The Hitchicker's Guide to Vector Databases

Tips and tricks not to get lost in vector space!

Clelia Astra Bertelli, 12/06/2025

Who am I?

Who am I?

Education

BSc in ***Biological Sciences*** at University of Pavia (almost completed!)

Working experience

- **Founding Engineer** at Criad Ltd
- **AI Engineer** at LegalForLandlords
- **Open Source Engineer** at LlamaIndex

LlamaIndex

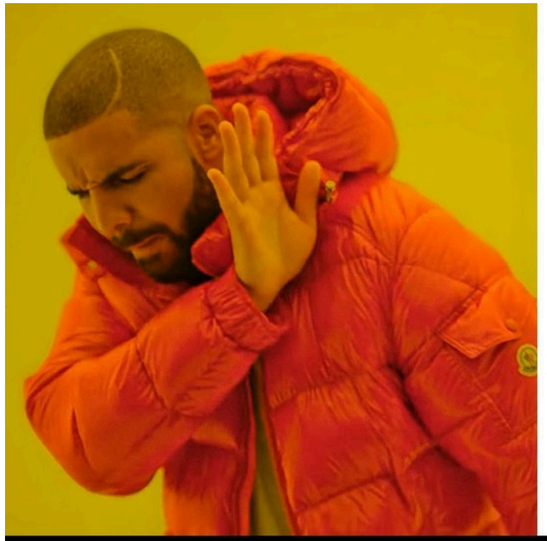
(serious pic)

Random Facts

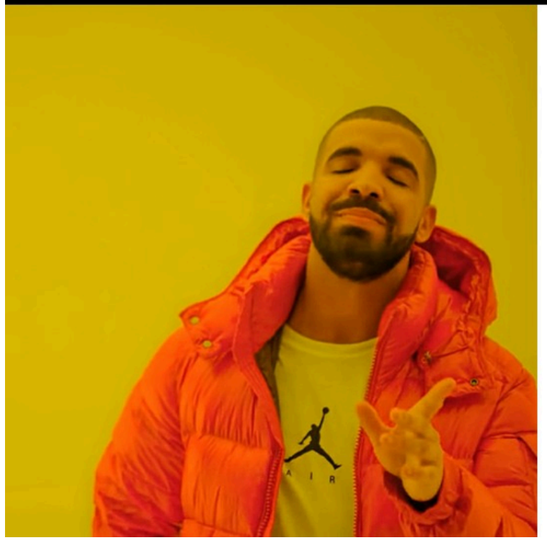
- I'm from Italy, but moving to Germany!
- I love cats, travelling, art and museums
- I'm in love with Vienna
- Coding is not only a job, it's a passion since I was 17

(happy pic)

