cisco TALOS

Build your RAG in 90 minutes

Who am I?

CyberSecurity Researcher at Cisco Talos

- Over 20 years in cybersecurity
- Mobile malware lover
- APT hunter
- Reverse engineer

Located in Portugal

Vitor Ventura

@_vventura

What is an LLM?

Large Language Model

- In simple words it’s a list of tokens with different weights
- Basically massive precomputed neural network of interconnected tokens.
- Current models are NOT open sourced, we only know the weigths we don't know the relation between the words

What is an LLM?

Large Language Model usage components

- Prompt
- Inference engine
- Interface

What is an LLM?

Large Language Model usage components

- Prompt – A template of communication with the LLM. Contains certain special tokens to mark positions with intent.
- Inference engine – The engine that will run the prompt on the model, Ollama, LLamaCPP,
- Interface – We need to send the data somehow. REST api, web page, app, etc

What is an LLM?

Were to get the models?

- HuggingFace – A hub for all of them

Interesting models:

- LLama the model we will be using
- MS Phi models are interesting because they are SLMS (Small Language models)
- Gemma Google open weight models family

What is an LLM?

How to improve an LLM?

- Retrain it
- Fine-tune it
- Use it in a RAG (not really improving the LLM, but improving the experience of using one)

What is an LLM?

How to improve an LLM?

- Retrain it
- Fine-tune it
- **Use it in a RAG (not really improving the LLM, but improving the experience of using one)**

What is a RAG?

- A process that allows usage of pretrained LLM models with new information.

- Along with a question the user provides the LLM with some context to "help" it to generate an answer.

What is a RAG?

Retrieval augmented generation components

- Indexer - Indexer tool that will ensure semantic similarity
- Information source - vector database, internet, internal documents, source code, etc
- Context generator – The component that will generate the information that is sent to the LLM inference engine along with with the question
- LLM system
  - Inference Engine
  - Prompt
  - Interface

What is a RAG?

Retrieval augmented generation components

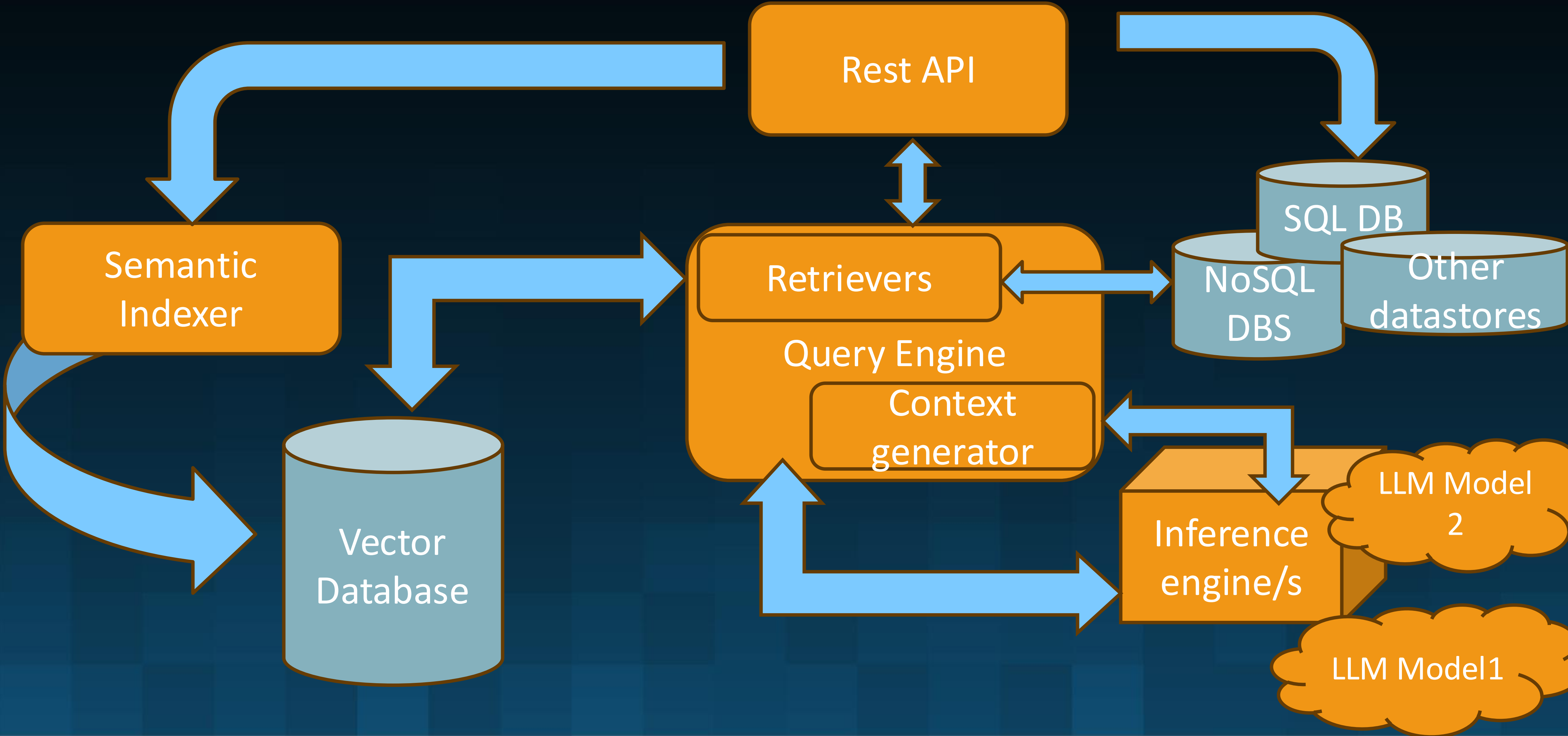
- Indexer – Indexer tool that will ensure semantic similarity

