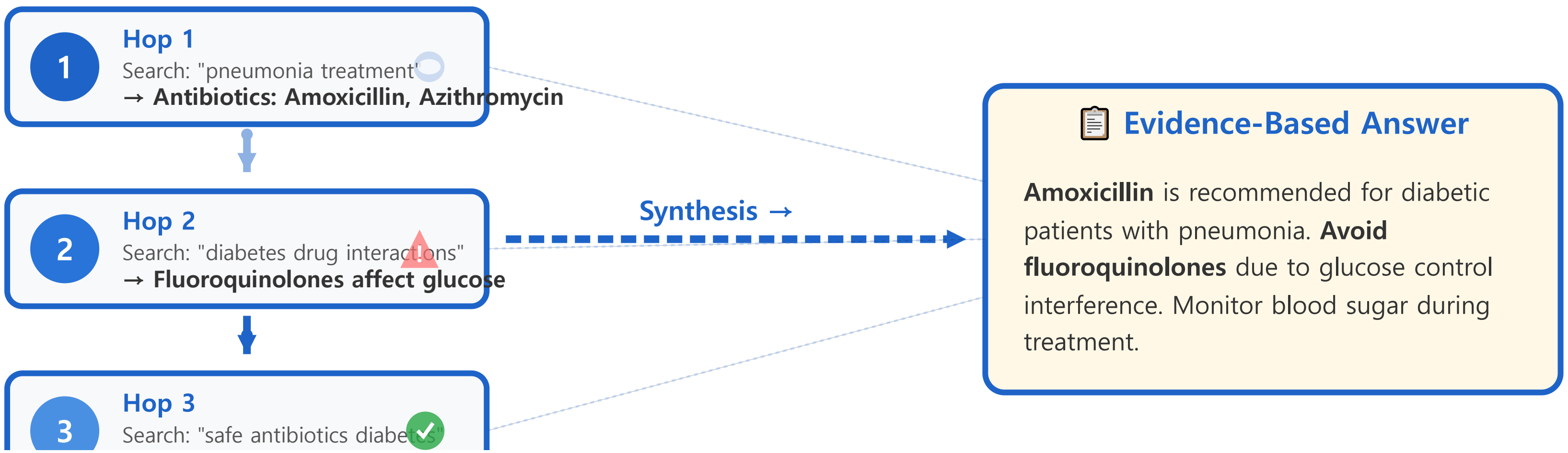


Hierarchical Retrieval
Level 1: Documents
Search 10,000 documents → Top 100
~50ms
↓ Coarse Filter
Level 2: Sections
Search 500 sections → Top 50
~30ms
↓ Fine Filter
Level 3: Sentences
Search 200 sentences → Top 10
~20ms
Efficiency
100ms total vs 500ms flat
5x faster
Accuracy
Precision@10: 94%
2% drop acceptable
Memory
Incremental loading
Lower RAM usage

Multi-hop Reasoning

🔍 Multi-hop Query Example

Q: "Treatment for pneumonia in diabetic patients?"



Citation Generation

📝 Citation Formats

APA Style

Smith, J. (2024). Title. Journal, 12(3), 45-67.

MLA Style

Smith, John. "Title." Journal 12.3 (2024): 45-67.

Vancouver

Smith J. Title. Journal. 2024;12(3):45-67.

📊 Evidence Strength Indicators

⭐⭐⭐ High: Systematic reviews, RCTs

⭐⭐ Medium: Cohort studies

⭐ Low: Case reports, expert opinion

💡 Inline Citation Example

> Aspirin reduces cardiovascular events by 25% [Smith et al., 2024 ⭐⭐⭐] in high-risk patients [Johnson, 2023 ⭐⭐].

Evidence Scoring System

🔺 Evidence Pyramid

Meta-analyses & Systematic Reviews	Score: 9-10
Randomized Controlled Trials (RCTs)	Score: 7-8
Cohort Studies	Score: 5-6
Case-Control Studies	Score: 3-4
Case Reports & Expert Opinion	Score: 1-2

📋 GRADE System Factors

- ✅ Study design quality
- ✅ Consistency of results
- ✅ Directness of evidence
- ✅ Precision (CI, p-value)
- ✅ Publication bias check

Confidence Calibration

🎯 What is Calibration?