An overview of (text) RAG

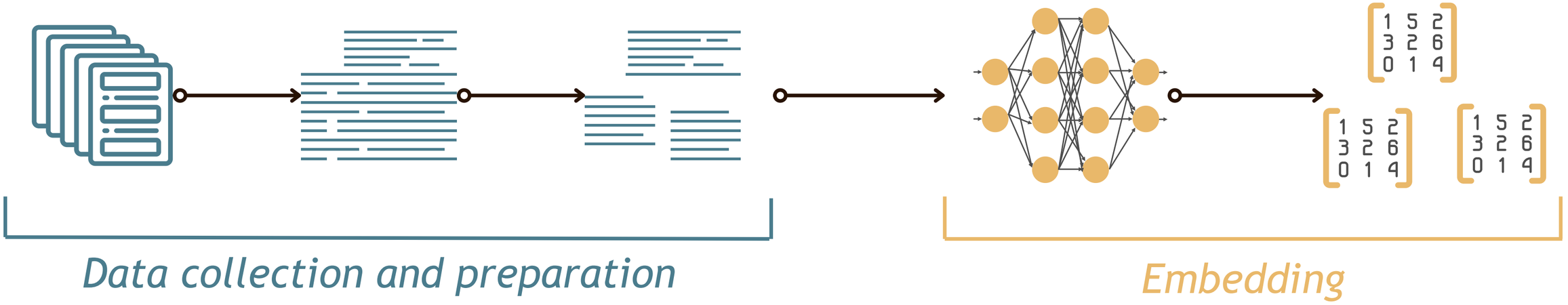


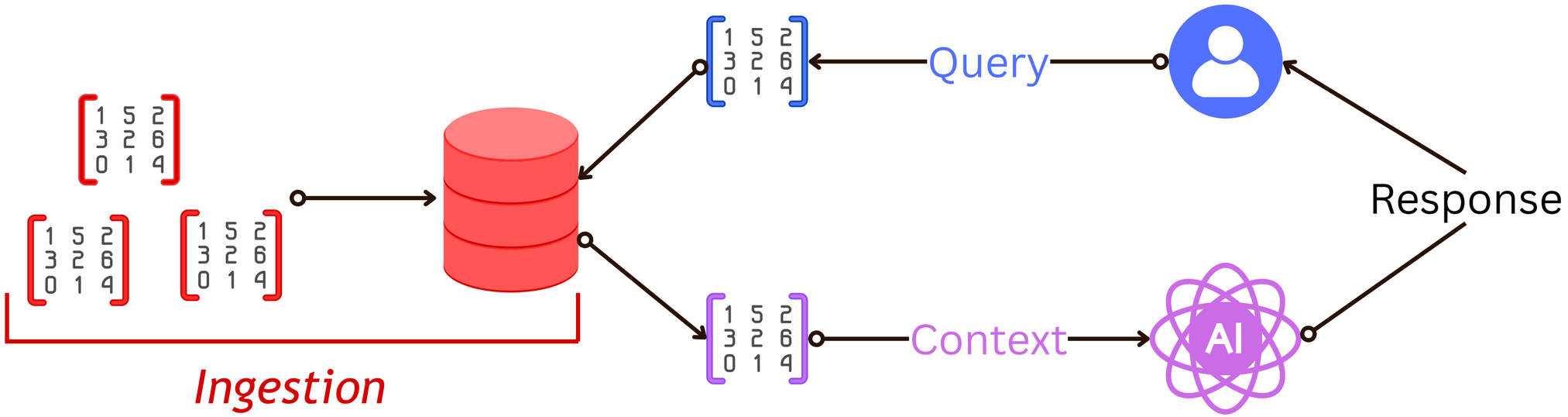
Talking
to ChatGPT



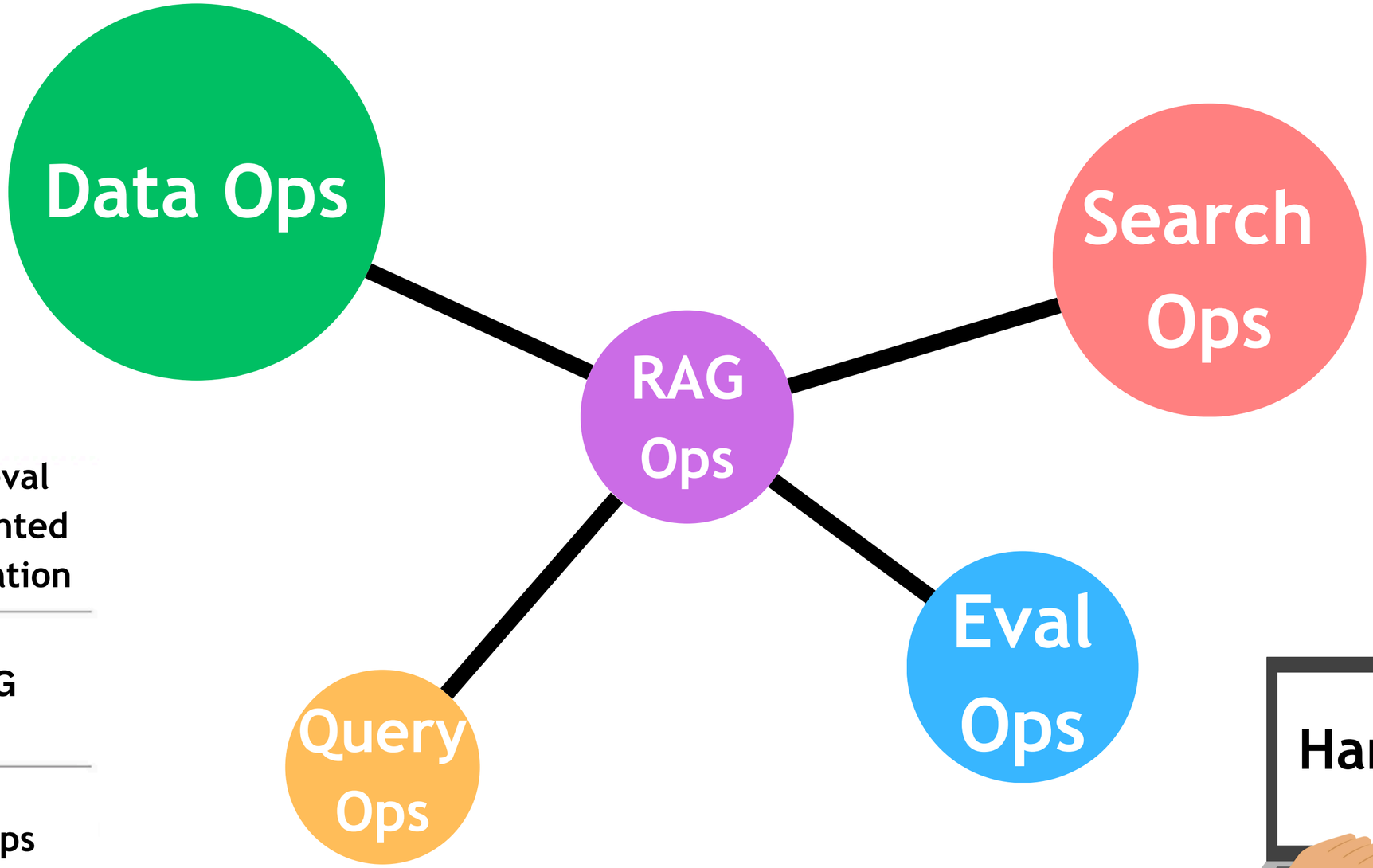
Collecting data, process them, spin up a vector db, ingest data into it, connect vector db to local LLM, ask questions about data

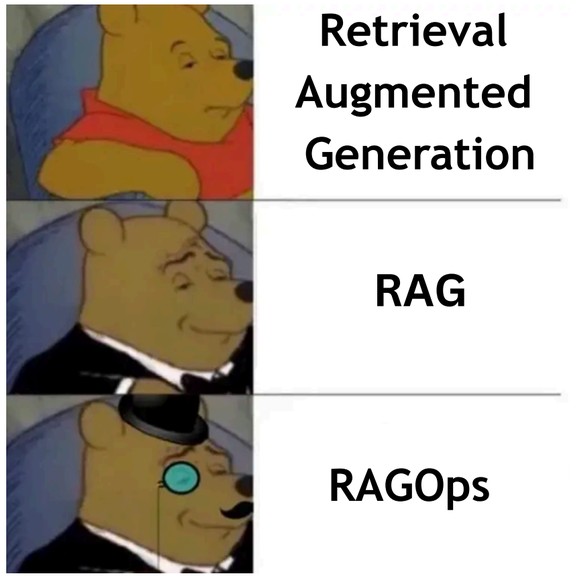
An overview of (text) RAG





What we'll work on





How we will work

QR code with link to the repo

An explanatory part

A sneak peek into the code!