- Information source –
  - Vector Stores
    - Qdrant, FAISS, Weaviate, Chroma

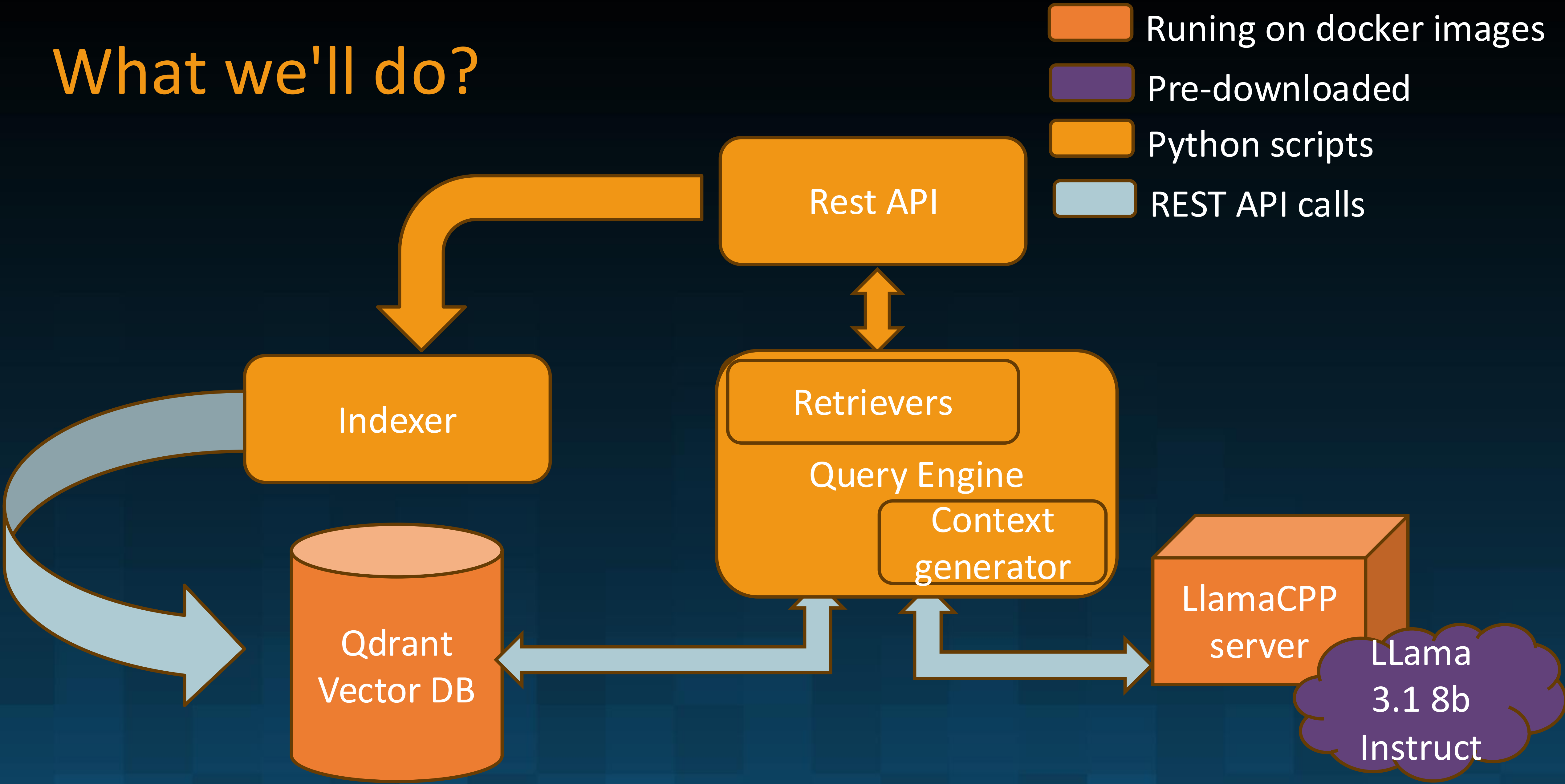
  - Other stores
    - Anything relevant data that can be added to the context

- Context generator – The component that will generate the information that is sent to the LLM inference engine along with with the question

Possible RAG architecture



What we'll do?



🟧 Runing on docker images

🟪 Pre-downloaded

🟧 Python scripts

🟦 REST API calls

Rest API

Indexer

Retrievers

Query Engine

Context generator

Qdrant Vector DB

LlamaCPP server

LLama 3.1 8b Instruct

How we'll do it?

1. First we start by indexing data
   - We need to index it in such a way that we keep the semantic similarity.
   - Then we store it on the vector database.

But what is semantic similarity and how do we achieve it?!

- It can't be just a word, otherwise you would lose the semantic meaning.
  (And for that we already have fuzzy logic and BM25 full text search algorithms)

How to semantically index data?

We need pieces of information to which we will call chunks.

Chunks can be created by splitter

- A splitter will split the information based on textual characteristics of the information

Examples:

Text splitter, Character splitter, Recursive Character Splitter, Sentence splitter, Semantic splitting

How to semantically index data?

We need pieces of information to which we will call chunks.

Chunks can be converted into tokens.

- By a LLM model which can then be embeeded
- It may be a word, a subword or even multiple words

Examples:

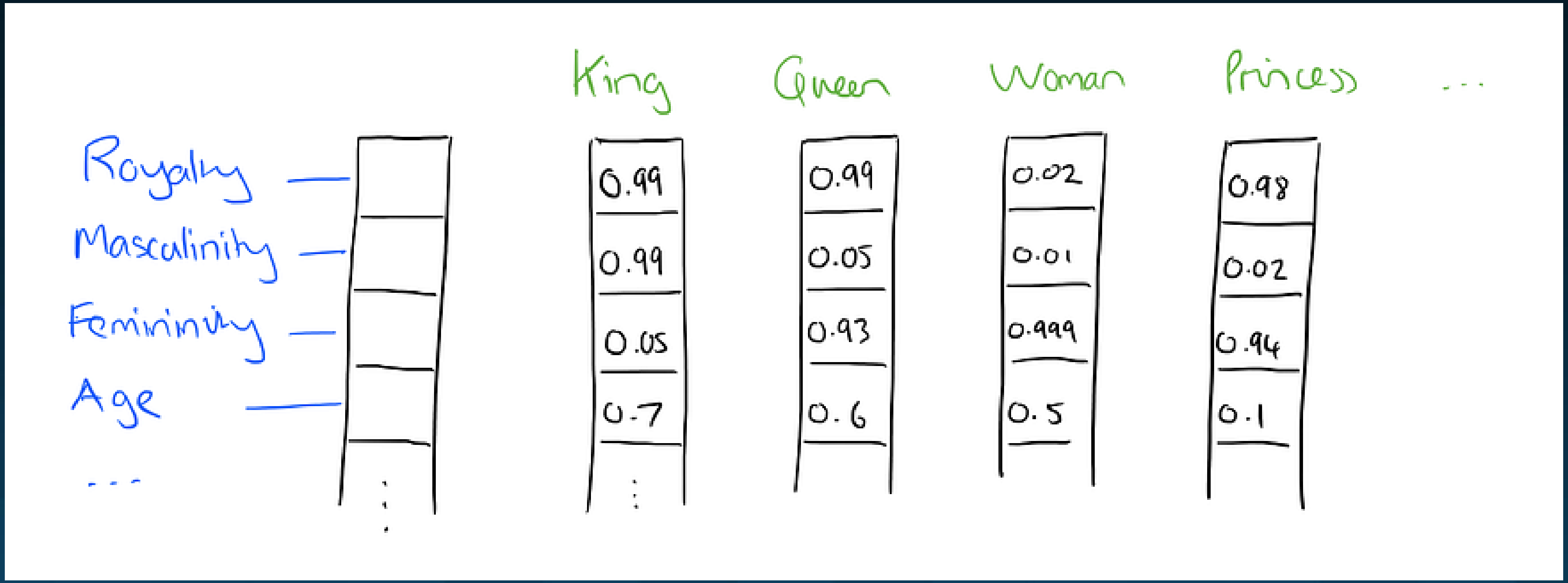
OpenAI Tiktoken

Huggingface Tokenizers

How to semantically index data?

