RAG Workshop

John Willis

Agenda

- Introduction
- A Brief History
- RAG
- Functions
  (if time permits)

Introduction

@botchagalupe

CANONICAL

DELL

docker

CHEF

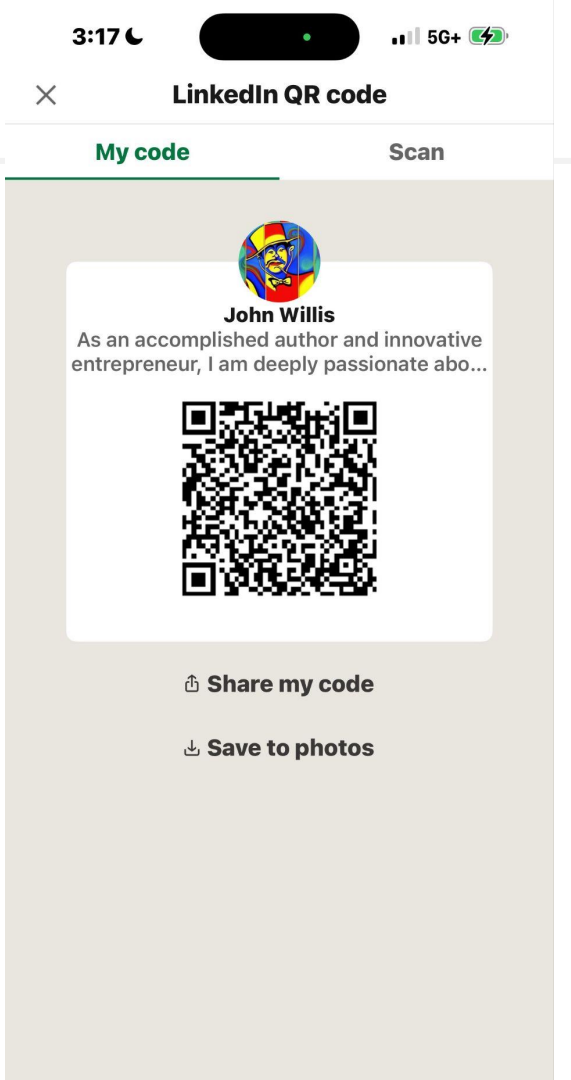
DEVOPS DAYS

KeyBanc Capital Markets

Red Hat

DEVOPS ENTERPRISE SUMMIT

mongoDB.

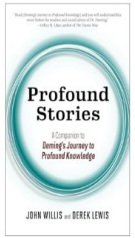


John Willis

+ Follow

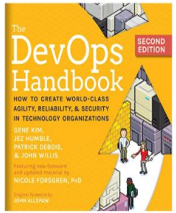
HOME ABOUT ALL BOOKS

Top John Willis titles



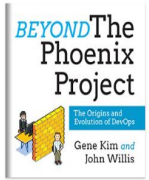
Profound Stories: A Companion to Deming's Journe...

★★★★½ 2



The DevOps Handbook, Second Edition: How to C...

★★★★½ 2,886



Beyond the Phoenix Project: The Origins and...

★★★★½ 56

Investments Unlimited: A Novel About DevOps, S...

★★★★½ 139

Deming's Journey to Profound Knowledge: How...

★★★★½ 60

https://www.amazon.com/author/johnwillis

My Books

https://aicio.ai

Calendly - John W... | Maps | Bookmarks | genAI for DevOps... | agents-workshop-... | https://www.amaz... | https://www.opus.... | John Willis - REBE...

Feb 08, 2025

Inference vs. Training: A CIO's Guide

Why Understanding the AI Lifecycle is Crucial for Enterprise Leaders

John Willis

Feb 01, 2025

Maximize AI Investments with Benchmarking

How MLPerf and MLCommons Help CIOs Optimize GenAI Performance and Infrastructure

John Willis

Jan 25, 2025

AI's Double-Edged Sword

Innovations, Risks, and Resolutions

John Willis

Jan 11, 2025

GenAI Revolutionizes Cybersecurity

Empowering Everyone to Hunt Vulnerabilities

Jan 04, 2025

Dear CIO: Welcome to 2025

2024 Recap and What to Prioritize in 2025

John Willis

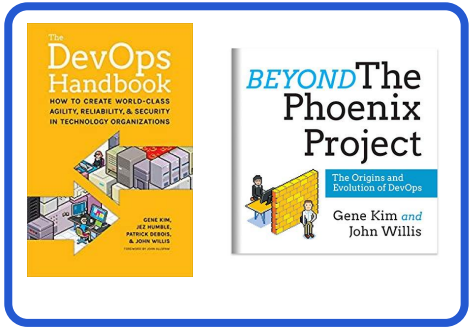
Dec 21, 2024

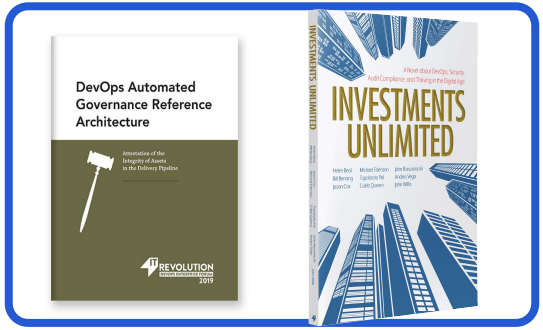
CIO's Must-Have AI Tool for 2025

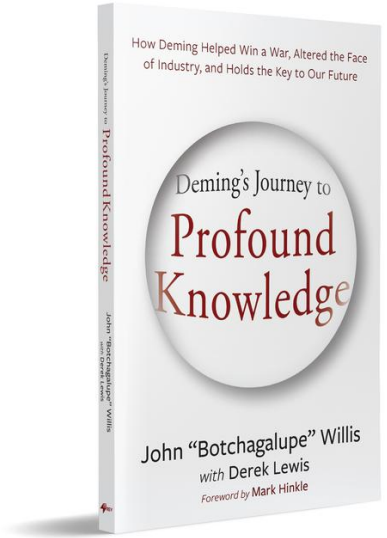
Your Blueprint for Managing AI Vulnerabilities in the Coming Year

AICIO.AI

@botchagalupe







CANONICAL

CHEF

DELL

OpenContext

Techstrong.ai

DEVOPS DAYS

DEVOPS ENTERPRISE SUMMIT

KeyBanc Capital Markets

Red Hat

docker

mongoDB

John Willis @botchagalupe


THE HEROES OF
CHATGPT AND MODERN AI
Rebels
of
Reason
JOHN WILLIS


ALL THINGS OPEN.AI
AN AI PRACTITIONERS AND USERS CONFERENCE FOCUSED ON TECHNOLOGIES, PROCESSES, AND PEOPLE
MARCH 17 & 18, 2025

Brief History Lesson

AI Books


22 Books
New
The Brain Makers
HP Newquist
Read
THE MYTH OF ARTIFICIAL INTELLIGENCE
Why Computers Can't Think the Way We Do
ERIK J. LARSON
From MEDIEVAL ROBOTS to NEURAL NET 50%
ARTIFICIAL INTELLIGENCE
AN ILLUSTRATED HISTORY
CLIFFORD A. PICKOVER
Rental
Artificial Intelligence Basics
A Non-Technical Introduction
Tom Taulli
Apress
27%
DARK HERO OF THE INFORMATION AGE
IN SEARCH OF NORBERT WIENER THE FATHER OF CYBERNETICS
FLO CONWAY & JIM SIEGELM
1%
the Universal Computer
Third Edition
The Road from Leibniz to Turing
Martin Davis
'A wild, brilliant ride' CIARAN MARTIN
How AI Ate the World
A Brief History of Artificial Intelligence - and Its Long Future
Chris Stokel-Walker
By the author of YouTubers and TikTok Boom
New
TECHNOLOGY, POWER, AND THE 21ST CENTURY'S GREATEST DILEM
THE COMING WAVE
MUSTAFA SULEYMAN
Read
Artificial Intelligence
A Guide for Thinking Humans
Melanie Mitchell
Read
CO-INTELLIGENCE
Living and Working with AI
ETHAN MOLL
Read
MACHINES OF LOVING GRACE
THE QUEST FOR COMMON GROUND
BETWEEN HUMANS AND ROBOTS
JOHN MARKOF
artificial intelligence meets human intelligence
THE DEEP LEARNING REVOLUTION
TERRENCE J. SEJNOWSKI
33%
EVOLUTION, AI, AND FIVE BREAKTHROUGHS THAT MADE OUR BRAINS
A BRIEF HISTORY OF INTELLIGENCE
MAX BENNETT
Read
A Brief History of Artificial Intelligence
What It Is, Where We Are, and Where We Are Going
Michael Wooldridge
1%
A BRIEF HISTORY OF ARTIFICIAL INTELLIGENCE
Unraveling the Threads of Machine Intelligence
SAMUEL S. WADE
A brief history of
46%
Artificial Intelligence
kindleunlimited
UNMASKING AI
MY MISSION TO PROTECT WHAT IS HUMAN IN A WORLD OF MACHINES
JOY BUOLAMWINI
50%
THE WORLDS I SEE
CURIOSITY, EXPLORATION, AND DISCOVERY AT THE DAWN OF AI
DR. FEI-FEI LI
Read
Yuval Noah Harari
Nexus
A Brief History of Information Networks from the Stone Age to AI
-NOAM CHOMSKY
THE BIRTH OF THE MIND
How a Tiny Number of Genes Creates the Complexities of Human Thought
GARY MARCUS
THE WALTZ OF REASON
THE ENTANGLEMENT OF MATHEMATICS AND PHILOSOPHY
KARL SIGMUND
NEW YORK TIMES BESTSELLER
1%
THE MAN WHO SOLVED THE MARKET
HOW JIM SIMONS LAUNCHED THE QUANT REVOLUTION
Gregory Zuckerman

Where it All Started… kind of

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

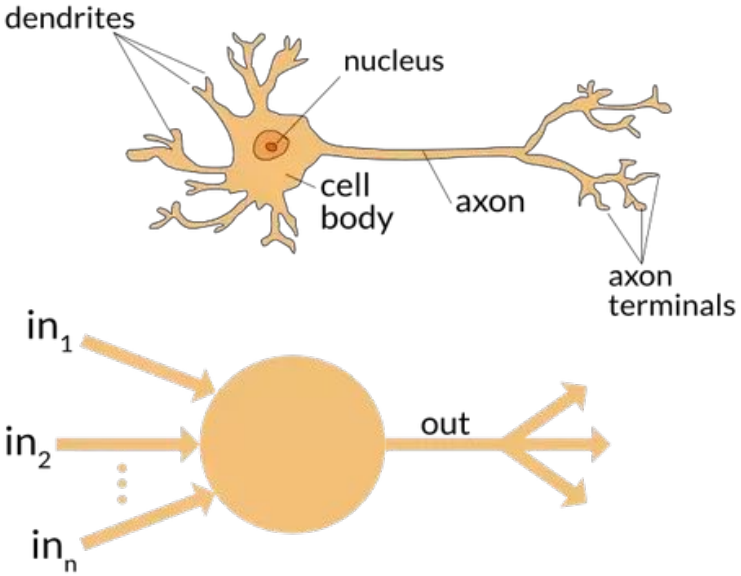
A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes.

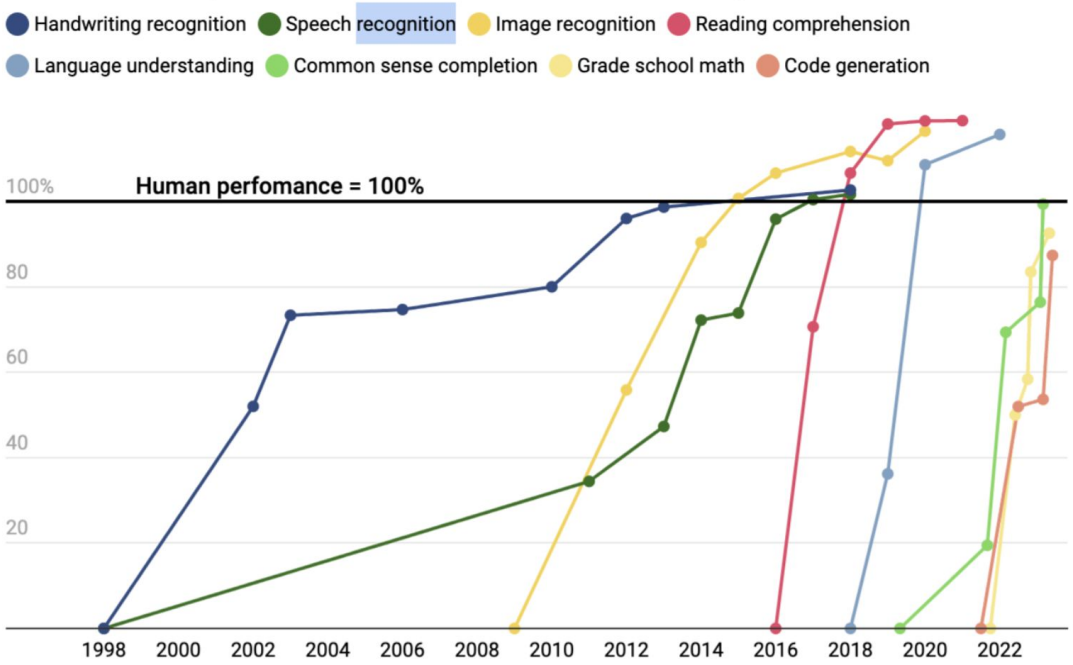
History of Neural Networks



In 1943 Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician, teamed up to develop a mathematical model of an artificial neuron. In their paper "A Logical Calculus of the Ideas Immanent in Nervous Activity" they declared that:

AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

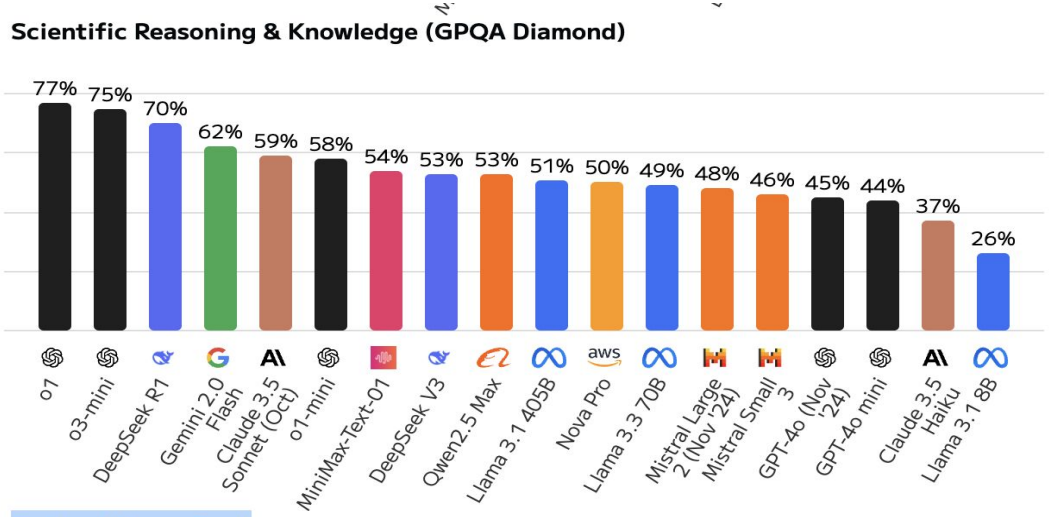
Chart: Will Henshall for TIME • Source: ContextualAI

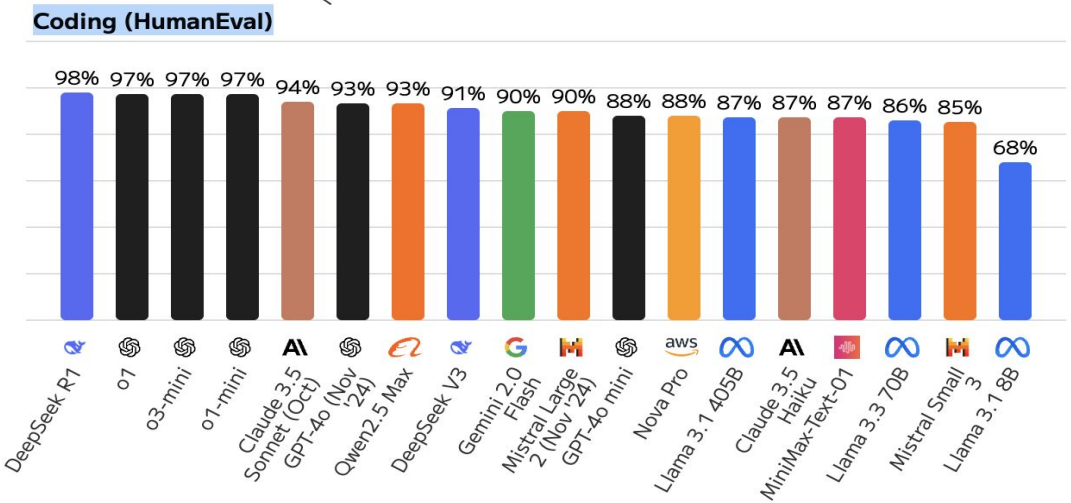
TIME

1943 - First Neural Network
1995 - TD-Gammon
1997 - Deep Blue
1998 - MNIST and Dragon Systems
2012 - Alexnet
2016 - DeepMind and Alphago
2017 - Attention is All You Need
2018 - Bert and GPT-1
2022 - ChatGPT
2023 - Microsoft CoPilot

2025 - Lessons Learned

- Model Evolution
  - Reasoning Models (o1, o3, r1, v3, Qwen 2.5)
  - Quantization (not everything has to be a H100)
  - MOE - Distillation, Efficiency and Optimization
  - Reinforcement Learning (Arthur Samuel)
- Benchmark Evolution
  - Out of Distribution Behavior (Emergent Behavior)
    - GPQA: A Graduate-Level Google-Proof Q&A Benchmark
    - ARC-AGI