Text extraction: your best friend and worst enemy

PDF

Note

PKI stands for Pyruvate Kinase 1 (produced by the pykF gene, so aka PykF). The isoform is PKII, or Pyruvate Kinase 2, produced by the pykA gene, so aka PykA

- What was the initial source of PKI in this experiment, and how was its presence initially confirmed?
- The initial source of PKI (Pyruvate Kinase I) in this experiment is an extract of centrifuged *E. coli*, grown overnight in a liquid terrain and engineered to overexpress the PKI gene. Once the proteins are extracted from the bacterial material, an electrophoretic approach (SDS PAGE) is taken to identify the presence of a band (around 40-50kDa) where the PKI should be placed. The presence of the band does not give the certainty, but strongly indicates that PKI is in the bacterial mix.

- What is the purpose of the heat treatment step performed at 50°C for 15 minutes, and what principle does it exploit?
- The purpose of the heat treatment step (performed at 50°C for 15 minutes) is to denaturate the proteins that are heat-sensitive, coarsely purifying the PKI (and other heat-insensitive proteins) from other proteins and especially from its (otherwhise difficult to distinguish) isoform, PKII

- How does centrifugation help in separating proteins after the heat treatment?
- Centrifugation (that happens at 4000 rm for 8 minutes) helps the separation of denaturated (heat-sensitive) proteins from the non-denaturated (heat-insensitive) proteins because the denaturated proteins clump together and precipitate, whereas the other ones remain in solution.

Extracted text

```
Note
PKI stands for Pyruvate Kinase 1 (produced by the pykF gene, so aka PykF). The isoform is PKII, or Pyruvate Kinase 2, pr
oduced by
the pykA  gene, so aka PykA
What was the initial source of PKI in this experiment, and how was its presence initially confirmed?
The initial source of PKI (Pyruvate Kinase I) in this experiment is an extract of centrifuged E. coli , grown overnight
in a liquid terrain and
engineered to overexpress the PKI gene. Once the proteins are extracted from the bacterial material, an electrophoretic
approach (SDS
P AGE) is taken to identify the presence of a band (around 40-50kDa) where the PKI should be placed. The presence of the
 band does
not give the certainty , but strongly indicates that PKI is in the bacterial mix.
What is the purpose of the heat treatment step performed at 50°C for 15 minutes, and what principle does it exploit?
The purpose of the heat treatment step (performed at 50°C for 15 minutes) is to denaturate the proteins that are heat-se
nsitive,
coarsely purifying the PKI (and other heat-insensitive proteins) from other proteins and especially from its (otherwise
 dif ficult to
distinguish) isoform, PKII
How does centrifugation help in separating proteins after the heat treatment?
Centrifugation (that happens at 4000 rm for 8 minutes) helps the separation of denaturated (heat-sensitive) proteins fro
m the non-
denaturated (heat-insensitive) proteins because the denaturated proteins clump together and precipitate, whereas the oth
er ones
remain in solution.
```

Markdown | Text | JSON | Images | Layout | XLSX | PDF

Note

PKI stands for Pyruvate Kinase 1 (produced by the pykF gene, so aka PykF). The isoform is PKII, or Pyruvate Kinase 2, produced by the pykA gene, so aka PykA.

What was the initial source of PKI in this experiment, and how was its presence initially confirmed?

The initial source of PKI (Pyruvate Kinase I) in this experiment is an extract of centrifuged E. coli, grown overnight in a liquid terrain and engineered to overexpress the PKI gene. Once the proteins are extracted from the bacterial material, an electrophoretic approach (SDS PAGE) is taken to identify the presence of a band (around 40-50kDa) where the PKI should be placed. The presence of the band does not give the certainty, but strongly indicates that PKI is in the bacterial mix.

What is the purpose of the heat treatment step performed at 50°C for 15 minutes, and what principle does it exploit?

The purpose of the heat treatment step (performed at 50°C for 15 minutes) is to denaturate the proteins that are heat-sensitive, coarsely purifying the PKI (and other heat-insensitive proteins) from other proteins and especially from its (otherwise difficult to distinguish) isoform, PKII.

How does centrifugation help in separating proteins after the heat treatment?

Trust your guts on text extraction

AstraBert/ingest-anything

Sometimes
it is worth to invest
some more
resource and
extract text better
to avoid
downstream
headaches!🤯

LlamaCloud

Markdown | Text | JSON | Images | Layout | XLSX | PDF

Note

PKI stands for Pyruvate Kinase 1 (produced by the pykF gene, so aka PykF). The isoform is PKII, or Pyruvate Kinase 2, produced by the pykA gene, so aka PykA.

What was the initial source of PKI in this experiment, and how was its presence initially confirmed?

The initial source of PKI (Pyruvate Kinase I) in this experiment is an extract of centrifuged E. coli, grown overnight in a liquid terrain and engineered to overexpress the PKI gene. Once the proteins are extracted from the bacterial material, an electrophoretic approach (SDS PAGE) is taken to identify the presence of a band (around 40-50kDa) where the PKI should be placed. The presence of the band does not give the certainty, but strongly indicates that PKI is in the bacterial mix.

What is the purpose of the heat treatment step performed at 50°C for 15 minutes, and what principle does it exploit?