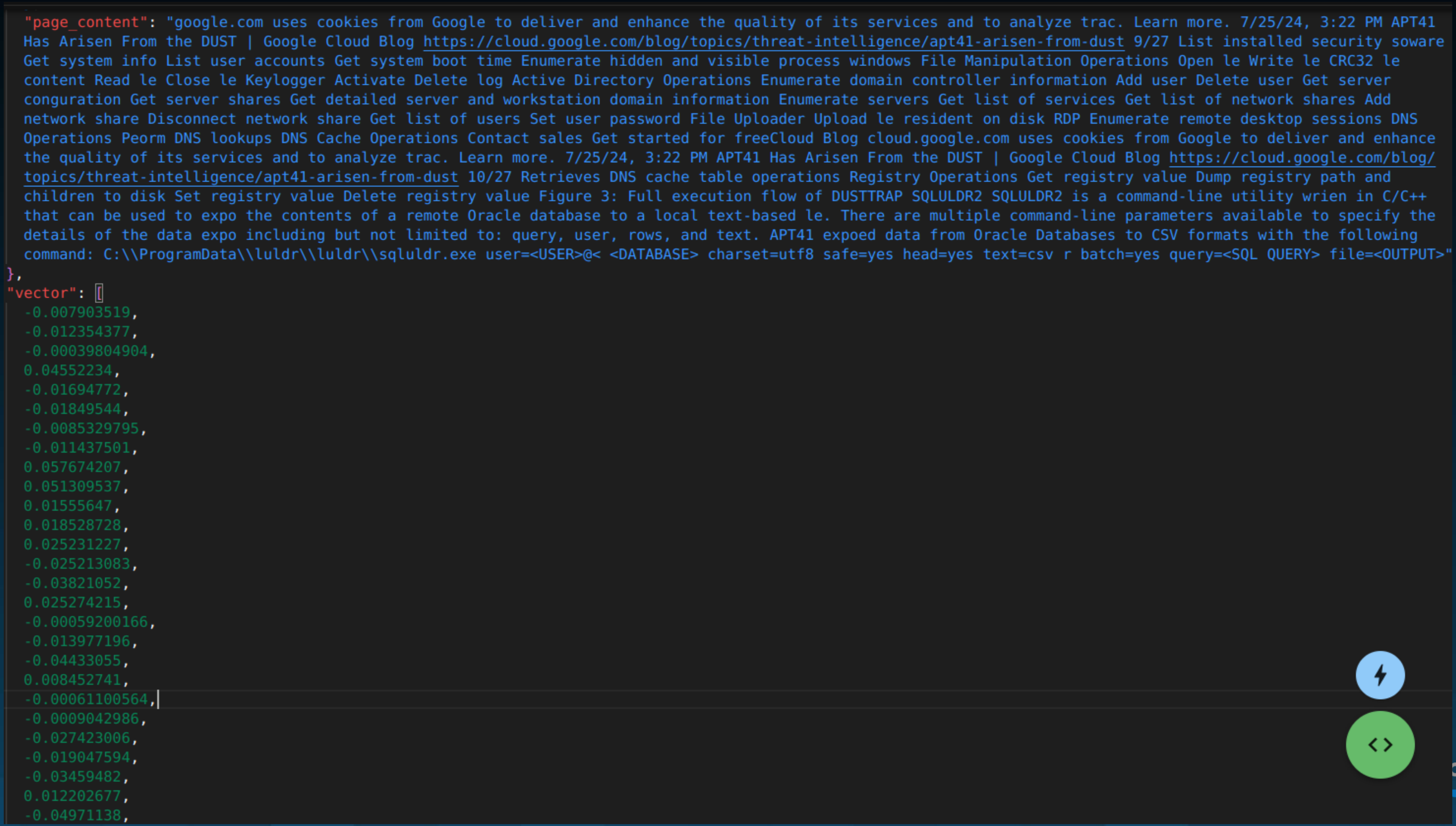
We need to embed the chunks or tokens

Data is represented vector of floats between –1 and 1



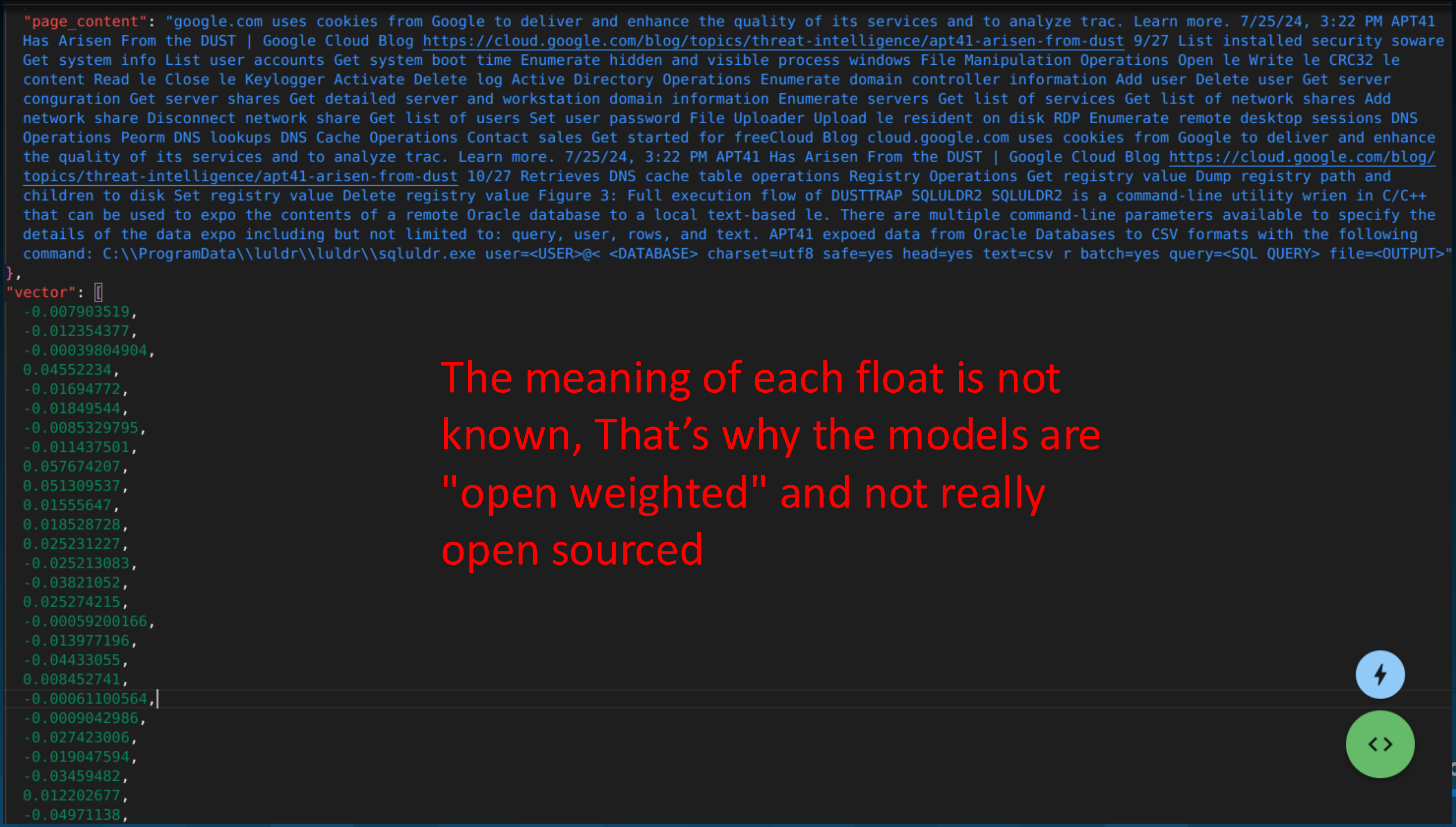
		King	Queen	Woman	Princess	...
Royalty		0.99	0.99	0.02	0.98	
Mascalinity		0.99	0.05	0.01	0.02	
Femininity		0.05	0.93	0.999	0.94	
Age		0.7	0.6	0.5	0.1	
...	⋮	⋮				