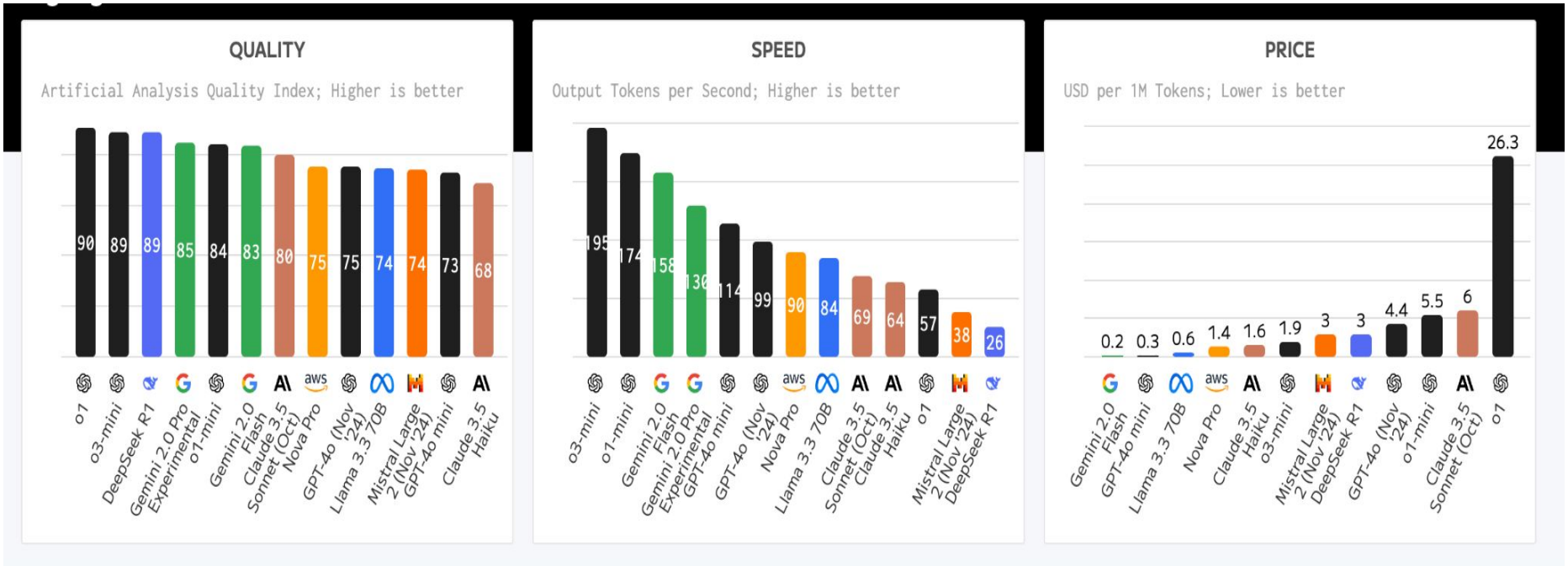




https://artificialanalysis.ai/models

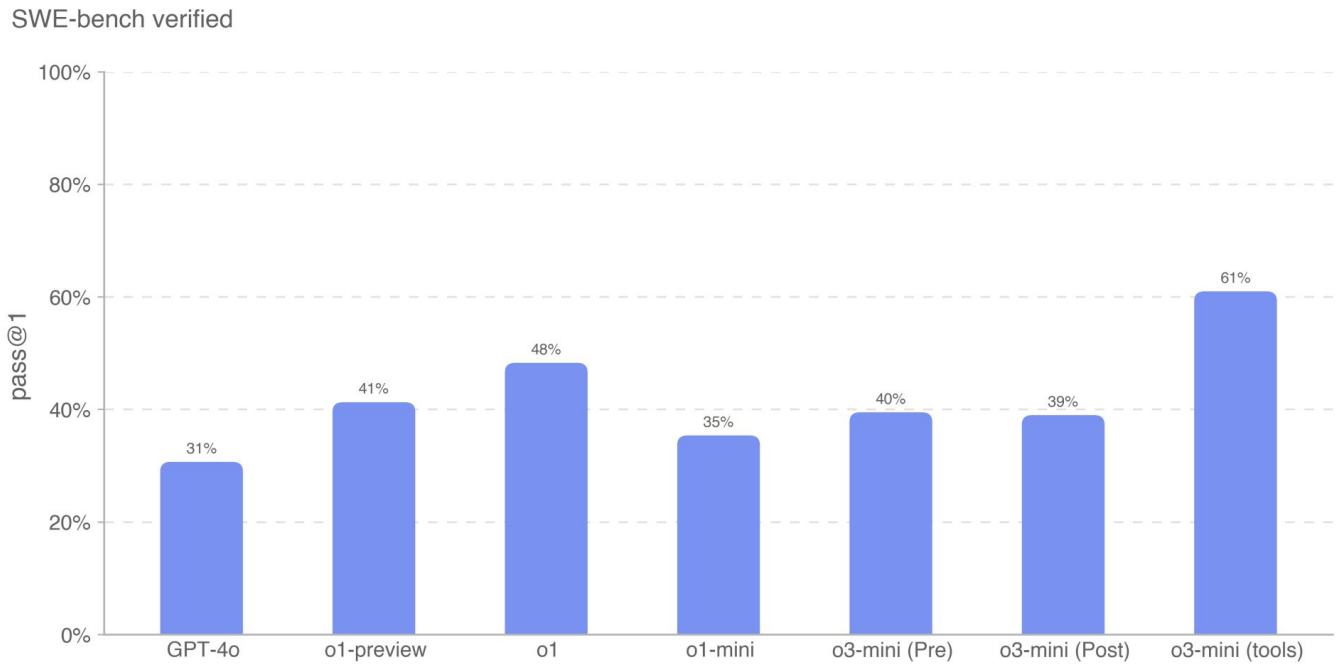
Average "State of the Art" in June 2024 was around 60% (Reasoning) and 70% (Coding)

2024 June - Claude 3.5 Sonnet
2024 July - Llama 3.1
2024 Sept - O1
2024 Dec - O3
2025 Jan - R1



https://artificialanalysis.ai/models

SWE-Bench



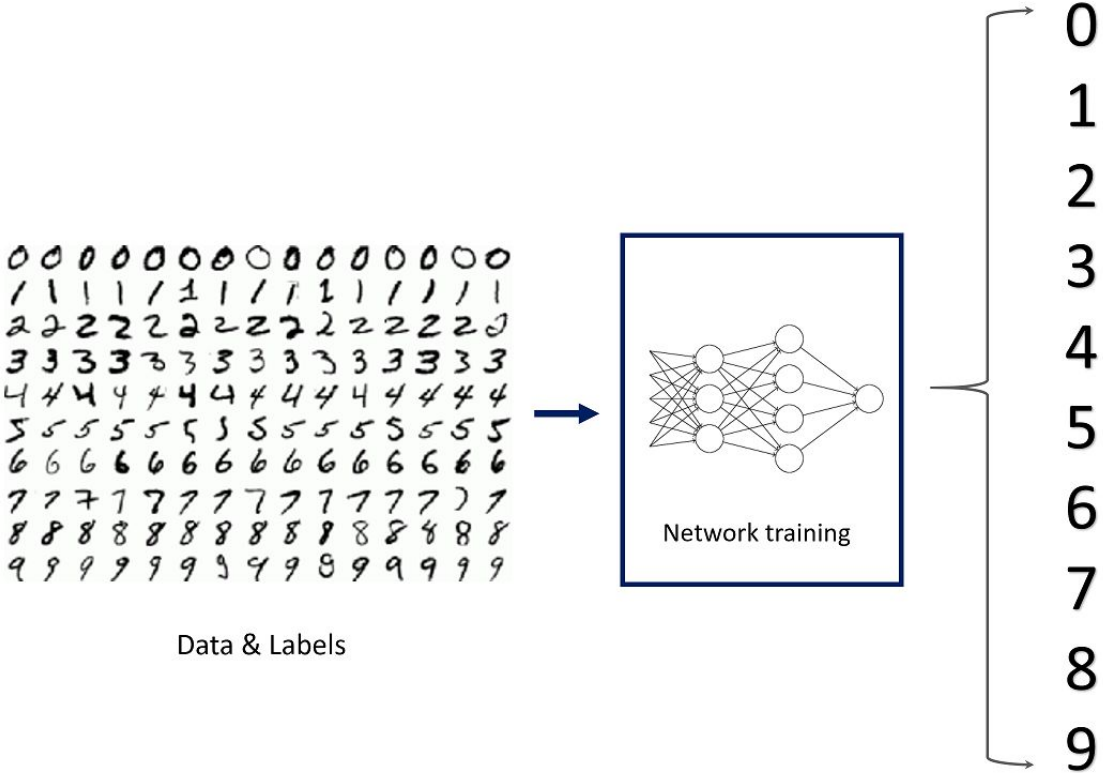
https://cdn.openai.com/o3-mini-system-card.pdf

GenAI Basics

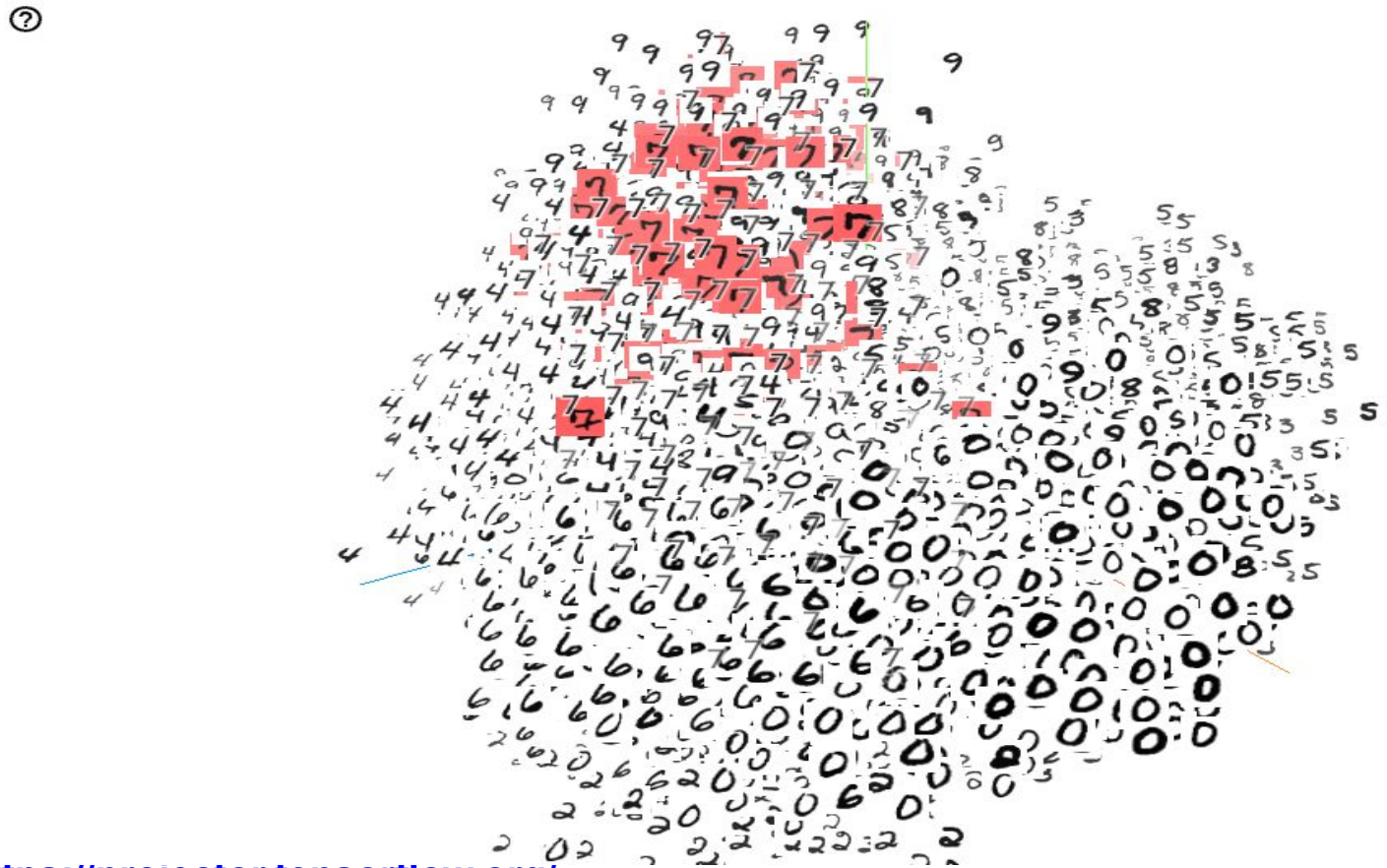
ML versus DL

- **Machine Learning**
  - A subset of AI that involves training algorithms to learn from and make predictions or decisions based on data.
- **Deep Learning**
  - A subset of Machine Learning, uses neural networks with many layers (hence "deep") to analyze vast amounts of data, enabling more complex and accurate predictions.
- **Generative AI (GenAI)**
  - GenAI creates content like text, images, audio, and video based on user prompts. It powers tools like chatbots and virtual assistants by recognizing patterns from large datasets. e.g.,, large language models (LLMs) use these patterns to mimic human writing styles.

MNIST Dataset

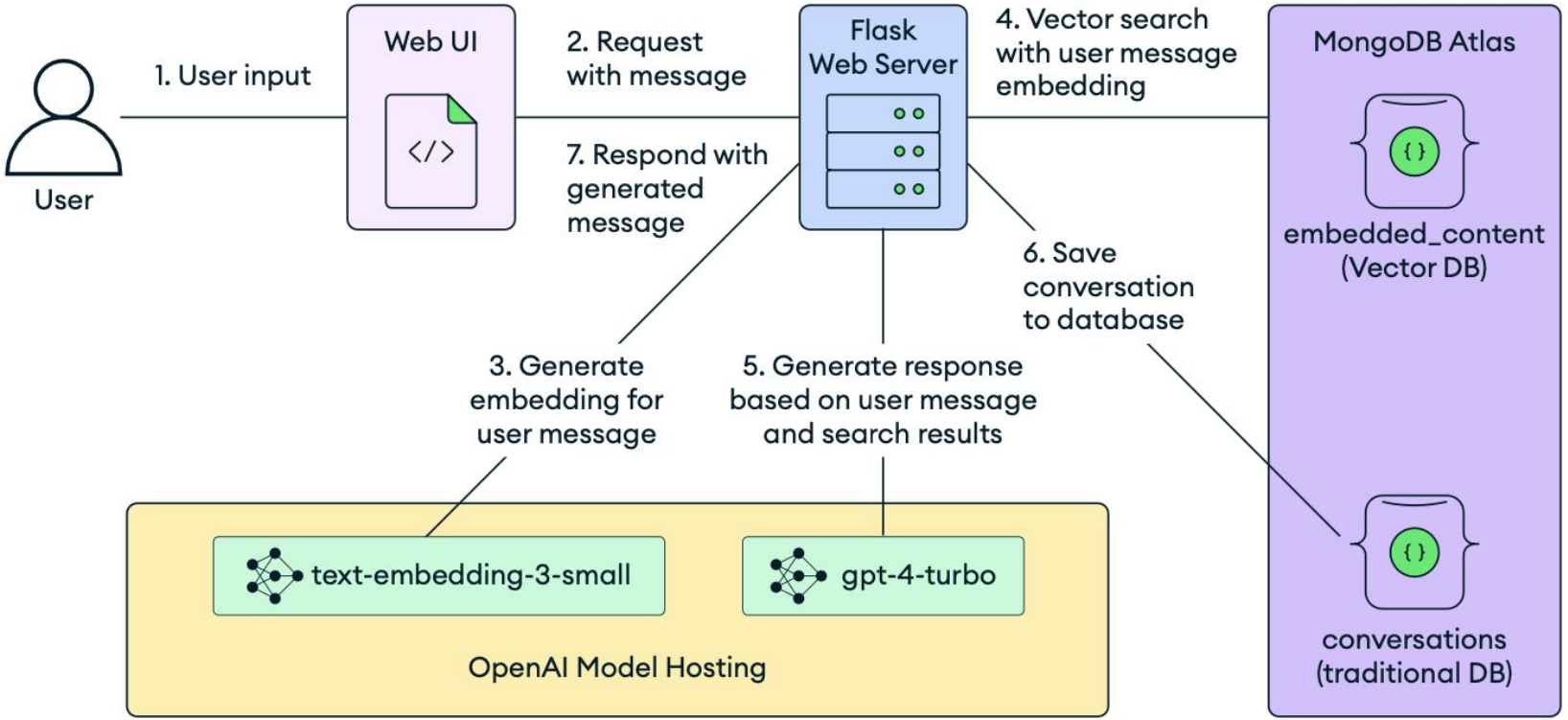


Vector Embeddings (MNIST)



https://projector.tensorflow.org/

Basic Chatbot



Building AI Intensive Python Applications - MongoDB Press

GenAI Basics - LORMA Stack

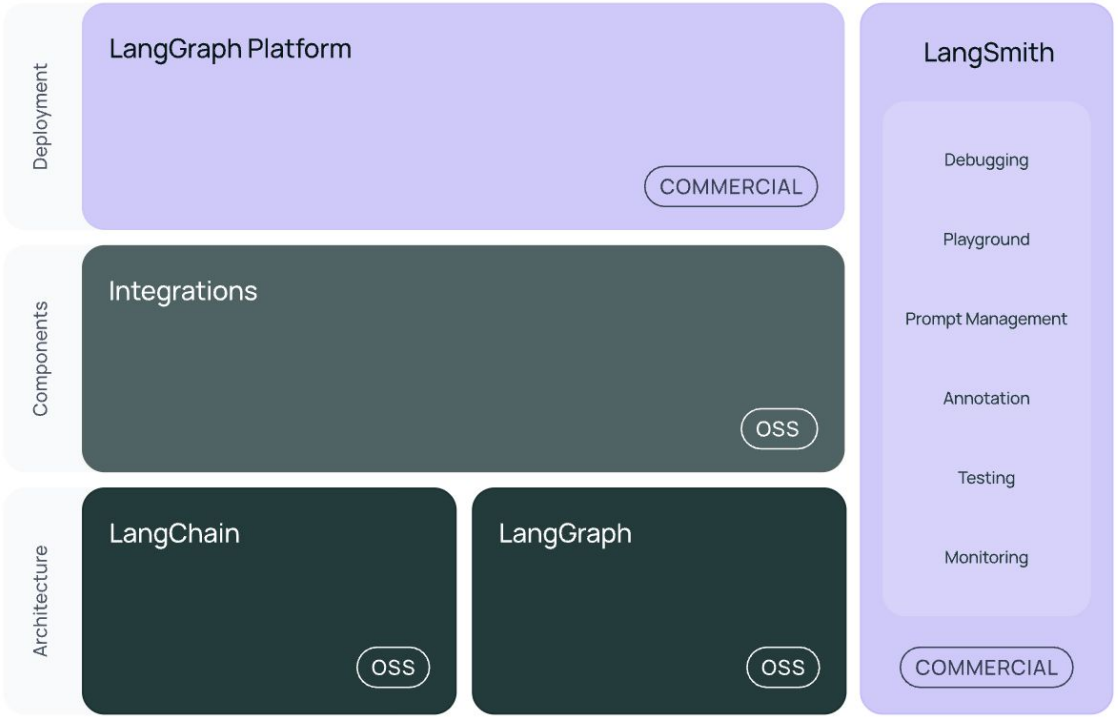
- **L -** Language Model Orchestration (LLM/SLM)
- **O -** Observability
- **R -** Retrieval Augmentation Generation
- **M -** Model Providers
- **A** - Autonomous Agents

L - Language Model Orchestration

GenAI Basics - LLM Orchestration

- Langchain
- LLamaIndex
- Haystack
- DSPy

GenAI Basics - Langchain Architecture



https://python.langchain.com/docs/introduction/

Comparison of Langchain and LlamaIndex

- **LangChain**
  - Flexible, modular framework suitable for a broad range of LLM-powered applications.
  - Excels in user interaction stages.
  - Preferred for complex, adaptable applications requiring detailed query processing.

- **LlamaIndex**
  - Optimized for the initial stages of data handling and indexing.
  - Focuses on data indexing and enhancing LLM performance.
  - Efficient data retrieval and query capabilities.
  - Better for search and retrieval tasks

O - Observability

GenAI Basics - Observability

- Arize
- LangSmith
- Galileo
- TrueEra
- WhyLabs

GenAI Basics - Observability

- **Key Requirements for AI System Design:**

- **Security**: Prevent exposure of private or confidential information from model weights or application.
- **Reputation**: Ensure output does not damage the business, e.g., no promotion of competitors.
- **Correctness**: Avoid mistakes or hallucinations; ensure accurate responses.
- **Style**: Adhere to tone/style guidelines, such as formal tone for legal applications.
- **Consistency**: Provide predictable, similar outputs for identical inputs.
- **Ethics**: Uphold ethical standards, avoiding biased or harmful content.

This simplified list covers essential considerations when evaluating and designing AI systems.

GenAI Basics - Observability

John Willis

As an accomplished author and innovative entrepreneur, I am deeply passionate about exploring and advancing the synergy between Generative AI technologies and the transformative principles of Dr. Edwards Deming.