The purpose of the heat treatment step (performed at 50°C for 15 minutes) is to denaturate the proteins that are heat-sensitive, coarsely purifying the PKI (and other heat-insensitive proteins) from other proteins and especially from its (otherwise difficult to distinguish) isoform, PKII.

How does centrifugation help in separating proteins after the heat treatment?

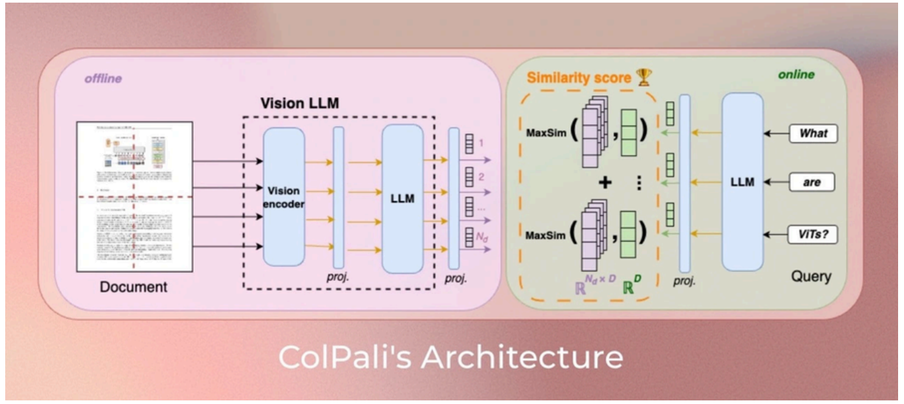
OCR+Agentic Parsing

pypdf

BACK TO THE ROOTS

```
Note
PKI stands for Pyruvate Kinase 1 (produced by the pykF gene, so aka PykF). The isoform is PKII, or Pyruvate Kinase 2, pr
oduced by
the pykA  gene, so aka PykA
What was the initial source of PKI in this experiment, and how was its presence initially confirmed?
The initial source of PKI (Pyruvate Kinase I) in this experiment is an extract of centrifuged E. coli , grown overnight
in a liquid terrain and
engineered to overexpress the PKI gene. Once the proteins are extracted from the bacterial material, an electrophoretic
approach (SDS
P AGE) is taken to identify the presence of a band (around 40-50kDa) where the PKI should be placed. The presence of the
 band does
not give the certainty , but strongly indicates that PKI is in the bacterial mix.
What is the purpose of the heat treatment step performed at 50°C for 15 minutes, and what principle does it exploit?
The purpose of the heat treatment step (performed at 50°C for 15 minutes) is to denaturate the proteins that are heat-se
nsitive,
coarsely purifying the PKI (and other heat-insensitive proteins) from other proteins and especially from its (otherwise
 dif ficult to
distinguish) isoform, PKII
How does centrifugation help in separating proteins after the heat treatment?
Centrifugation (that happens at 4000 rm for 8 minutes) helps the separation of denaturated (heat-sensitive) proteins fro
m the non-
denaturated (heat-insensitive) proteins because the denaturated proteins clump together and precipitate, whereas the oth
er ones
remain in solution.
```

Object-based parsing



Multi-modal extraction

How to make vectors make sense

Default

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvails` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Sloppy

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvailabilityDidChangeNotification` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Ragged right

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvailabilityDidChangeNotification` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Default

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvails` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Sloppy

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvailabilityDidChangeNotification` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Ragged right

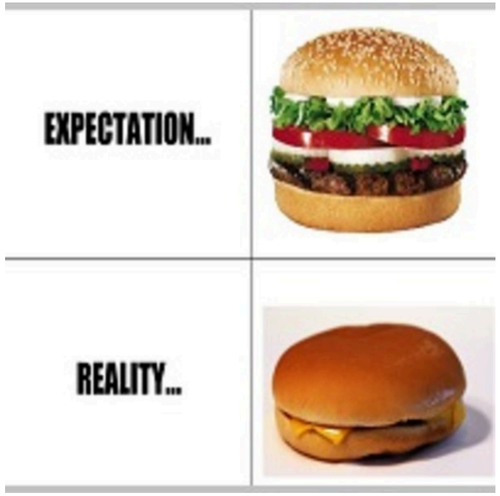
Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvailabilityDidChangeNotification` (yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

(yeech!). It is unacceptable to hyphenate these very long names or for them to disappear off the side of the page. Beyond that, the output should not look too ugly: there should be no extra-long white spaces, and long text paragraphs tend to look bad without right justification.

Ragged right

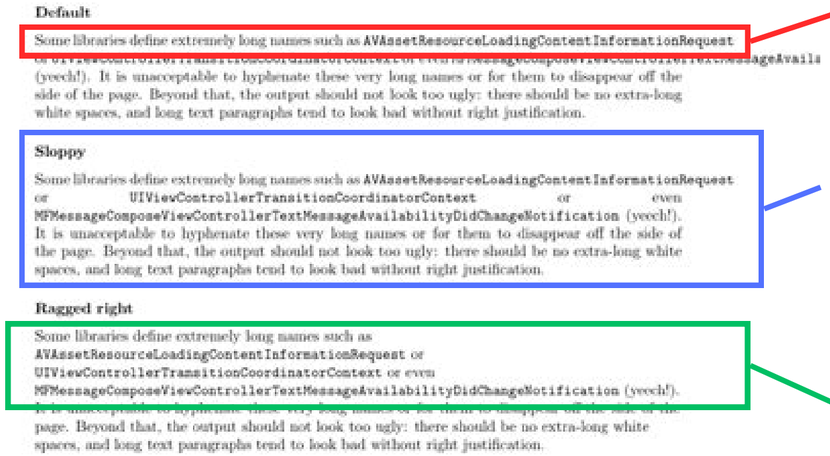
Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or

Some libraries define extremely long names such as `AVAssetResourceLoadingContentInformationRequest` or `UIViewControllerTransitionCoordinatorContext` or even `MFMessageComposeViewControllerTextMessageAvailabilityDidChangeNotification` (yeech!).