If model says 80% confidence, it should be correct 80% of the time

Temperature Scaling

Adjust logits with temperature T
$p' = \mathrm{softmax}(\mathrm{logits} / T)$

Platt Scaling

Logistic regression on outputs
$p' = 1 / (1 + \exp(A \cdot p + B))$

Isotonic Regression

Non-parametric calibration
Monotonic function fitting

📊 Calibration Metrics

- **ECE** (Expected Calibration Error): |confidence - accuracy|
- **MCE** (Maximum Calibration Error): max|confidence - accuracy|
- **Brier Score**: Mean squared error of probabilities

Query Decomposition

🔍 Complex Query

"What are the contraindications for prescribing metformin in elderly patients with chronic kidney disease?"

↓ Decompose ↓

1. Metformin contraindications
2. Elderly patients drug considerations
3. Chronic kidney disease drug safety
4. Metformin + CKD interactions

↓ Integrate Results ↓

📋 Synthesized Answer

Metformin is contraindicated in CKD stage 4-5 (eGFR <30) due to lactic acidosis risk. In elderly CKD stage 3, dose reduction to 500mg BID with careful monitoring is recommended.

Vector Database Selection

🎯 Pinecone
- ✅ Fully managed
- ✅ Excellent scalability
- ❌ Proprietary, costly

🌊 Weaviate
- ✅ Open source
- ✅ Built-in vectorization
- ⚠️ Self-hosting required

⚡ Milvus
- ✅ High performance
- ✅ Trillion-scale
- ⚠️ Complex setup

🚀 Qdrant
- ✅ Rust-based speed
- ✅ Easy deployment
- ✅ Good for medical

🔍 Selection Criteria

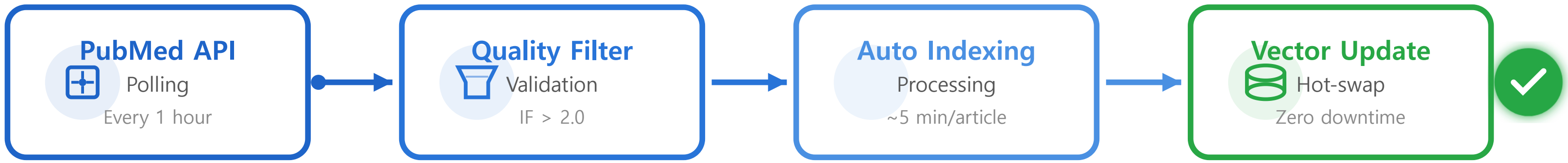
- Data volume: **>10M vectors**
- QPS: **1000+**
- Latency: **<100ms**
- HIPAA compliance: **Required**

Reserved Slot (L03_20)

추후 내용이 추가될 자리입니다. 강의 흐름의 연속성을 위해 번호를 보존합니다.

Real-time Literature Updates

Real-time Update Pipeline



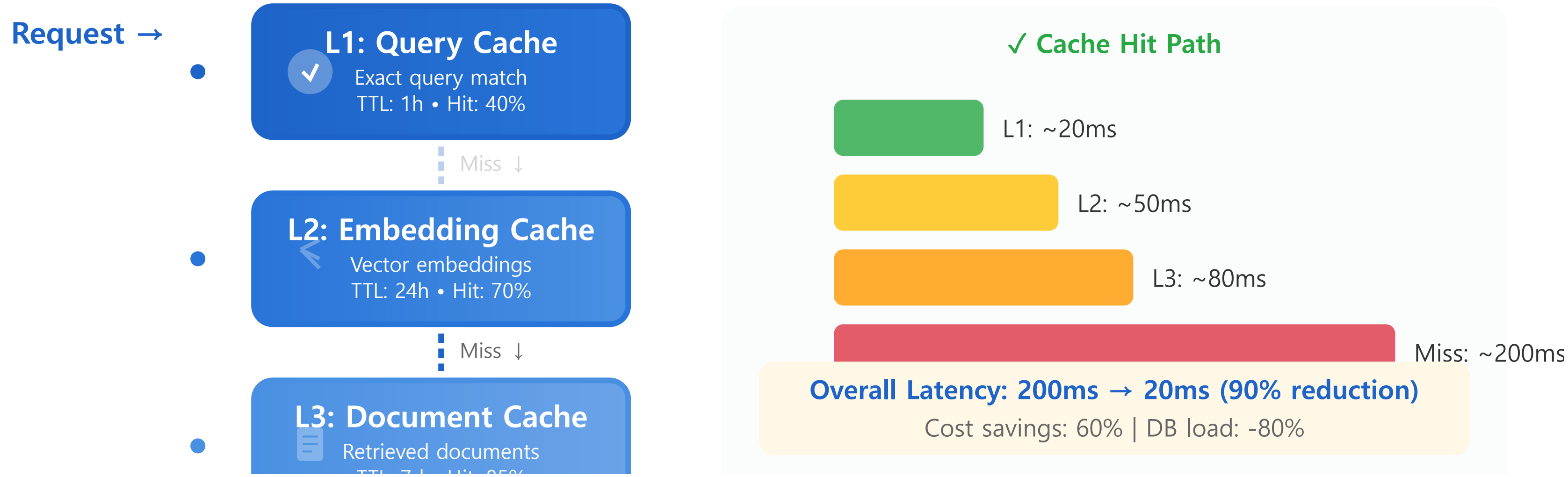
T+0	T+2min	T+7min	T+8min
~4,000 articles	~600 selected	600 indexed	Live in <30min

