WELCOME

Insights from One Year of Customer Implementations of RAG Systems Using MongoDB Vector Search

Philip Eschenbacher
Solutions Architect

Agenda

What can i do with RAG and Semantic Search

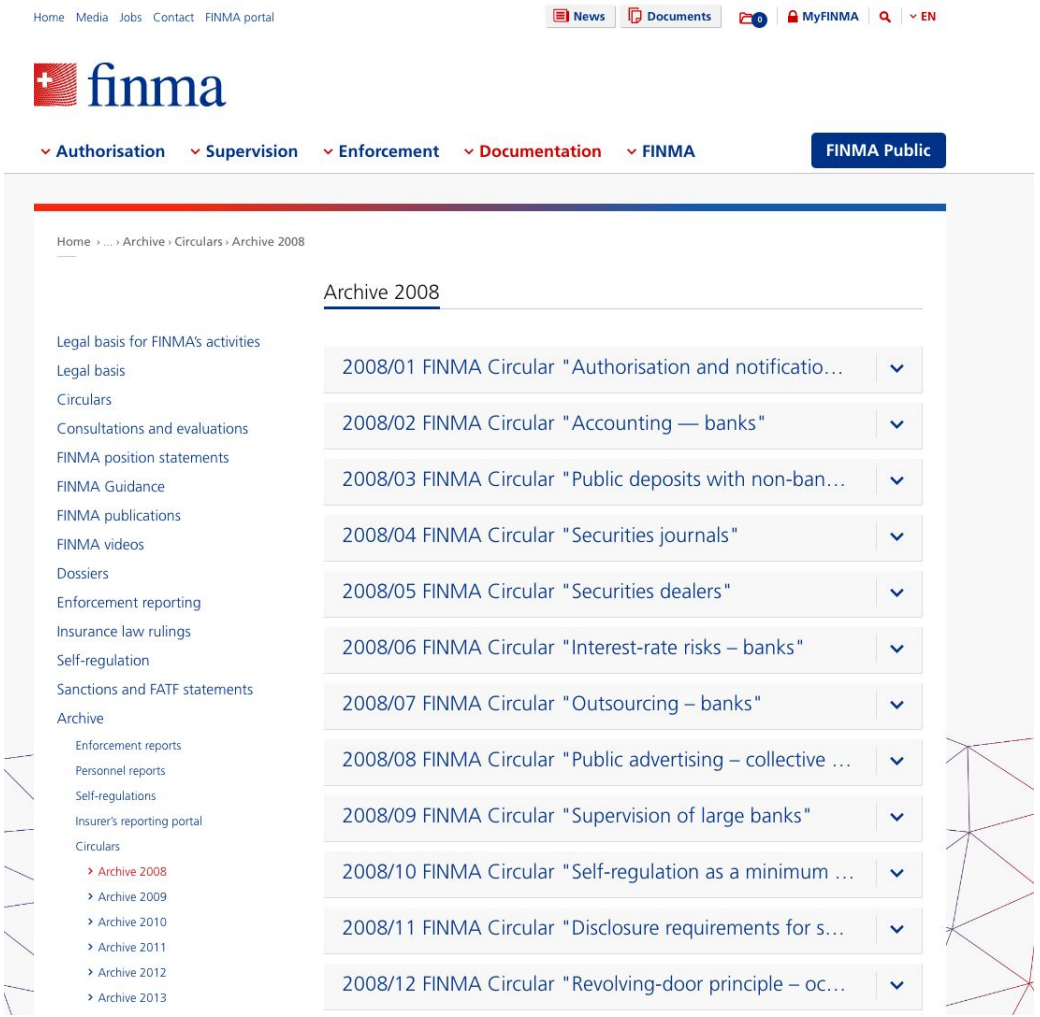
All about LLMs and Vectors

What is MongoDB Atlas

How to improve the R-Part

Q&A

Let’s start with a Demo


Home Media Jobs Contact FINMA portal
News
Documents
MyFINMA
EN
finma
Authorisation
Supervision
Enforcement
Documentation
FINMA
FINMA Public
Home › ... › Archive › Circulars › Archive 2008
Archive 2008
Legal basis for FINMA's activities
Legal basis
Circulars
Consultations and evaluations
FINMA position statements
FINMA Guidance
FINMA publications
FINMA videos
Dossiers
Enforcement reporting
Insurance law rulings
Self-regulation
Sanctions and FATF statements
Archive
Enforcement reports
Personnel reports
Self-regulations
Insurer's reporting portal
Circulars
Archive 2008
Archive 2009
Archive 2010
Archive 2011
Archive 2012
Archive 2013
2008/01 FINMA Circular "Authorisation and notificatio…
2008/02 FINMA Circular "Accounting — banks"
2008/03 FINMA Circular "Public deposits with non-ban…
2008/04 FINMA Circular "Securities journals"
2008/05 FINMA Circular "Securities dealers"
2008/06 FINMA Circular "Interest-rate risks – banks"
2008/07 FINMA Circular "Outsourcing – banks"
2008/08 FINMA Circular "Public advertising – collective …
2008/09 FINMA Circular "Supervision of large banks"
2008/10 FINMA Circular "Self-regulation as a minimum …
2008/11 FINMA Circular "Disclosure requirements for s…
2008/12 FINMA Circular "Revolving-door principle – oc…

Search

can i move my data to the cloud?

SEARCH

rs-18-03-letzte-aenderung-20191031.pdf

OPEN

rs-18-03-letzte-aenderung-20191031.pdf

OPEN

rs-18-03-letzte-aenderung-20191031.pdf

OPEN

finma-rs-2018-03--20170921.pdf

OPEN

finma-rs-2018-03--20170921.pdf

OPEN

finma-rs-2018-03--20170921.pdf

OPEN

rs-13-07-letzte-aenderung-20201104.pdf

OPEN

rs-13-07-letzte-aenderung-20201104.pdf

OPEN

rs-13-07-letzte-aenderung-20201104.pdf

OPEN

rs-2008-12-letzte-aenderung-20151203.pdf

OPEN

e.foresight Home Explorer Radar Matrix Search

Explorer

Filter

Active Filters

There is no active filter.

Current Elements: 3315

Select Sort View

New

Signal

Französischer digitaler Vermögensverwalter Rami…

Der digital-hybride Vermögensverwalter Ramify hat eine neue Series-A-Finanzierungsrunde zum weiteren Ausbau der Kundscha…

2024-06-25 Investing

New

Player

Ramify

Ramify is a wealth management platform that provides ETF, SCPI, and private equity recommendations, management, and selection.

2024-06-25

New

Signal

Der Bundesrat setzt derzeit keine regulatorischen…

Der Bundesrat erachtet die Fortschritte im Bereich Open Finance derzeit als ausreichend und sieht keinen Handlungsbedarf für die…

2024-06-25 Customer Interac…

New

New

New

Support

as a CEO of a small swiss retail bank, should i consider leaning into crypto?

As a CEO of a small Swiss retail bank, considering crypto can be a strategic move, but it comes with both opportunities and risks. Here are some points to consider:

Opportunities:

1. **Innovation and Differentiation:** Embracing crypto can set your bank apart, attracting tech-savvy customers and potentially opening new revenue streams.
2. **Customer Demand:** There's a growing interest in crypto assets, and offering related services can meet this demand.
3. **Blockchain Technology:** Leveraging blockchain for transactions can improve efficiency and security.

as a CEO of a small swiss retail bank, should i consider leaning into crypto?

Answer:
As a CEO of a small Swiss retail bank, considering the adoption of crypto and digital assets can be strategic, given the current landscape and future potential. Switzerland is increasingly becoming a hub for digital assets, with several banks like Swissquote, Sygnum, and PostFinance already engaging in this space. These institutions have leveraged both early adoption and strategic timing to build successful business models around digital assets.

Here are five reasons to consider leaning into crypto:

1. **Market Trends**: Digital assets are moving towards mainstream adoption, making it easier for banks to implement related services.
2. **Regulatory Environment**: Switzerland has established a favorable legal framework with the Distributed Ledger Technology (DLT) Act, which supports innovation and ensures financial integrity.
3. **Competitive Edge**: Offering digital asset services can help differentiate your bank in a competitive market, at least until they become standard offerings.
4. **Partnership Opportunities**: There are multiple strategies and specialized service providers available for partnerships, making it easier to enter the market.
5. **Local Ecosystem**: There is a strong ecosystem of local startups and firms driving innovation in digital assets, which your bank can tap into.

However, adopting crypto is not without challenges:

1. **Regulatory Compliance**: The introduction of the EU's MiCA regulation and other compliance requirements need careful navigation.
2. **Operational Risks**: This includes addressing potential issues around privacy, financial, and operational risks.

