
powered by
Cheshire Cat AI
Don't Get Lost
in
Vector Space

WHO I AM

Nicola Procopio

Senior Data Scientist

Cheshire Cat - Core Contributor

Contacts

https://it.linkedin.com/in/nicolaprocopio

https://github.com/nickprock

https://huggingface.co/nickprock

Actually @ FINCONS GROUP

Previous Main Experiences

healthware international

integris

ICAR CNR

UNIVERSITÀ DELLA CALABRIA

Communities and Open Source Projects

AI DEMOCRATIZE

tag Talent Garden Cosenza

iaml ITALIAN ASSOCIATION FOR MACHINE LEARNING

Cheshire Cat
The AI Assistant Framework

PRODUCTION-READY AI ASSISTANT FRAMEWORK

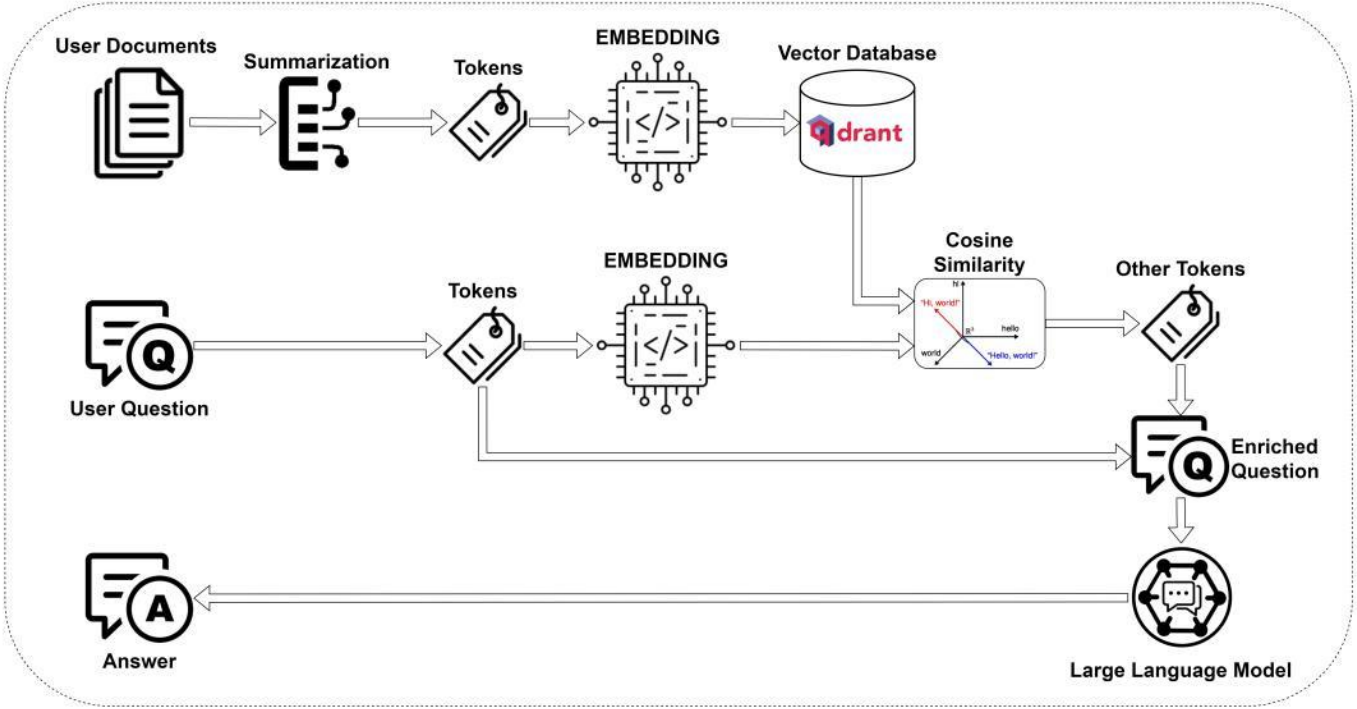
- **ready to fight**
  dockerized
  model agnostic

- **RAG + action agent**
  (docs, convos and procedures)

- **plugin system and registry**
  (hooks, tools and forms)

