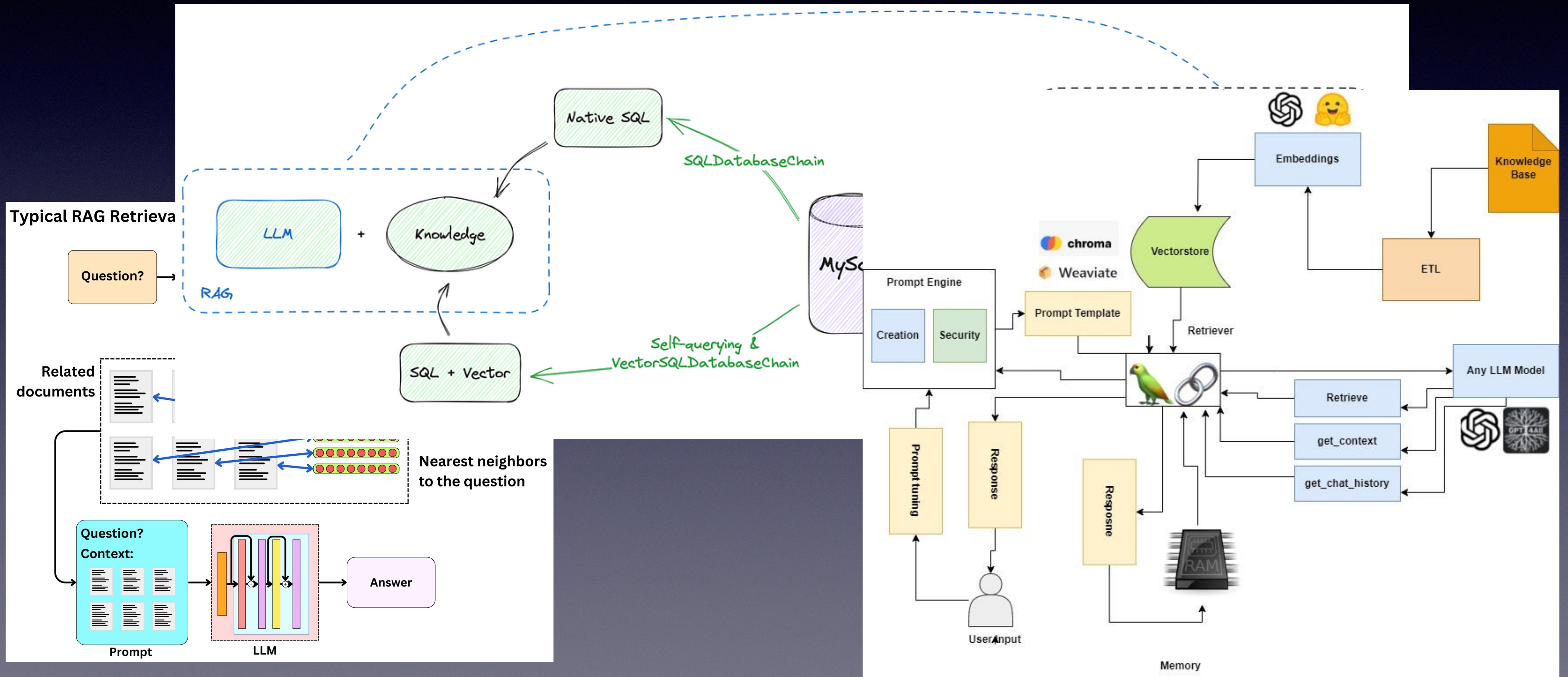


# vlite (2)

simple vector database made in numpy



# Problem





# Motivation

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! \*\*\*\*\* Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! \*\*\*\*\* Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! \*\*\*\*\* Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! \*\*\*\*\* Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! \*\*\*\*\* Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

☐

Railway

Deployment

...Crashed! Uh oh. Your deployment for chatpdf in [chatwithpdf](#)

 <https://github.com> > [chroma-core](#) > [chroma](#) > [issues](#) 

## Bug: querying collection when no data added causes crash ...

Apr 12, 2023 — chroma-core / chroma Public. Notifications ... Bug: querying collection when no data added causes crash #341 ... When there is no data, don't error ...