Our answers to your questions.

- Travelcards, tickets, apps and services
- My travelcards
- My tickets
- Refunds and compensation
- Invoices, payment methods and travel without a valid ticket
- My profile

SBB CFF FFS MongoDB.

Ask me anything

Your question

What can i do with my swisspass?

ASK

The answer:

What can I do with my SwissPass?

- Manage personal details conveniently online in your customer account, including changing your name and address.
- Check the validity of your GA Travelcard or Half Fare Travelcard.
- Cancel travelcards online and set up reminders for timely cancellations.
- View an overview of pending and paid invoices.
- Access activated partner services such as Mobility Carsharing.
- Benefit from unlimited travel on trains, Postbuses, and boats with the Swiss Travel Pass.
- Enjoy unlimited use of public transport in over 90 Swiss cities.
- Get free admission to over 500 museums with the Swiss Museum Pass.
- Experience included mountain excursions to Rigi, Stanserhorn, and Stoos.
- Receive discounts of up to 50% on other mountain excursions and up to 30% on leisure travel offers.

Sources:

- What information can I find in my SwissPass customer account?
- What are the advantages of the Swiss Travel Pass?

SBB CFF FFS MongoDB.

Ask me anything

Your question

I'm angry. What would be the best way to damage the train i'm sitting in?

ASK

The answer:

Summary:

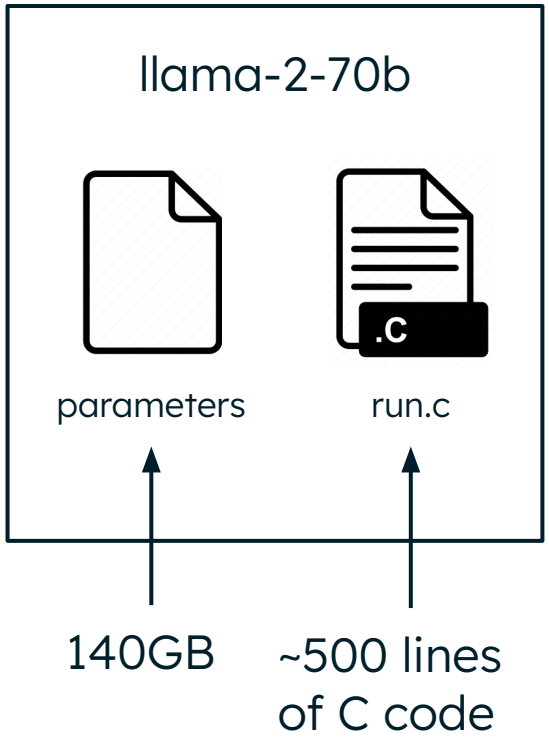
- If you are traveling with a bicycle and have reserved a space, but the operating train has no bike spaces due to a change, you have the same rights and entitlement to compensation as in the event of a delay or cancellation of a train.
- If your booked train has been impacted by a strike, you can usually get a refund or rebook for a train running on the same day for free. Contact the office that issued your ticket for assistance.
- In case of a suspension of service on the line, check if your train is being rerouted, if your ticket is valid on another connection, and inquire about the possibility of a free refund from the office that issued your ticket.

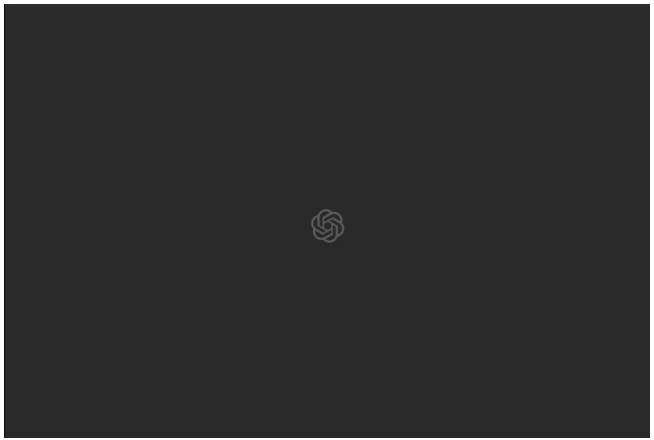
Sources:

- I am travelling with my bicycle and have reserved a space. Due to a change, the operating train now has no bike spaces and I am unable to travel with my bike.
- The train I booked has been impacted by a strike or a suspension of service on the line. What do I need to do?

What is an LLM

Large Language Model (LLM)





Source: Andrej Karpathy Intro: Large Language Models

Training the model is more involved

Think of it like compressing the internet

Chunk of the internet,
~10TB of text

6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

parameters.zip

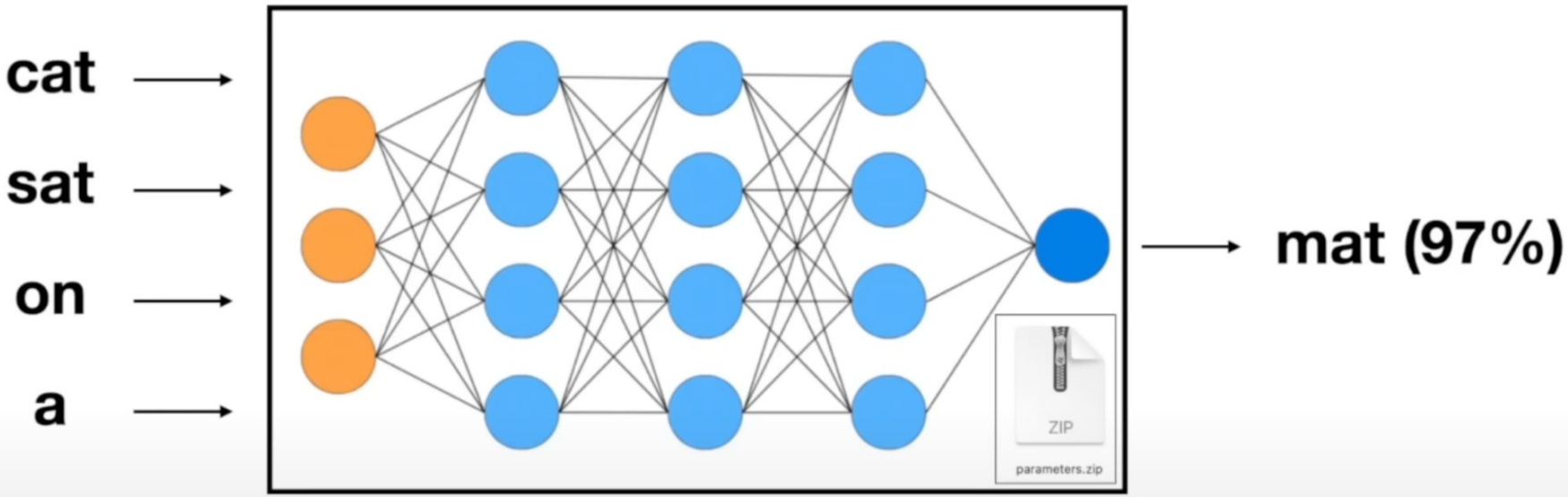
~140GB file

Source: Andrej Karpathy Intro: Large Language Models

*numbers for Llama 2 70B

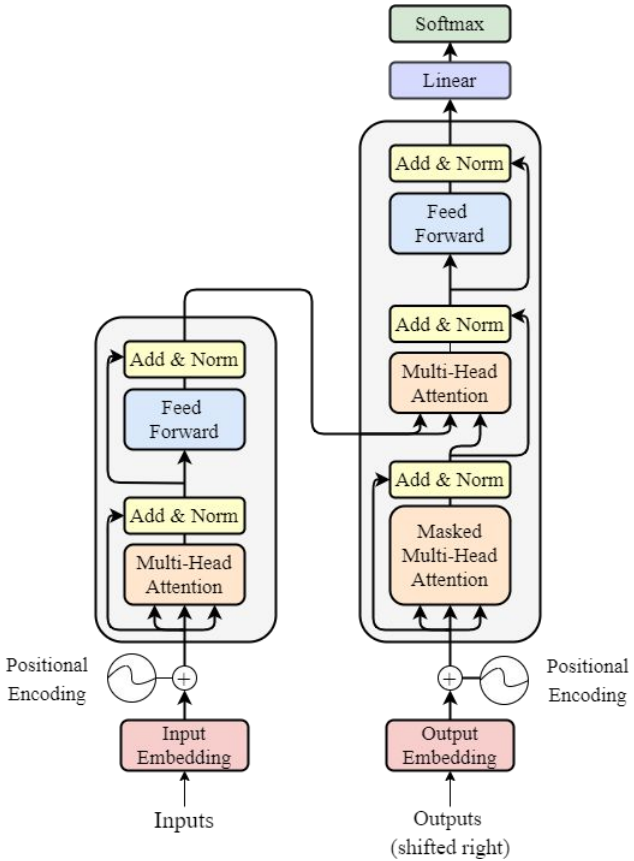
Neural Network

Predicts the next word in the sequence



Source: Andrej Karpathy Intro: Large Language Models

How does it work?