Cheshire Cat
The AI Assistant Framework

CHESHIRE CAT's "VANILLA" RAG PIPELINE

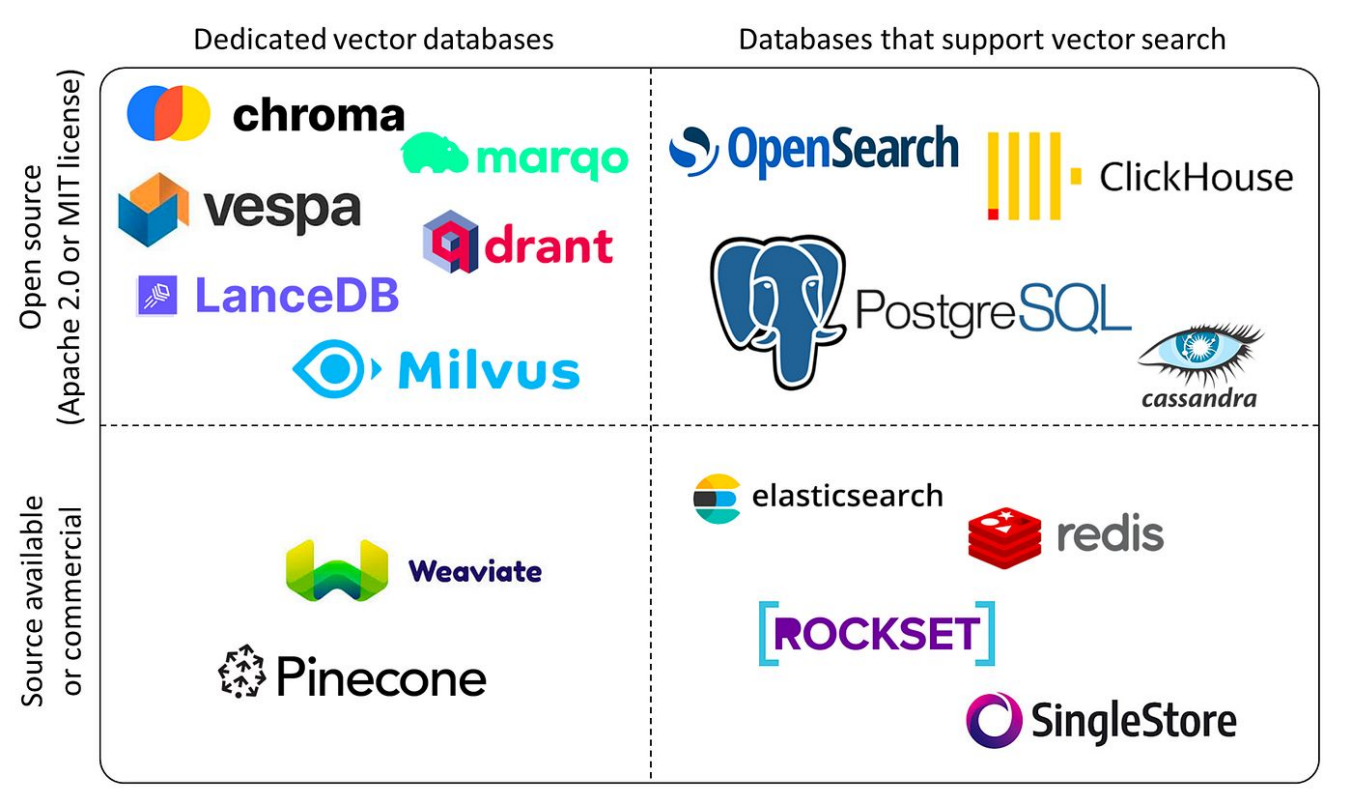


Customize it using **hooks, tools** and **plugins!**

[documentation]

Cheshire Cat
The AI Assistant Framework

WHY QDRANT?



[Vector Database Benchmarks]