Missing: ~~db~~ | Show results with: **db**

github.com  
https://github.com › chroma-core › chroma › issues

## [Bug]: Illegal instruction (core dumped) · Issue #506

May 9, 2023 — Trying databases other than **Chroma** as **Chroma DB crashes** with certain types of GPU PromtEngineer/localGPT#468. Closed. @catriri. Copy link ...

github.com  
https://github.com › hwchase17 › langchain › issues

## Chroma DB Error · Issue #7260 · langchain-ai / langchain

Jul 6, 2023 — Answer. From your description, it appears that you're encountering two main **issues**. The first is that the chromadb package is not installed, and ...

Missing: ~~erashing~~ | Show results with: **crashing**

github.com  
https://github.com › urig › chroma-crash ⋮

urig/chroma-crash

## Reproducing a **crash** of **Chroma DB** when running in Google Cloud Run - [urig/chroma-crash](#).

github.com  
https://github.com › hwchase17 › langchain › issues

## Error while loading saved index in chroma db · Issue #2491

Apr 6, 2023 — RuntimeError: **Chroma** is running in http-only client mode, and can only be run with 'chromadb.api.fastapi.FastAPI' as the chroma\_api\_impl. see ...

```

2023-05-22 23:26:13 DEBUG chromadb.db.index.hnswlib time to pre process our knn query: 7.128715515136719e-05
2023-05-22 23:26:13 DEBUG chromadb.db.index.hnswlib time to run knn query: 0.806598078201293945
2023-05-22 23:26:13 INFO uvicorn.access 172.25.0.1:33282 - "POST /api/v1/collections/chatpdf_collection/query HTTP/1.1" 200
2023-05-22 23:26:16 ERROR chromadb.server.fastapi dictionary changed size during iteration
Traceback (most recent call last):
  File "/usr/local/lib/python3.10/site-packages/anyio/streams/memory.py", line 94, in receive
    return self.receive_nowait()
  File "/usr/local/lib/python3.10/site-packages/anyio/streams/memory.py", line 89, in receive_nowait
    raise WouldBlock
anyio.WouldBlock

```

```
During handling of the above exception, another exception occurred:
```

```
Traceback (most recent call last):
  File "/usr/local/lib/python3.10/site-packages/starlette/middleware/base.py", line 43, in call_next
    message = await recv_stream.receive()
  File "/usr/local/lib/python3.10/site-packages/anyio/streams/memory.py", line 114, in receive
    raise EOFStream
anyio.EOFStream
```

```
During handling of the above exception, another exception occurred:
```

```
Traceback (most recent call last):
  File "/chroma/.chromadb/server/fastapi/_init_.py", line 47, in catch_exceptions_middleware
    return await call_next(request)
  File "/usr/local/lib/python3.10/site-packages/starlette/middleware/base.py", line 46, in call_next
    raise app_exc
  File "/usr/local/lib/python3.10/site-packages/starlette/middleware/base.py", line 36, in coro
    await self.app(scope, request.receive, send_stream.send)
  File "/usr/local/lib/python3.10/site-packages/starlette/middleware/exceptions.py", line 75, in __call__
    raise exc
  File "/usr/local/lib/python3.10/site-packages/starlette/middleware/exceptions.py", line 64, in __call__
    await self.app(scope, receive, sender)
  File "/usr/local/lib/python3.10/site-packages/fastapi/middleware/asyncexitstack.py", line 21, in __call__
    raise e
  File "/usr/local/lib/python3.10/site-packages/fastapi/middleware/asyncexitstack.py", line 18, in __call__
    await self.app(scope, receive, send)
  File "/usr/local/lib/python3.10/site-packages/starlette/routing.py", line 680, in __call__
    await route.handle(scope, receive, send)
  File "/usr/local/lib/python3.10/site-packages/starlette/routing.py", line 275, in handle
    await self.app(scope, receive, send)
  File "/usr/local/lib/python3.10/site-packages/starlette/routing.py", line 65, in app
    response = await func(request)
  File "/usr/local/lib/python3.10/site-packages/fastapi/routing.py", line 231, in app
    raw_response = await run_endpoint_function(
  File "/usr/local/lib/python3.10/site-packages/fastapi/routing.py", line 162, in run_endpoint_function
    return await run_in_threadpool(dependant.call, **values)
  File "/usr/local/lib/python3.10/site-packages/starlette/concurrency.py", line 41, in run_in_threadpool
    return await anyio.to_thread.run_sync(func, *args)
  File "/usr/local/lib/python3.10/site-packages/anyio/to_thread.py", line 31, in run_sync
    return await get_asyncio_loop().run_sync_in_worker_thread(
  File "/usr/local/lib/python3.10/site-packages/anyio/_backends/_asyncio.py", line 937, in run_sync_in_worker_thread
    return await future
  File "/usr/local/lib/python3.10/site-packages/anyio/_backends/_asyncio.py", line 867, in run
    result = context.run(func, *args)
  File "/chroma/.chromadb/server/fastapi/_init_.py", line 158, in add
    result = self._api._add(
  File "/chroma/.chromadb/api/local.py", line 140, in _add
    self._db.add_incremental(collection_uuid, added_uuids, embeddings)
  File "/chroma/.chromadb/db/clickhouse.py", line 542, in add_incremental
    index.add(uuids, embeddings)
  File "/chroma/.chromadb/db/index/hnswlib.py", line 148, in add
    self._save()
  File "/chroma/.chromadb/db/index/hnswlib.py", line 187, in _save
    pickle.dump(self._label_to_id, f, pickle.HIGHEST_PROTOCOL)
RuntimeError: dictionary changed size during iteration
2023-05-22 23:26:11 INFO uvicorn.access 172.25.0.1:39262 - "POST /api/v1/collections/chatpdf_collection/add HTTP/1.1"
```