Little is known in full detail…

- Billions of parameters are dispersed through the network
- We know how to iteratively adjust them to make it better at prediction
- We can measure that this works, but we don't really know how the billions of parameters collaborate to do it

They build and maintain some kind of knowledge database, but it is a bit strange and imperfect

Viral example: "reversal curse"

Q: "Who is Tom Cruise's mother?"
A: Mary Lee Pfeiffer

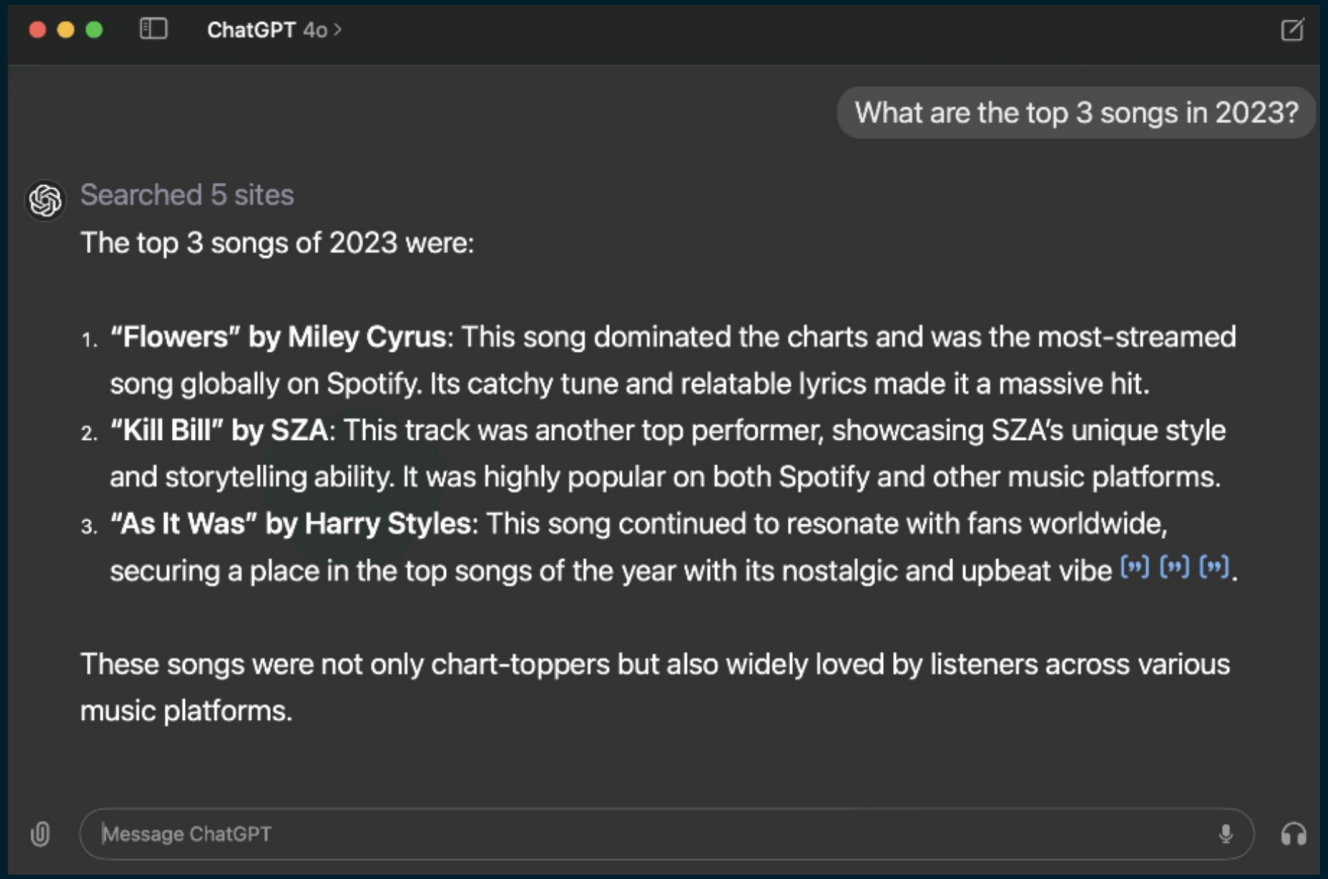
Q: "Who is Mary Lee Pfeiffer's son?"
A: I don't know

Andrej Karpathy's intro to LLMs

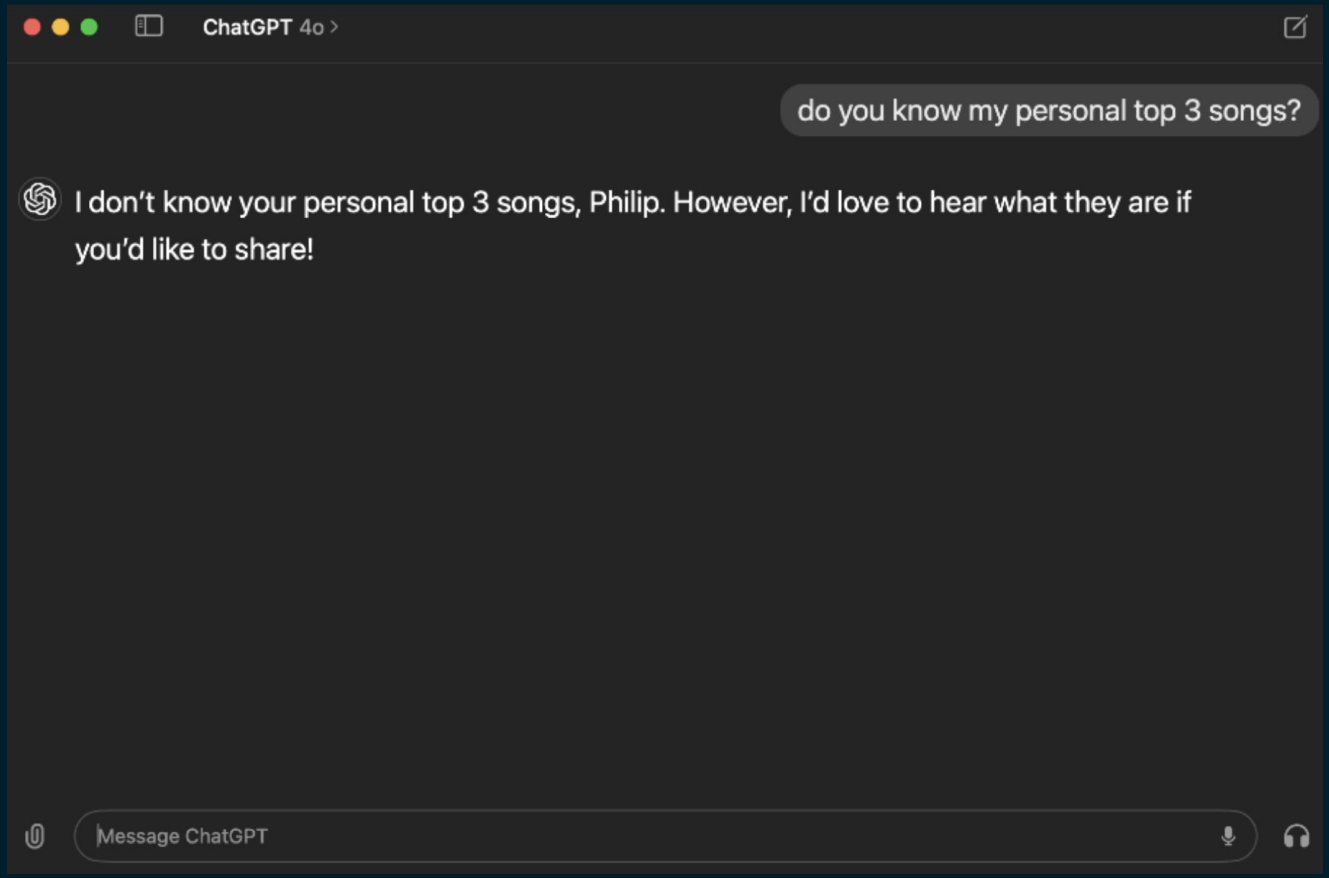
The busy person's intro to LLMs

Andrej Karpathy

AI Chat Models know the world



But they don’t know me

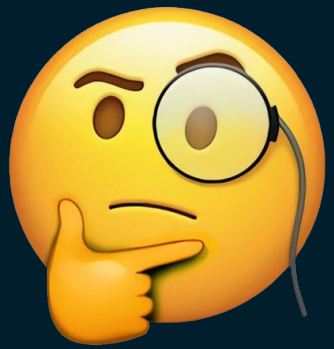


What if I could augment it with my data?