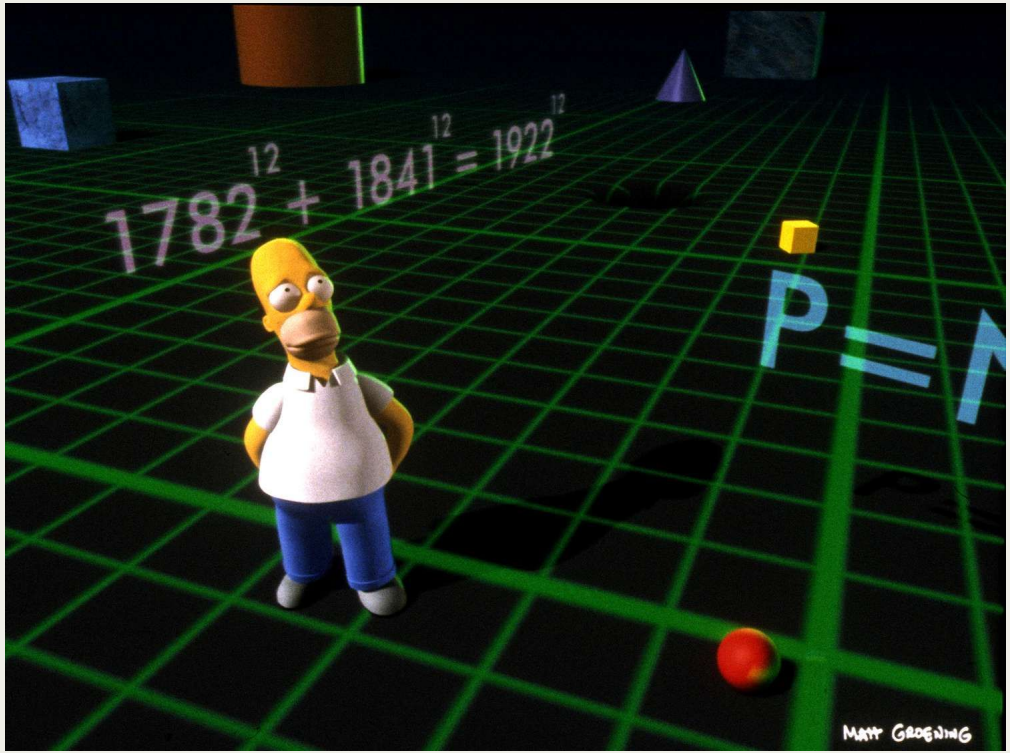
Cheshire Cat
The AI Assistant Framework

WHAT'S VECTOR SPACE?

A **multi-dimensional continuous space** where the **objects** are **represented** as **vectors.**

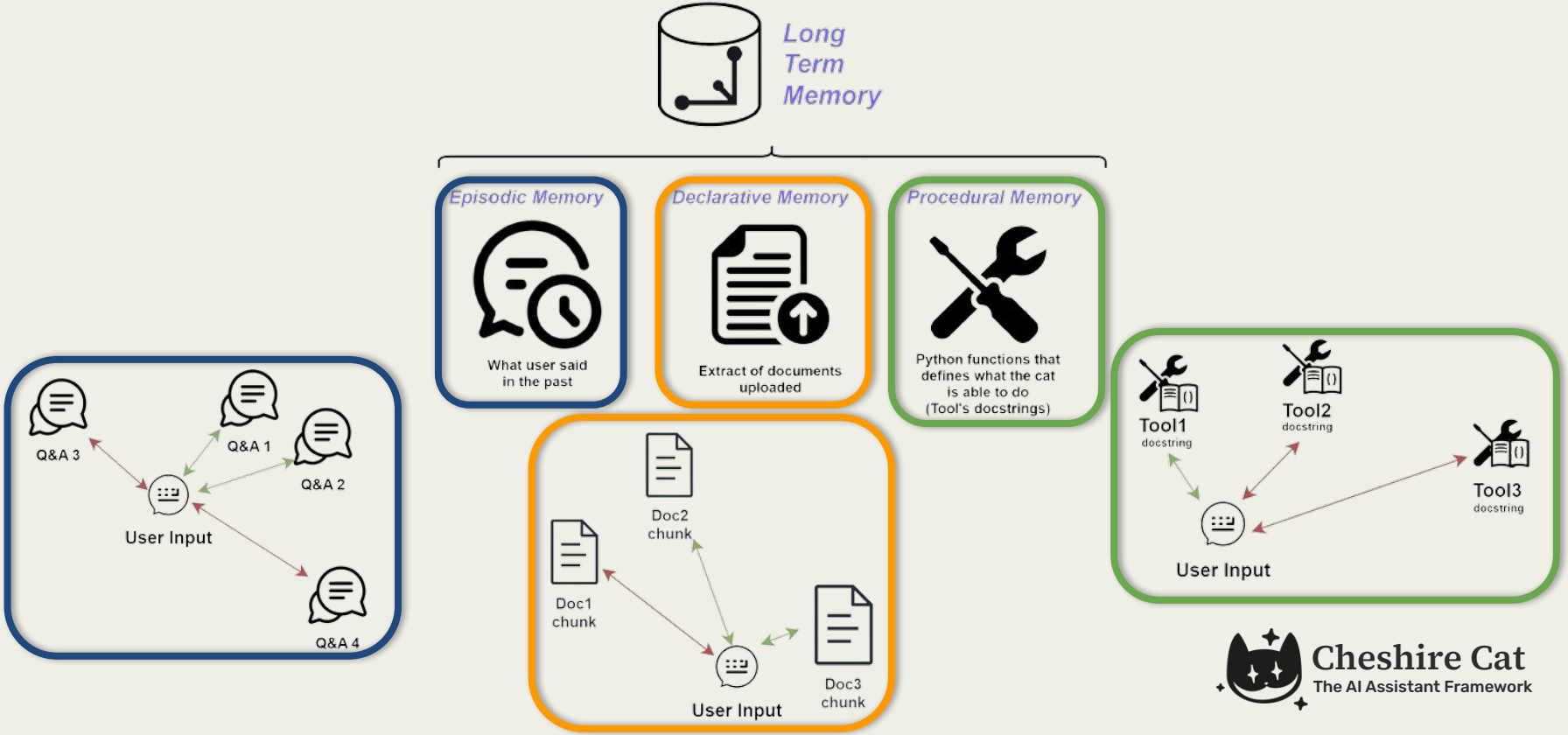
In NLP also called **semantic space** and the objects are words, sentences, documents.