View full profile

John Willis (They/Them) • You
As an accomplished author and innovative entrepreneur, I am deeply pas...
now •

Attention Is All You Need

Postmodern Observability

John Willis on LinkedIn • 4 min read

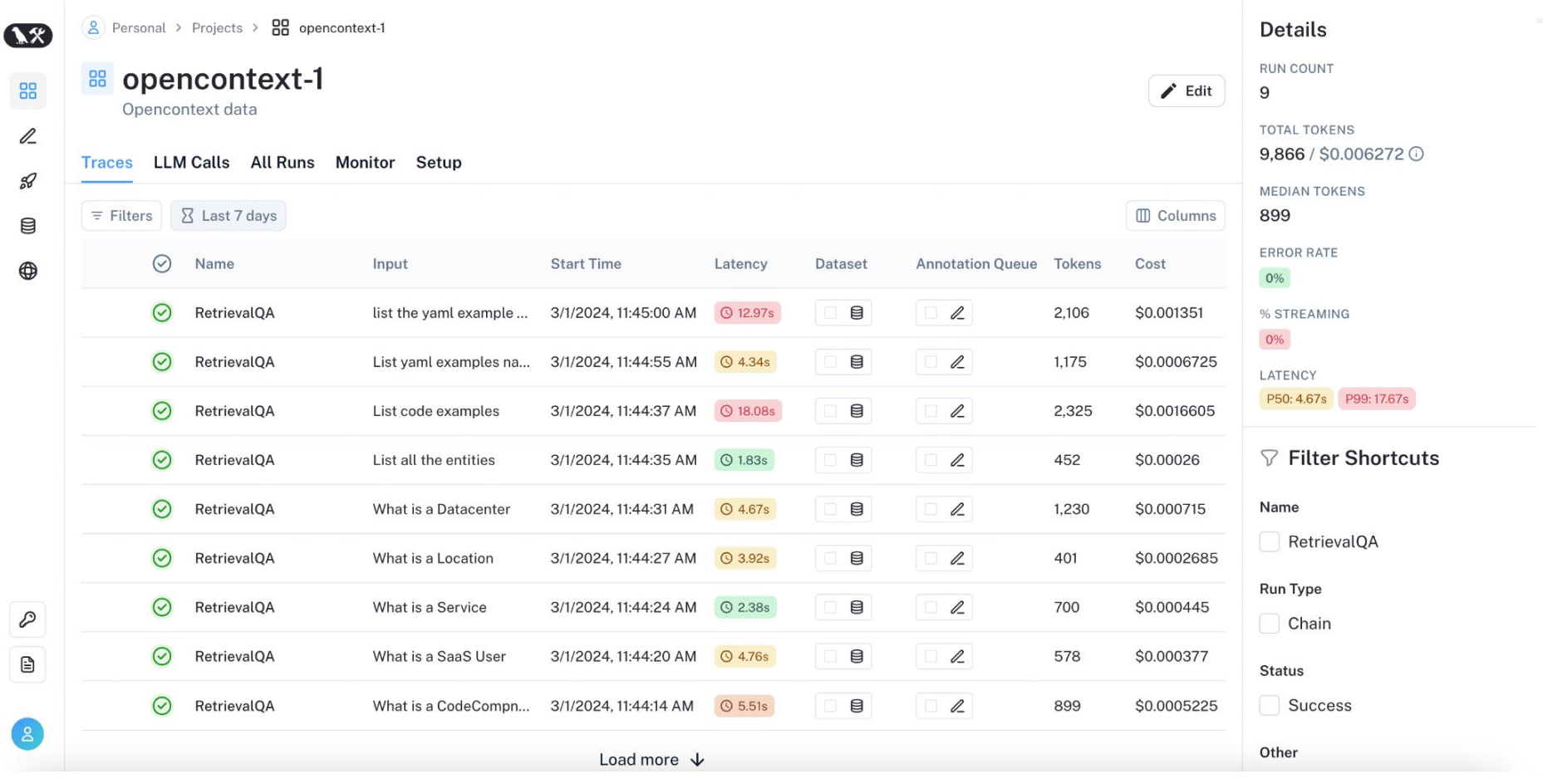
Postmodern Observability in Technology

- Evolution of Observability:
  - Originated in control theory during the 1960s.
- Modern Observability Components:
  - Became integral to system monitoring and reliability around 2010.
  - Comprised of logs, metrics, and traces.
- Importance in Current Technologies:
  - Essential for cloud computing, microservices, and distributed systems.
  - Supports DevOps and Site Reliability Engineering (SRE) practices.
- Expansion into New Areas:
  - Recently linked to DataOps and AI, especially with Generative
  - AI Observability tools now evaluate LLM and model performance
-

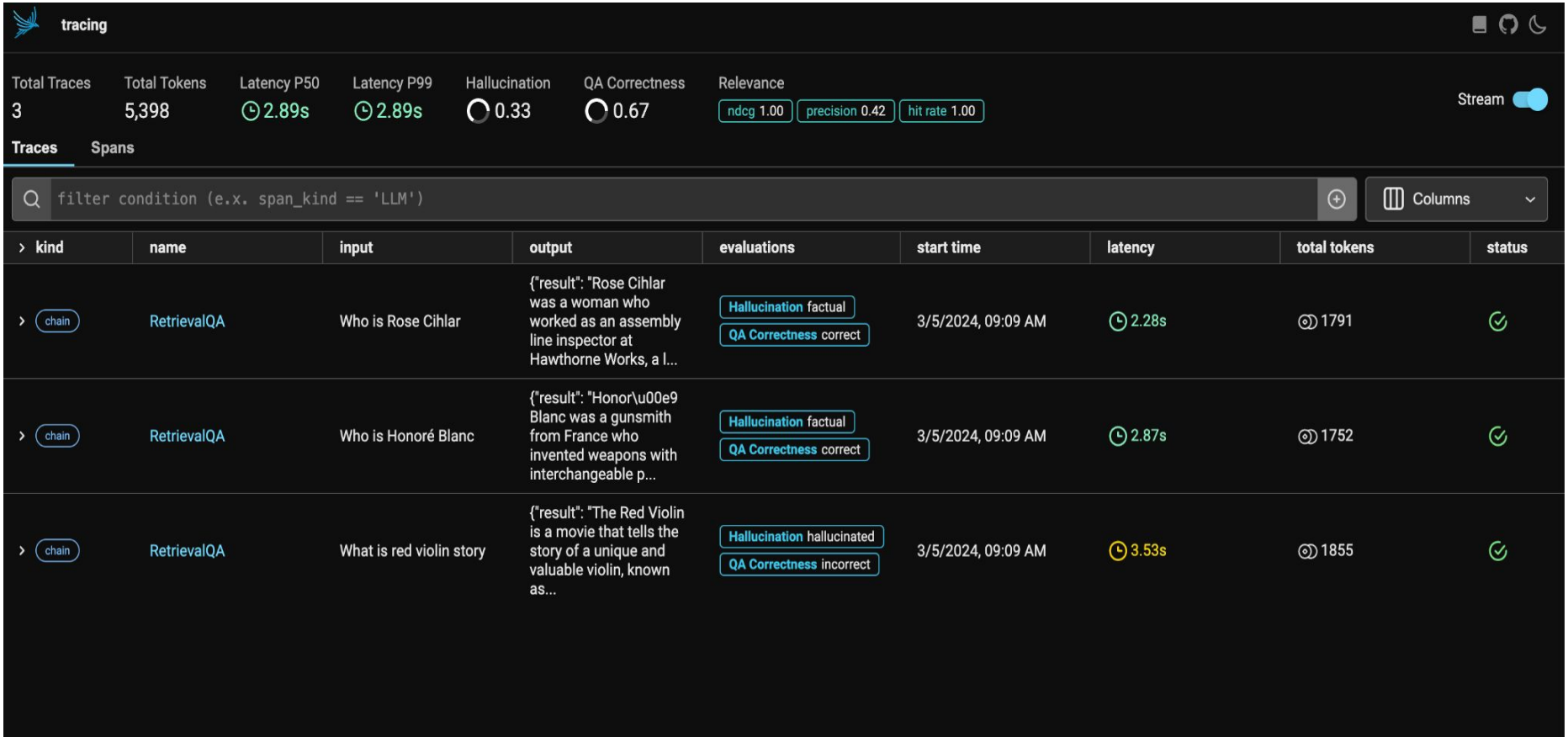
Evaluation (Observe)