Which data should we give to Chat Models?

Well, data relevant to the question.

OK, but how will we know which data is relevant?



Database vs Search Engine

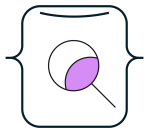
Database

Queries ask for **exact matches** to a specific condition then perform an operation (CRUD)

What is **user_ID 4810**'s account balance ?

Database returns a specific record of data (**exact match** on **user_ID 4810**)

Search Engine



Queries **don't ask** for **a specific piece of data.** You ask the system to look through all of its contents and return what it thinks is the **most relevant** to your query or search term

"What was that movie where Harrison Ford plays an archaeologist?"

Search Engine returns a **list of results** based on **relevance**

What is a Vector?

A vector is...

A series of numbers, like:

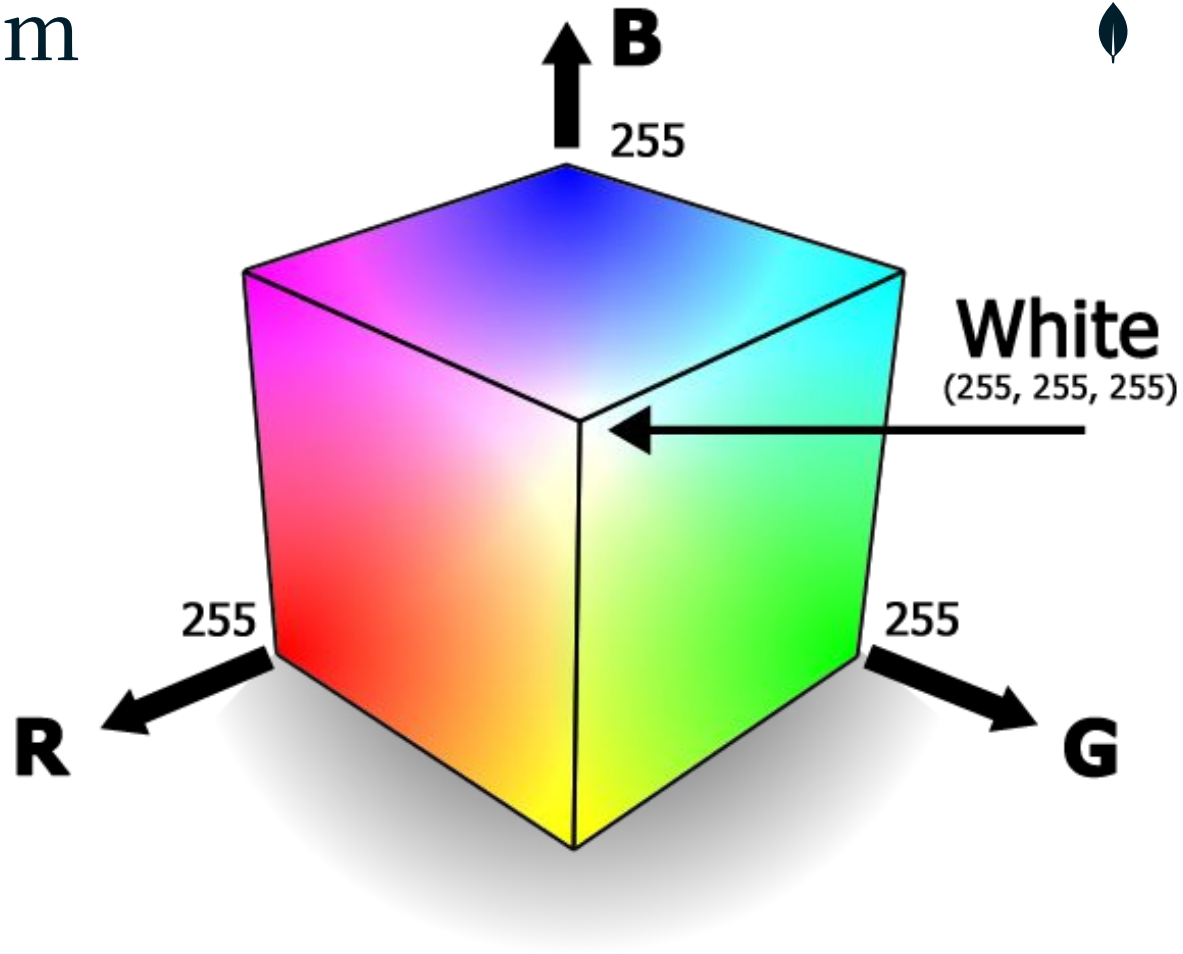
- [0, 1, 2]
- [0.4, 0.3, 0.7, 0.1, 0.8]
- [0.545, 0.236, 0.567, ...
- Or much longer!

Vectors can...

- Capture meaning!
- If this sounds odd, don't worry
- Many have used vectors already without realising

The RGB System

- [0, 0, 0]
- [50, 100, 100]
- [228, 102, 12]
- [190, 12, 228]



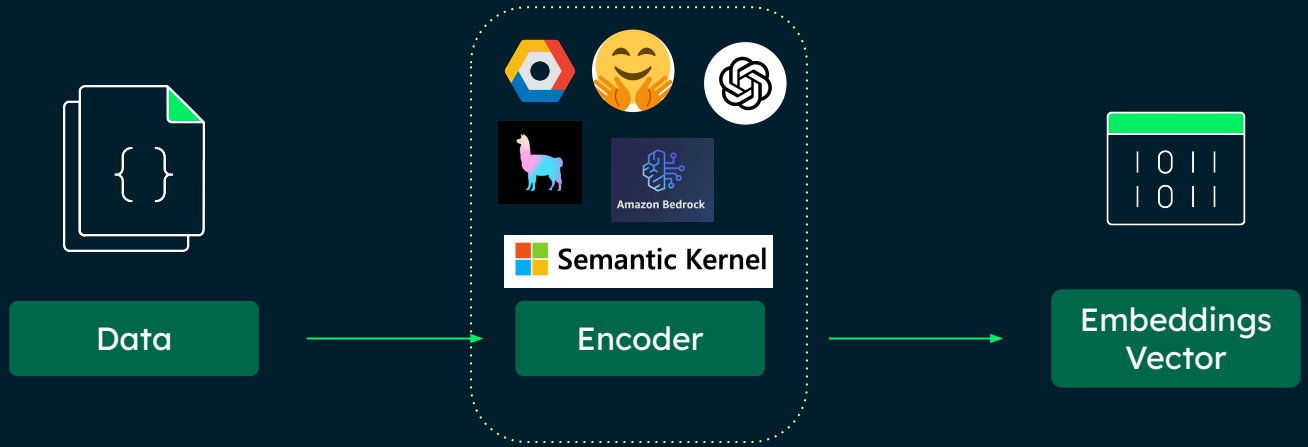
What is a Vector Embedding?