# Solution

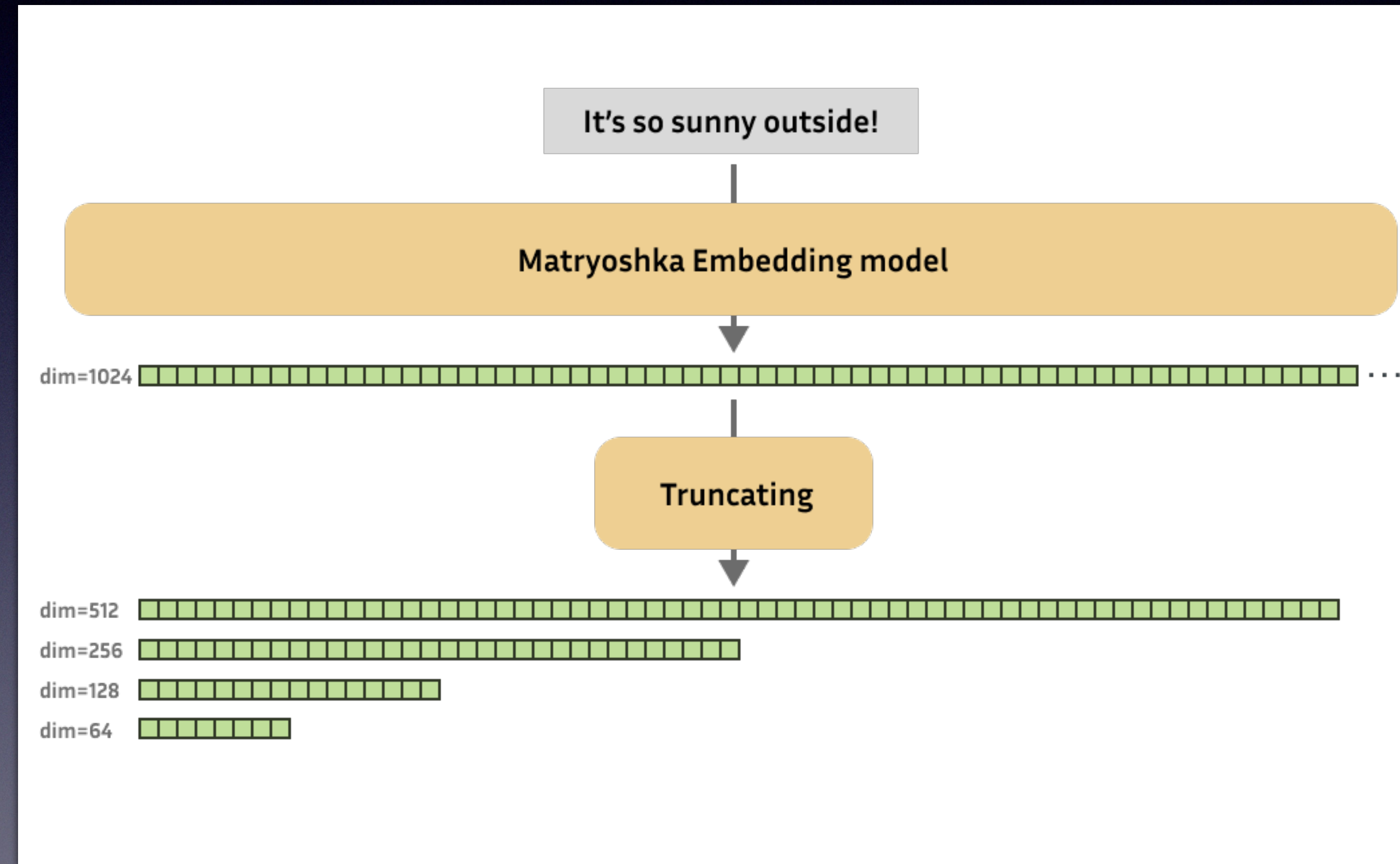