The dimension of the space is defined using **embedders.**



Cheshire Cat
The AI Assistant Framework

CHESHIRE CAT's MEMORY



THE DRUNKEN CAT EFFECT

Cheshire Cat
The AI Assistant Framework

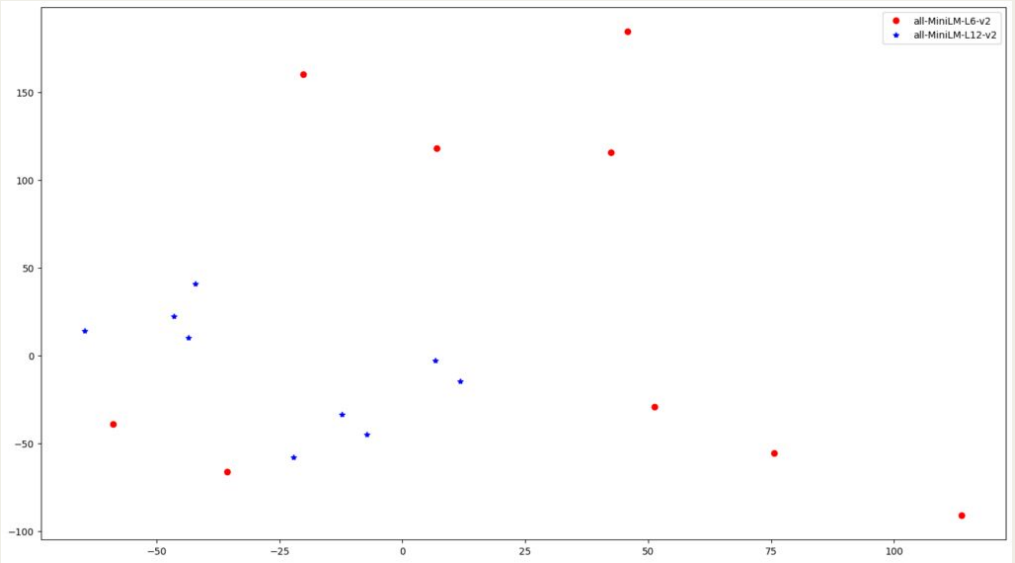
PROBLEM

SAME SENTENCES

- 'A man is eating food.',
- 'A man is eating a piece of bread.',
- 'The girl is carrying a baby.',
- 'A man is riding a horse.',
- 'A woman is playing violin.',
- 'Two men pushed carts through the woods.',
- 'A man is riding a white horse on an enclosed ground.',
- 'A monkey is playing drums.',
- 'Someone in a gorilla costume is playing a set of drums.'