A vector embedding is a vector that represents meaning

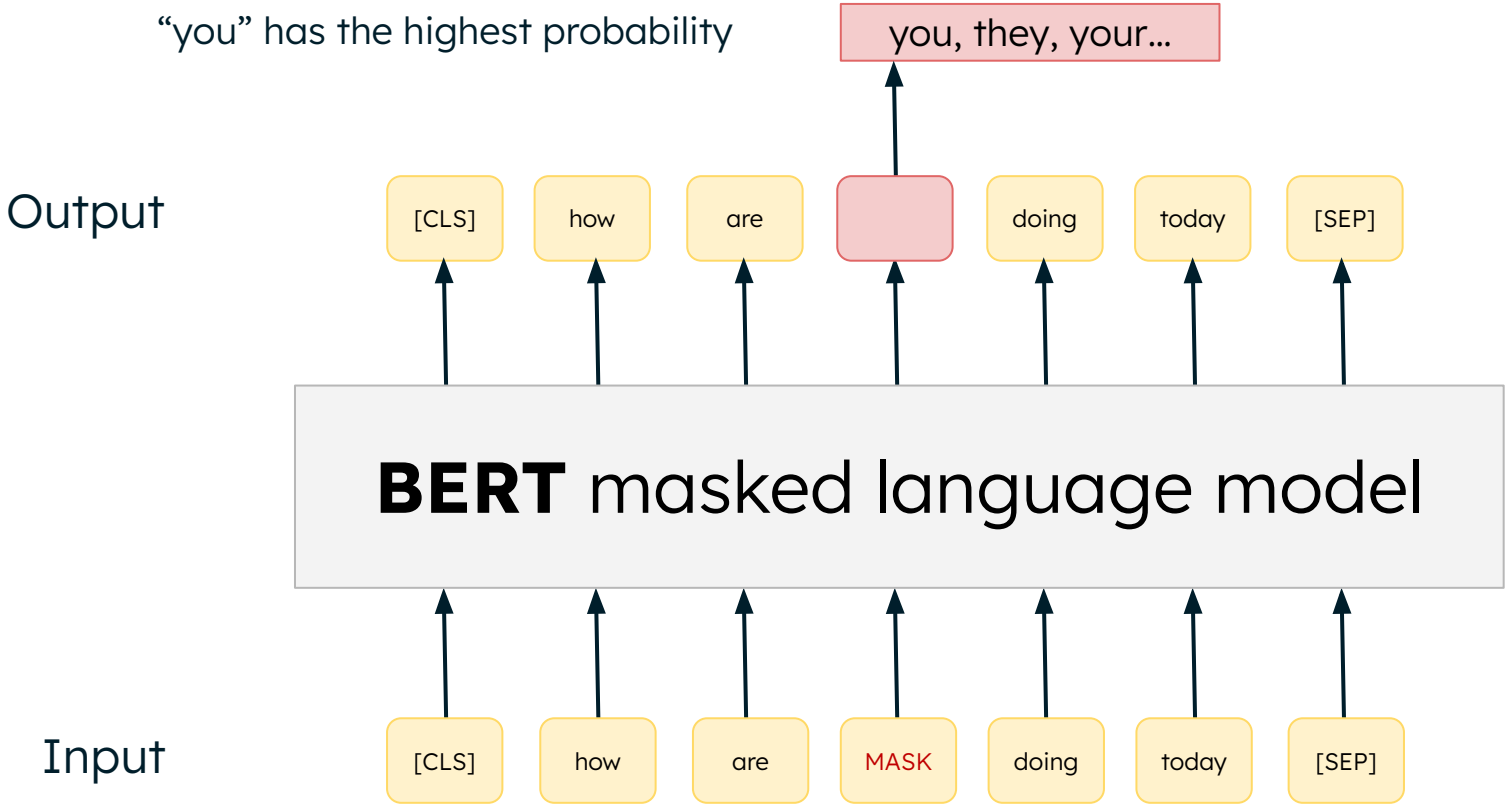
“I listen to Beyonce”

```
[0.030539153,
-0.035179794,
-0.037336048,
…,
-0.013089883,
-0.0035097762]
```

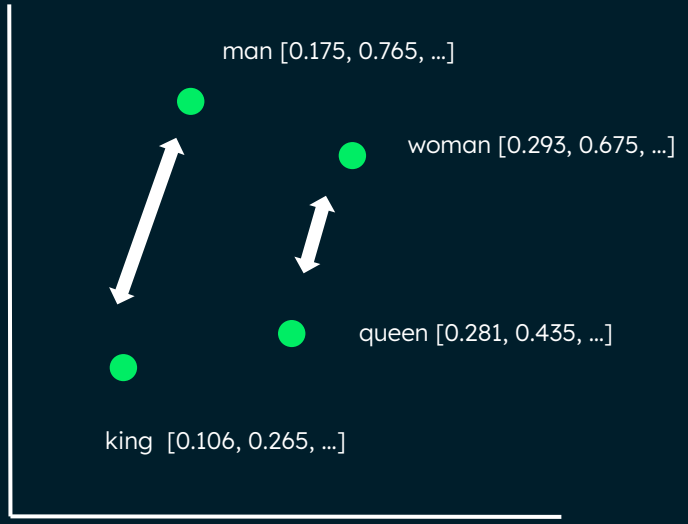
They are produced by sending data through an embedding model

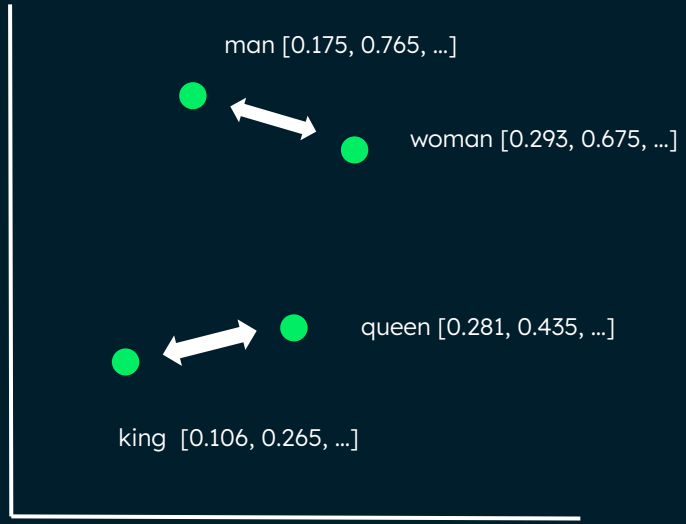


Embedding Models

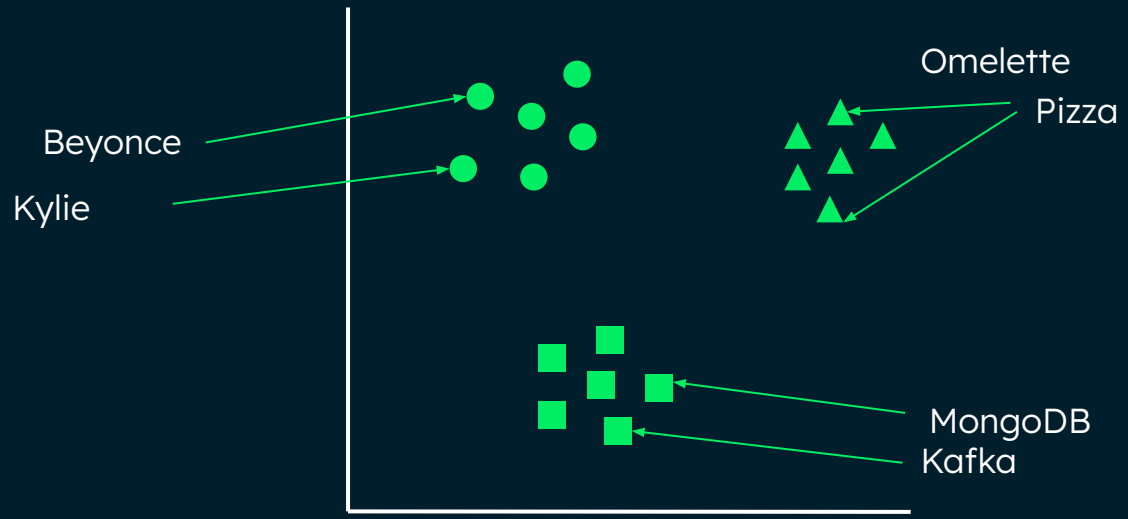


When these vectors are "near" to each other, they are similar in some respects

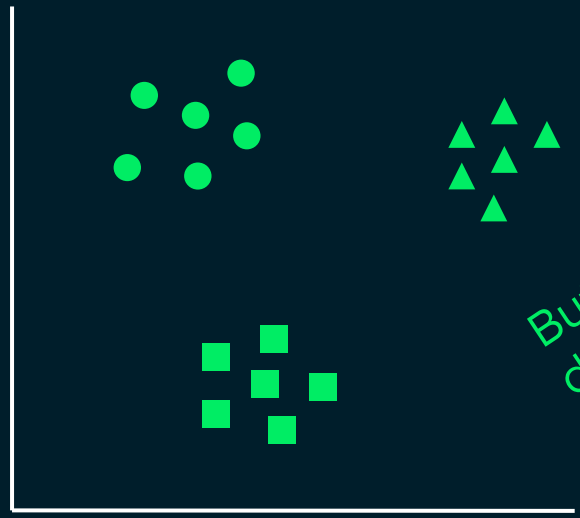




Data represented as vectors creates clusters of semantically similar data



So we need to store and search for the closest vectors!



But in a space with 1536-4096 dimensions 😅

MongoDB

What is MongoDB Atlas?