Chunking is all you need!

Sentence/token-based



Neural net or LLM-based

Semantic or embedding-based

It all boils down to the complexity
of your text and to the nuances that you
want to capture within it!

An example with code chunking

AstraBert/code-ragent

```python
class FileExistsWarning(Warning):
    """Warns you that a file exists"""

class DirPath(BaseModel):
    path: str
    @model_validator(mode="after")
    def validate_dir_path(self) -> Self:
        if Path(self.path).is_dir():
            if len(os.listdir(self.path)) == 0:
                raise ValueError("You should provide a non-empty directory")
            else:
                return self
        else:
            raise ValueError("You should provide the path for an existing directory")

class OutputPath(BaseModel):
    file: str
    @field_validator("file")
    def file_exists_warning(cls, file: str):
        if os.path.splitext(file)[1] != ".pdf":
            raise ValueError("Output file must be a PDF")
        p = Path(file)
        if p.is_file():
            warnings.warn(f"The file {file} already exists, you are about to overwrite it", FileExistsWarning)
        return file

class MultipleFileConversion(BaseModel):
    input_files: List[FilePath]
    output_files: List[str] | List[OutputPath] | None
    @model_validator(mode="after")
    def validate_multiple_file_conversion(self) -> Self:
        if self.output_files is not None and len(self.input_files) != len(self.output_files):
            raise ValueError("Input and output files must be lists of the same length")
        else:
            if self.output_files is None:
                self.output_files = [OutputPath(file=(fl.file.replace(os.path.splitext(fl.file)[1],".pdf"))) for fl in self.input_files]
            else:
                if isinstance(self.output_files[0], str):
                    self.output_files = [OutputPath(file=fl) for fl in self.output_files]
        return self
```

```python
from chonkie import CodeChunker

# Basic initialization with default parameters
chunker = CodeChunker(
    language="python",                  # Specify the programming language
    tokenizer_or_token_counter="gpt2", # Tokenizer to use
    chunk_size=512,                     # Maximum tokens per chunk
    include_nodes=False                 # Optionally include AST nodes in o
)
```

Use Abstract Syntax Trees (ABTs) to produce meaningful code chunks.

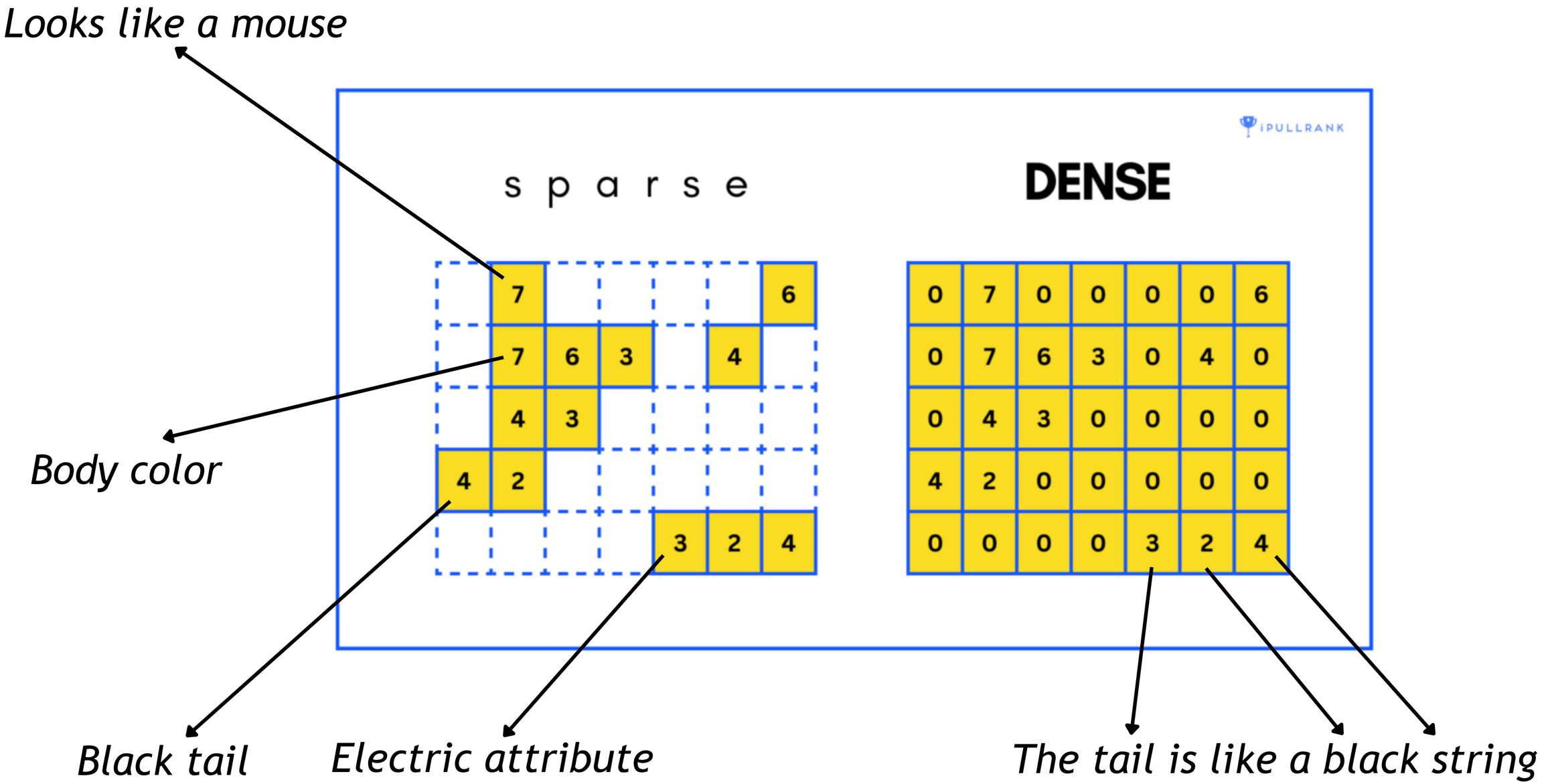
Gotta catch ‘em(beddings) all!