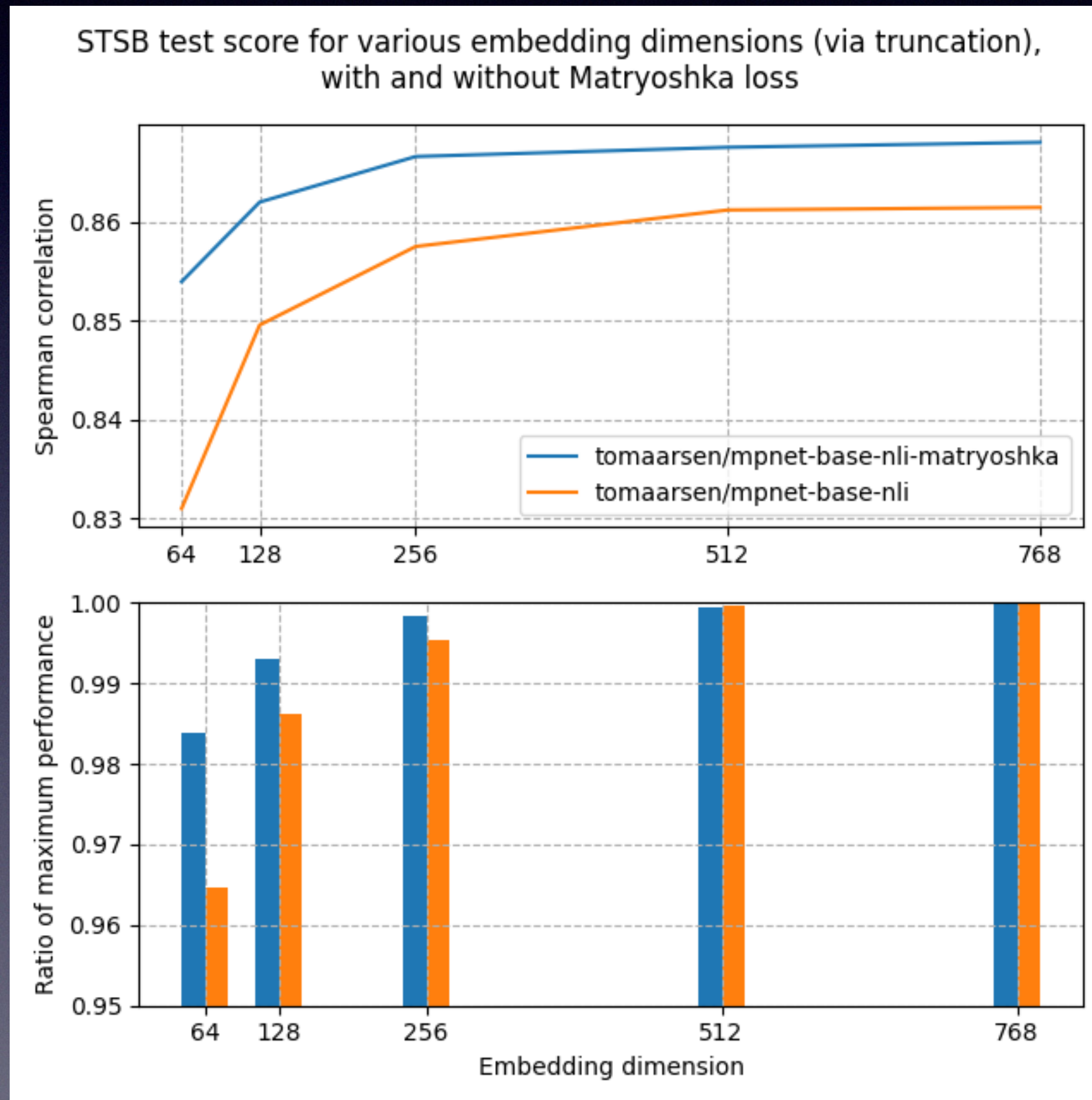