4,000+	<30min	99.5%
New articles/day	Detection latency	Indexing success

🔍 Quality Filters

- ✓ Peer-reviewed journals only
- ✓ Impact factor > 2.0
- ✓ Full-text availability required
- ✓ Duplicate detection algorithm

Caching Optimization



⚙️ Redis Configuration

Memory: 16GB with LRU eviction

Persistence: RDB + AOF for durability

Cluster: 3 nodes with replication

📊 Performance Impact

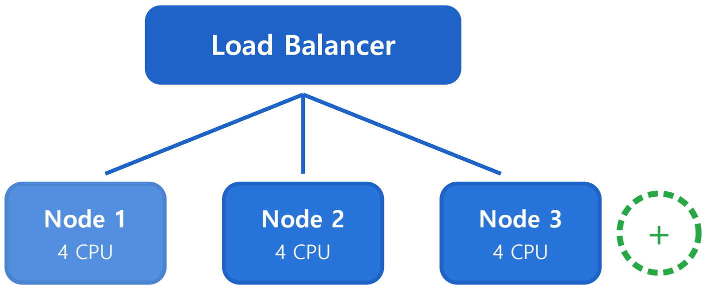
Latency reduction: **200ms → 20ms**

Cost savings: **60% reduction**

DB load: **-80%**

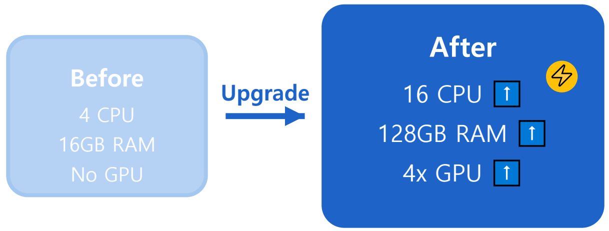
Scalability Patterns

↔ Horizontal Scaling



Capacity: Linear Growth ↑

↕ Vertical Scaling



Capacity: Hardware Limited 🔒

Faster per-node but can't scale infinitely

Sharding Strategies

Hash-based	Range-based	Geo-based
Uniform distribution	Date/category sharding	Regional data locality