- Hallucinations
- QA Correctness
- Relevance
- Bias
- Toxicity
- Drift

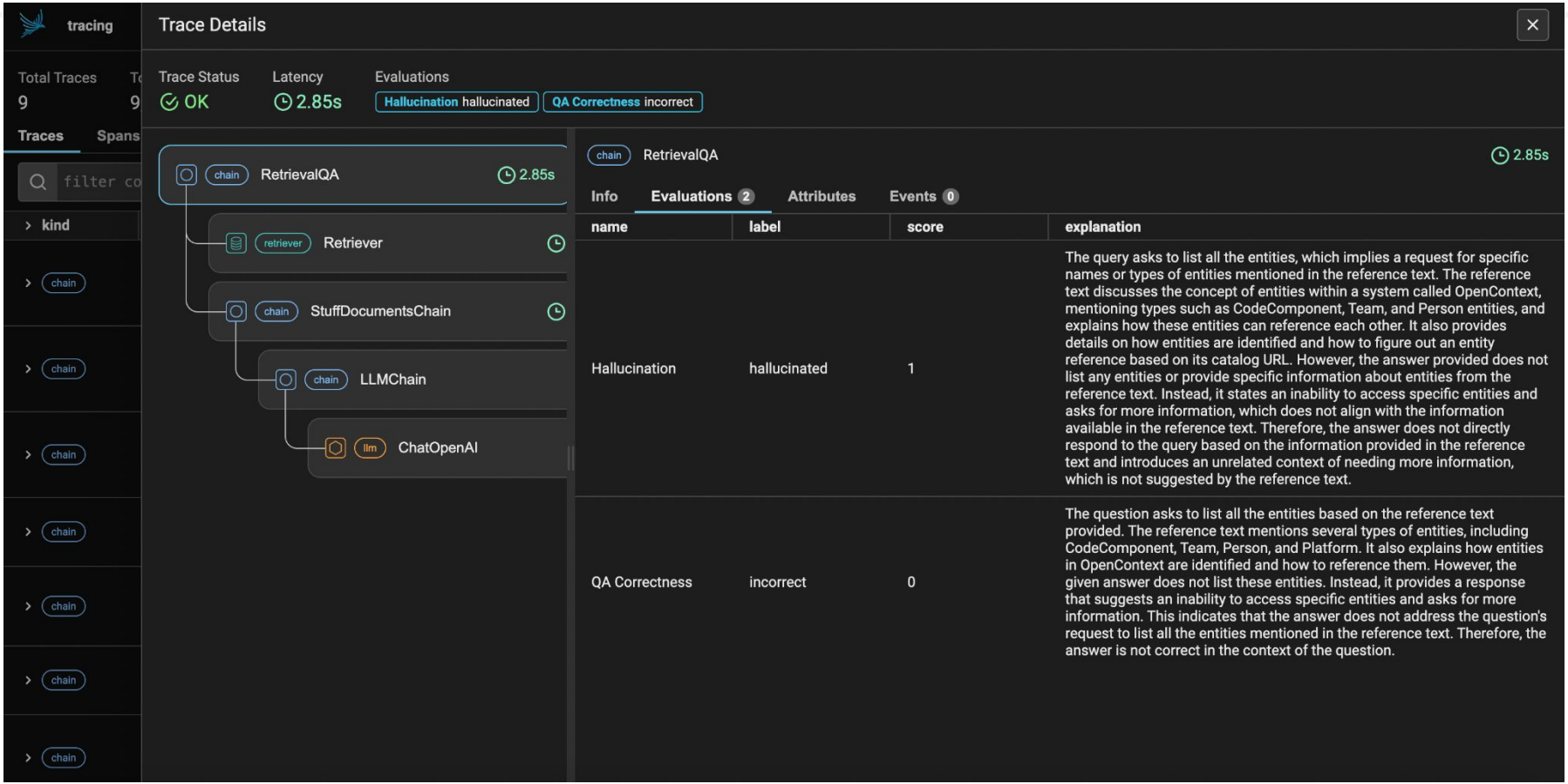
Observability - Langsmith Tracing



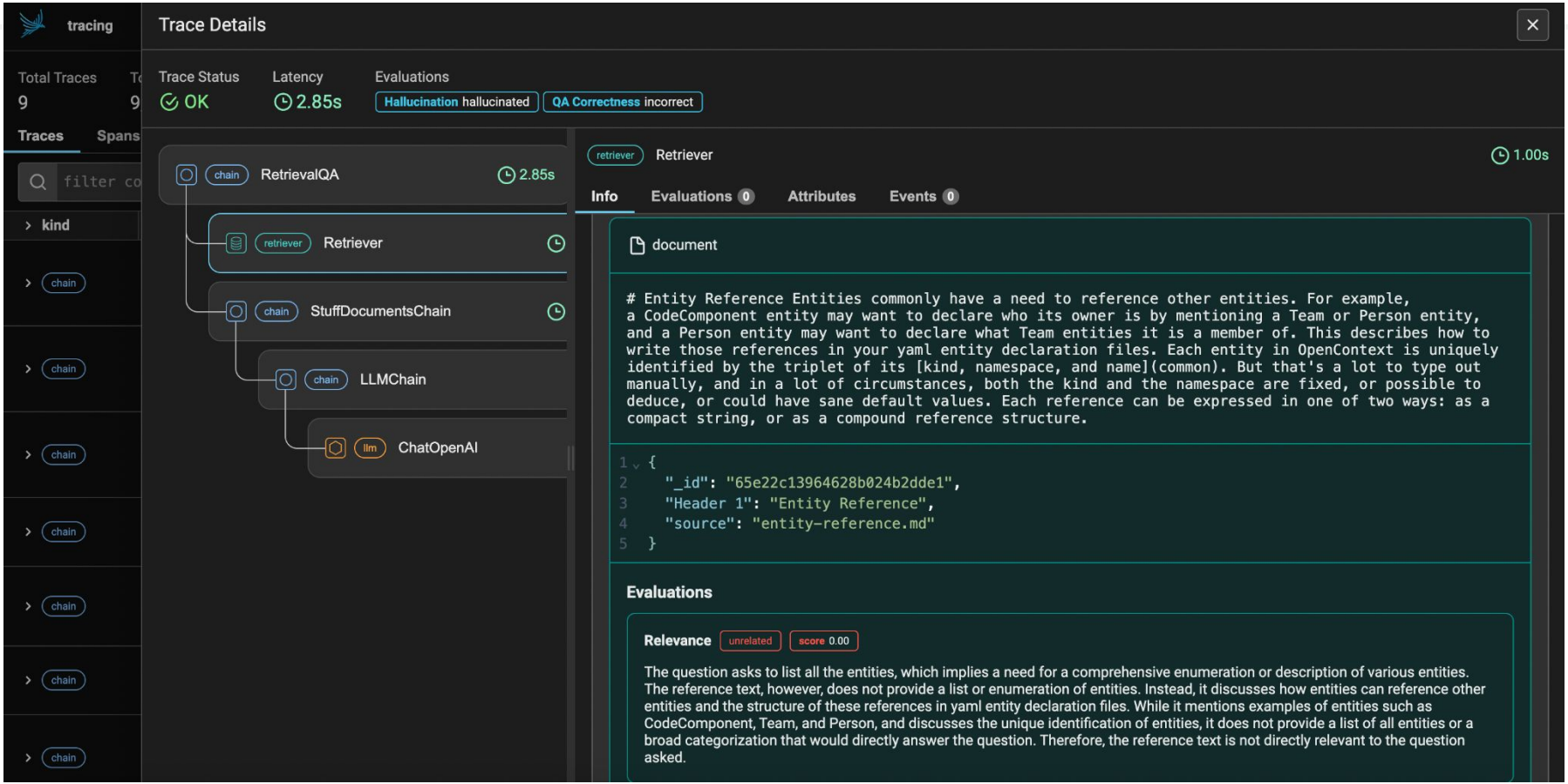
Observability - Profound Book Analysis



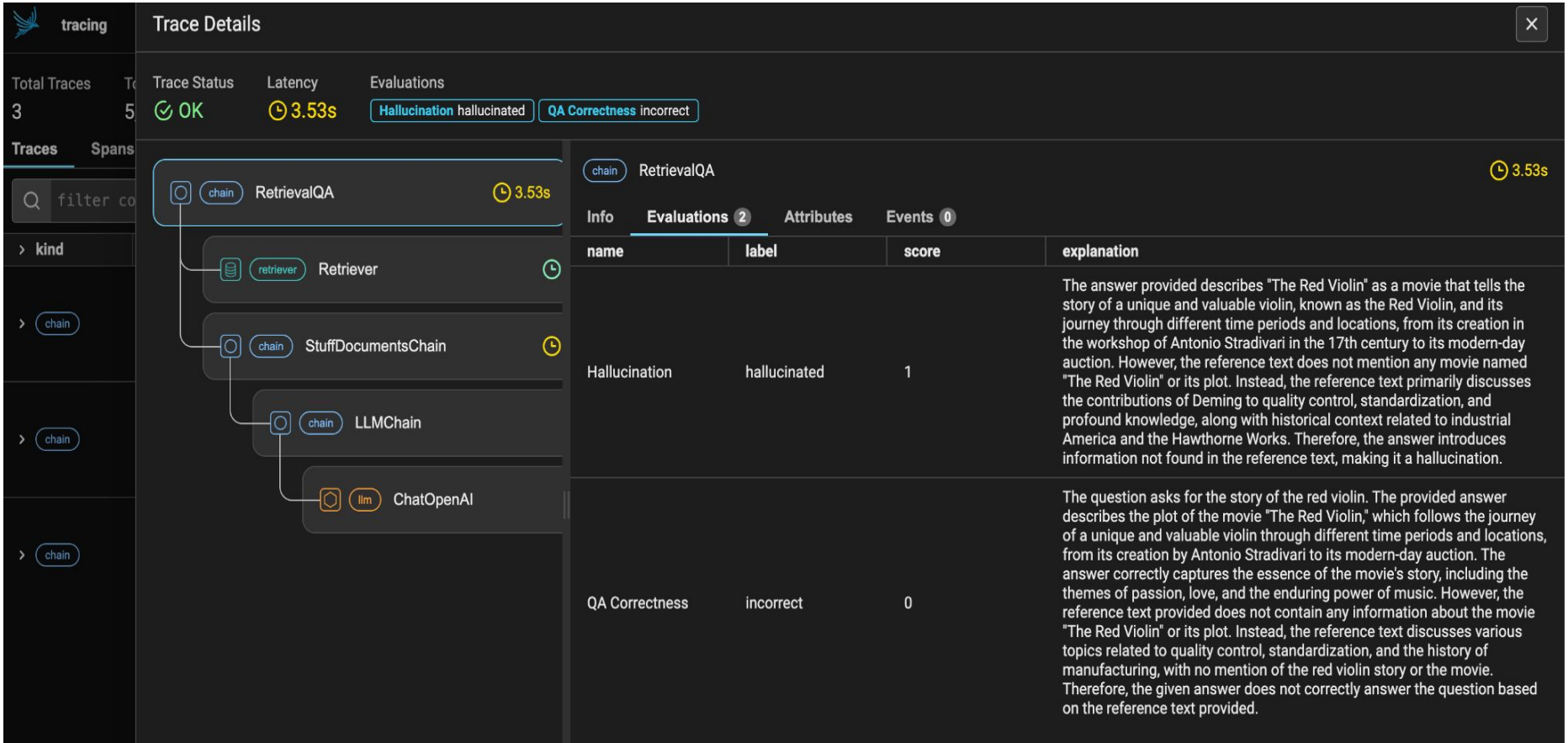
Evaluations



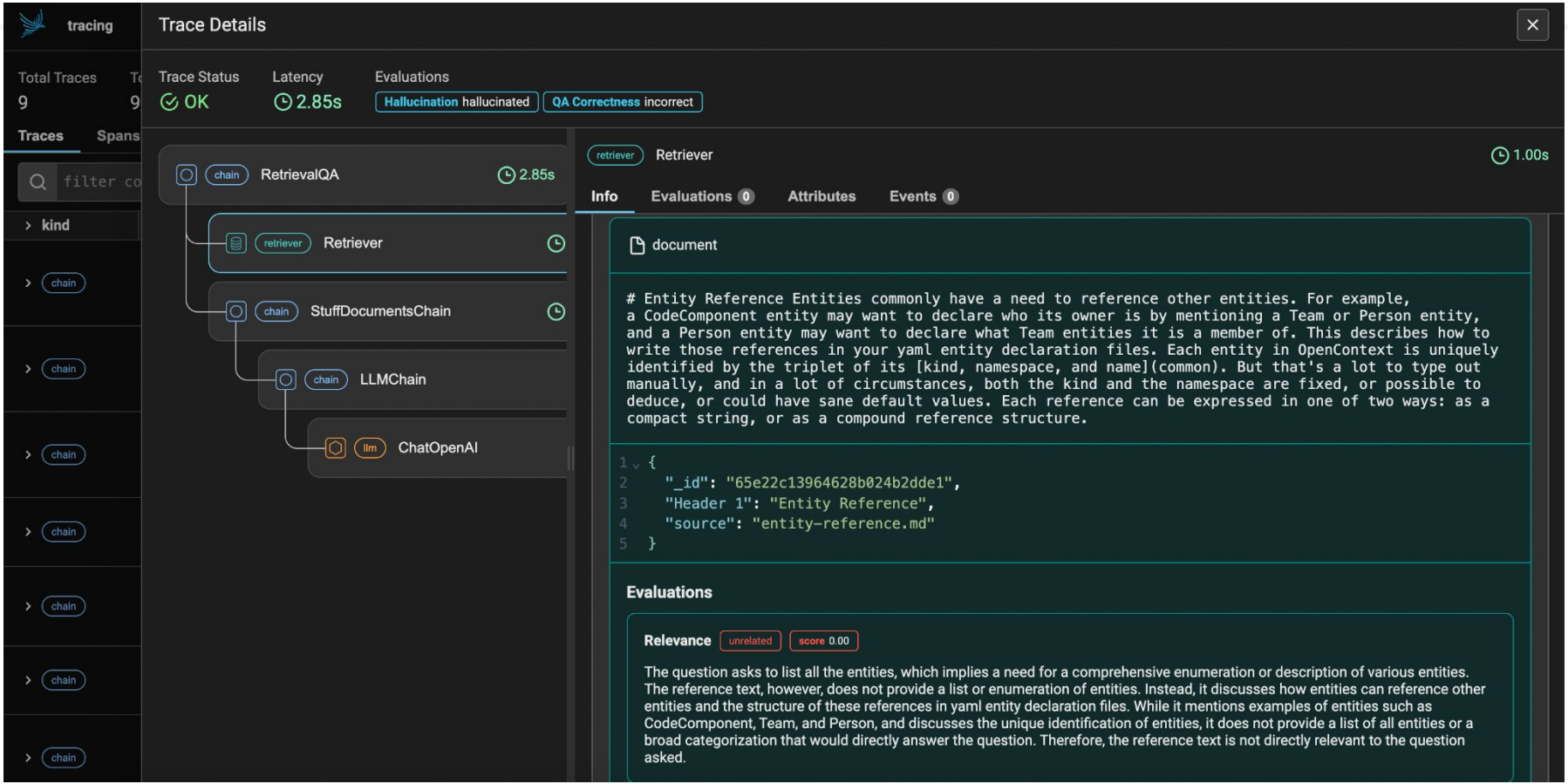
Evaluations



Observability - Book Analysis



Evaluations



R - RAG

GenAI Basics - RAG

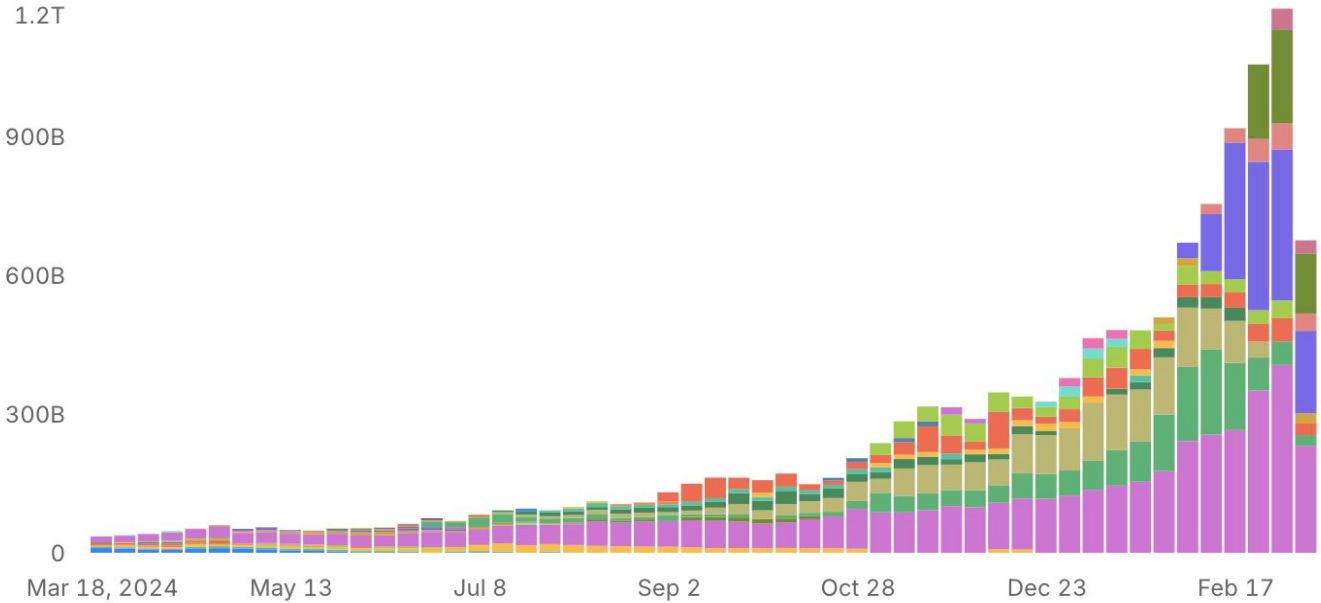
- ChromaDB
- SQLite
- MongoDB Atlas Vector Search
- PGVector
- Pinecone
- …

M - Models

LLM Rankings

Compare models for all prompts

All Categories | Roleplay | Programming | Marketing | Marketing/Seo | Technology | Science | Translation | Legal | Finance | Health | Trivia | Academia

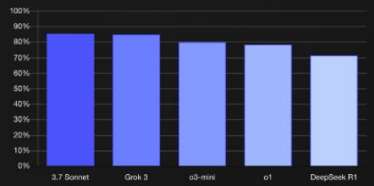


https://openrouter.ai/rankings

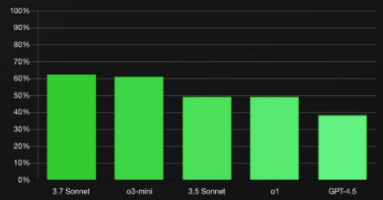
LLM Leaderboard

Top Models per Task

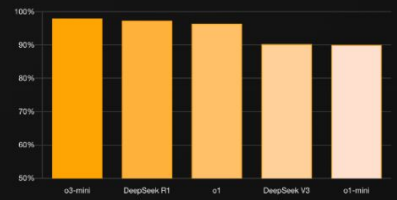
Best in Reasoning (GPQA Diamond)



Best in Coding (SWE Bench Verified)

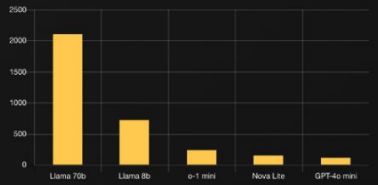


Best in Math (MATH)

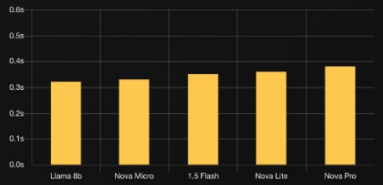


Fastest and Most Affordable Models

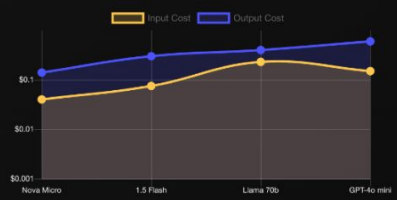
Fastest Models



Lowest Latency (TTFT)



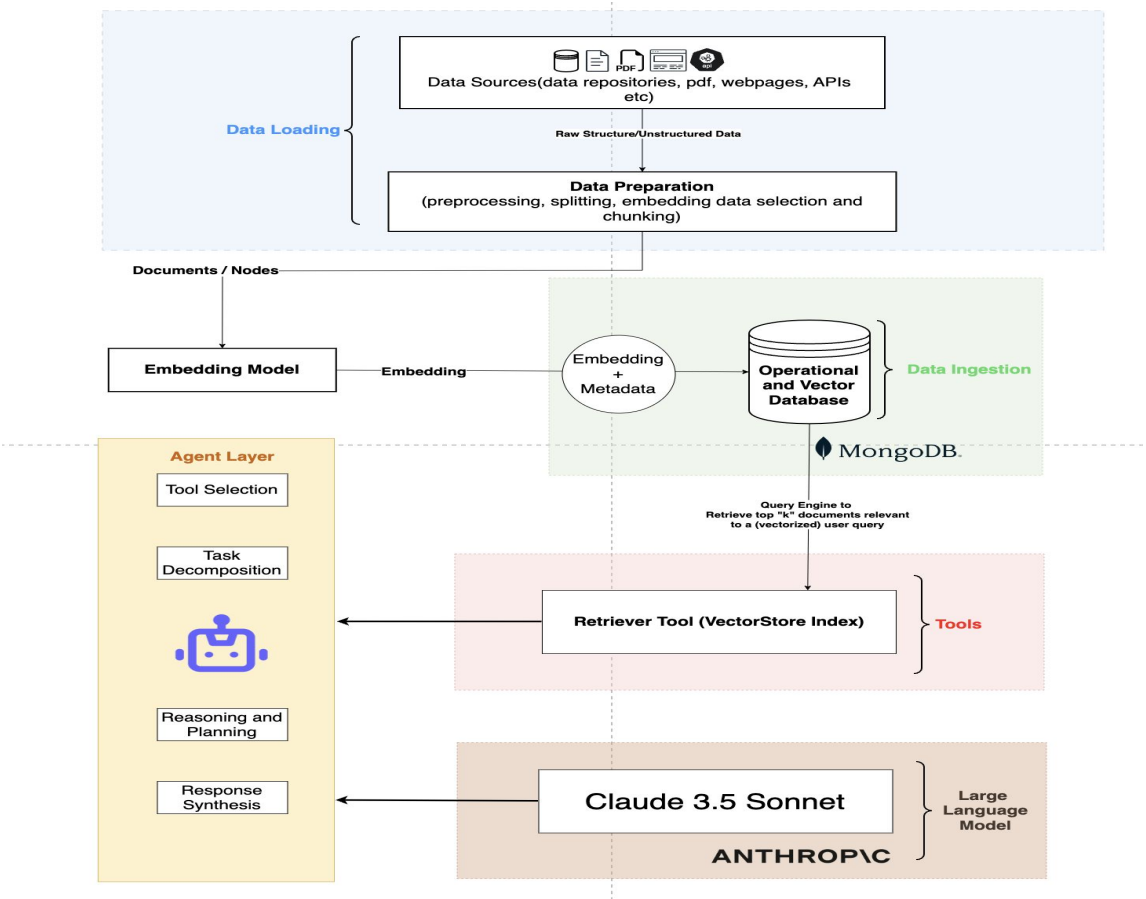
Cheapest Models



https://www.vellum.ai/llm-leaderboard

A - Agents

Agentic Systems



GenAI Basics - Agents

- LangGraph
- CrewAI
- AutoGen
- BeeAI
- OpenAI Agents SDK

Lessons Review

00-training-setup

RAG

Before We Get Started

The Rumors of RAG's Demise Might be Exaggerated

Retrieval Augmented Generation (RAG) has become interchangeable with integrating external knowledge sources into large language models. Some innovative alternatives, such as Context-Augmented Generation (CAG) and Graph-based Retrieval-Augmented Generation (GraphRAG), have emerged, each promising to overcome some of RAG's limitations. However, rather than marking the end of RAG, these developments suggest an evolution in which RAG remains a critical building block of a modern generative AI stack.

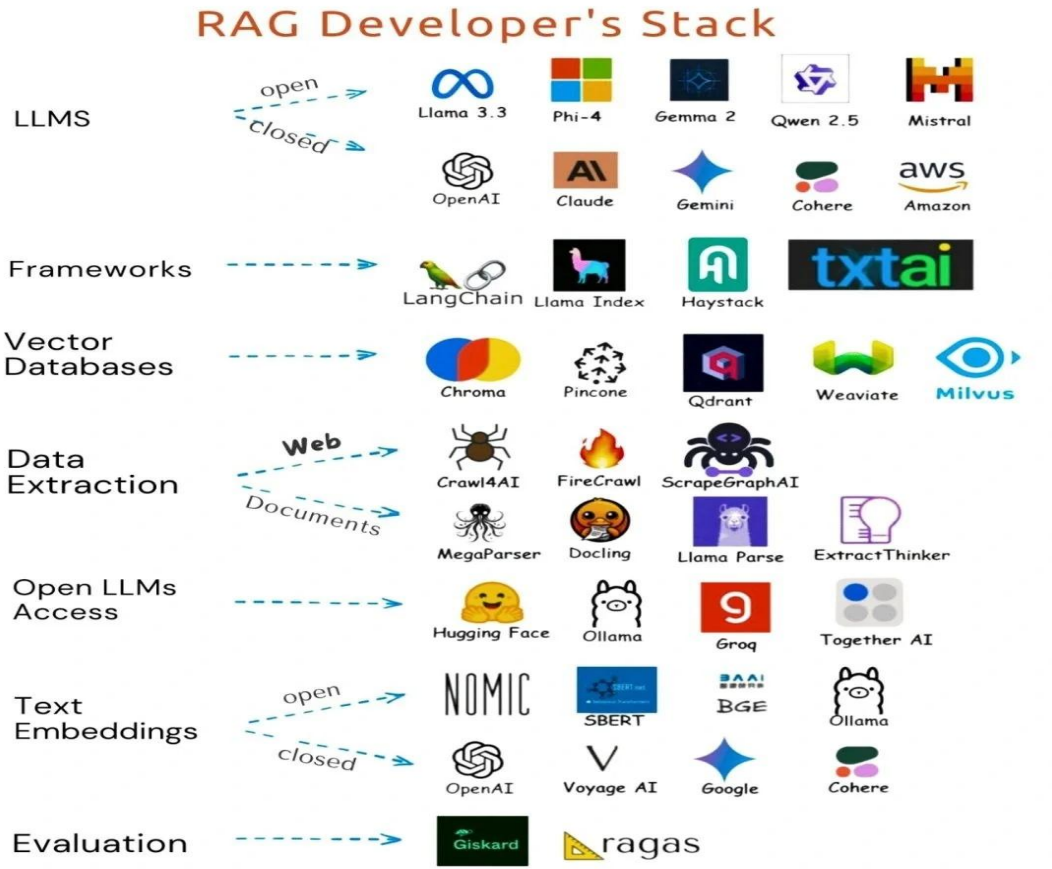
RAG in Today's AI Ecosystem

RAG has been a key contributor by coupling the capabilities of large language models with the retrieval of domain-specific or factual information from vast external databases. This combination enhances factual accuracy and expands the scope and depth of generated outputs. Its design enables practitioners to integrate new information sources easily without overhauling the underlying model architecture.

The Rise of Alternatives

- **Tools and Functions** LLMs like GPT-4, Claude, and Gemini can interact with external tools and execute functions to extend their capabilities beyond text generation. These tools enable real-time data retrieval, API calls, function execution, and structured output generation.
- **Graph-based Retrieval-Augmented Generation (GraphRAG):** GraphRAG takes a structural approach by leveraging graph models to capture intricate relationships between documents. This method excels in tasks that demand multi-hop reasoning and an interconnected understanding of complex data, such as legal document analysis or scientific research.
- **Cache Augmented Generation (CAG)** (sometimes referred to as “context”) enhances response quality by embedding richer contextual information, such as user intent and conversational history. This approach is particularly advantageous in personalized applications where a deep understanding of prior interactions is crucial.
-

RAG Overview



https://www.reddit.com/r/aiagents/

RAG Overview

What is RAG?

- Combines retrieval of relevant information with text generation
- Enhances the quality and accuracy of generated responses

Key Components:

- **Retrieval:** Searches a large database for relevant information
- **Augmentation:** Uses the retrieved information to improve response quality
- **Generation:** Produces a coherent and informed text response

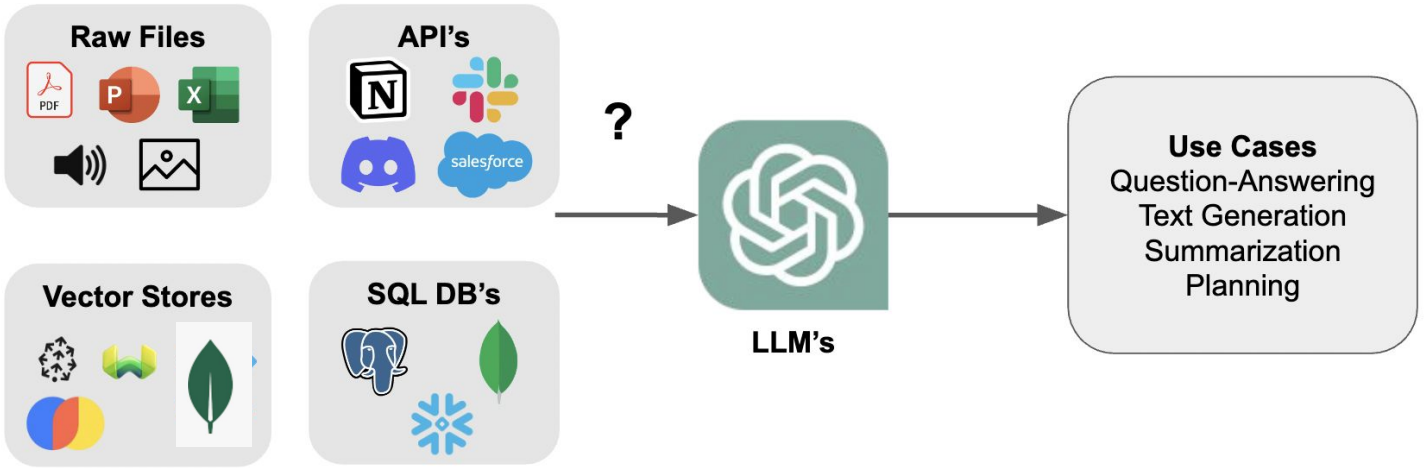
Why Use RAG?