AstraBert/Pokemon-Bot

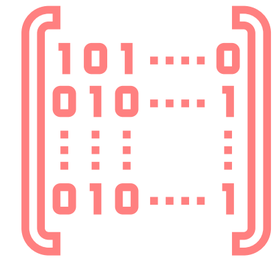
Electric attributes, light yellow little mouse Pokémon, with pink electric storage pouches in the cheeks, black patterns around the ears, and black tail

It has an electric attribute, looks like a mouse, its body is orange, its tail is like a black string, its beard is like an electric sac, and it has tentacles, and its expression is very cute.

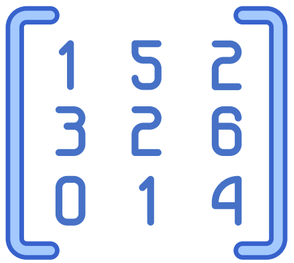
Gotta catch 'em(beddings) all!



Gotta catch ‘em(beddings) all!



Sparse



Dense

What mouse-like electric Pokemon has a black string tail?

1- Dedenne

...

10- Pichu

1- Dedenne

2- Pichu

...

The real (reran)king!

AstraBert/Pokemon-Bot

1- Dedenne

...

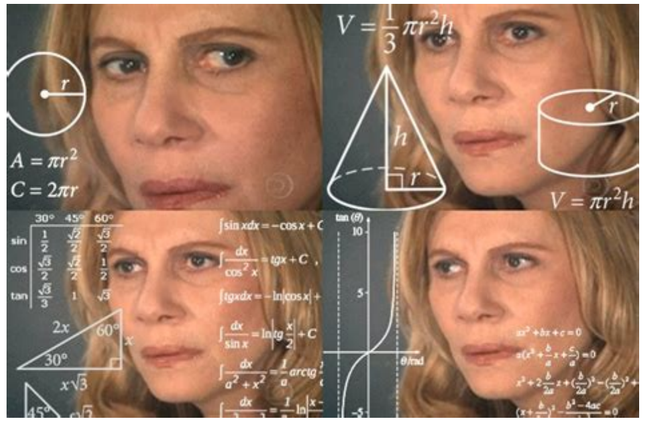
10- Pichu

1- Dedenne

2- Pichu

...

What is the right answer?



The right answer is Reranking!

The real (reran)king!

AstraBert/Pokemon-Bot

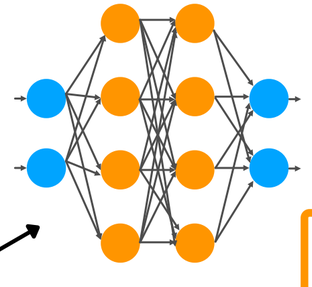
1- Dedenne
...
10- Pichu

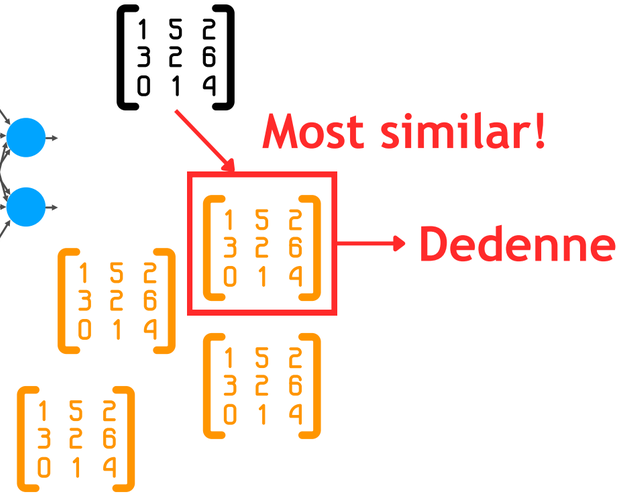
1- Dedenne
2- Pichu
...

Rescore

Original query

1- Dedenne
...
20- Pichu





Fuse the results

1- Dedenne
...
10- Pichu

+

1- Dedenne
2- Pichu
...