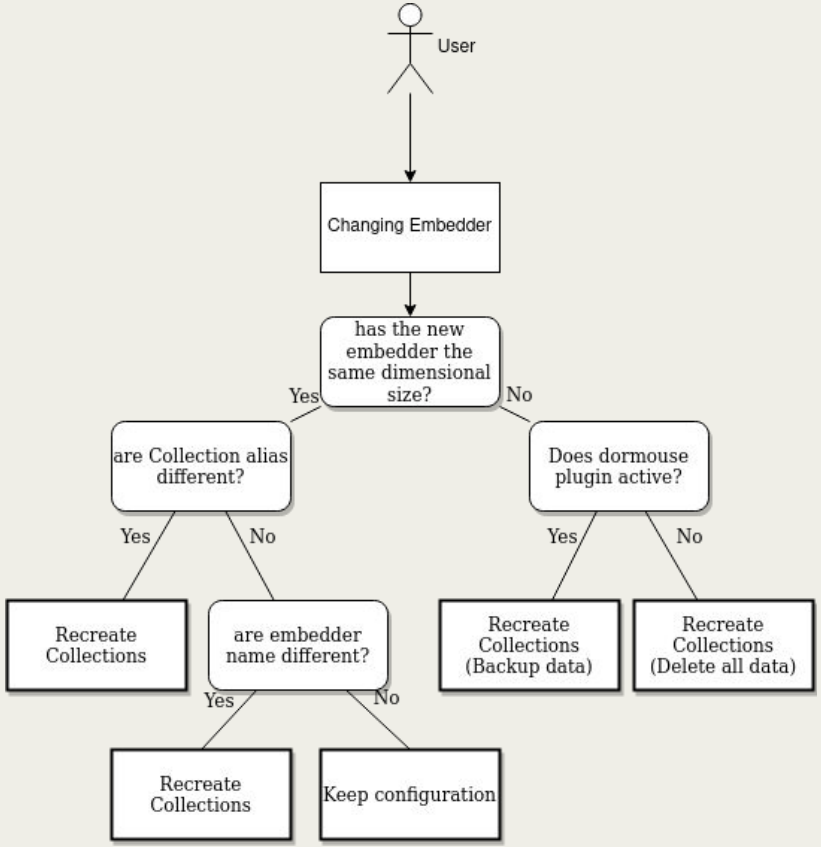
DIFFERENT EMBEDDERS WITH SAME SIZE

- all-MiniLM-L6-v2
- all-MiniLM-L12-v2



Cheshire Cat
The AI Assistant Framework

USE QDRANT ALIASES AND DON'T MIX EMBEDDINGS!



Cheshire Cat
The AI Assistant Framework

ACCURACY AND PERFORMANCE USING QUANTIZATION

High-dimensional vector embeddings can be memory-intensive, the formula to estimate memory size:

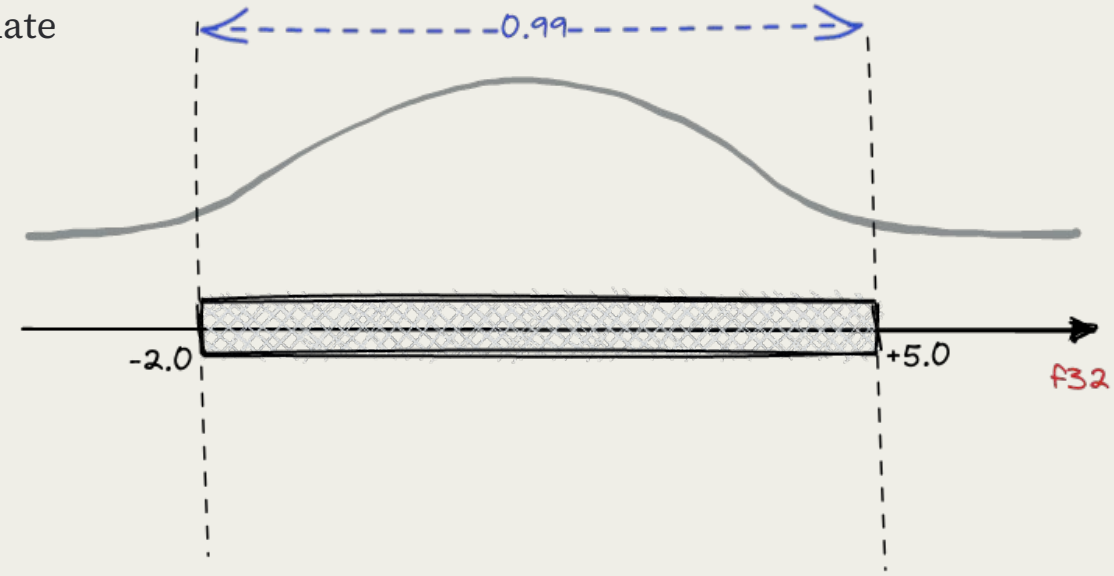
*memory_size = 1.5 * number_of_vectors * vector_dimension * 4 bytes*

Cheshire Cat uses **Scalar Quantization** to use up less memory.

Trick 1:

Hybrid mode:

- original vector on Disk
- quantized vector in RAM

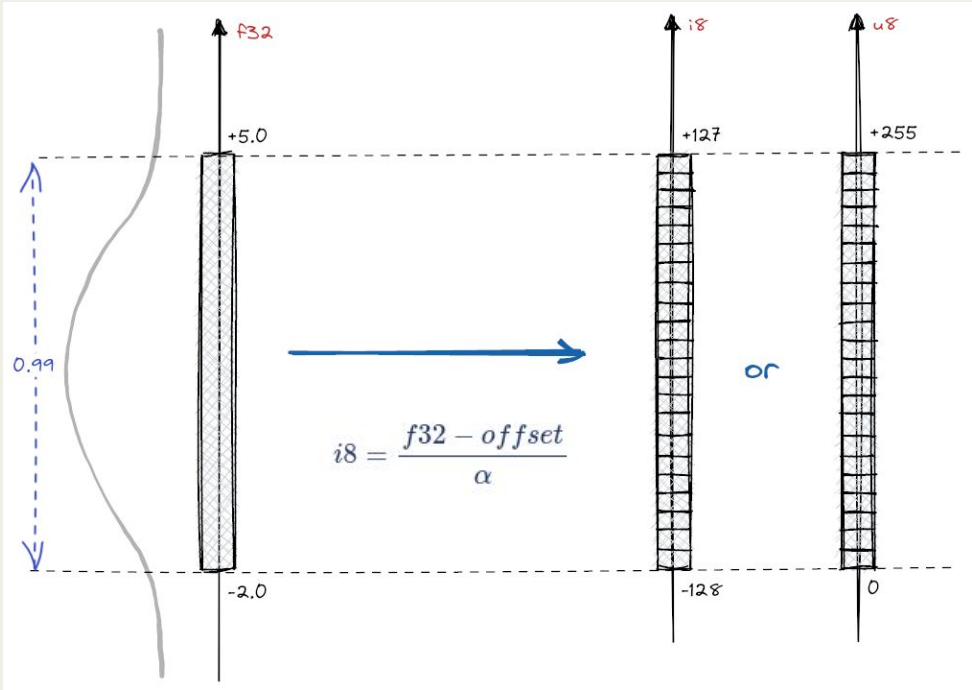


https://qdrant.tech/articles/scalar-quantization/

Cheshire Cat
The AI Assistant Framework

WHAT'S QUANTIZATION?

- **Scalar Quantization** converts the ***float32*** embeddings into ***int8***
- from a range of infinite value to 256 elements
- the calibration dataset greatly influences performance since it defines the quantization buckets



Trick 2:
The **quantile parameter** in scalar quantization determines the quantization bounds.
In Cheshire Cat it's set to 0.95, we exclude the 5% outliers.

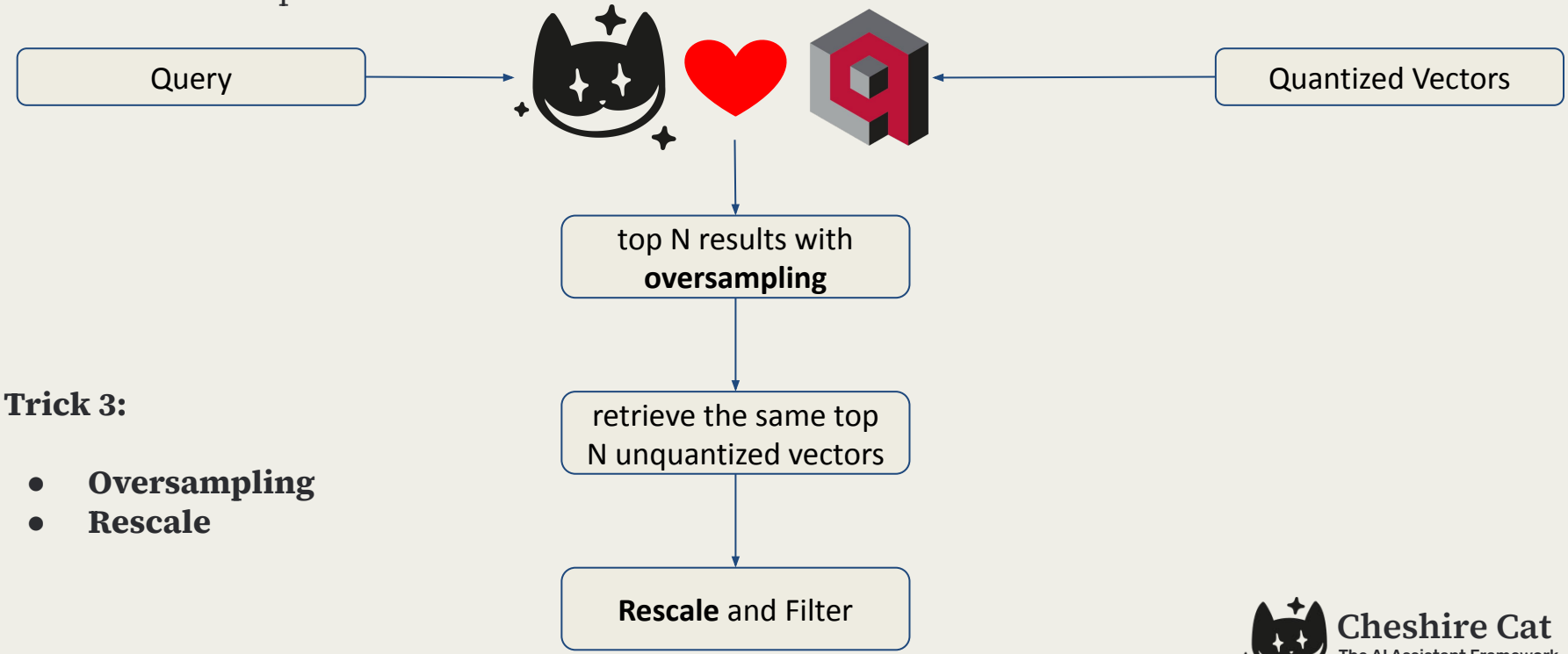
https://qdrant.tech/articles/scalar-quantization/