High Availability

- ✓ 3x replication factor
- ✓ Auto-failover in <5s

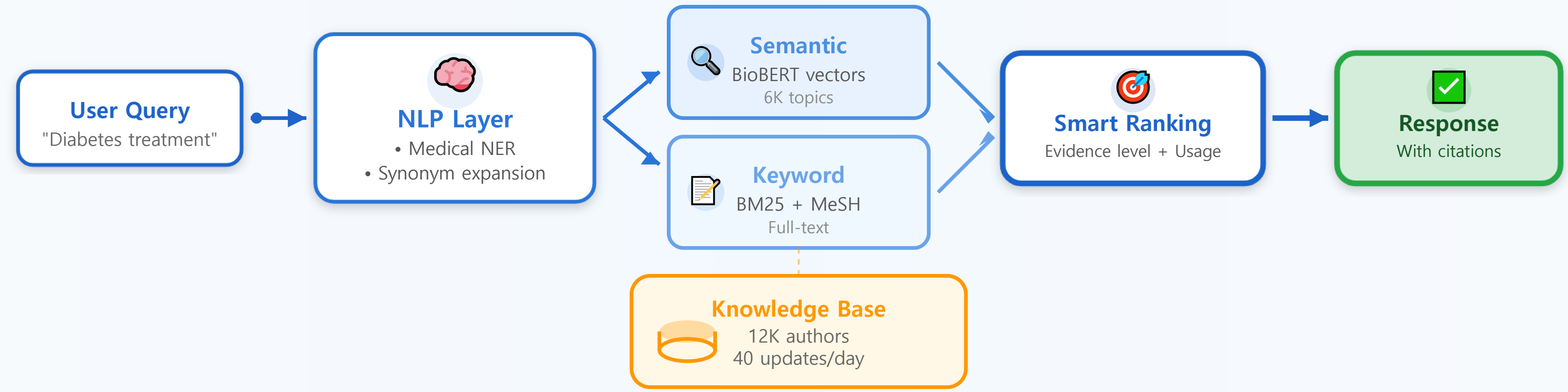
✓ 99.99% uptime SLA

Case Study: UpToDate Integration

World's Leading Clinical Decision Support System

6,000+	12,000+	130+	40+
Clinical Topics	Expert Authors	Countries	Updates/day

📚 UpToDate RAG Architecture



📊 Clinical Impact (Evidence-Based Outcomes)

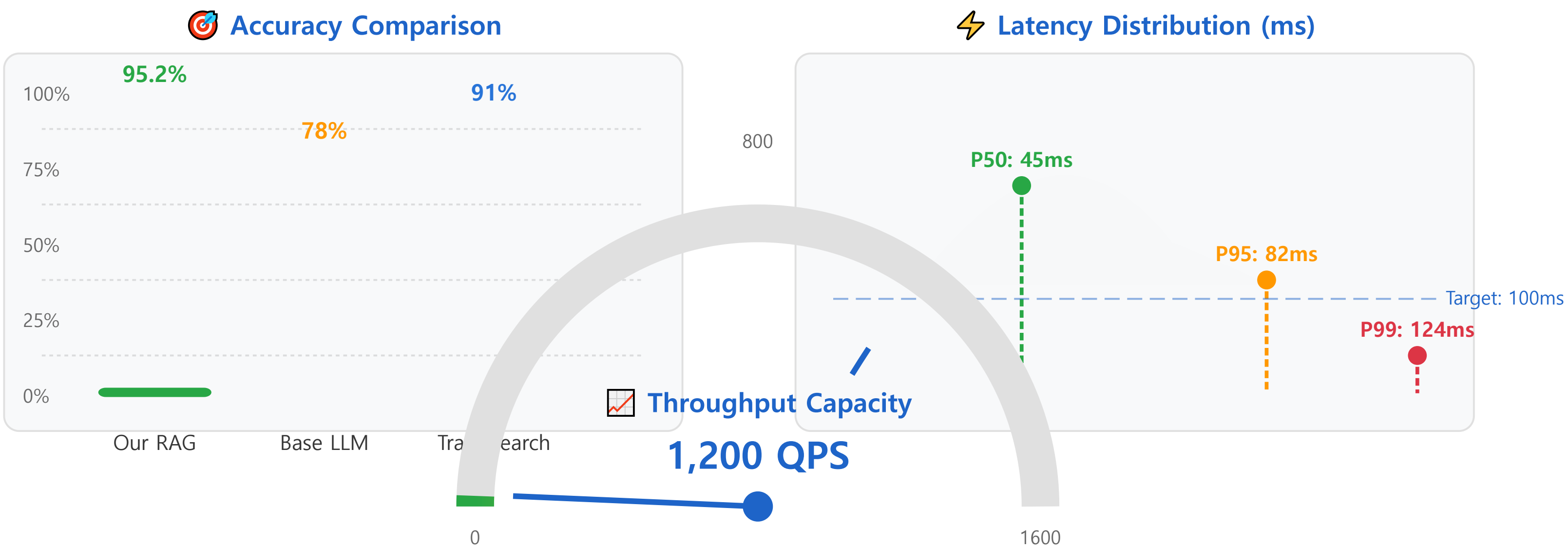
↓ 6% mortality reduction
(NEJM 2012)