How to semantically index data?



```
  "page_content": "google.com uses cookies from Google to deliver and enhance the quality of its services and to analyze trac. Learn more. 7/25/24, 3:22 PM APT41
  Has Arisen From the DUST | Google Cloud Blog https://cloud.google.com/blog/topics/threat-intelligence/apt41-arisen-from-dust 9/27 List installed security soware
  Get system info List user accounts Get system boot time Enumerate hidden and visible process windows File Manipulation Operations Open le Write le CRC32 le
  content Read le Close le Keylogger Activate Delete log Active Directory Operations Enumerate domain controller information Add user Delete user Get server
  conguration Get server shares Get detailed server and workstation domain information Enumerate servers Get list of services Get list of network shares Add
  network share Disconnect network share Get list of users Set user password File Uploader Upload le resident on disk RDP Enumerate remote desktop sessions DNS
  Operations Peorm DNS lookups DNS Cache Operations Contact sales Get started for freeCloud Blog cloud.google.com uses cookies from Google to deliver and enhance
  the quality of its services and to analyze trac. Learn more. 7/25/24, 3:22 PM APT41 Has Arisen From the DUST | Google Cloud Blog https://cloud.google.com/blog/
  topics/threat-intelligence/apt41-arisen-from-dust 10/27 Retrieves DNS cache table operations Registry Operations Get registry value Dump registry path and
  children to disk Set registry value Delete registry value Figure 3: Full execution flow of DUSTTRAP SQLULDR2 SQLULDR2 is a command-line utility wrien in C/C++
  that can be used to expo the contents of a remote Oracle database to a local text-based le. There are multiple command-line parameters available to specify the
  details of the data expo including but not limited to: query, user, rows, and text. APT41 expoed data from Oracle Databases to CSV formats with the following
  command: C:\\ProgramData\\luldr\\luldr\\sqluldr.exe user=<USER>@< <DATABASE> charset=utf8 safe=yes head=yes text=csv r batch=yes query=<SQL QUERY> file=<OUTPUT>"
},
"vector": [
  -0.007903519,
  -0.012354377,
  -0.00039804904,
  0.04552234,
  -0.01694772,
  -0.01849544,
  -0.0085329795,
  -0.011437501,
  0.057674207,
  0.051309537,
  0.01555647,
  0.018528728,
  0.025231227,
  -0.025213083,
  -0.03821052,
  0.025274215,
  -0.00059200166,
  -0.013977196,
  -0.04433055,
  0.008452741,
  -0.00061100564,
  -0.0009042986,
  -0.027423006,
  -0.019047594,
  -0.03459482,
  0.012202677,
  -0.04971138,
```

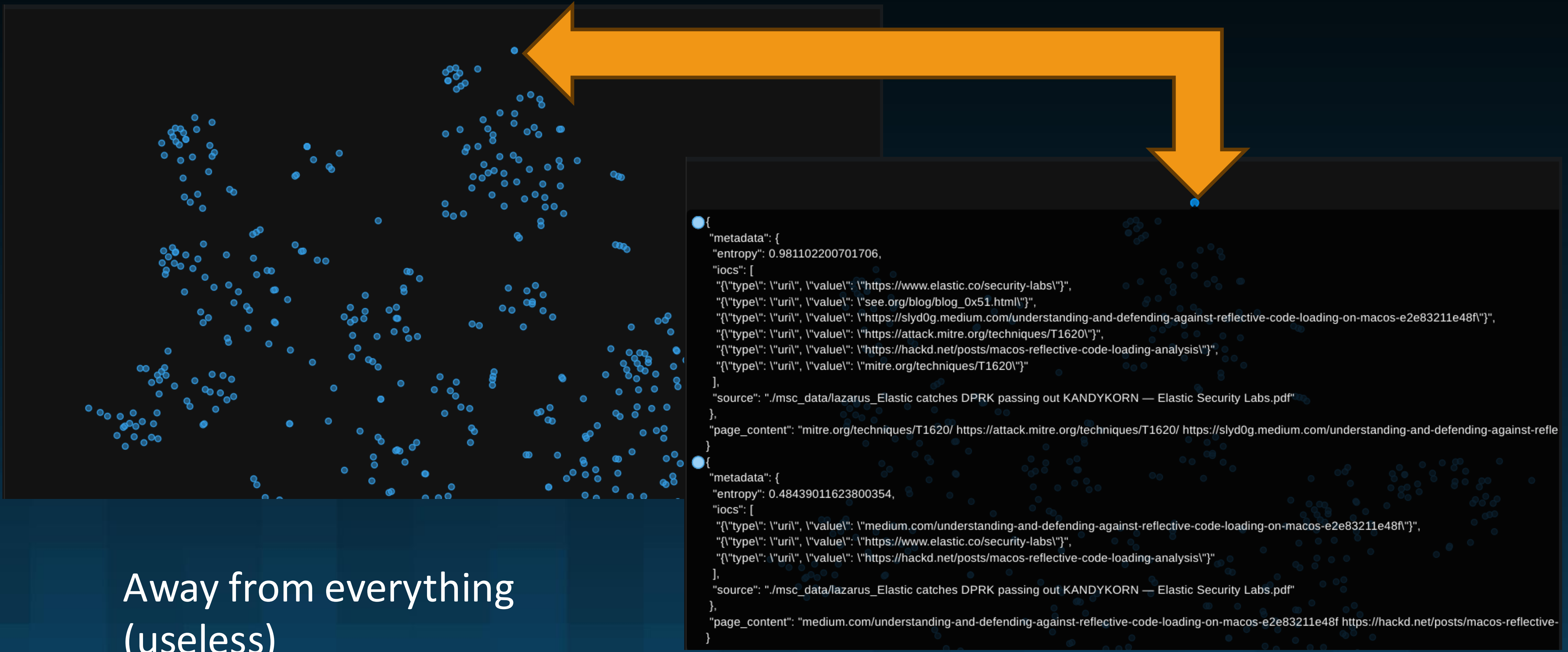
How to semantically index data?



```
  "page_content": "google.com uses cookies from Google to deliver and enhance the quality of its services and to analyze trac. Learn more. 7/25/24, 3:22 PM APT41
  Has Arisen From the DUST | Google Cloud Blog https://cloud.google.com/blog/topics/threat-intelligence/apt41-arisen-from-dust 9/27 List installed security soware
  Get system info List user accounts Get system boot time Enumerate hidden and visible process windows File Manipulation Operations Open le Write le CRC32 le
  content Read le Close le Keylogger Activate Delete log Active Directory Operations Enumerate domain controller information Add user Delete user Get server
  conguration Get server shares Get detailed server and workstation domain information Enumerate servers Get list of services Get list of network shares Add
  network share Disconnect network share Get list of users Set user password File Uploader Upload le resident on disk RDP Enumerate remote desktop sessions DNS
  Operations Peorm DNS lookups DNS Cache Operations Contact sales Get started for freeCloud Blog cloud.google.com uses cookies from Google to deliver and enhance
  the quality of its services and to analyze trac. Learn more. 7/25/24, 3:22 PM APT41 Has Arisen From the DUST | Google Cloud Blog https://cloud.google.com/blog/
  topics/threat-intelligence/apt41-arisen-from-dust 10/27 Retrieves DNS cache table operations Registry Operations Get registry value Dump registry path and
  children to disk Set registry value Delete registry value Figure 3: Full execution flow of DUSTTRAP SQLULDR2 SQLULDR2 is a command-line utility wrien in C/C++
  that can be used to expo the contents of a remote Oracle database to a local text-based le. There are multiple command-line parameters available to specify the
  details of the data expo including but not limited to: query, user, rows, and text. APT41 expoed data from Oracle Databases to CSV formats with the following
  command: C:\\ProgramData\\luldr\\luldr\\sqluldr.exe user=<USER>@< <DATABASE> charset=utf8 safe=yes head=yes text=csv r batch=yes query=<SQL QUERY> file=<OUTPUT>"
},
"vector": [
  -0.007903519,
  -0.012354377,
  -0.00039804904,
  0.04552234,
  -0.01694772,
  -0.01849544,
  -0.0085329795,
  -0.011437501,
  0.057674207,
  0.051309537,
  0.01555647,
  0.018528728,
  0.025231227,
  -0.025213083,
  -0.03821052,
  0.025274215,
  -0.00059200166,
  -0.013977196,
  -0.04433055,
  0.008452741,
  -0.00061100564,
  -0.0009042986,
  -0.027423006,
  -0.019047594,
  -0.03459482,
  0.012202677,
  -0.04971138,
```