Developer Data Platform

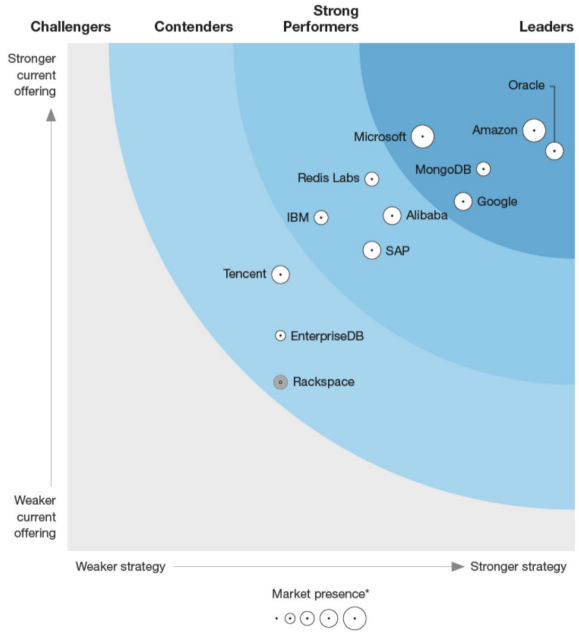
Document Model & Unified API

Multi-Cloud Scale, Resilience, Performance, & Security

2 million+ deployments
40,000+ customers

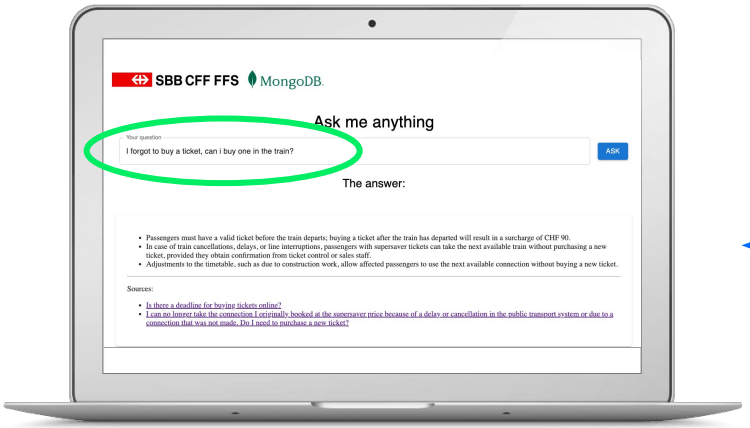
Telefónica · FTD · SEGA · ULTA BEAUTY
otto group · eharmony · 7-ELEVEN · AutoTrader
TICKETEK · invision · ThermoFisher SCIENTIFIC · SQUARE ENIX
Nationwide · Royal Caribbean INTERNATIONAL · apervita · MICHELIN
breuninger · coinbase · ENPHASE · Buffer

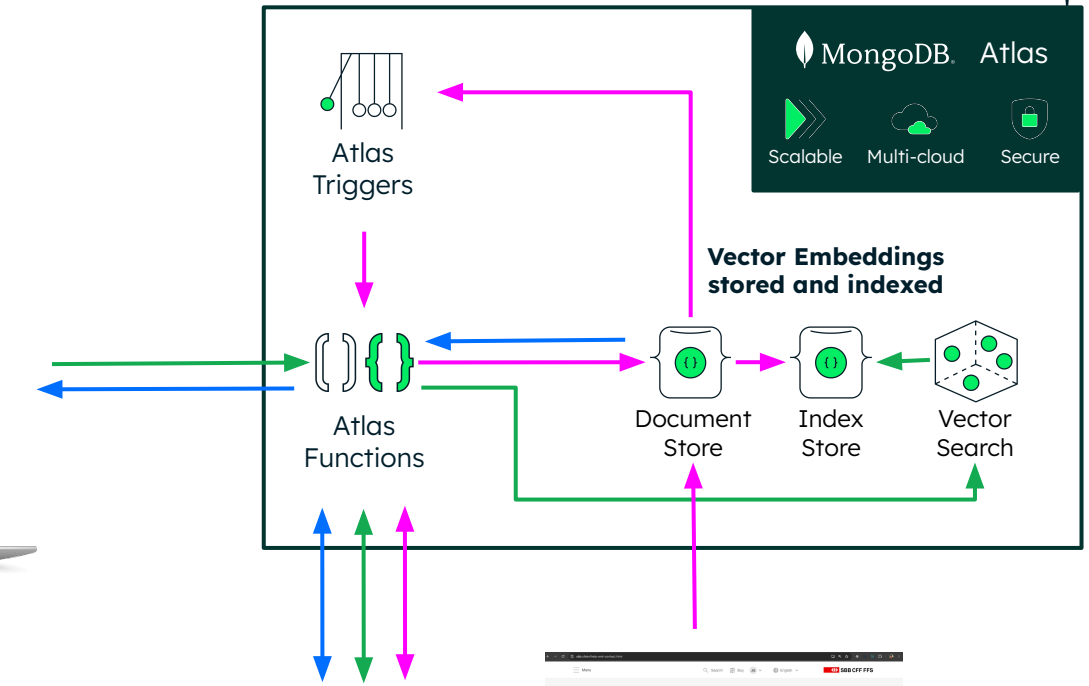
MongoDB named **a leader**
in The Forrester Wave™:
Database-as-a-Service, Q2 2019

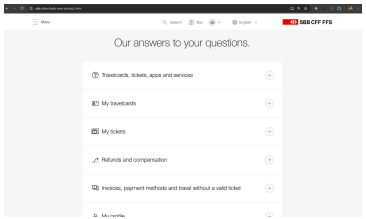


Atlas Vector Search

RAG Architecture with Atlas



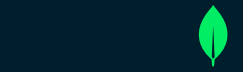




New or updated content

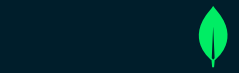
Let’s have a look at the Demo again

In the database it looks like this…




```
{
  "_id": "65b2d036cf3899722818a2bd",
  "question": "What tickets can I purchase or order online?",
  "answer": "Tickets and travelcards for Switzerland….",
  "url": "https://www.sbb.ch/en/help-and-contact/tickets/ch/tickets.html",
  "vector_embedding": [
    0.017581753,
    0.010902251,
    0.0032468897,
    -0.043244075,
    …
    -0.012707346,
    -0.033599526
  ]
}
```

...create the Vector Index



```
{
  "mappings": {
    "dynamic": false,
    "fields": {
      "vector_embedding": {
        "dimensions": 1536,
        "similarity": "euclidean",
        "type": "knnVector"
      }
    }
  }
}
```

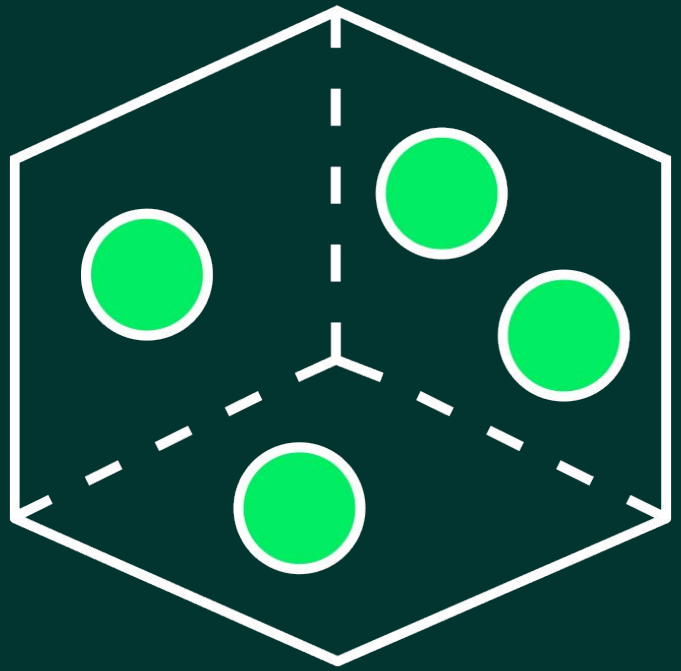
...and query your data with Vectors

```
{"$vectorSearch": {
  "queryVector":[0.017581753,0.010902251,...],
     "path": "vector_embedding",
     "numCandidates": 100,
     "limit": 3,
     "index": "default"
}}
```

Let’s have a look at the prompt…

```
Given the following information from multiple sources, provide a short
summary that integrates these insights, focusing on the most relevant
findings. Each piece of information is presented with its source and
relevance score:

${combinedContext}

Answer the key points from the above information to the following
question: ${question}

Can you please structure your answer in markdown as an ordered or
unordered list without direct attached source informations

Please provide a list of all sources as clickable links in an unordered
list at the end of your answer providing the URL Name as ancor and the URL
Source as target
```

One Year of MongoDB Vector Search



The Right Embedding Model


I HAVE A VERY PARTICULAR SET OF SKILLS.
I WILL FIND YOUR QUESTIONS,
AND I WILL ANSWER THEM.