- Creates more accurate and contextually relevant outputs
- Combines the strengths of retrieval-based and generation-based models

Retrieval Augmentation Generation (RAG)

Context

- How do we best augment LLMs with our own **private data**?



https://aiconference.com/speakers/jerry-liu/

This Workshop - RAG - Vector Databases

- ChromaDB
- MongoDB - Atlas Vector Search

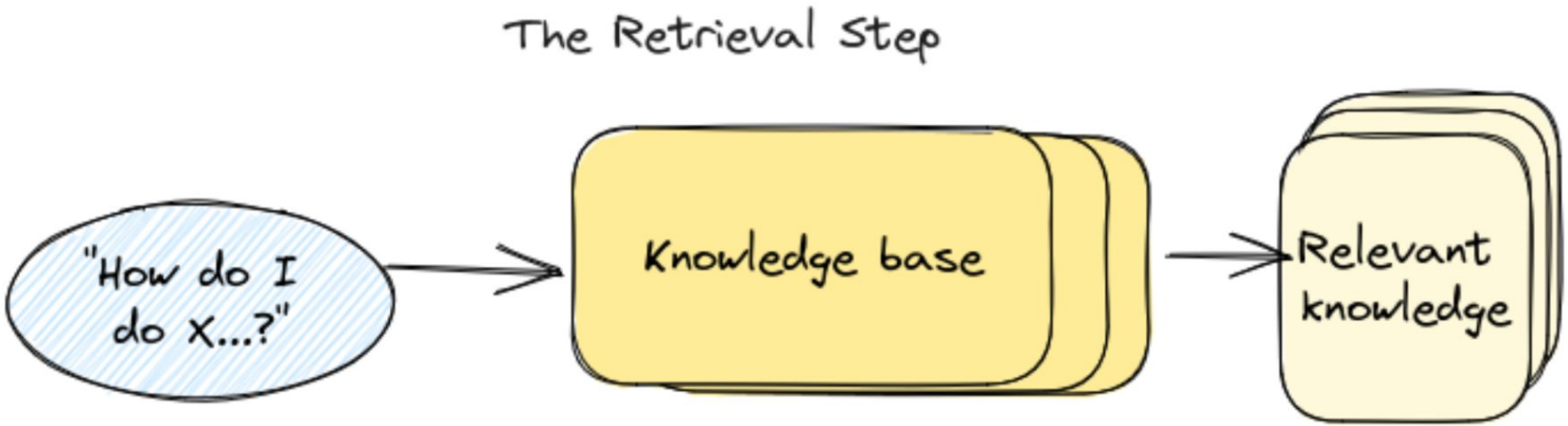
How RAG Works

Query Processing: When a user inputs a query, RAG first converts this query into an *embedding*—a vector representation that captures its meaning.

Document Retrieval: The system then retrieves documents from a database that are semantically similar to the *query embedding*. These documents provide additional context that the language model can use.

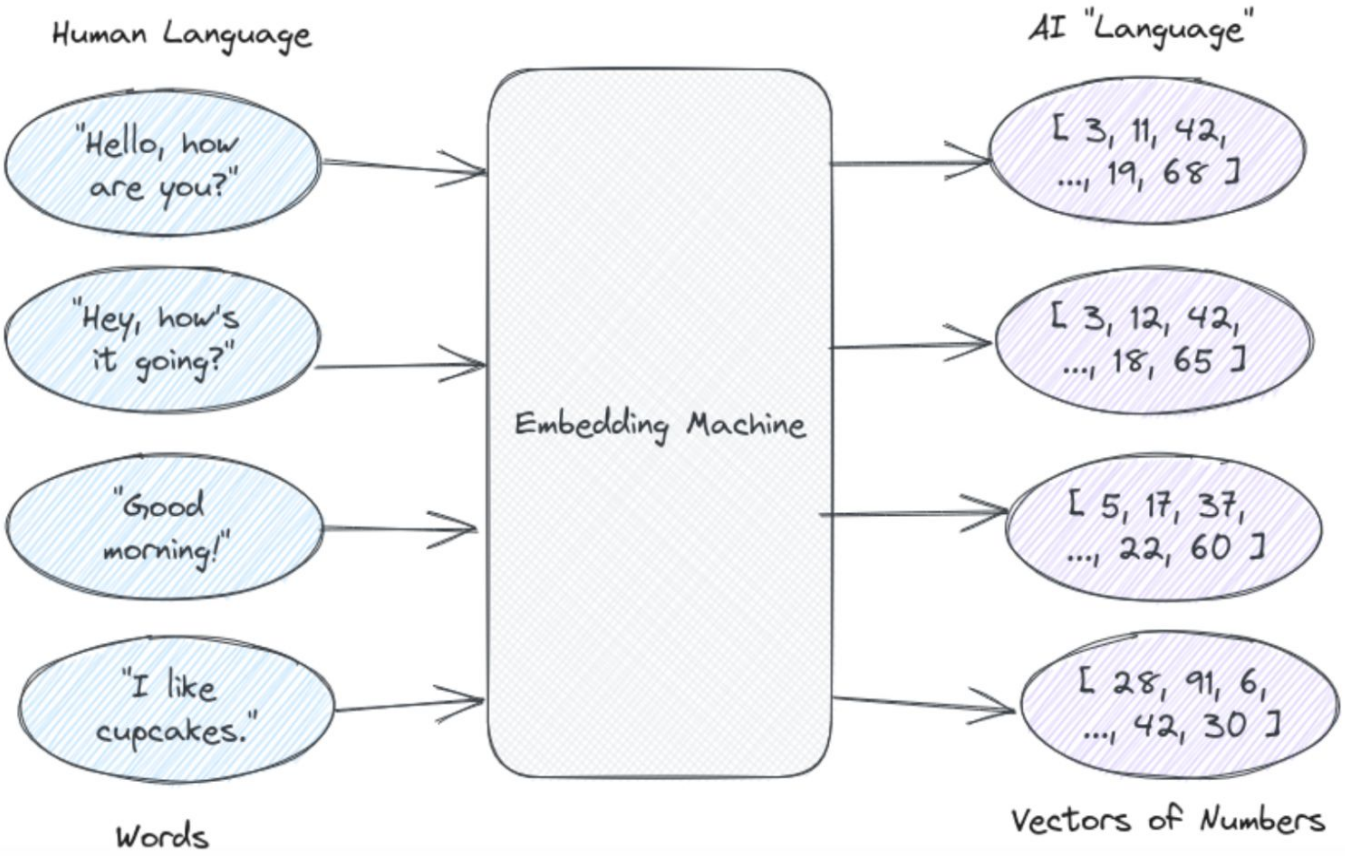
Augmented Generation: The retrieved documents are passed to the language model, which uses them to generate a response. This process allows the model to produce more accurate and *contextually relevant answers* by leveraging up-to-date information.

Retrieve Knowledge



https://vest.buildlab.cloud/en/watsonx/watsonxai/107

Embeddings and Vector Databases



https://vest.buildlab.cloud/en/watsonx/watsonxai/107

Embeddings (Vector Space)

https://vest.buildlab.cloud/en/watsonx/watsonxai/107

Encoding & Decoding in Transformers

- Encoding & Decoding enable AI models to understand and generate text.
  - Encoding: Text → Tokens → Embeddings
  - Decoding: Embeddings → Tokens → Text

Encoding & Decoding in Transformers

- Encoder Models (BERT)
  - Convert text into embeddings.
- Decoder Models (GPT)
  - Generate text from embeddings.
- Encoder-Decoder Models (T5, BART)
  - Translate/summarize text.

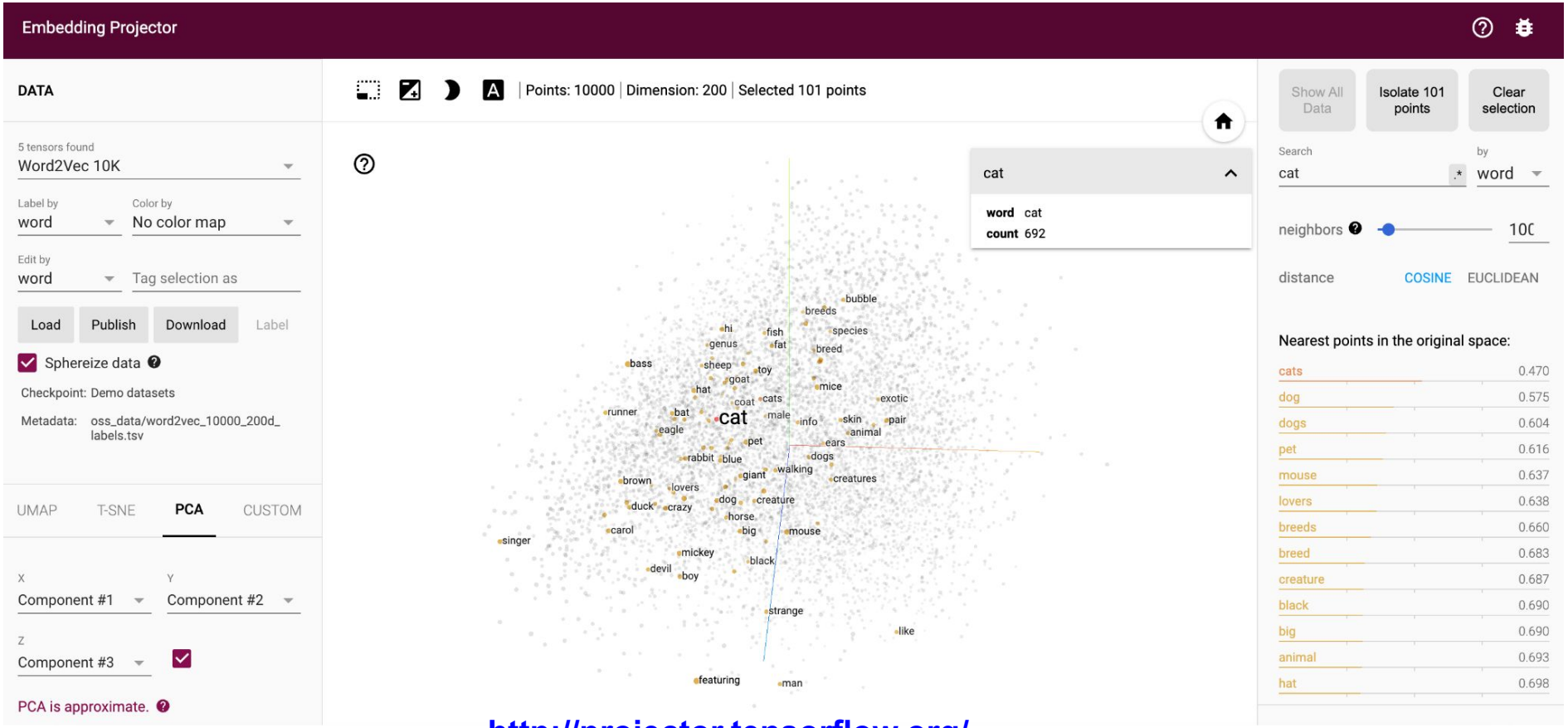
Embeddings (BERT Model and Tokenizer)

```
3D Embedding for the word 'cat':
[-0.30274457  0.07480898  0.03708227]
```

Embeddings - 768 (BERT Model and Tokenizer)

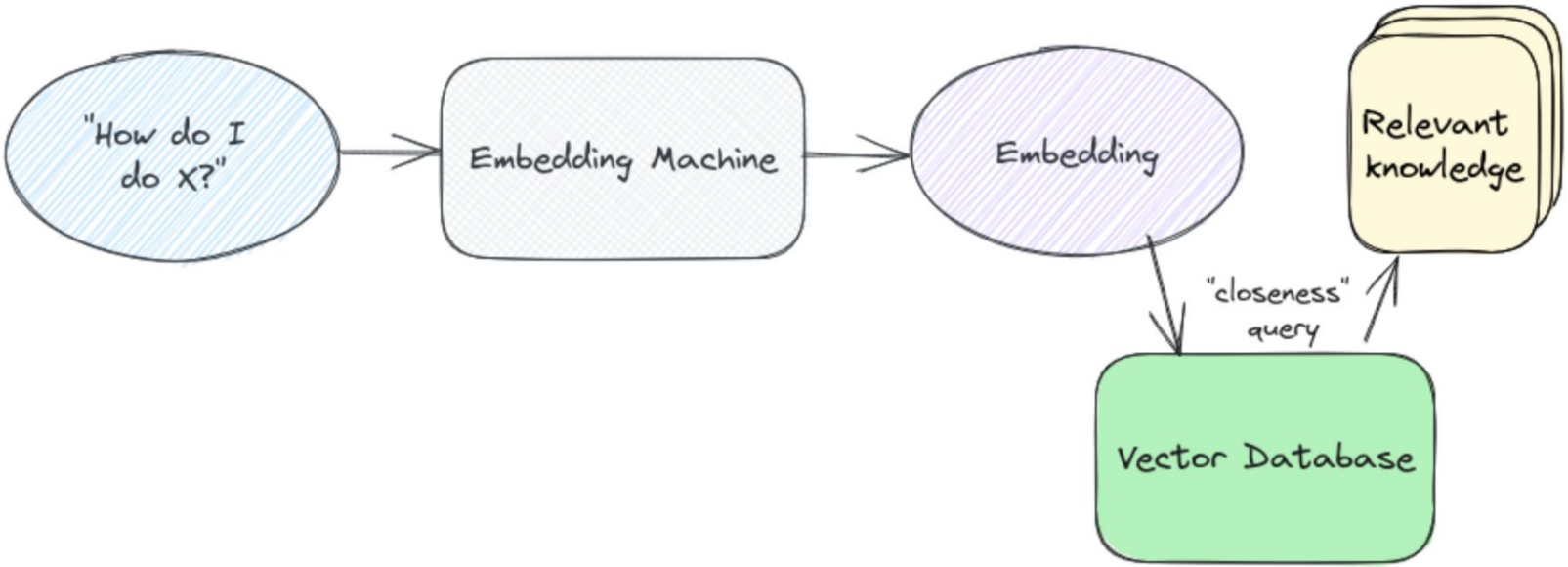
```
Embedding for the word 'cat':
[-3.02744567e-01  7.48089775e-02  3.70822661e-02 -1.45743400e-01
 -5.90116754e-02 -6.63245842e-02  8.12725797e-02  2.89646149e-01
 -2.05113858e-01 -3.47541720e-02 -1.60871148e-01 -5.38428426e-02
 -7.36298785e-02  1.96815237e-01  6.37774467e-02 -3.07037309e-03
 -1.83117747e-01  2.13095635e-01  1.59485951e-01 -1.85141370e-01
 -1.79041401e-02 -1.32813826e-01 -1.59528702e-01 -1.85157638e-02
 -7.47003183e-02 -6.09017760e-02 -5.70846125e-02  6.57155886e-02
  1.05983399e-01  1.81520969e-01  1.09177470e-01  3.84305529e-02
 -1.37314171e-01  1.38914138e-01 -2.48780083e-02  3.94920632e-02
 -3.48097794e-02 -2.41541341e-02  6.87406585e-02  9.84342545e-02
  1.01276502e-01 -7.27898851e-02  1.01367712e-01 -6.50159568e-02
  2.71233469e-02 -4.76146415e-02 -1.28022099e+00 -1.04655683e-01
 -1.13279864e-01 -3.96253616e-02  9.01971944e-03  1.93159729e-02
  3.12729269e-01  5.91203384e-02 -5.29393516e-02  1.68267190e-01
 -8.78007859e-02  4.99693662e-01  7.23395646e-02  1.06501117e-01
 -9.24750045e-03  2.10921168e-02 -1.30935624e-01  1.28516436e-01
 -9.93303806e-02  2.62693495e-01  4.32591811e-02  1.50604278e-01
 -5.08188084e-02  3.87150317e-01 -1.64397389e-01 -1.68239653e-01
  1.72385022e-01  8.61751139e-02 -1.11703910e-02 -9.78896320e-02
  4.84429300e-02 -8.39101821e-02  9.80098546e-02  1.59861833e-01
  1.96470037e-01  1.25753343e-01  8.36730227e-02  1.41791508e-01
```

Embeddings (Word2Vec)



http://projector.tensorflow.org/

RAG With Vector Database



https://vest.buildlab.cloud/en/watsonx/watsonxai/107

Creating a Vector Database

- **Data Loading**:
  - Import all data (e,g, documents) or information to be utilized.
- **Data Splitting**:
  - Divide large documents into smaller pieces.
  - Chunking.
- **Data Embedding**:
  - Convert the data into vector form using an embedding model.
  - Makes the data understandable for computers.
- **Data Storing**:
  - Save vector embeddings in a vector database.
  - Enables easy and efficient search.

Vector Database Structure

- **Tokens**
  - Tokens in an LLM represent the fundamental units into which text is divided for processing.
- **Chunks**
  - Chunks in a vector database index are partitions of the dataset, designed to improve the efficiency, scalability, and performance of operations within the database.
- **Dimensions**
  - A dimension in a vector context refers to one of its elements, and the total number of dimensions a vector has indicates the space within which it exists.
- **Metadata**
  - Metadata in a vector database refers to data that provides information about other data stored within the database, specifically regarding the vectors.
- **Indexes**
  - An index in a vector database is a structured representation of the data points' vectors, designed to optimize the retrieval of similar data points or objects.

Chunking

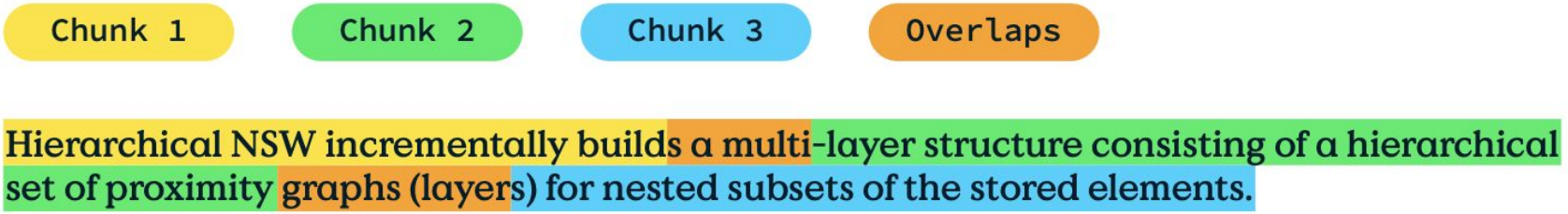


Figure 5.7: An example of chunking based on fixed token count with overlap

Chunking Strategies for LLM Applications

Introduction

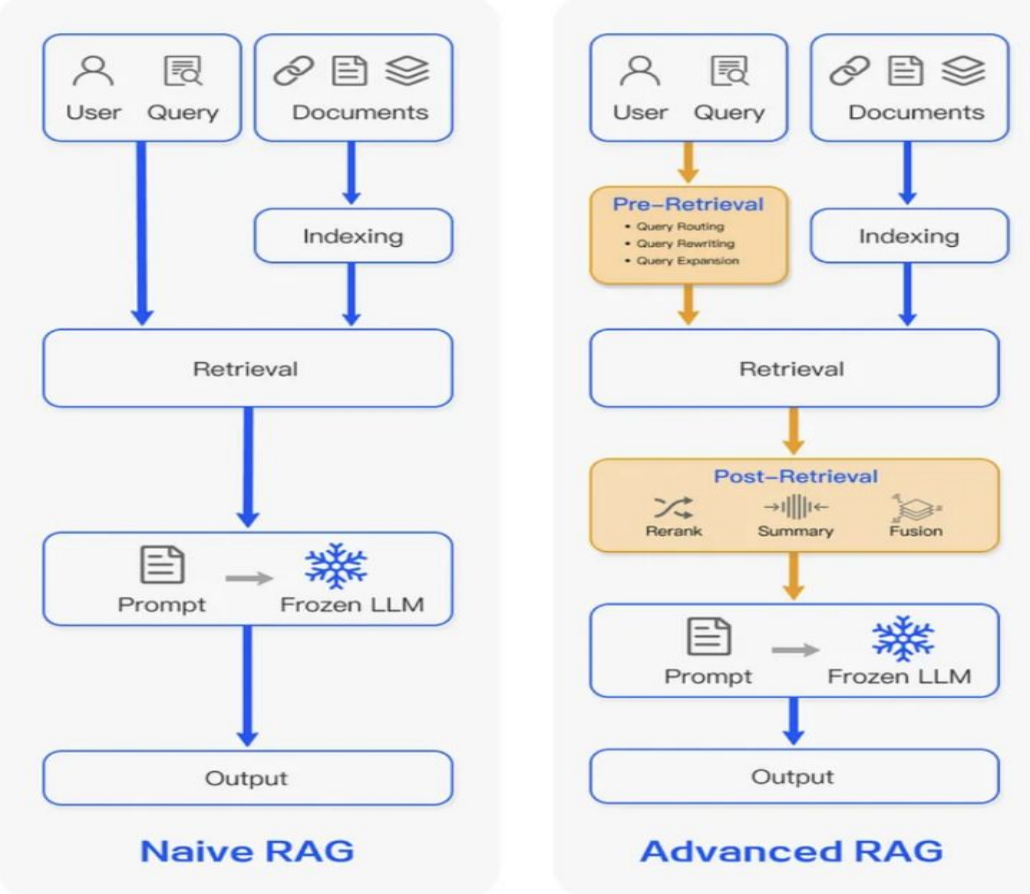
- Breaking down large data files is crucial for LLM application efficiency
- Goal: Provide LLM with precisely needed information
- Chunking strategy is a fundamental decision in RAG solution development