The meaning of each float is not known, That’s why the models are "open weighted" and not really open sourced

How to semantically index data?

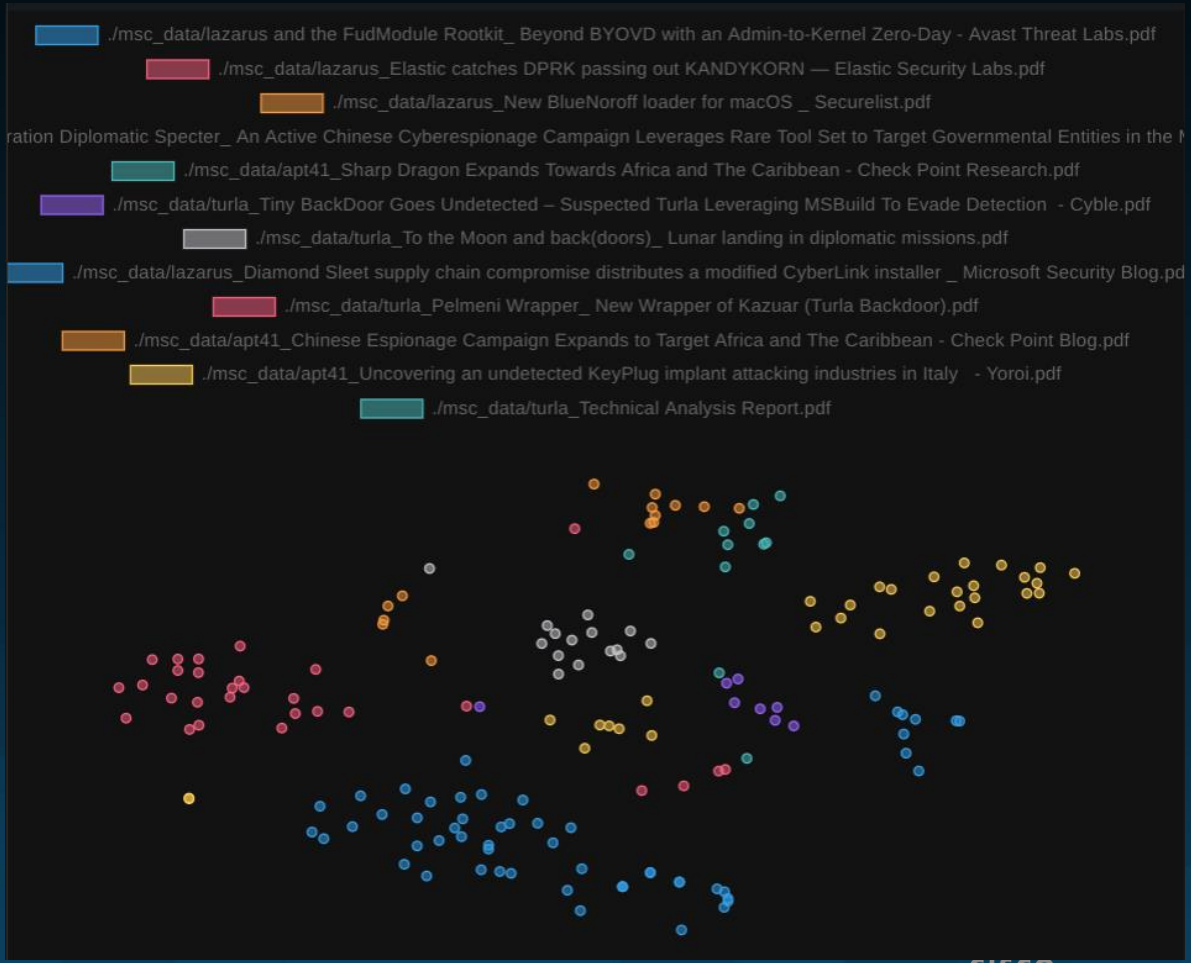


```
{
 "metadata": {
  "entropy": 0.981102200701706,
  "iocs": [
   "{\"type\": \"uri\", \"value\": \"https://www.elastic.co/security-labs\"}",
   "{\"type\": \"uri\", \"value\": \"see.org/blog/blog_0x51.html\"}",
   "{\"type\": \"uri\", \"value\": \"https://slyd0g.medium.com/understanding-and-defending-against-reflective-code-loading-on-macos-e2e83211e48f\"}",
   "{\"type\": \"uri\", \"value\": \"https://attack.mitre.org/techniques/T1620\"}",
   "{\"type\": \"uri\", \"value\": \"https://hackd.net/posts/macos-reflective-code-loading-analysis\"}",
   "{\"type\": \"uri\", \"value\": \"mitre.org/techniques/T1620\"}"
  ],
  "source": "./msc_data/lazarus_Elastic catches DPRK passing out KANDYKORN — Elastic Security Labs.pdf"
 },
 "page_content": "mitre.org/techniques/T1620/ https://attack.mitre.org/techniques/T1620/ https://slyd0g.medium.com/understanding-and-defending-against-refle
}
{
 "metadata": {
  "entropy": 0.48439011623800354,
  "iocs": [
   "{\"type\": \"uri\", \"value\": \"medium.com/understanding-and-defending-against-reflective-code-loading-on-macos-e2e83211e48f\"}",
   "{\"type\": \"uri\", \"value\": \"https://www.elastic.co/security-labs\"}",
   "{\"type\": \"uri\", \"value\": \"https://hackd.net/posts/macos-reflective-code-loading-analysis\"}"
  ],
  "source": "./msc_data/lazarus_Elastic catches DPRK passing out KANDYKORN — Elastic Security Labs.pdf"
 },
 "page_content": "medium.com/understanding-and-defending-against-reflective-code-loading-on-macos-e2e83211e48f https://hackd.net/posts/macos-reflective-
}
```