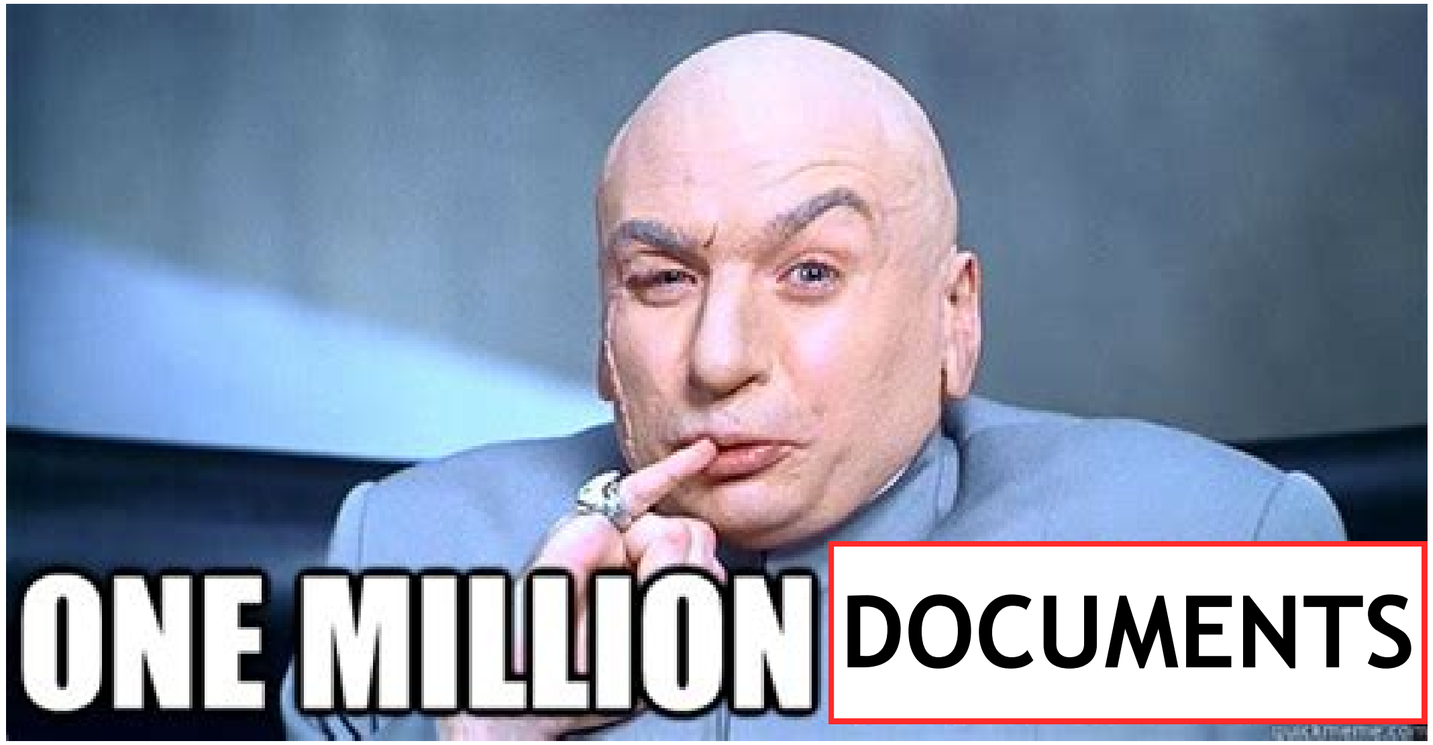
→

1- Dedenne
...
6- Pichu
...

Compression, baby!

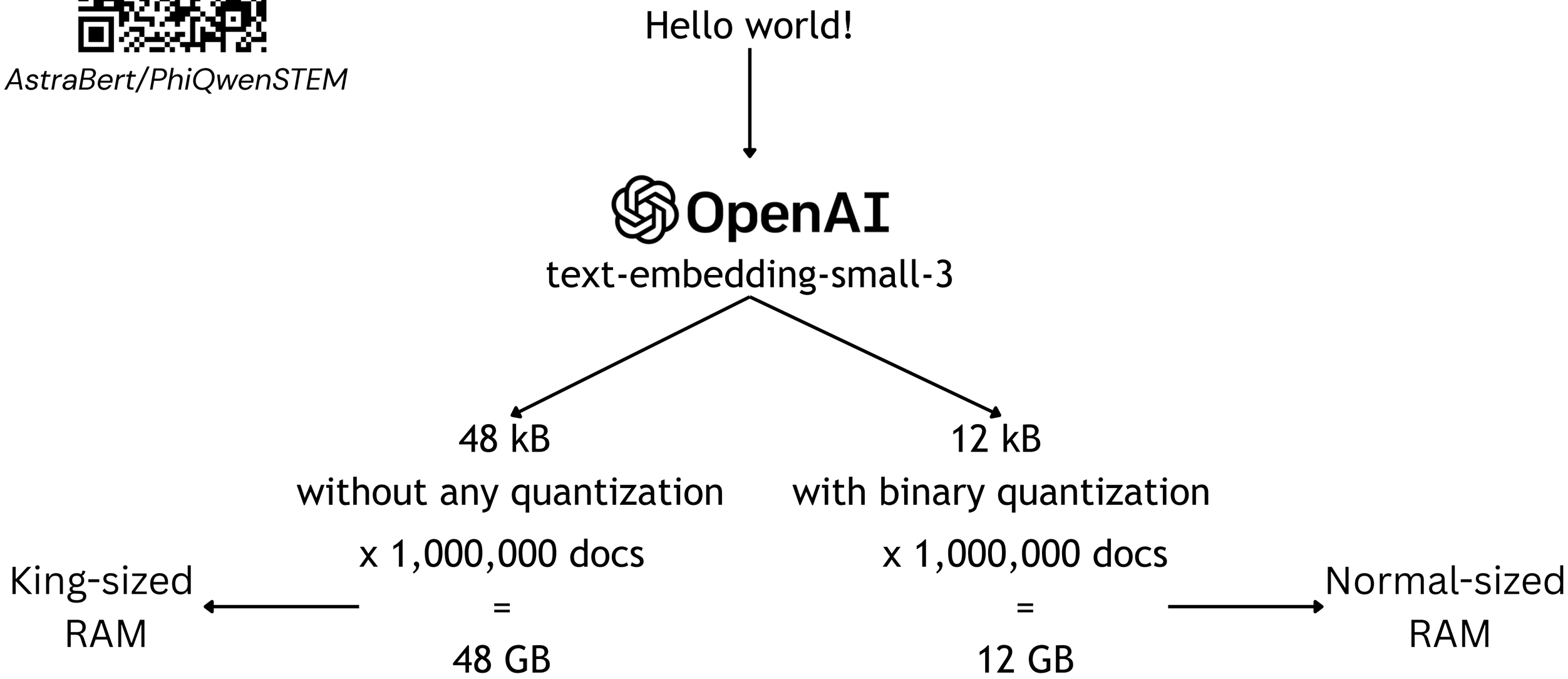
AstraBert/PhiQwenSTEM

Or how do you deal with...