# vlite (2)

- in memory db
- Just uses numpy arrays, matrices
- 🔥 *Fastest* vector db retrieval with binary embeddings, less than 1.1s to search 500k documents
- >77.95% faster than Chroma on indexing, >422% faster on retrieval, and >3.6x smaller on disk
- 🦜 in langchain 0.1.17 (today)
- Ingest text, PDF, CSV, PPTX, webpages, OCR out of the box using surya-ocr



# MRL



From: <https://huggingface.co/blog/matryoshka>



# Binary Quantized Embeddings

- Most embedding models: 1024 dim at float32, 4 bytes per dim
- Mixedbread model : 1024 dim at binary, 128 dim at int8

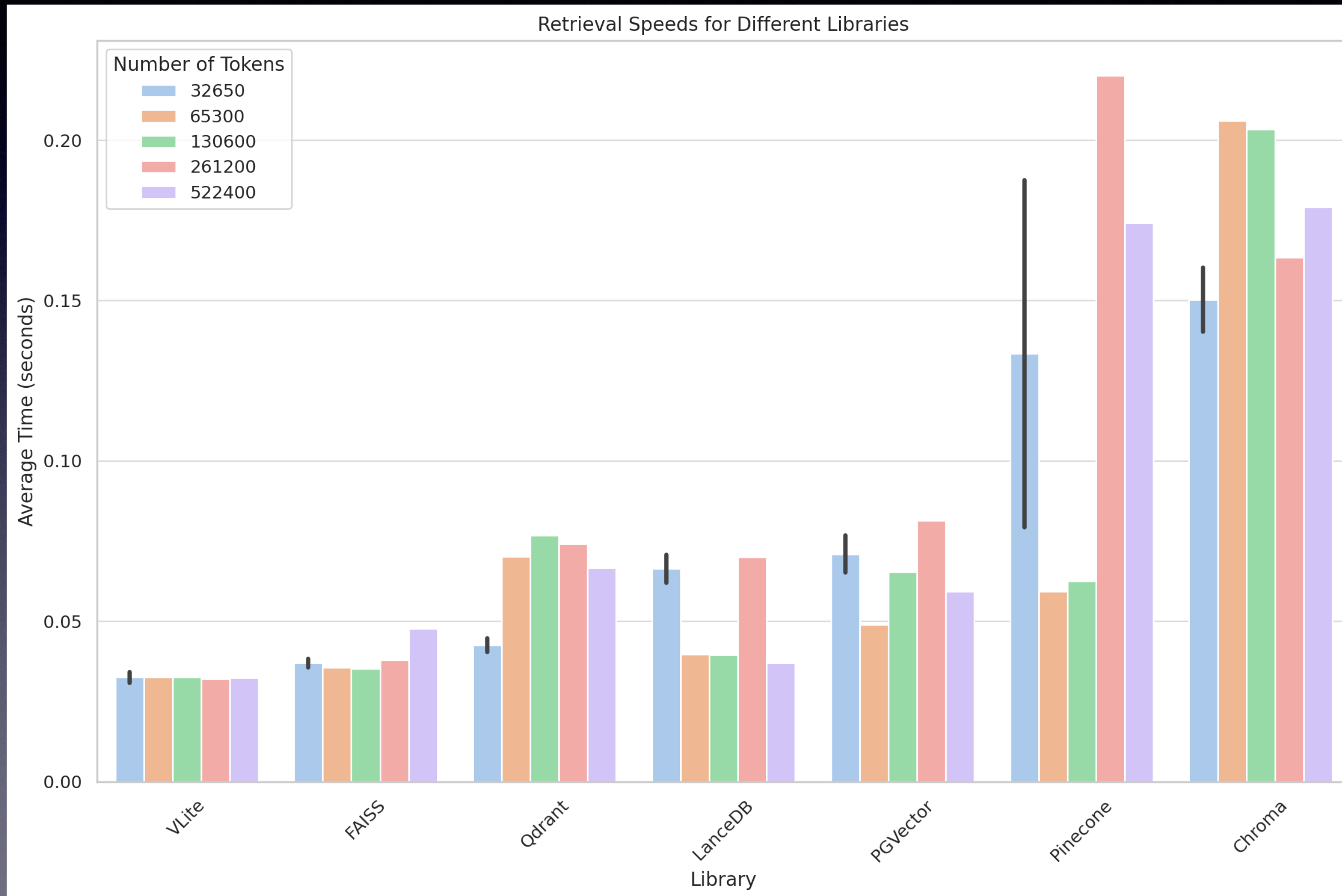


# Docs

- `add(data: List[str], metadata: Dict)`
  - Long, short, multiple or just one piece of text
    - Supports 128 dims at int8 w/ sota mixedbread large run locally
- `retrieve(string: str, top_k: int)`
  - Something you want to recall
    - Returns similarity scores and matching texts from corpus



# Results





# CTX files

## CTX

ID: CTXF

Version: 1

Header: {"embedding\_size": 64, "context\_length": 512 ...}

Embeddings: [-157 -139 -108 ...]

Contexts: ["birds aren't real"]

Metadata: {"created\_at": "2024-04-18", ...}

```
huggingface/tokenizers: The current
To disable this warning, you can e
    - Avoid using `tokenizers`
    - Explicitly set the environme
560K data/attention.pdf
huggingface/tokenizers: The current
To disable this warning, you can e
    - Avoid using `tokenizers`
    - Explicitly set the environme
220K contexts/vlite-500.ctx
```



# Create dumb tools

As interfaces need more data from you to make their algorithms better, embeddings allow for higher fidelity representations on your data and can be used with llms directly — creating a standard for embedding cookies, similar to browser cookies in browsers is important for composability and portability.

This could exist via a sign in with embeddings option or by importing your data or through a new browser made for LLMs