Away from everything (useless)
content similar

How to semantically index data?

Not perfect but we can see some clustering



Embeddings

Let's look at the code

Retrievers

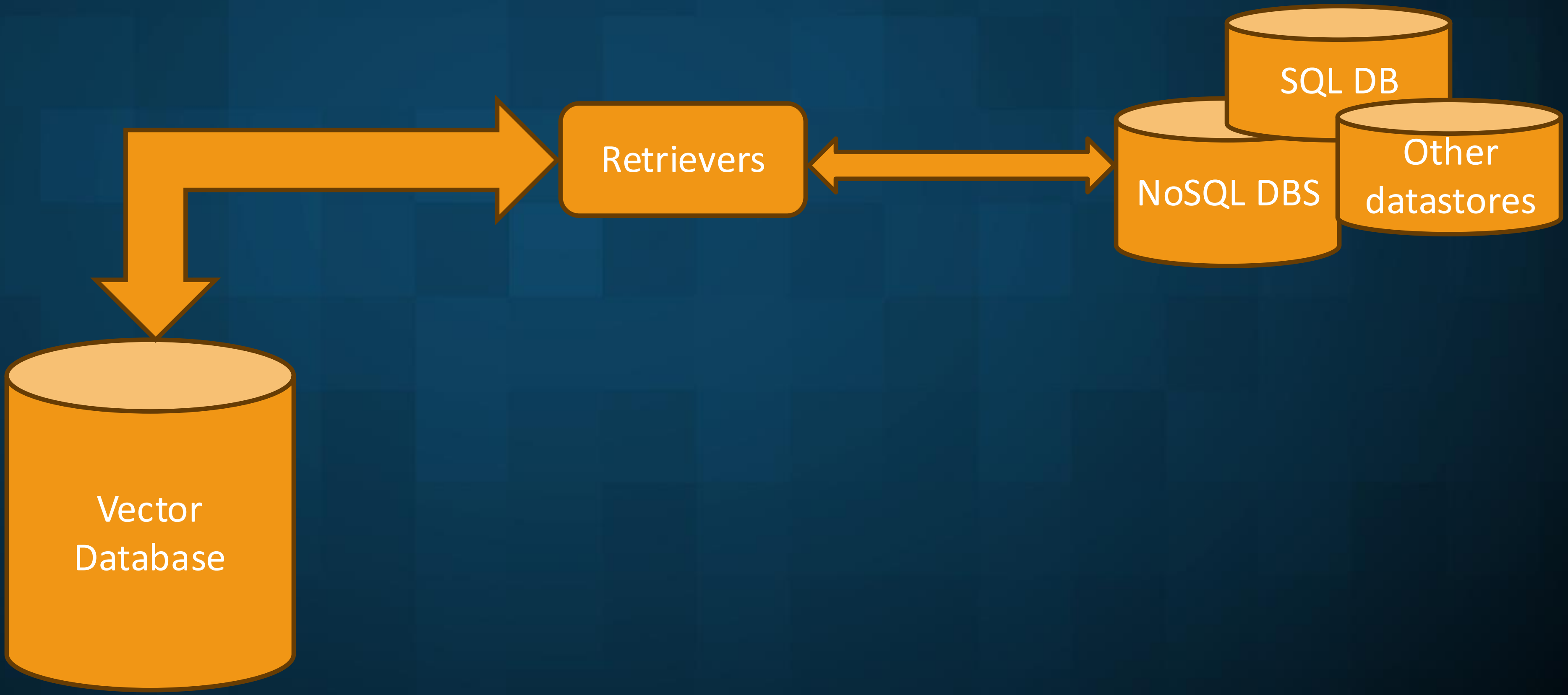
We want to retrieve the information relevant for our question.

Vector search

Hybrid search

Fulltext search

Others



Retreiver

Let's look at the code

Context generators

What is context?

Information that is given to the LLM to generate the answer.

This information is collected from the available sources

It increases the LLM "knowledge" beyond the contents of the training dataset.

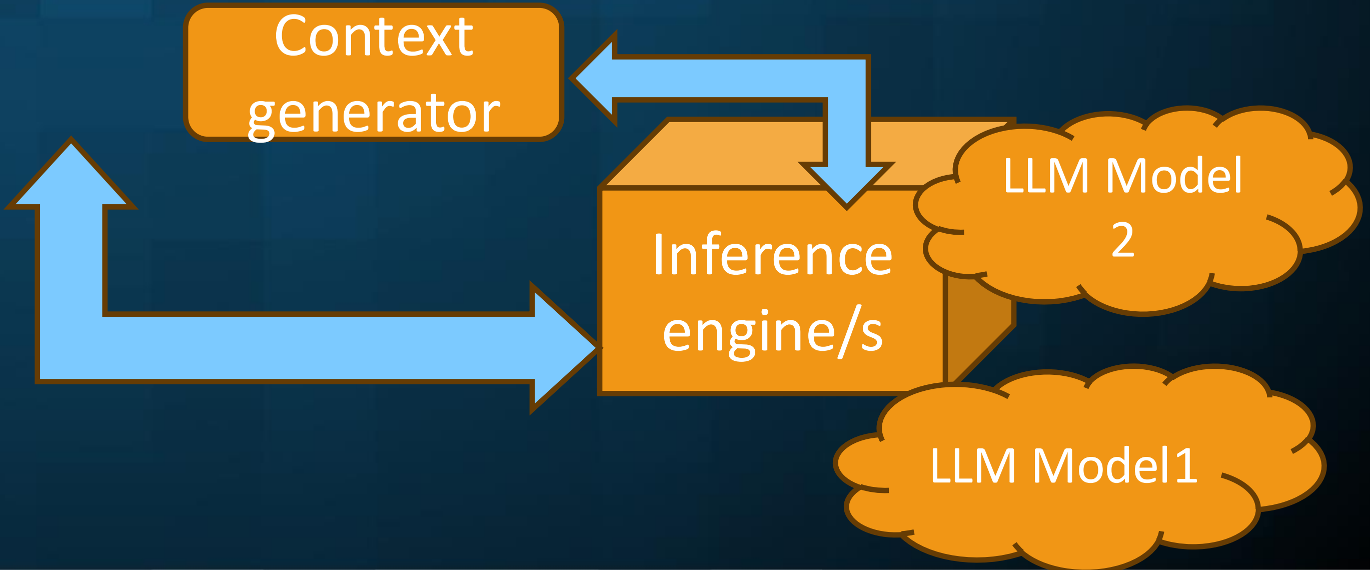
Context generators

What is context generator?