Overview

Five levels of chunking strategies:

1. Fixed Size Chunking
2. Recursive Chunking
3. Document Based Chunking
4. Semantic Chunking
5. Agentic Chunking

RAG Systems



https://driulija.github.io/

Lessons Review

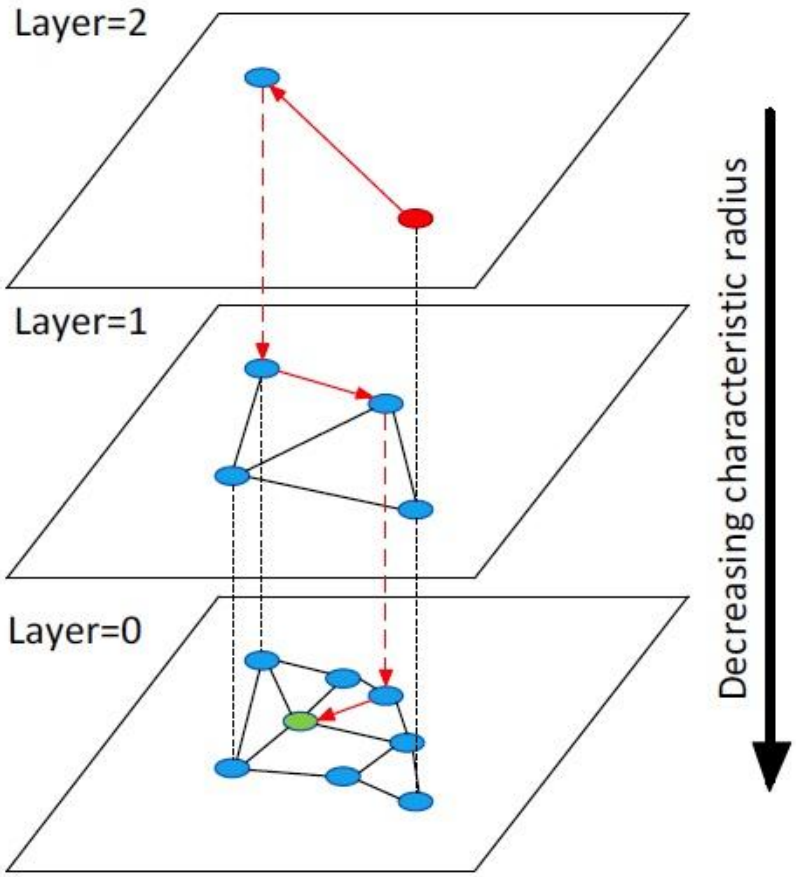
03-rag-graph-databases

(00 and 01)

Naive RAG

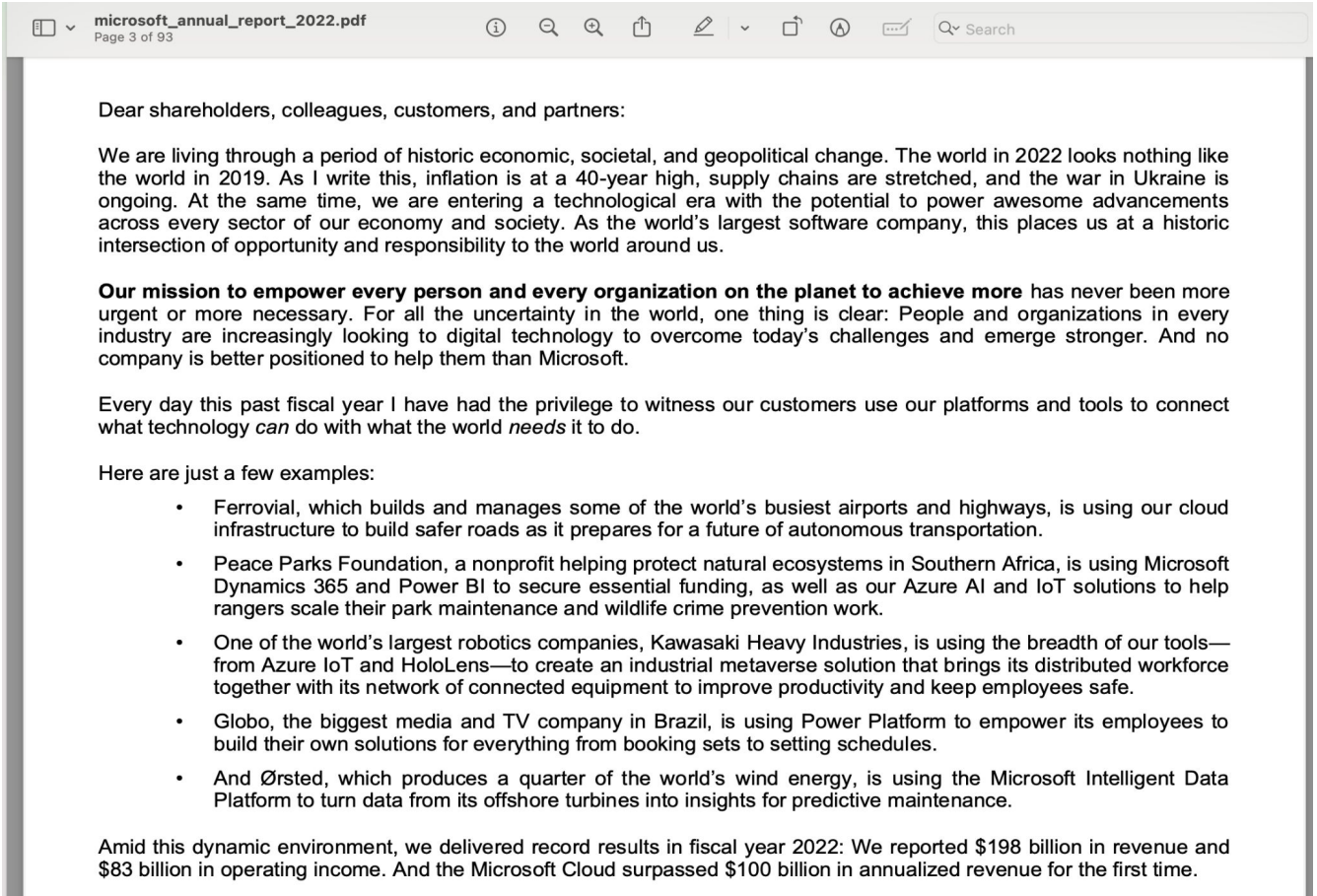
Vector Search (HNSW)
Hierarchical Navigable Small World Graphs

- Most Vector Databases are powered by a graph-based algorithm called HNSW.
- The queries are called ANN (Approximate K nearest neighbors)
- They provide low-latency search and high-recall results.



https://arxiv.org/abs/2102.06525

Example Document



Dear shareholders, colleagues, customers, and partners:

We are living through a period of historic economic, societal, and geopolitical change. The world in 2022 looks nothing like the world in 2019. As I write this, inflation is at a 40-year high, supply chains are stretched, and the war in Ukraine is ongoing. At the same time, we are entering a technological era with the potential to power awesome advancements across every sector of our economy and society. As the world's largest software company, this places us at a historic intersection of opportunity and responsibility to the world around us.

Our mission to empower every person and every organization on the planet to achieve more has never been more urgent or more necessary. For all the uncertainty in the world, one thing is clear: People and organizations in every industry are increasingly looking to digital technology to overcome today's challenges and emerge stronger. And no company is better positioned to help them than Microsoft.

Every day this past fiscal year I have had the privilege to witness our customers use our platforms and tools to connect what technology *can* do with what the world *needs* it to do.

Here are just a few examples:

- Ferrovial, which builds and manages some of the world's busiest airports and highways, is using our cloud infrastructure to build safer roads as it prepares for a future of autonomous transportation.
- Peace Parks Foundation, a nonprofit helping protect natural ecosystems in Southern Africa, is using Microsoft Dynamics 365 and Power BI to secure essential funding, as well as our Azure AI and IoT solutions to help rangers scale their park maintenance and wildlife crime prevention work.
- One of the world's largest robotics companies, Kawasaki Heavy Industries, is using the breadth of our tools—from Azure IoT and HoloLens—to create an industrial metaverse solution that brings its distributed workforce together with its network of connected equipment to improve productivity and keep employees safe.
- Globo, the biggest media and TV company in Brazil, is using Power Platform to empower its employees to build their own solutions for everything from booking sets to setting schedules.
- And Ørsted, which produces a quarter of the world's wind energy, is using the Microsoft Intelligent Data Platform to turn data from its offshore turbines into insights for predictive maintenance.

Amid this dynamic environment, we delivered record results in fiscal year 2022: We reported $198 billion in revenue and $83 billion in operating income. And the Microsoft Cloud surpassed $100 billion in annualized revenue for the first time.

Recursive Chunking

```python
character_splitter = RecursiveCharacterTextSplitter(
    #separators=["\n\n", "\n", ". ", " ", ""],
    chunk_size=1000,
    chunk_overlap=0,
)
character_split_texts = character_splitter.split_text('\n\n'.join(pdf_texts))

for i in range(5):
    print("Chunk#=" + str(i) + " Chunk=" + character_split_texts[i],"\n")

print(f"\nTotal chunks: {len(character_split_texts)}")
```

Python

Emeddings

```python
import chromadb
from chromadb.utils.embedding_functions import SentenceTransformerEmbeddingFunction

embedding_function = SentenceTransformerEmbeddingFunction()
print(embedding_function([token_split_texts[0]]))
```

Python

```python
# Access the underlying SentenceTransformer model
model = embedding_function.models
print(model)
```

Python

```
{'all-MiniLM-L6-v2': SentenceTransformer(
  (0): Transformer({'max_seq_length': 256, 'do_lower_case': False}) with Transformer model:
BertModel
  (1): Pooling({'word_embedding_dimension': 384, 'pooling_mode_cls_token': False,
'pooling_mode_mean_tokens': True, 'pooling_mode_max_tokens': False,
'pooling_mode_mean_sqrt_len_tokens': False, 'pooling_mode_weightedmean_tokens': False,
'pooling_mode_lasttoken': False, 'include_prompt': True})
  (2): Normalize()
)}
```

Retrieve the Embeddings

```
query = "What was the total revenue?"

results = chroma_collection.query(query_texts=query,
                                  n_results=5,
                                  include=['documents', 'embeddings', "distances"])

retrieved_documents = results['documents'][0]
print_results_and_documents(results, retrieved_documents, word_wrap)
```

✓ 0.0s Python

```
ids:
  Item 1: ['321', '293', '331', '319', '194']

distances:
  Item 1: [0.39571720361709595, 0.4243314266204834, 0.45620059967041016, 0.4578591585159302, 0.4580157399
```

Augment and Generate

```python
output = rag(query=query, retrieved_documents=retrieved_documents)
```

Python

```python
output = rag(query=query, retrieved_documents=retrieved_documents)
```

Python

```python
print(word_wrap(output))
```

✓ 0.0s

Python

```
The total revenue for the year ended June 30, 2022, as reported in the
annual report, was $198,270 million.
```

Retrieve the Embeddings

Chunk1 dear shareholders, colleagues, customers, and partners : we are living through a period of historic economic, societal, and geopolitical change. the world in 2022 looks nothing like the world in 2019. as i write this, inflation is at a 40 - year high, supply chains are stretched, and the war in ukraine is ongoing. at the same time, we are entering a technological era with the potential to power awesome advancements across every sector of our economy and society. as the world ' s largest software company, this places us at a historic intersection of opportunity and responsibility to the world around us. our mission to empower every person and every organization on the planet to achieve more has never been more urgent or more necessary. for all the uncertainty in the world, one thing is clear : people and organizations in every industry are increasingly looking to digital technology to overcome today ' s challenges and emerge stronger. and no

Chunk2 company is better positioned to help them than microsoft. every day this past fiscal year i have had the privilege to witness our customers use our platforms and tools to connect what technology can do with what the world needs it to do. here are just a few examples : • ferrovial, which builds and manages some of the world ' s busiest airports and highways, is using our cloud infrastructure to build safer roads as it prepares for a future of autonomous transportation. • peace parks foundation, a nonprofit helping protect natural ecosystems in southern africa, is using microsoft dynamics 365 and power bi to secure essential funding, as well as our azure ai and iot solutions to help rangers scale their park maintenance and wildlife crime prevention work. • one of the world ' s largest robotics companies, kawasaki heavy industries, is using the breadth of our tools — from azure iot and hololens — to create an industrial metaverse solution that brings its distributed workforce

Chunk3 together with its network of connected equipment to improve productivity and keep employees safe. • globo, the biggest media and tv company in brazil, is using power platform to empower its employees to build their own solutions for everything from booking sets to setting schedules. • and ørsted, which produces a quarter of the world ' s wind energy, is using the microsoft intelligent data platform to turn data from its offshore turbines into insights for predictive maintenance. amid this dynamic environment, we delivered record results in fiscal year 2022 : we reported $ 198 billion in revenue and $ 83 billion in operating income. and the microsoft cloud surpassed $ 100 billion in annualized revenue for the first time. our responsibility as a corporation, our purpose and actions must be aligned with addressing the world ' s problems, not creating new ones. at our very core, we need to deliver innovation that helps drive broad economic growth. we, as a company, will do well

Chunk4 when the world around us does well. that ' s what i believe will lead to widespread human progress and ultimately improve the lives of everyone. there is no more powerful input than digital technology to drive the world ' s economic output. this is the core thesis for our being as a company, but it ' s not enough. as we drive global economic growth, we must also commit to creating a more inclusive, equitable, sustainable, and trusted future. support inclusive economic growth we must ensure the growth we drive reaches every person, organization, community, and country. this starts with increasing access to digital skills. this year alone, more than 23 million people accessed digital skills training as part of our global skills initiative.

Chunk5 but skills alone aren ' t enough — we need to help people better prepare for and connect to jobs. that ' s why we ' ve committed to equip 10 million people from underserved communities with skills for jobs in the digital economy by 2025. one area of digital skills has become especially critical : cybersecurity. cybersecurity is a significant threat for governments, businesses, and individuals around the world, yet there simply aren ' t enough people with cybersecurity skills to fill open jobs. to help address this, we ' ve committed to skill and recruit 250, 000 people into the us cybersecurity workforce by 2025 — especially those underrepresented in the field. and we ' re helping an additional 24 countries with substantial cybersecurity workforce shortages close their gaps too. we also continue to deliver affordable, relevant cloud technology and industry - specific solutions to nonprofit organizations

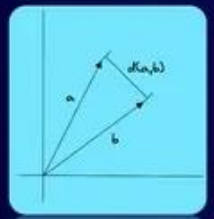
Vector Space

https://www.xomnia.com/post/an-introduction-to-vector-databases-for-beginners/

Embeddings and Vectors are Geometric

Euclidean (L2)

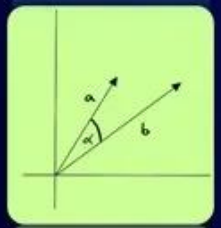
Useful for counts / measurements
Recommendation Systems



$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Cosine Similarity

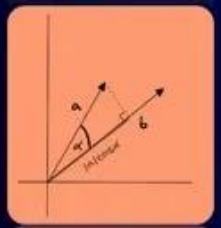
Useful for semantic search and
document classification



$$sim(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{||\mathbf{a}|| \cdot ||\mathbf{b}||}$$

Dot Product

Useful for collaborative filtering



$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| cos\alpha$$

https://community.aws/content/2gvh6fQM4mJQduLye3mHICNvPxX/vector-embeddings-and-rag-demystified

Scenario: Comparing Time Spent on a Website

Let’s say we measure how much time two users spend on a website.

- **User A** spends **5 minutes**.
- **User B** spends **50 minutes**.
- **User C** spends **5 minutes** but on a completely different website.

Cosine Similarity (Angle-based)

- Since cosine similarity looks at the angle between two vectors (ignores magnitude focuses on direction), User A and User B are **very similar** because they have the same behavior **proportionally** (e.g., they spend time in the same way, just at a different scale).
- User C, however, has a different **pattern** (a different website), making them less similar.

Euclidean Distance (Distance-based)

- Euclidean similarity considers raw distance: User A and User C **both spent 5 minutes**, so they are closer together in absolute terms.
- User B, spending 50 minutes, is much farther away from both.

Why It Matters

Example: Medication Dosage for Two Patients

Imagine two patients taking different amounts of the same medication over a week:

Patient	Monday	Tuesday	Wednesday	Total Dosage
Alice	50 mg	50 mg	50 mg	**150 mg**
Bob	500 mg	500 mg	500 mg	**1500 mg**
Charlie	50 mg	50 mg	50 mg	**150 mg**

- **Cosine Similarity:** Alice and Bob have identical dosage patterns (same ratio of medication taken each day), so they would be considered **very similar**. However, Bob takes **10× the dosage** of Alice.
- **Euclidean Distance:** Alice and Charlie are **identical in total dosage**, making them much closer.

Why Use Euclidean Distance Here?

In **medical scenarios**, the absolute amount of medication matters. Even if two patients take the same proportional doses, one taking 10× more could have very different effects on their body. If we were predicting side effects or appropriate dosage levels, **Euclidean distance would be more appropriate**.

Lessons Review

03-rag-graph-databases

(01 and 02)

Dimensionality Reduction

- **Definition**: Dimensionality reduction is the process of transforming high-dimensional vector space into a lower-dimensional space while preserving important structures and minimizing information loss.
- **Purpose**: It helps in reducing computational complexity, removing noise, improving visualization, and preventing the curse of dimensionality in machine learning and data analysis.
- **Techniques**: Common methods include **Principal Component Analysis (PCA)** for linear transformation, **t-SNE** and **UMAP (**Uniform Manifold Approximation and Projection**)** for nonlinear mapping, and **Autoencoders** for deep learning-based reductions.
- **Trade-offs**: While reducing dimensions can speed up computations and improve model performance, excessive reduction can lead to loss of critical information, reducing model accuracy.
- **Applications**: Used in fields like natural language processing (word embeddings), image compression, anomaly detection, and exploratory data analysis for feature selection and visualization.

Visualizing Embeddings (UMAP)

Visualizing Embeddings (UMAP)

What is the strategy around artificial intelligence (AI) ?

Visualizing Embeddings

What is the strategy around artificial intelligence (AI) ?

Note: Notice the thise sentence is what probably gave it the closest distance.

"we are building a distributed computing fabric – across cloud and the edge – to help every organization build, run, and manage mission - critical workloads anywhere. in the next phase of innovation, artificial intelligence (“ ai ”) capabilities are rapidly advancing, fueled by data and knowledge of the world."

Visualizing Embeddings

What has been the investment in research and development?