↓ 19% shorter hospital stays
(Systematic review)

92% change clinical decisions
(User survey)

Performance Benchmarks

📊 RAG System Performance Metrics



95.2% Retrieval Accuracy	**82ms** P95 Latency	**98.5%** Citation Accuracy	**2%** Hallucination Rate

Hallucination Mitigation

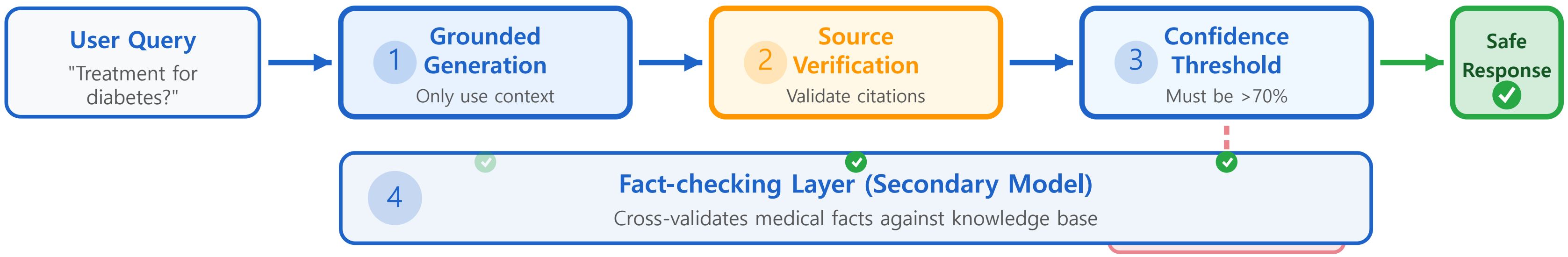
⚠ Types of Hallucinations