https://huggingface.co/blog/embedding-quantization#scalar-int8-quantization

Cheshire Cat
The AI Assistant Framework

ACCURACY AND PERFORMANCE USING QUANTIZATION

In retrieve step:



Trick 3:

- **Oversampling**
- **Rescale**

https://qdrant.tech/articles/scalar-quantization/

Cheshire Cat
The AI Assistant Framework

ACCURACY AND PERFORMANCE USING QUANTIZATION

Using these Tips & Tricks the Cheshire Cat's vector search can achieve up to 4x lower memory footprint and even up to 2x performance increase!

Cheshire Cat
The AI Assistant Framework

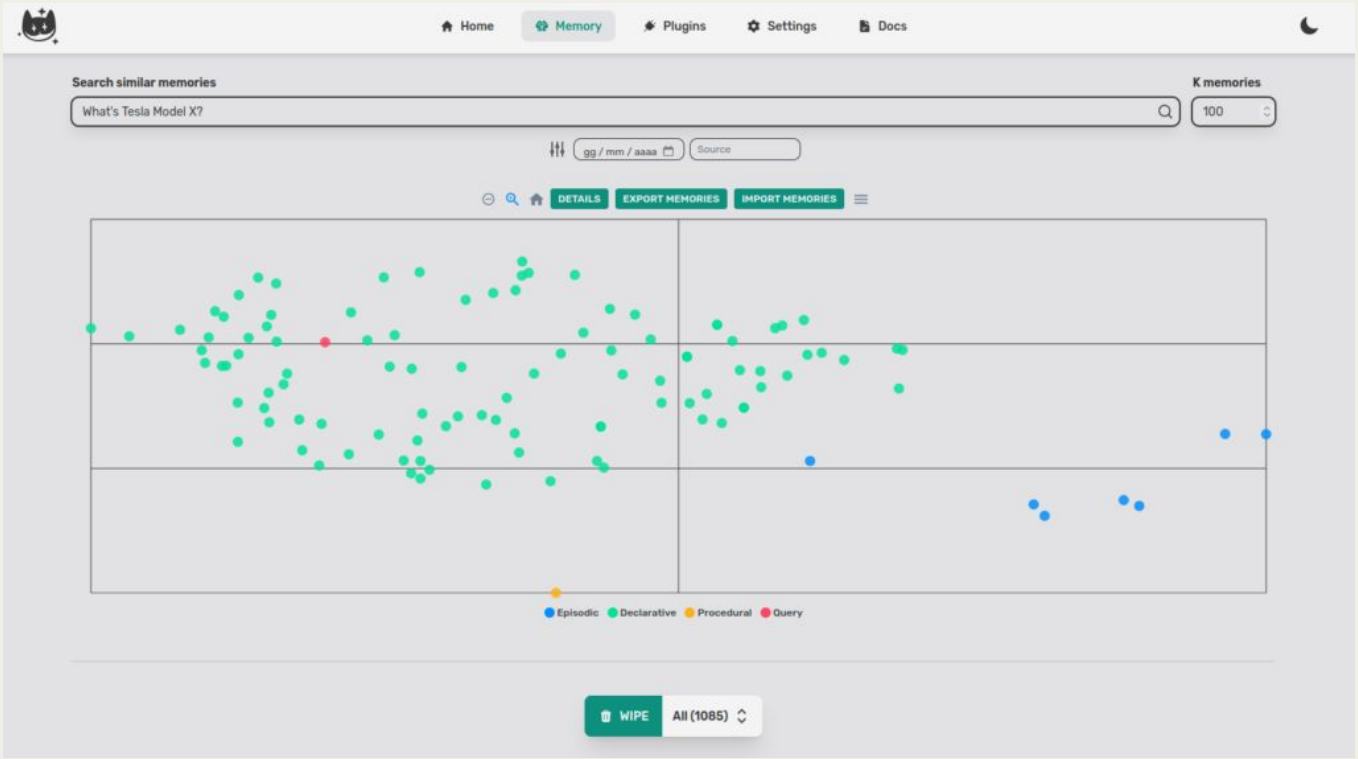
MATRYOSHKA EMBEDDERS

Matryoshka Representation Learning (MRL) is an advanced machine learning approach that **encodes data at multiple levels of granularity within a single vector representation**.