This requires a ridiculously lightweight, simple vectordb not focused on precision or enterprise customers but focused on “just working” without dependency hell or slow retrieval times

vlite 2 is all of that and more. It just works. no databases(its just a dictionary), lightweight (~25kb), and entirely built on a clean and small file system called CTX which saves all your vlite data in .ctx files





# Future

- Swift/JS libraries
  - Use CTX files to be used between all different libraries and stuff them into a browser cookie / browser SQL
- Run embedding models using Web GPU

Questions, comments, concerns:

<https://github.com/sdan>

@sdand



A small list of short-term, high-impact projects for hackers, builders, and anime pfp X accounts.

BOUNTIES LISTED

vlite compatible chroma client

python

10d 21h 53m 29s \$1000

More

Build a silent autonomous leaf blowing or collecting robot

11d 8h 25m 54s \$15000

More

Lowest Total Cost Controller for Comma Challenge

Python

Control Th...

12d 23h 46m 48s \$250

More

Most Interesting Controller for Comma Challenge (RL?)

Python

Reinforcem...

12d 23h 46m 48s \$250

More

Translate diffusion models from scratch

diffusion ...

math termi...

244d 23h 46m 49s \$50

More

Build a zipline from Bernal heights to the ferry building


zipline

constructi...

28d 3h 7m 2s \$1000

More



	find god	
Stop texting Go find God Come back after you found God		
Trust God		2018-12-14
FOLLOW GOD <a href="https://t.co/JlXaaO6Ps5">https://t.co/JlXaaO6Ps5</a> <a href="https://t.co/PR2LXa4Sa9">https://t.co/PR2LXa4Sa9</a>		2019-11-08
replace pride with love		Posted July 29, 2018
stay inspired		2018-08-21
stop trying start doing		Posted July 27, 2018
question everything		Posted April 22, 2018
Keep moving forward		2019-01-01
Should I name more?		Posted July 22, 2020
Raise confidence and consciousness		2018-11-08