Factual Errors
Wrong dosage, incorrect diagnosis

Fabricated Citations
Made-up study references

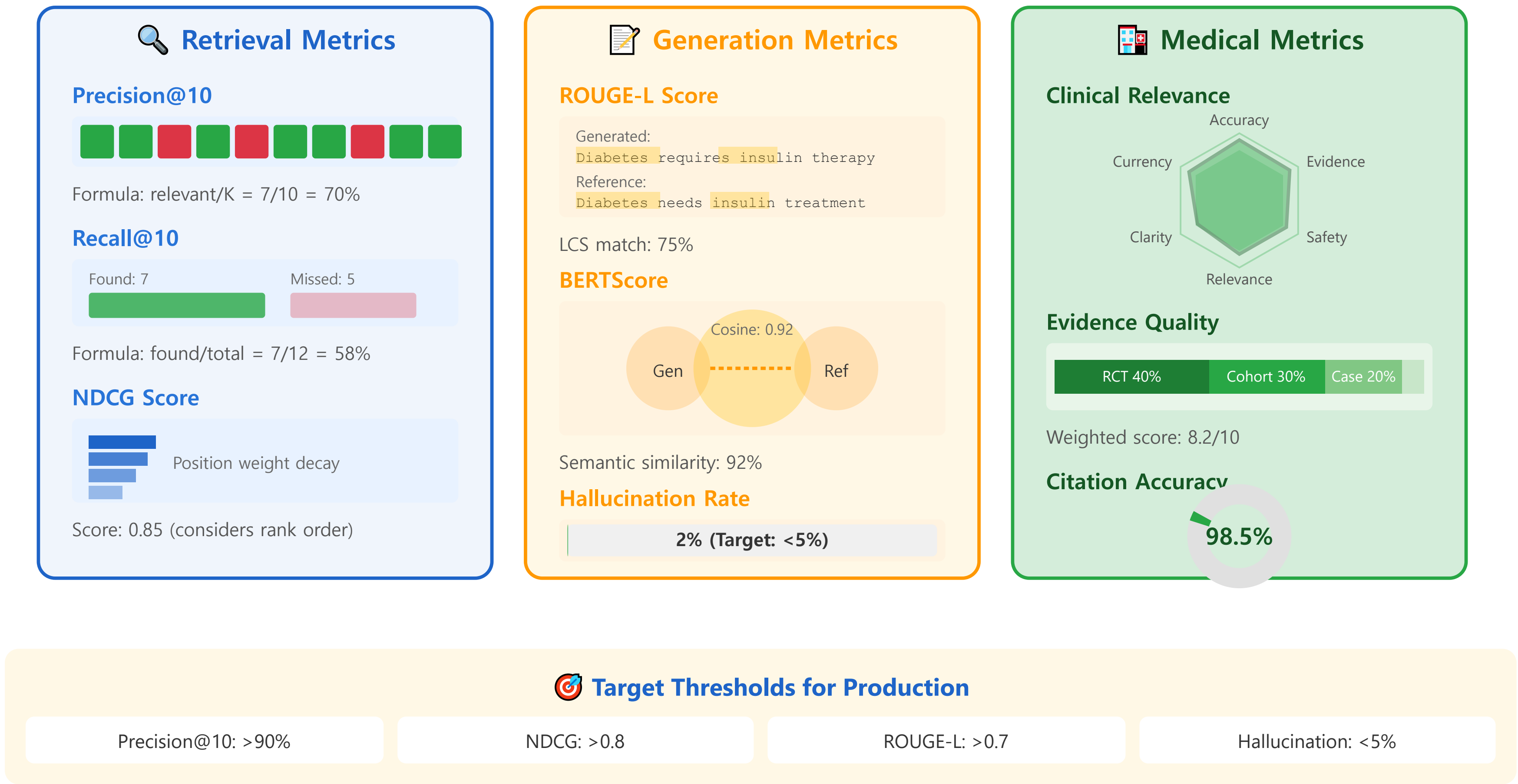
Outdated Info
Using obsolete guidelines

🛡 Mitigation Pipeline



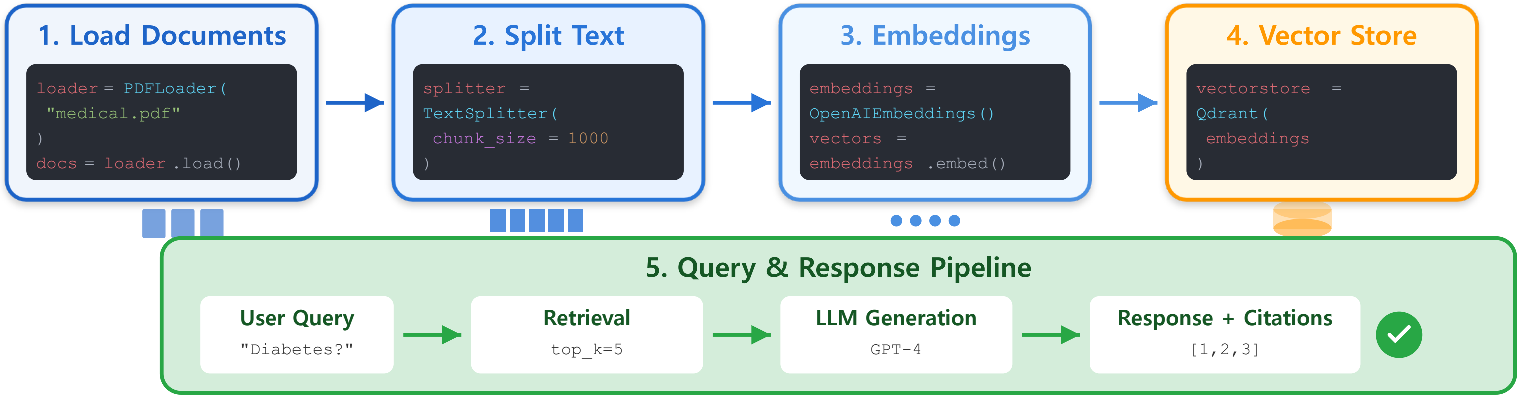
📊 **Results: Hallucination rate reduced from 12% → 2% | Citation accuracy: 98.5%**