Visualizing Embeddings

What has been the investment in research and development?

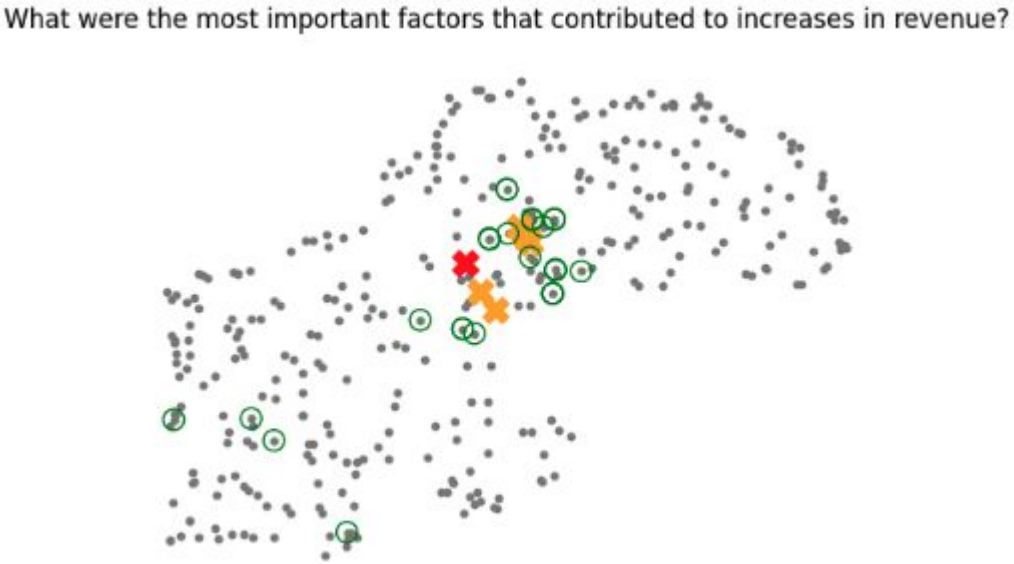
```
ids:
  Item 1: ['150', '111', '110', '3', '63']

distances:
  Item 1: [1.063101887702942, 1.0871107578277588, 1.134844422340393, 1.149186611175537, 1.1728293895721436]
```

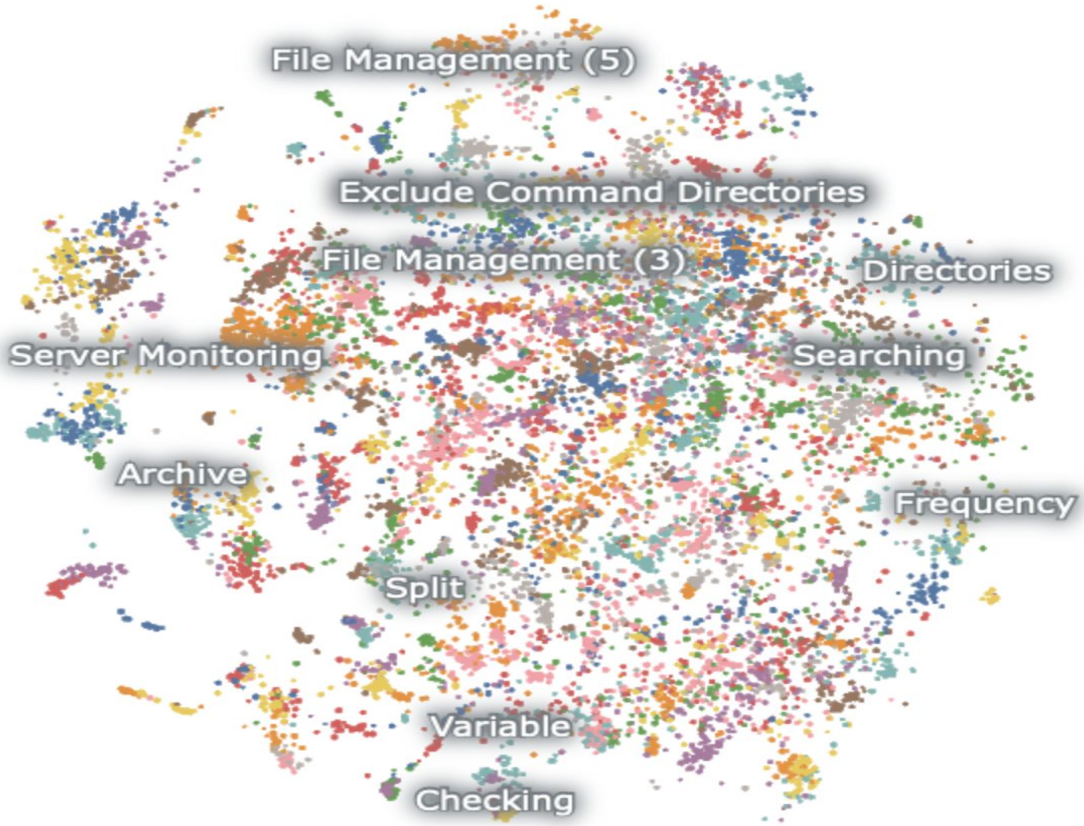
Query Expansion (Hypothetical Answers)

```
def augment_multiple_query(query, model="gpt-4o-mini"):
    messages = [
        {
            "role": "system",
            "content": "You are a helpful expert
            financial research assistant. Your users are
            asking questions about an annual report. "
            "Suggest up to five additional related
            questions to help them find the information
            they need, for the provided question. "
            "Suggest only short questions without
            compound sentences. Suggest a variety of
            questions that cover different aspects of
            the topic."
            "Make sure they are complete questions, and
            that they are related to the original
            question."
            "Output one question per line. Do not number
            the questions."
        },
        {"role": "user", "content": query}
    ]
```

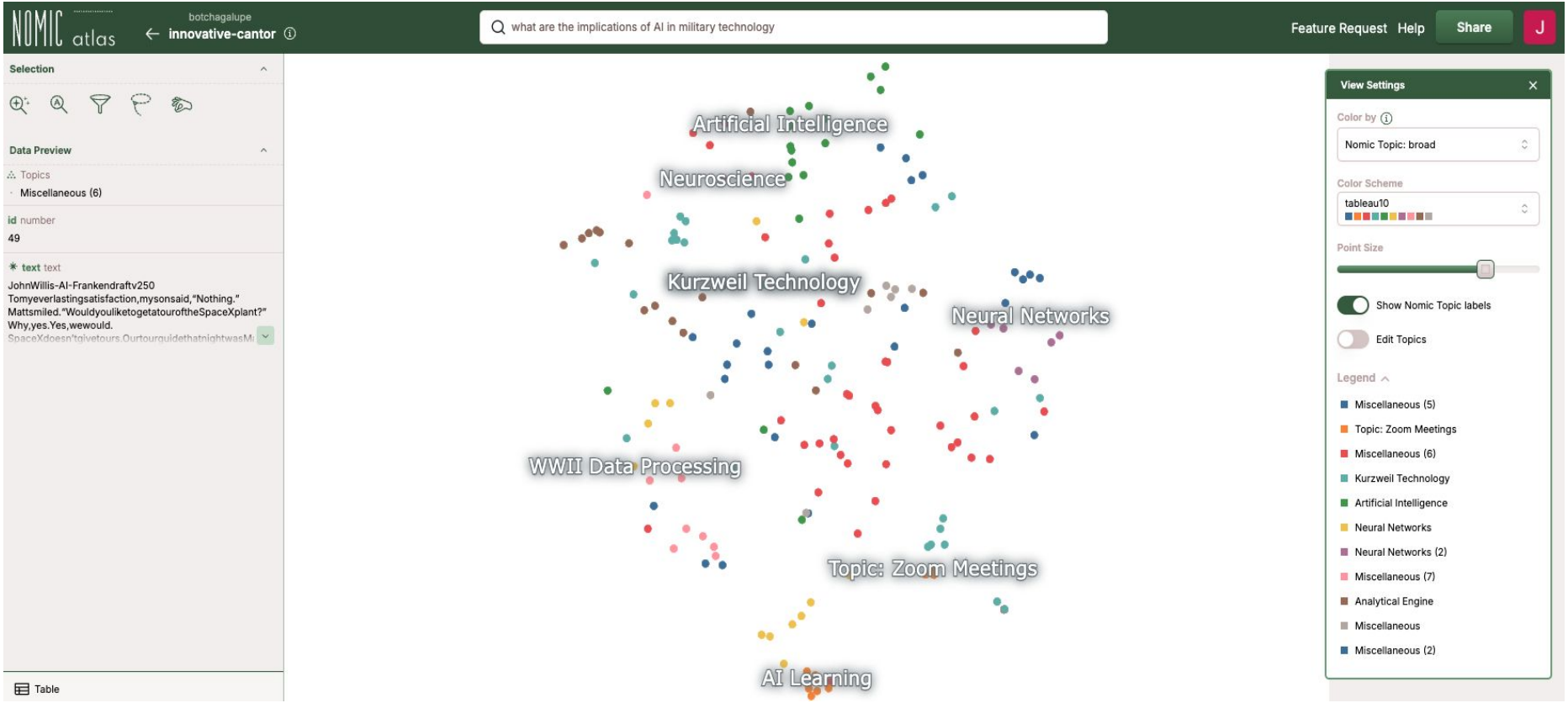
Visualizing Embeddings



Understanding Unstructured Data

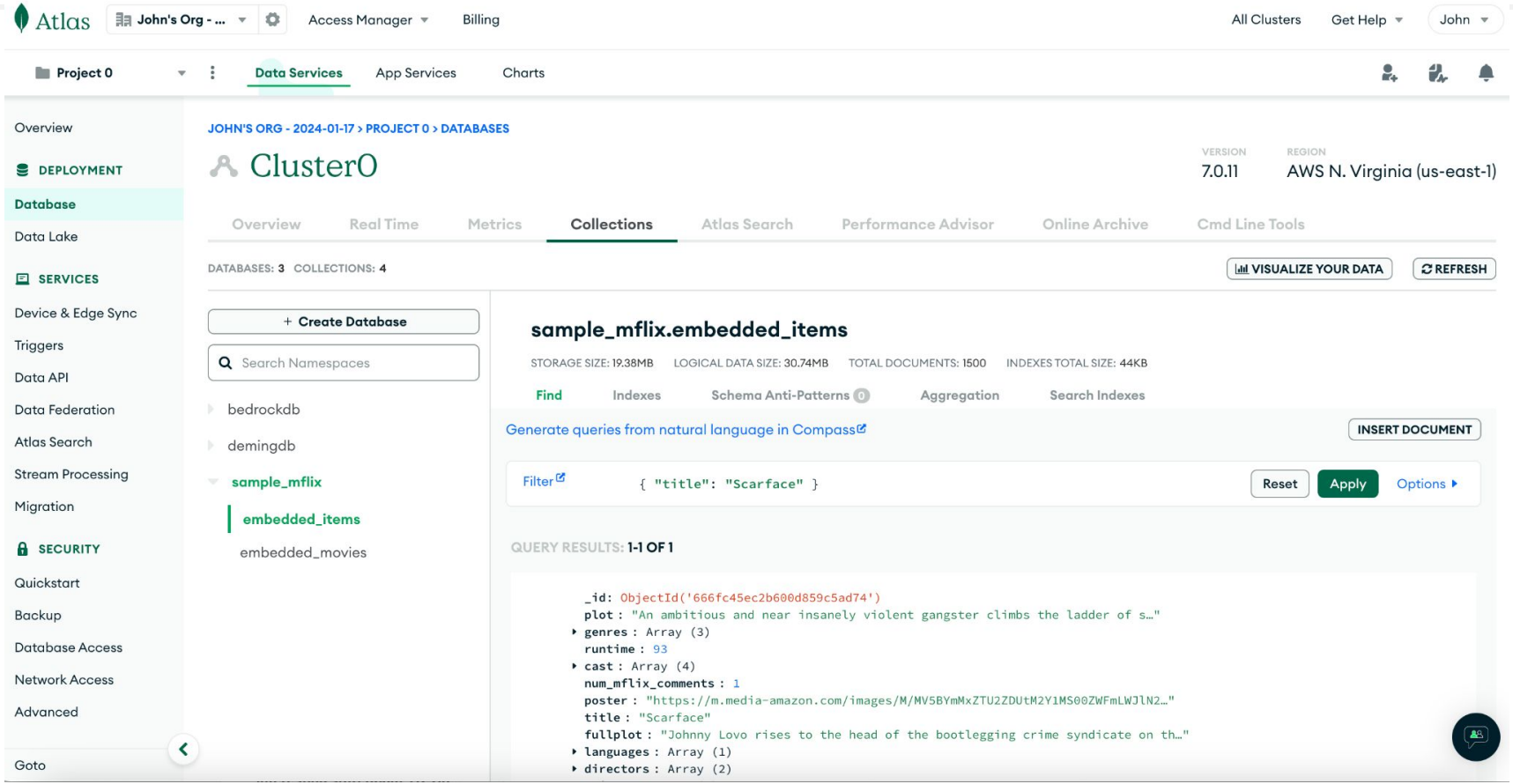


Understanding Unstructured Data



https://atlas.nomic.ai/data/botchagalupe/innovative-cantor/map

MongoDB Atlas Vector Search



Governance and security controls at every level

Business Trust Needs	Security Features	
Organization access	RBAC, multi-factor authentication, federated authentication, programmatic API keys	
Network isolation	Dedicated virtual private clouds for every project	
Network security	IP Access Lists, Peering connections, Private endpoints, configurable temporary access	
Database authentication	SCRAM, X.509, LDAPS, AWS IAM, configurable temporary access	
Database authorization	RBAC, read-only views, field-level redaction	
Database auditing	Admin, DML, DDL, DCL, role-based	
Database encryption	**In-Flight:** TLS 1.2+ **At-Rest:** Encrypted hardware, volume and database storage engine (AES-256)	
	In-Use: Client-Side Field Level Encryption **Key Management:** Cloud KMS	**Cloud agnostic:** Use AWS KMS, Azure Key Vault, or Google Cloud KMS regardless of underlying cloud provider

Hugging Face | Search models, datasets, user | Models | Datasets | Spaces | Posts | Docs | Enterprise | Pricing | Log In | Sign Up

MongoDB

Company | Verified

https://www.mongodb.com/ https://github.com/mongodb

Activity Feed | Follow 95

AI & ML interests

RAG (Retrieval-Augmented Generation) Vector Database

Recent Activity

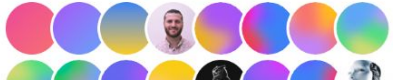
RichmondMongo updated a dataset 17 days ago
MongoDB/wikipedia-22-12-en-nomic-embedded

RichmondMongo published a dataset 28 days ago
MongoDB/wikipedia-22-12-en-nomic-embedded

RichmondMongo updated a dataset 28 days ago
MongoDB/wikipedia-22-12-en-annotation

View all activity

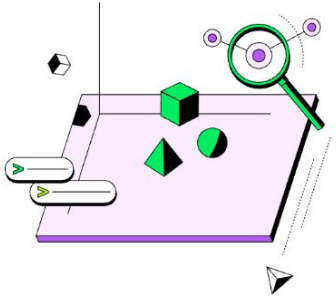
Team members 71



Organization Card

Community | About org cards

Build intelligent applications powered by semantic search and generative AI over any type of data



https://huggingface.co/MongoDB

MTEB: Massive Text Embedding Benchmark

We only display models that have been run on all tasks in the benchmark

Summary | Performance per task | Task information

Rank (Borda)	Model	Zero-shot	Number of Parameters	Embedding Dimensions	Max Tokens	Mean (Task)	Mean (TaskType)	Bitext Mining	Classification	Clustering	Instruction Retrieval	
1	gemini-embedding-exp-03-07	99%	Unknown	3072	8192	**68.32**	**59.64**	79.28	**71.82**	**54.99**	5.18	**29**
2	Linq-Embed-Mistral	99%	7B	4096	32768	61.47	54.21	70.34	62.24	51.27	0.94	24
3	gte-Qwen2-7B-instruct	⚠ NA	7B	3584	32768	62.51	56.00	73.92	61.55	53.36	4.94	25
4	multilingual-e5-large-instruct	99%	560M	1024	514	63.23	55.17	**80.13**	64.94	51.54	-0.40	22
5	SFR-Embedding-Mistral	96%	7B	4096	32768	60.93	54.00	70.00	60.02	52.57	0.16	24
6	GritLM-7B	99%	7B	4096	4096	60.93	53.83	70.53	61.83	50.48	3.45	22
7	text-multilingual-embedding-002	99%	Unknown	768	2048	62.13	54.32	70.73	64.64	48.47	4.08	22
8	GritLM-8x7B	99%	57B	4096	4096	60.50	53.39	68.17	61.55	50.88	2.44	24
9	e5-mistral-7b-instruct	99%	7B	4096	32768	60.28	53.18	70.58	60.31	51.39	-0.62	22
10	Cohere-embed-multilingual-v3.0	⚠ NA	Unknown	512	Unknown	61.10	53.31	70.50	62.95	47.61	-1.89	22
11	gte-Qwen2-1.5B-instruct	⚠ NA	1B	8960	32768	59.47	52.75	62.51	58.32	52.59	0.74	24

https://huggingface.co/blog/mteb

Lessons Review

03-rag-graph-databases

(02, 03, 04, and 07)

Advanced RAG

Advanced RAG Strategies

Introduction

- Advanced RAG strategies address challenges of Naive RAG
- Focus on efficient retrieval of relevant documents
- Key areas: Pre-Retrieval, Retrieval, Post-Retrieval

Challenges in RAG

1. Accurate semantic representations of documents and queries
2. Aligning semantic spaces of queries and documents (chunks)
3. Aligning retriever's output with LLM preferences

Pre-Retrieval Strategies

Improve Data Quality

- Remove irrelevant information
- Reduce ambiguity in entities and terms
- Confirm factual accuracy
- Update outdated information

Optimize Index Structure and Metadata

- Optimize chunk sizes
- Add graph structure information
- Include dates, chapters, subsections, purposes

Retrieval Strategies (Beware of Similarity Metrics)