Like a Matryoshka doll the levels are nested in one embedding, the greater the number of levels, the more detail the embedding maps.

Cheshire Cat
The AI Assistant Framework

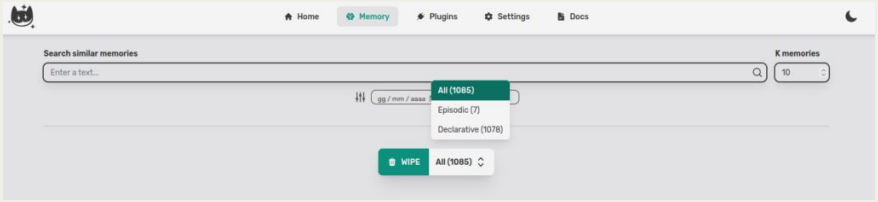
TAKE A LOOK INTO THE CAT's MEMORY



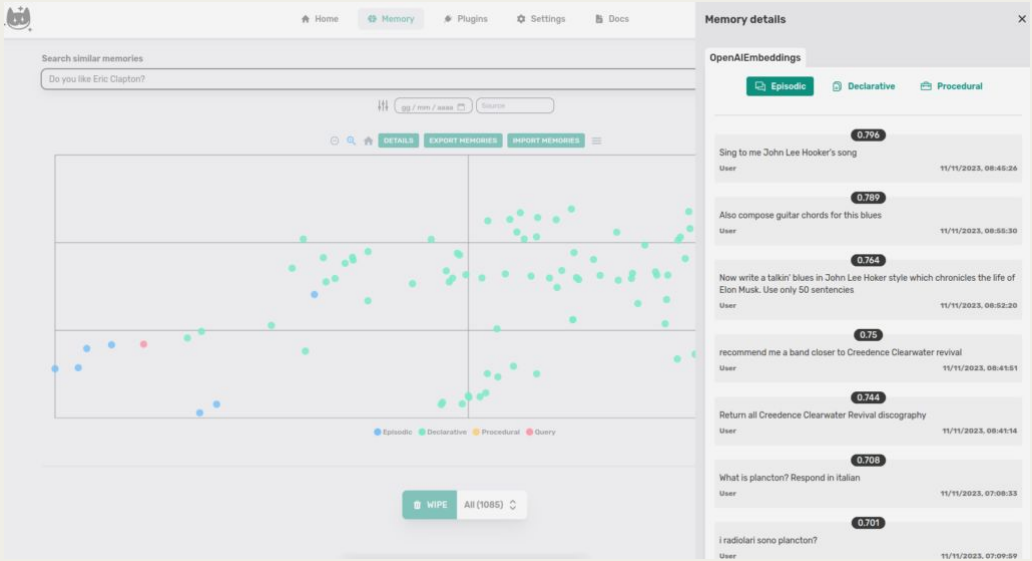
Search

Cheshire Cat
The AI Assistant Framework

TAKE A LOOK INTO THE CAT's MEMORY



Filter & Wipe



Go Deep using Details

Cheshire Cat
The AI Assistant Framework

TAKE A LOOK INTO THE CAT's MEMORY


```
{
  "export_time": 1699709740967,
  "embedder": "OpenAIEmbeddings",
  "collections": {
    "episodic": [
      {
        "page_content": "Sing to me John Lee Hooker's song",
        "metadata": {
          "source": "user",
          "when": 1699688726.9407504
        },
        "type": "Document",
        "id": "f412f0cc-d164-458f-93b9-6bf58861390a",
        "score": 0.7969544,
        "vector": [
          -0.009110927,
          -0.022545898,
          0.007836538,
          -0.016915765,
          -0.013453993,
          0.027719537,
```


Export using Json

Wake up the Dormouse!

Cheshire Cat
The AI Assistant Framework

Any Questions?

CHESHIRE CAT AI

Cheshire Cat
The AI Assistant Framework

Thank you!

CHESHIRE CAT AI

NICOLA PROCOPIO,

CORE CONTRIBUTOR

Cheshire Cat
The AI Assistant Framework

Keep in touch!

CHESHIRE CAT AI

https://cheshirecat.ai

https://github.com/cheshire-cat-ai

https://www.linkedin.com/company/cheshire-cat-ai

https://medium.com/mad-chatter-tea-party

https://discord.gg/cheshire-cat

Cheshire Cat
The AI Assistant Framework