Multiple sources of context

Quality assurance

Context merging and reranking



Context generators - advanced

- Merge all context and summarize
- Remove noise from the context
- Recursively search for more information
- Extract entities to increase focus
- Translate context for normalization
- Along with the question determine the best model to generate answers
- Enrich context with external metadata

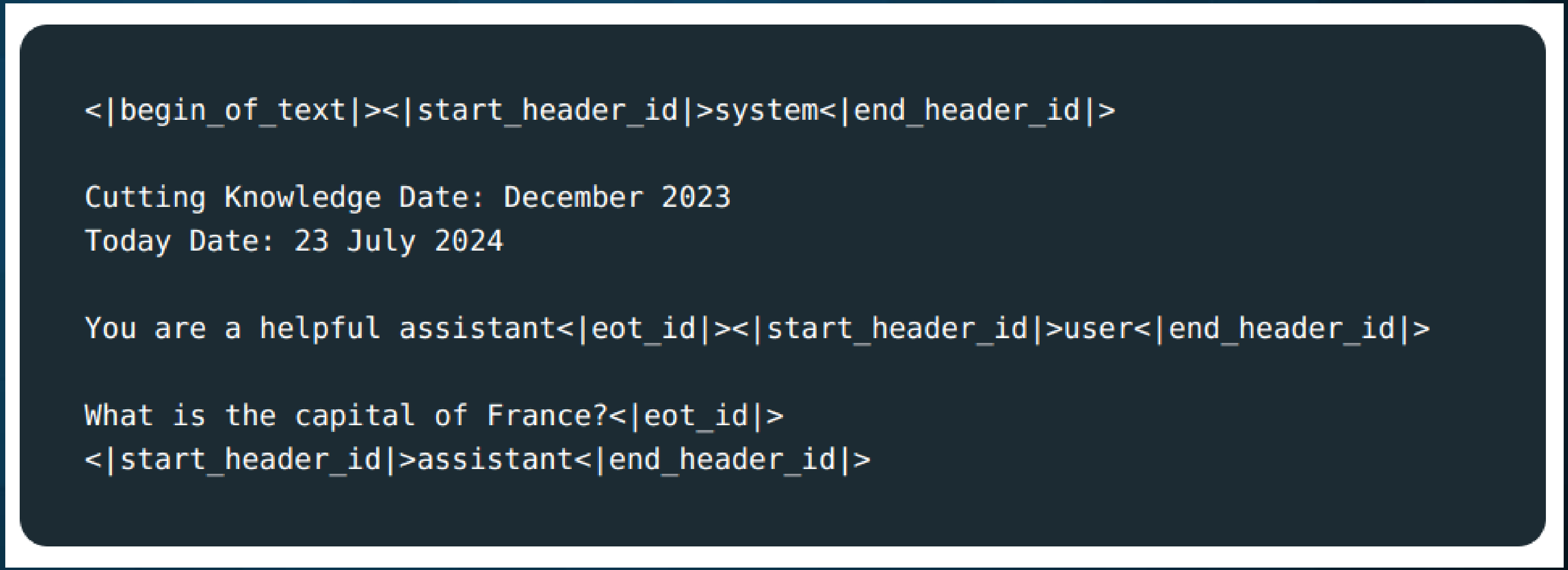
Context Generator

Let's look at the code

Prompts

The way to convey the LLM the context and make the question
Different models have different prompt formats.

Roles
- system
- assistant
- user
- ipython

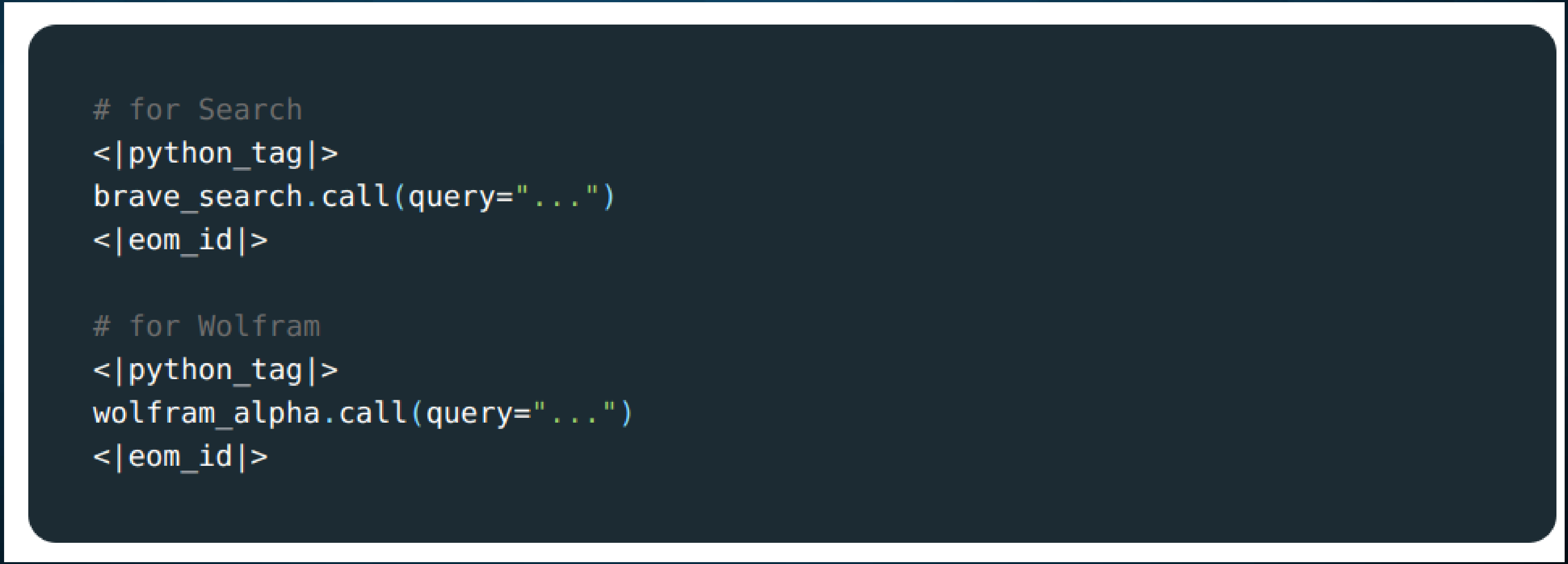


```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 23 July 2024

You are a helpful assistant<|eot_id|><|start_header_id|>user<|end_header_id|>

What is the capital of France?<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Prompts - advanced

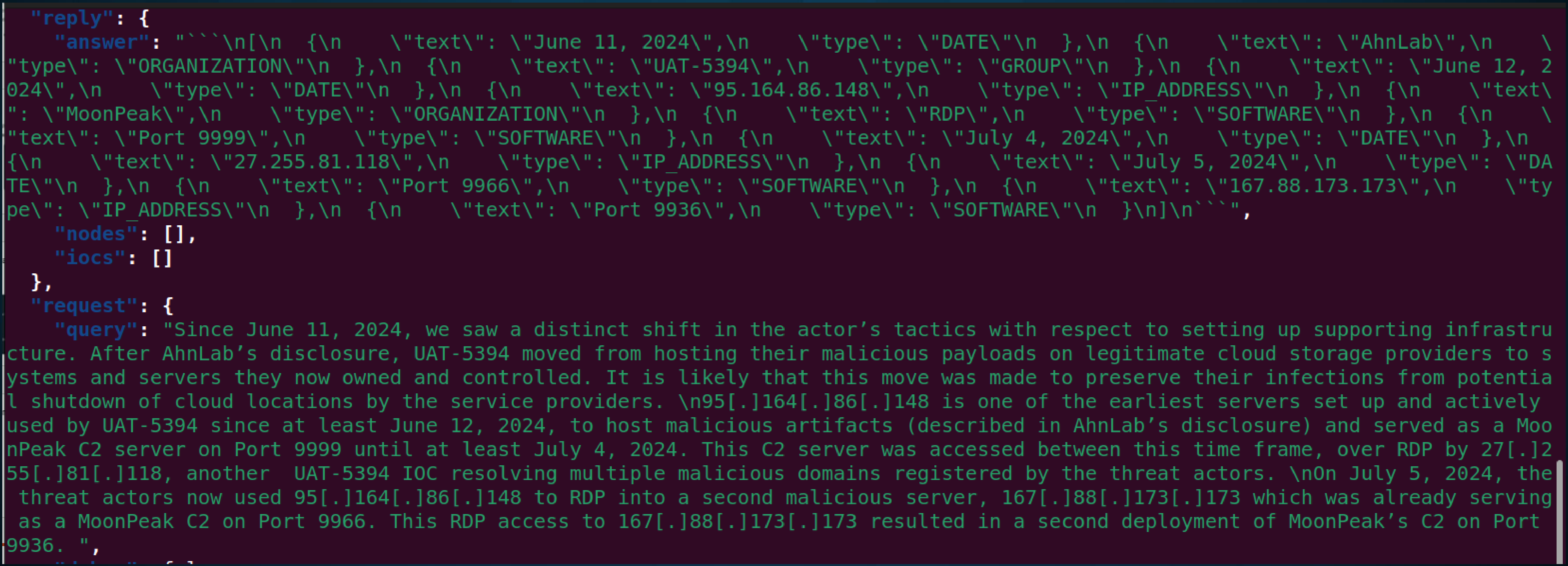
Run call tools

Run python code