Evaluation Metrics



Hands-on: RAG Pipeline

🔗 LangChain RAG Implementation Flow



💻 Complete Implementation Code

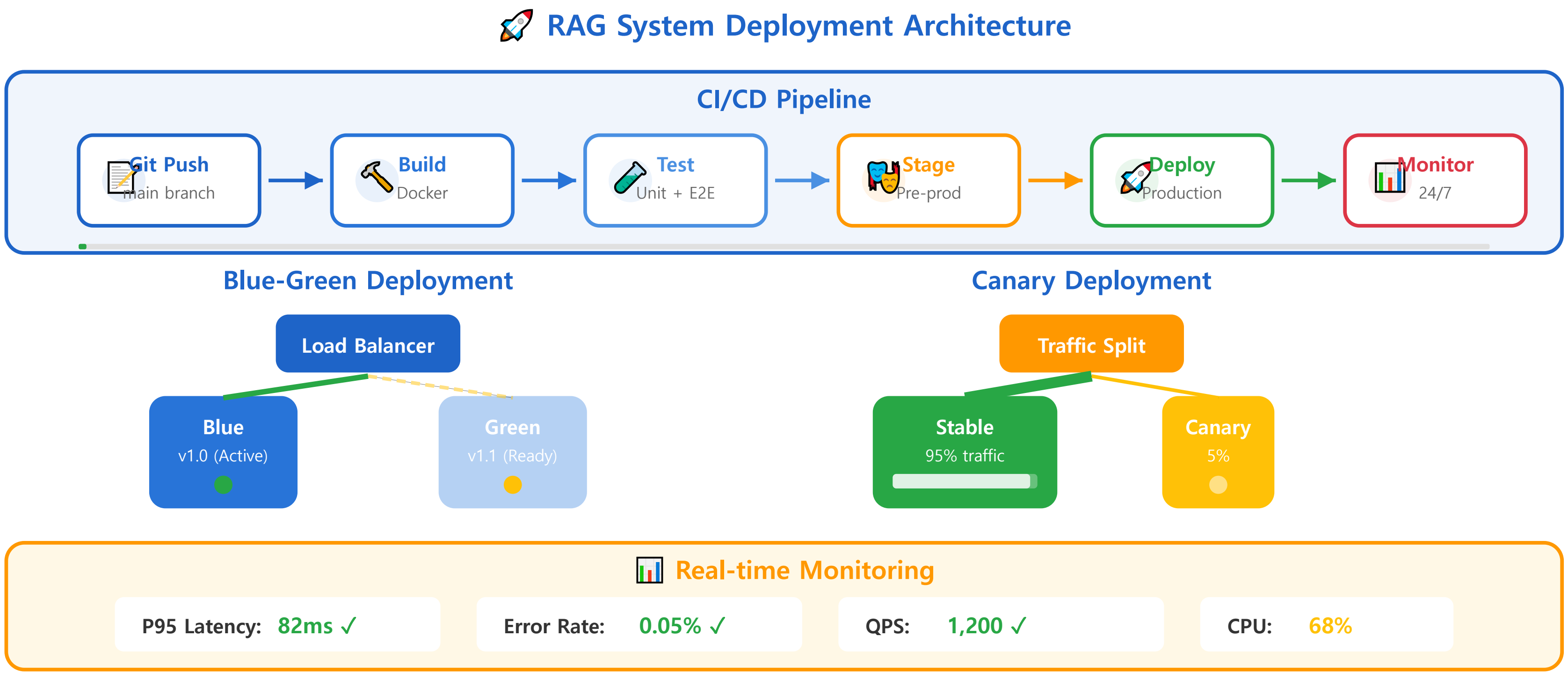
```
from langchain.vectorstores import Qdrant
from langchain.embeddings import OpenAIEmbeddings
from langchain.chains import RetrievalQA

# Setup Vector Store
vectorstore = Qdrant(embeddings=OpenAIEmbeddings(), collection="medical_kb")

# Create RAG Chain
qa_chain = RetrievalQA.from_chain_type(
    llm=OpenAI(temperature=0),
    retriever=vectorstore.as_retriever(search_kwargs={"k": 5}),
    return_source_documents=True
)

# Query with Citations
result = qa_chain({"query": "Treatment for Type 2 Diabetes?"})
print(result['answer'], result['source_documents'])
```

Deployment Strategies



82ms	**0.05%**	**1,200**	**68%**
P95 Latency	Error Rate	QPS	CPU Usage

Thank You

🎓 Key Takeaways

✅

✅

✅

✅

💡

📚

Vector DBs: Pinecone, Weaviate, Qdrant

Research: arXiv.org (cs.CL, cs.IR)

Questions & Answers