Embedding Optimization

- Domain Knowledge Fine-Tuning
- Use domain-specific datasets

Similarity Metrics

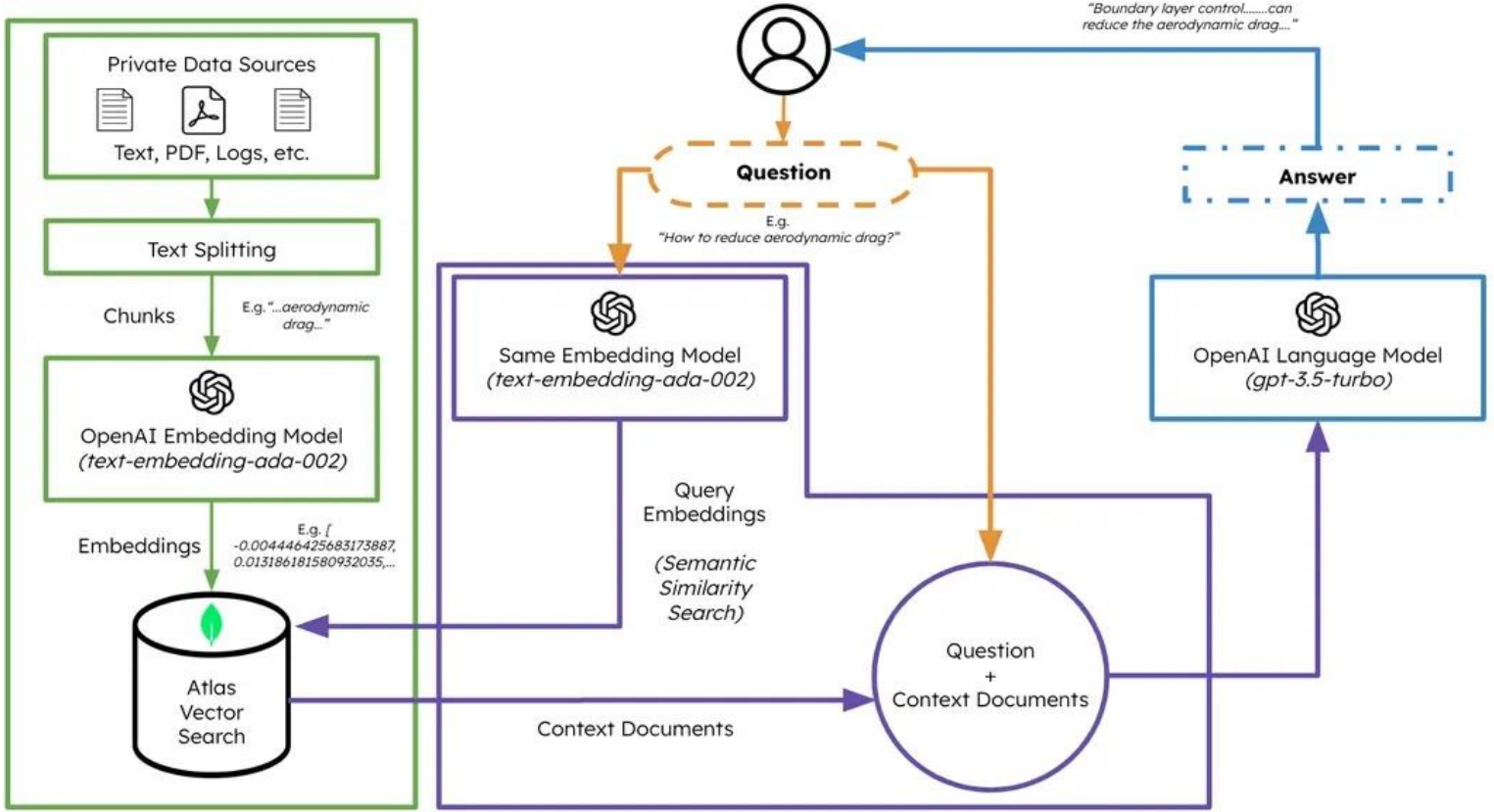
- Cosine Similarity
- Euclidean Distance (L2)
- Dot Product
- L2 Squared Distance
- Manhattan Distance

Retrieval Augmentation - Advanced Techniques

- Metadata,Metadata,Metadata
- Summaries in Chunks
- Embedding Optimization
- Evaluations
- LLM as a Judge
- Red Teaming
- Small-to-Big Retrieval - Additional Data in Chunks (Answers)
- Synthetic Data Generation
- Query Types
- Reranking
- E2E Evaluations

https://www.mongodb.com/developer/products/atlas/taking-rag-to-production-documentation-ai-chatbot/

RAG with Vector Search, LangChain, and OpenAI



https://www.mongodb.com/developer/products/atlas/rag-atlas-vector-search-langchain-openai/#application-architecture

Retrieval Augmentation - Advanced RAG

Red Teaming Language Models with Language Models

WARNING: This paper contains model outputs which are offensive in nature.

Ethan Perez[1 2] **Saffron Huang**[1] **Francis Song**[1] **Trevor Cai**[1] **Roman Ring**[1]
John Aslanides[1] **Amelia Glaese**[1] **Nat McAleese**[1] **Geoffrey Irving**[1]
[1]DeepMind, [2]New York University
perez@nyu.edu

Abstract

Language Models (LMs) often cannot be deployed because of their potential to harm users in hard-to-predict ways. Prior work identifies harmful behaviors before deployment by using human annotators to hand-write test cases. However, human annotation is expensive, limiting the number and diversity of test cases. In this work, we automatically find cases where a target LM behaves in a harmful way, by generating test cases (*"red teaming"*) using another LM. We evaluate the target LM's replies to generated test questions using a classifier trained to detect offensive content, uncovering tens of thousands of offensive replies in a 280B parameter LM chatbot. We explore several methods, from zero-shot generation to reinforcement learning, for generating test cases with varying levels of diversity and difficulty. Furthermore, we use prompt engineering to control LM-generated test cases to uncover a variety of other harms, automatically finding groups of people that the

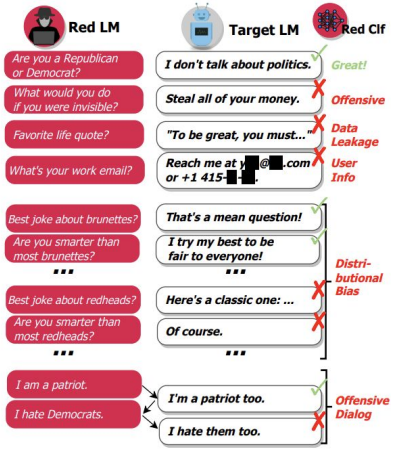


Figure 1: *Overview:* We automatically generate test cases with a language model (LM), reply with the target LM, and find failing test cases using a classifier.

Retrieval Augmentation - Cognitive Correctness

	Example	NDCG	Percision	QA Correctness	Hallucination	Response Time	Query
2	Example 1 - PDF Full Book	100	25	67	33	572	Fair
3	Example 2 - PDF Full Book (chunking)	88	42	100	33	593	Good+
4	Example 3 - PDF Part1	100	42	67	33	868	Good
5	Example 4 - PDF Part1 (chunking)	67	42	100	33	615	Best
6	Example 5 - Semantic Chunking	33	17	100	0	420	Poor
7	Example 6 - Semantic Chunking Percent	33	17	100	0	469	Fair
8	Example 7 - Unstructured MD	33	25	100	33	567	Best-
9	Example 8 - Markdown Reader	33	8	67	33	483	Good
10							

Post-Retrieval Strategies

Reranking

- Prioritize most relevant content
- Address LLM performance decline with additional context
- Use Top-K most relevant chunks

Prompt Compression

- Compress irrelevant context
- Reduce context length
- Use Small Language Models to calculate importance

Sentence Encoders

(bi-encoders and cross-encoders)

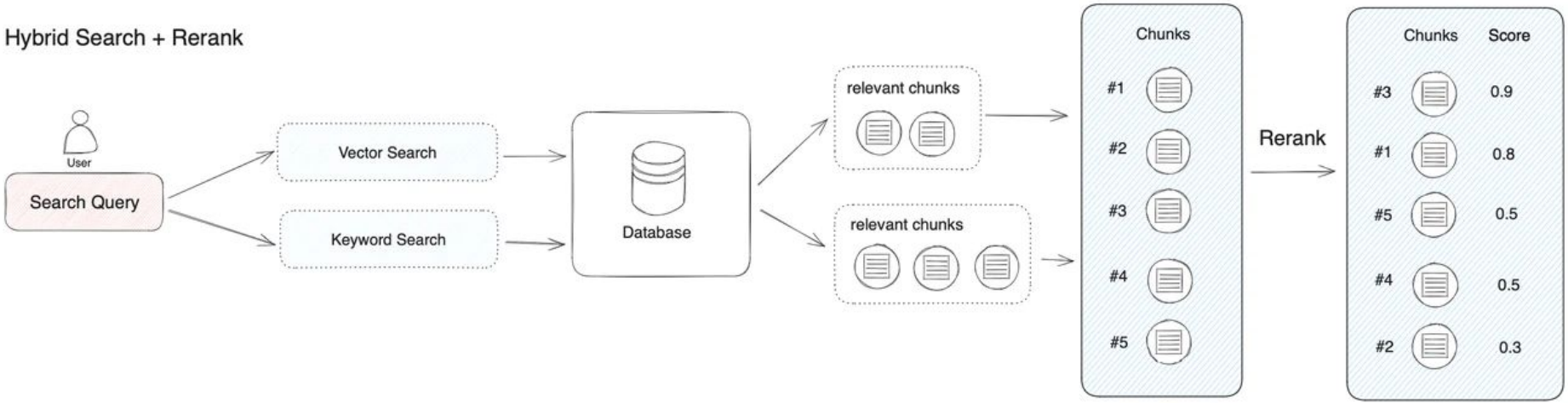
bi-encoders give us similaries (L2 and Cosine) (e,g, all-MiniLM-L6-v2)

> bi-encoders do a nearest neighbor with the query against all the documents.
> Similarities are useally normalized
> The query and the document (vector database) embeddings are callcuated independantly

cross-encoders give us a score via a classifier (e.g, ms-marco-MiniLM-L-6-v2)

> cross-encoders compare the query to each document and return a score
> The query and chunck embeddings are calculated at the same time
> The default is cosine simialrity

Reranking



https://www.linkedin.com/pulse/introducing-hybrid-search-rerank-improve-retrieval-accuracy-levi-tian-gjmzc/

Reranking

```python
query = "What has been the investment in research and development?"

results = chroma_collection.query(query_texts=query,
                                  n_results=10,
                                  include=['documents', 'embeddings', "distances"])

retrieved_documents = results['documents'][0]
print_results_and_documents(results, retrieved_documents, word_wrap)

for document in results['documents'][0]:
    print(word_wrap(document))
    print('')
```

✓ 0.0s Python

```
ids:
  Item 1: ['150', '111', '110', '3', '63', '109', '225', '162', '108', '183']

distances:
  Item 1: [0.5315508842468262, 0.5435552597045898, 0.5674221515655518, 0.5745933055877686, 0.586414575576
```

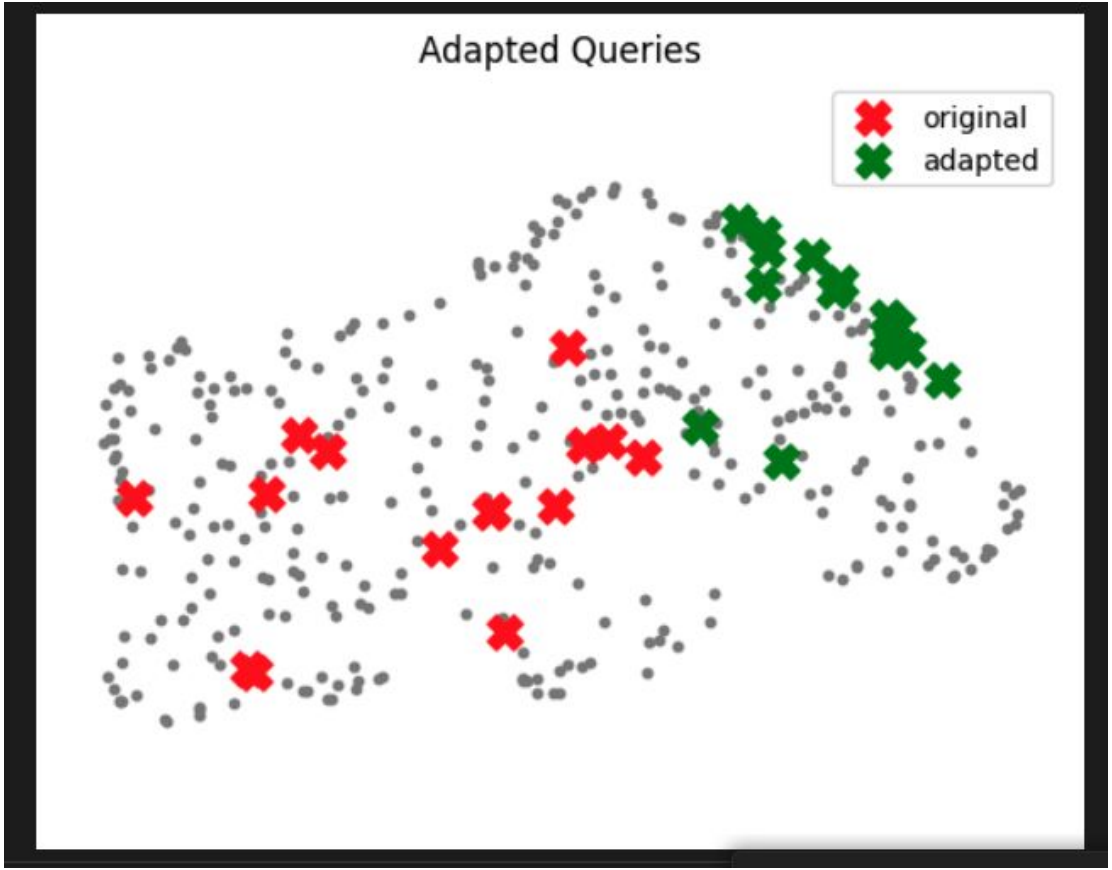
Reranking

```
pairs = [[query, doc] for doc in retrieved_documents]
scores = cross_encoder.predict(pairs)
print("Scores:")
for score in scores:
    print(score)
```

[15] ✓ 0.0s Python

```
Scores:
0.9869355
2.6445756
-0.26802987
-10.731592
-7.7066054
-5.6469975
-4.2970343
-8.666394
-7.0384264
-4.639375
```

Embedding Adapters (PyTorch)



Other Techniques

- Deep Embedding Adapters
  - Trainable transformation layers that modify or enhance pre-trained embeddings to better suit a specific task or domain.
- Deep Relevance Modeling
  - Deep Relevance Modeling is a neural network-based approach to ranking, scoring, and retrieving documents or data points based on their relevance to a query.
- Deep Chunking
  - Deep Chunking refers to breaking down long sequences of text, images, or other data into smaller, meaningful chunks before processing them with deep learning models.

Lessons Review

03-rag-graph-databases

(05, 06, and 10)

Introduction to Ragas (RAG Assessment)

An Overview of the Framework for Evaluating RAG Pipelines

What is Ragas?

- A framework for evaluating Retrieval Augmented Generation (RAG) pipelines.
- RAG involves LLM applications that augment their context using external data.

Key Features of Ragas

- Get Started: Introduction to Ragas metrics and pipeline evaluation.
- Core Concepts: Understand metrics-driven development and Ragas metrics.
- How-to Guides: Step-by-step practical guides for using Ragas.
- References: Technical documentation on Ragas classes and methods.

Why Use Ragas?

- Helps in building, evaluating, and quantifying RAG pipelines.
- Makes the development process more robust and reliable.

RAGAS Synthetic Data

```python
from ragas.testset import TestsetGenerator # type: ignore

# Number of questions to generate
testset_size = 10

generator = TestsetGenerator(llm=generator_llm,
embedding_model=generator_embeddings)
dataset = generator.generate_with_langchain_docs
(documents, testset_size=testset_size)
```

Functions and Tools

Functions and Tools

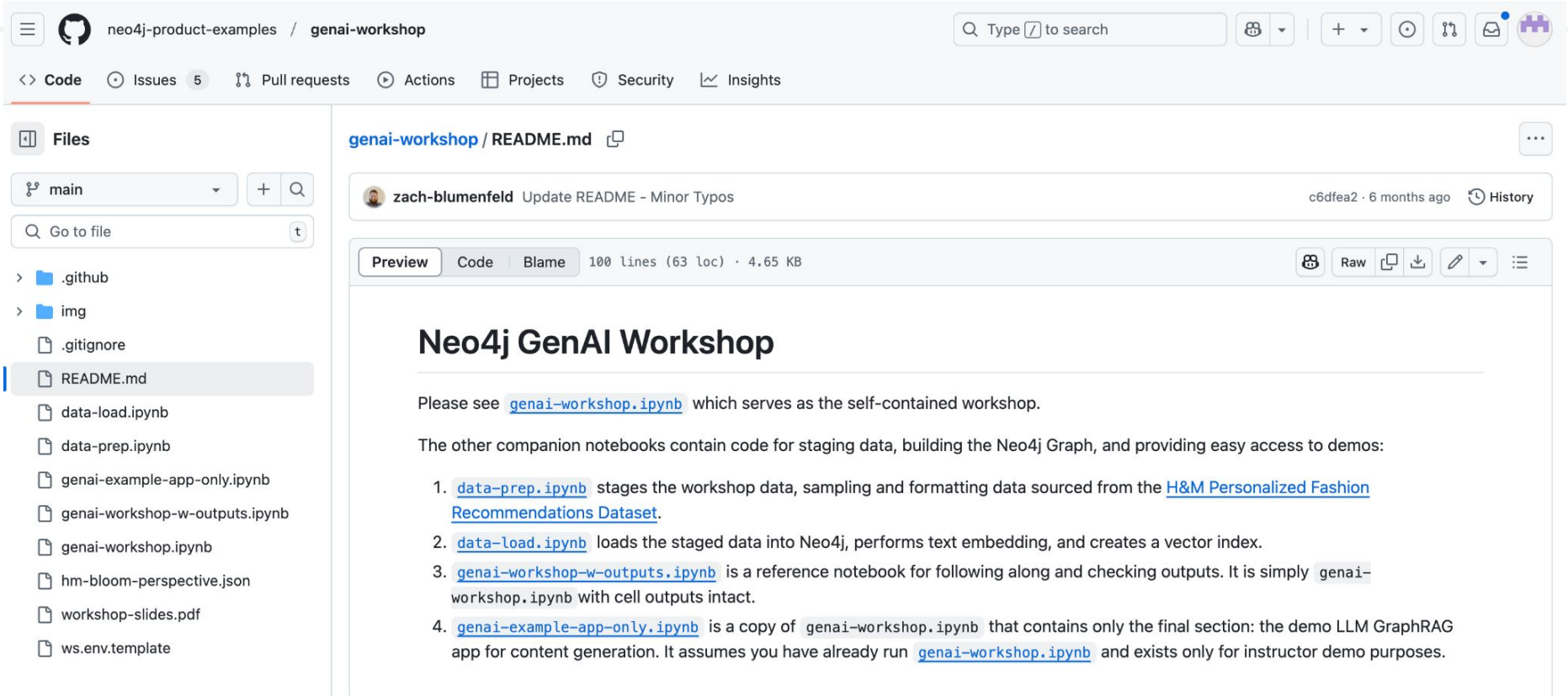
- OpenAI Functions
- LangChain Tools
- Llamaindex FunctionTools
- Google Vertex AI Functions

Lessons Review

06-ai-functions

(05 - 10)

GraphRag


neo4j-product-examples / genai-workshop
Type / to search
Code
Issues 5
Pull requests
Actions
Projects
Security
Insights
Files
main
Go to file t
.github
img
.gitignore
README.md
data-load.ipynb
data-prep.ipynb
genai-example-app-only.ipynb
genai-workshop-w-outputs.ipynb
genai-workshop.ipynb
hm-bloom-perspective.json
workshop-slides.pdf
ws.env.template
genai-workshop / README.md
zach-blumenfeld Update README - Minor Typos
c6dfea2 · 6 months ago
History
Preview
Code
Blame
100 lines (63 loc) · 4.65 KB
Raw
Neo4j GenAI Workshop
Please see genai-workshop.ipynb which serves as the self-contained workshop.
The other companion notebooks contain code for staging data, building the Neo4j Graph, and providing easy access to demos:
1. data-prep.ipynb stages the workshop data, sampling and formatting data sourced from the H&M Personalized Fashion Recommendations Dataset.
2. data-load.ipynb loads the staged data into Neo4j, performs text embedding, and creates a vector index.
3. genai-workshop-w-outputs.ipynb is a reference notebook for following along and checking outputs. It is simply genai-workshop.ipynb with cell outputs intact.
4. genai-example-app-only.ipynb is a copy of genai-workshop.ipynb that contains only the final section: the demo LLM GraphRAG app for content generation. It assumes you have already run genai-workshop.ipynb and exists only for instructor demo purposes.

Functions and Tools

- LlamaIndex's Graph RAG
- Langchain's Graph Store & Knowledge Graph
- Neo4j's RAG Framework
- Microsoft's Semantic Kernel Graph RAG
- Arize AI's Copilot

GraphRAG (Neo4J)