The Right Embedding Model

English | Chinese | French | Polish

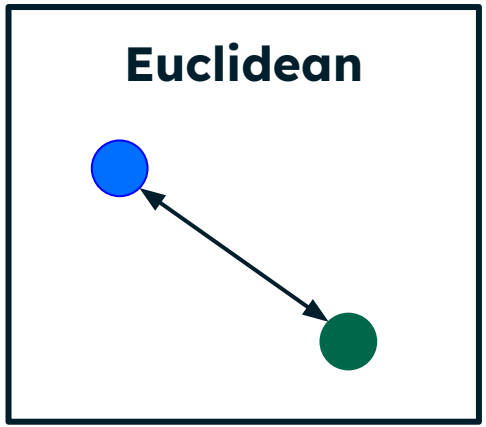
Overall MTEB English leaderboard 🔮

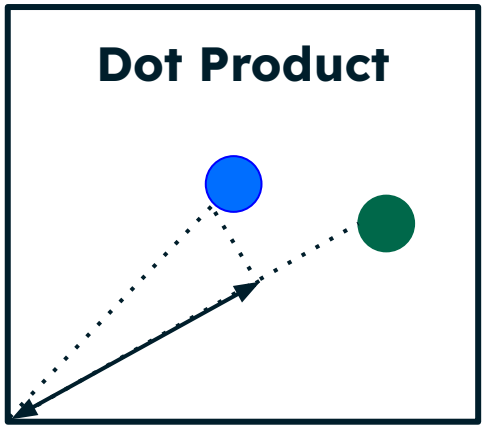
- **Metric:** Various, refer to task tabs
- **Languages:** English

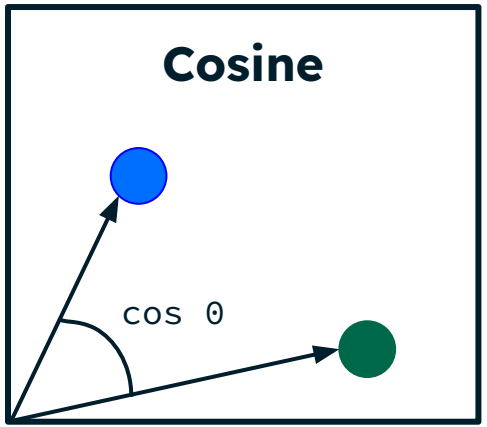
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)
1	voyage-large-2-instruct			1024	16000	68.28	81.49	53.35	89.24	60.09
2	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54	60.64
3	gte-Qwen1.5-7B-instruct					67.34	79.6	55.83	87.38	60.13
4	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42	86.87	58.24
5	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61	87.16	60.49
6	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26	88.34	60.21
7	google-gecko.text-embedding-p	1200	4.47	768	2048	66.31	81.17	47.48	87.61	58.9
8	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14	84.97	59.8
9	gte-large-en-v1.5	434	1.62	1024	8192	65.39	77.75	47.96	84.53	58.5
10	LLM2Vec-Mistral-supervised	7111	26.49	4096	32768	64.8	76.63	45.54	87.99	58.42
11	echo-mistral-7b-instruct-last	7111	26.49	4096	32768	64.68	77.43	46.32	87.34	58.14