```
# for Search
<|python_tag|>
brave_search.call(query="...")
<|eom_id|>

# for Wolfram
<|python_tag|>
wolfram_alpha.call(query="...")
<|eom_id|>
```

Prompts - examples



```
  "reply": {
    "answer": "```\n[\n  {\n    \"text\": \"June 11, 2024\",\n    \"type\": \"DATE\"\n  },\n  {\n    \"text\": \"AhnLab\",\n    \
"type\": \"ORGANIZATION\"\n  },\n  {\n    \"text\": \"UAT-5394\",\n    \"type\": \"GROUP\"\n  },\n  {\n    \"text\": \"June 12, 2
024\",\n    \"type\": \"DATE\"\n  },\n  {\n    \"text\": \"95.164.86.148\",\n    \"type\": \"IP_ADDRESS\"\n  },\n  {\n    \"text\
": \"MoonPeak\",\n    \"type\": \"ORGANIZATION\"\n  },\n  {\n    \"text\": \"RDP\",\n    \"type\": \"SOFTWARE\"\n  },\n  {\n    \
"text\": \"Port 9999\",\n    \"type\": \"SOFTWARE\"\n  },\n  {\n    \"text\": \"July 4, 2024\",\n    \"type\": \"DATE\"\n  },\n
{\n    \"text\": \"27.255.81.118\",\n    \"type\": \"IP_ADDRESS\"\n  },\n  {\n    \"text\": \"July 5, 2024\",\n    \"type\": \"DA
TE\"\n  },\n  {\n    \"text\": \"Port 9966\",\n    \"type\": \"SOFTWARE\"\n  },\n  {\n    \"text\": \"167.88.173.173\",\n    \"ty
pe\": \"IP_ADDRESS\"\n  },\n  {\n    \"text\": \"Port 9936\",\n    \"type\": \"SOFTWARE\"\n  }\n]\n```",
    "nodes": [],
    "iocs": []
  },
  "request": {
    "query": "Since June 11, 2024, we saw a distinct shift in the actor’s tactics with respect to setting up supporting infrastru
cture. After AhnLab’s disclosure, UAT-5394 moved from hosting their malicious payloads on legitimate cloud storage providers to s
ystems and servers they now owned and controlled. It is likely that this move was made to preserve their infections from potentia
l shutdown of cloud locations by the service providers. \n95[.]164[.]86[.]148 is one of the earliest servers set up and actively
used by UAT-5394 since at least June 12, 2024, to host malicious artifacts (described in AhnLab’s disclosure) and served as a Moo
nPeak C2 server on Port 9999 until at least July 4, 2024. This C2 server was accessed between this time frame, over RDP by 27[.]2
55[.]81[.]118, another  UAT-5394 IOC resolving multiple malicious domains registered by the threat actors. \nOn July 5, 2024, the
 threat actors now used 95[.]164[.]86[.]148 to RDP into a second malicious server, 167[.]88[.]173[.]173 which was already serving
 as a MoonPeak C2 on Port 9966. This RDP access to 167[.]88[.]173[.]173 resulted in a second deployment of MoonPeak’s C2 on Port
9936. ",
```

Prompts

Let's look at the code

Links

https://docs.llamaindex.ai/en/stable/
https://qdrant.tech/
https://github.com/spotify/annoy
https://github.com/facebookresearch/faiss
https://python.langchain.com/v0.2/docs/integrations/platforms/
https://weaviate.io/rag
https://azure.microsoft.com/en-us/blog/introducing-phi-3-redefining-whats-possible-with-slms/
https://llama.meta.com/ -
https://www.rungalileo.io/blog/mastering-rag-advanced-chunking-techniques-for-llm-applications
https://python.langchain.com/v0.2/docs/integrations/document_loaders/
https://github.com/openai/tiktoken
https://github.com/huggingface/tokenizers
https://medium.com/@prudhviraju.srivatsavaya/embedding-layer-vs-tokenizer-a1e4ade764e3
https://llama.meta.com/docs/how-to-guides/prompting/
https://llama.meta.com/docs/model-cards-and-prompt-formats/llama3_1

CISCO

TALOS

TALOSINTELLIGENCE.COM