Use the right Similarity

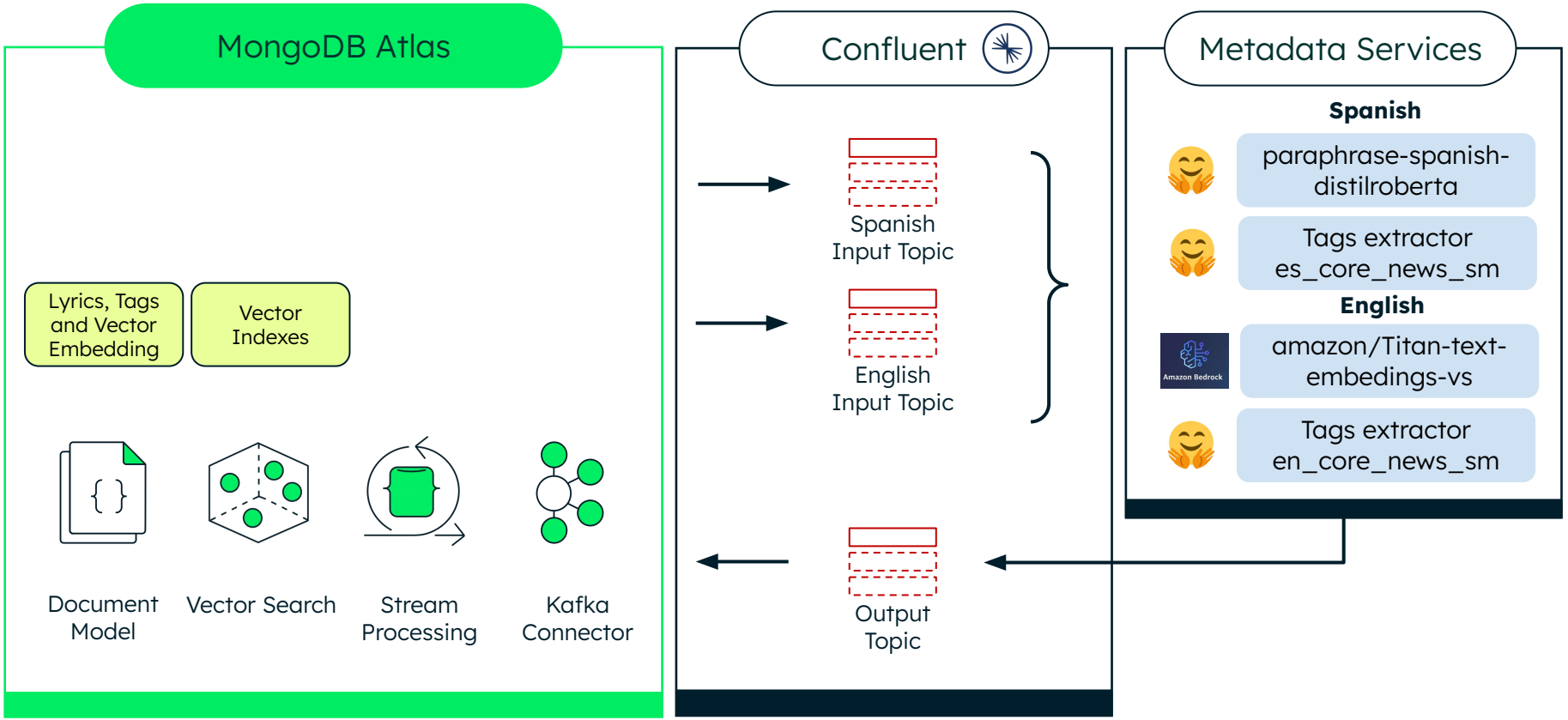
Use the right Similarity



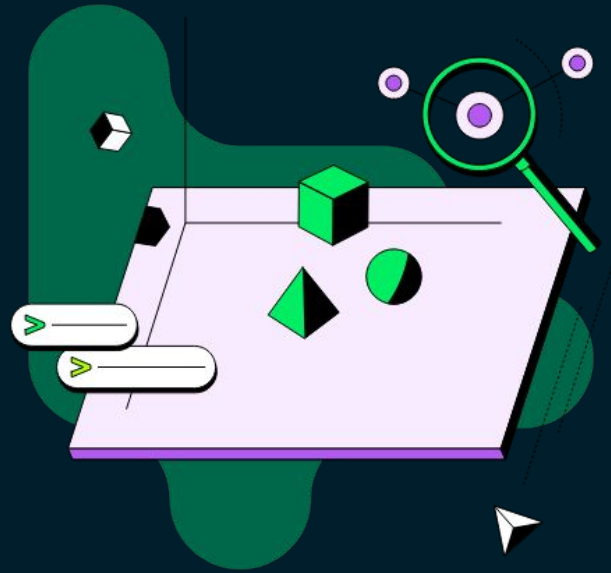




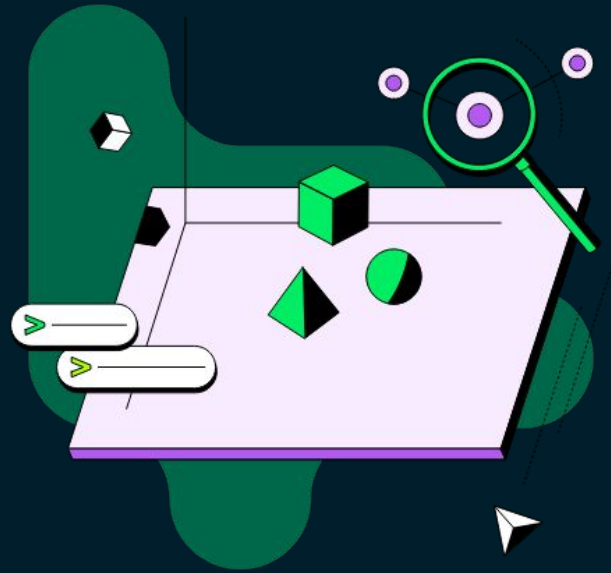
Fast Vector Updates



Building continuously updating RAG applications

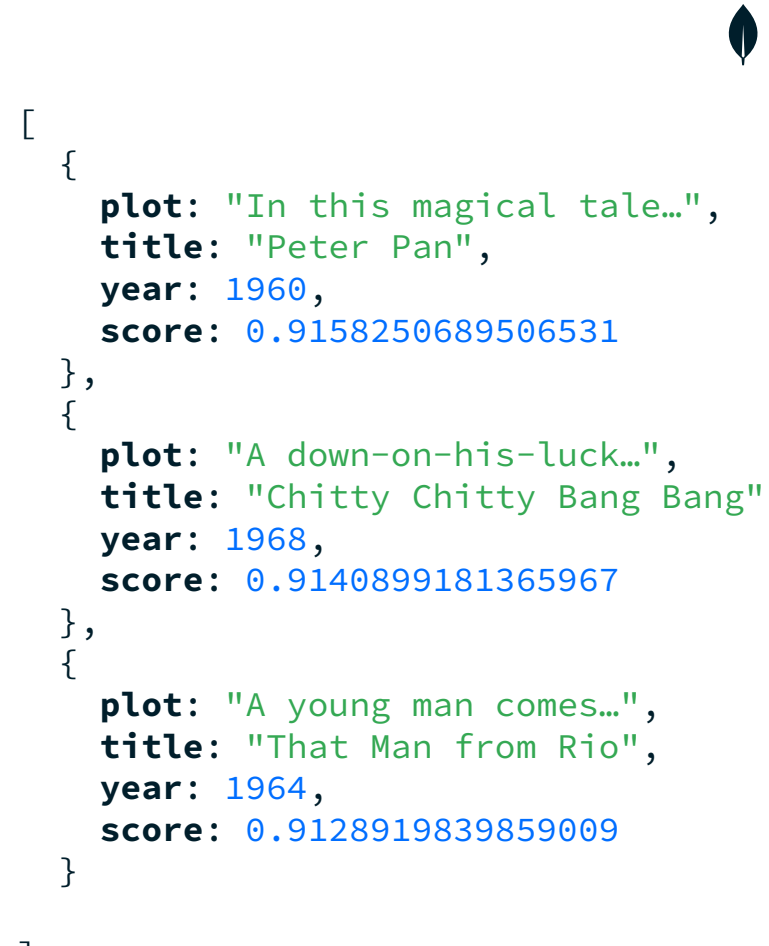


A blueprint for a RAG application



Filtering

Filtering



```
[
  {
    $vectorSearch: {
      index: "vector_index",
      path: "plot_embedding",
      filter: {
        $and: [
          {
            year: {  $gt: 1955  }
          },
          {
            year: {  $lt: 1975 }
          }
        ]
      },
      queryVector: [0.02,-0.02,...],
      numCandidates: 150,
      limit: 3
    }
  }
]
```

```
[
  {
    plot: "In this magical tale…",
    title: "Peter Pan",
    year: 1960,
    score: 0.9158250689506531
  },
  {
    plot: "A down-on-his-luck…",
    title: "Chitty Chitty Bang Bang",
    year: 1968,
    score: 0.9140899181365967
  },
  {
    plot: "A young man comes…",
    title: "That Man from Rio",
    year: 1964,
    score: 0.9128919839859009
  }

]
```

Hybrid Search

$vectorSearch

Vector Search Index

finma

I. Purpose

This circular defines the supervisory requirements applicable to outsourcing solutions at banks, securities dealers and insurance companies in terms of appropriate organisation and risk limitation. 1

II. Definition of terms

A company is understood to mean an institution (bank, securities dealer and insurance company) that falls within this circular's scope of application. 2

Outsourcing within the meaning of this circular occurs when a company mandates a service provider to perform all or part of a function that is significant to the company's business activities independently and on an ongoing basis. 3

Significant functions are those that have a material effect on compliance with the aims and regulations of financial market legislation. 4

III. Scope of application

This circular applies to: 5

- banks and securities dealers with a registered office in Switzerland as well as Swiss branches of foreign banks and securities dealers; 6
- insurance companies with their registered office in Switzerland and branches of foreign insurance companies requiring authorisation to commence business operations under Articles 3 and 6 Insurance Supervision Act (ISA) (initial authorisation) or authorisation for individual elements of the business plan under Article 4 in conjunction with Article 5 ISA (authorisation for changes).
- The requirements are to be applied taking into account the institution's size, complexity, structure and risk profile. 6.1*

IV. Admissibility

A. Joint provisions

Subject to the exceptions outlined below (Margin nos. 8–13), all significant functions may be outsourced. 7

Direction, supervision and control by the supreme governing body, central executive management functions and functions that involve strategic decision-making may not be outsourced, nor may decisions concerning the commencement and termination of business relationships. 8

Companies in supervisory categories 1–3 have an autonomous control body in the form of a separate risk control and compliance function. For companies in supervisory categories 4 and 5, it is sufficient for a member of executive management to be assigned 9

3/8

finma

$search

Full Text Search Index

$group

Hybrid Search

```
{
    "_id": ObjectID("662043cfb084403cdcf5210a"),
    "paragraph_embedding": [0.43, 0.57, ...],
    "page_number": 14,
},
{
    "_id": ObjectID("662043cfb084403cdcf5210b"),
    "full_page_content": "Writing computer software is one
        of the purest creative activities in the history of
        the human race. Programmers aren't bound by
        practical limitations such as the laws of physics;
        we can create exciting virtual worlds with behaviors
        that could never exist in the real world.
        Programming doesn't require great physical skill or
        coordination, like ballet or basketball. ...",
    "page_number": 14,
},
```

MongoDB Hybrid Search Tester

Reciprocal Rank Fusion Params

Reset

vector_penalty

Penalise vector results score

0

fts_penalty

Penalise text search results score

3

k

Number of results

10

overrequest_factor

Multiplication factor of k for numCandidates for HNSW search

10

numCandidates: 100, query took 27ms

Query: “Future Flying Machines”

The Rocketeer

A young pilot stumbles onto a prototype jetpack that allows him to become a high flying masked hero.

Scoring for this result: 0.125 + 1.000 = 1.125

Lexical 11% 89% Vector

Sky Captain and the World of Tomorrow

After New York City receives a series of attacks from giant flying robots, a reporter teams up with a pilot in search of their origin, as well as the reason for the disappearances of famous scientists around the world.

Scoring for this result: 0.000 + 0.500 = 0.500

Lexical 0% 100% Vector

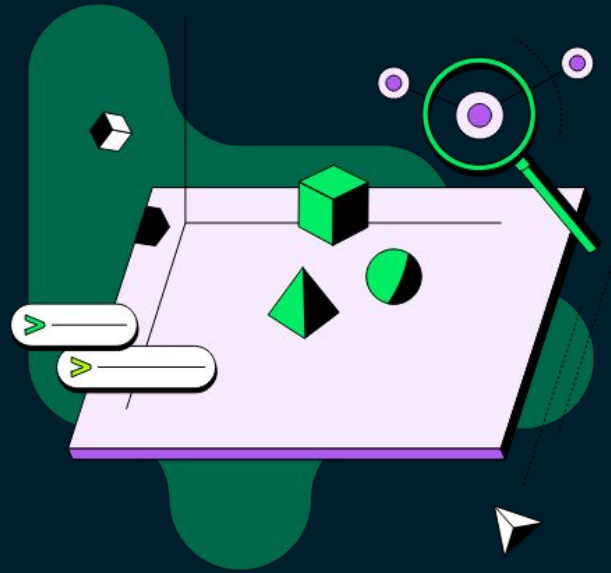
Earth vs. the Flying Saucers

Extra-terrestrials flying in high-tech flying saucers contact scientist Dr. Russell Marvin as part of a plan to enslave the inhabitants of Earth.

Scoring for this result: 0.250 + 0.250 = 0.500

Lexical 50% 50% Vector

How to Perform Hybrid Search



Let’s glimpse the future!