Compression, baby!

AstraBert/PhiQwenSTEM



Compression, baby!

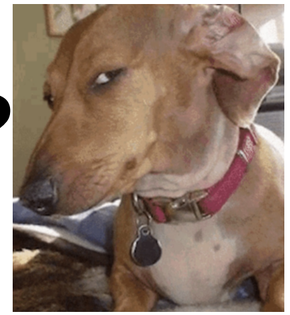
AstraBert/PhiQwenSTEM

✅ Binary quantization is memory-safe and efficient

✅ Binary quantization makes search easier and faster

✅ You can quantize *all* embeddings, being them text, image, multimodal...

But are the results good?



They are good, although not *as* good as if you didn't quantize

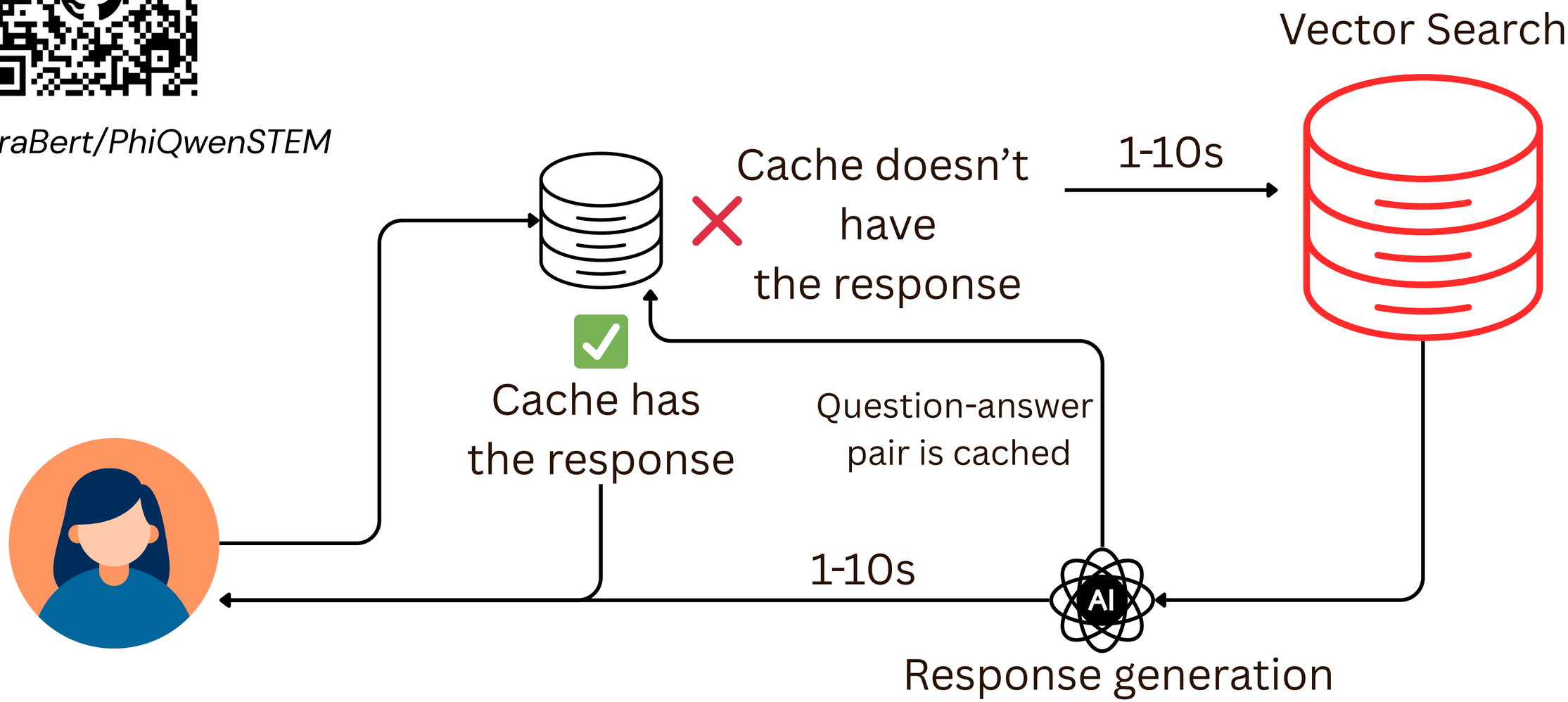
➡️

That's why you **re-rank after the first retrieval**, without quantization! And/or do **hybrid search** :)

Boost your search without searching

with semantic cache!

AstraBert/PhiQwenSTEM



To decompose or to expand? This is the problem!

AstraBert/ragcoon

(this is a RAGcoon)

Specific question

Generic Vectorized Documents

Generic, off-target answer

How do we fix this?

To decompose or to expand? This is the problem!

AstraBert/ragcoon

Generic question

No modifications

Specific question

Expand (HyDE)

Hypothetical Document

Complex question

Decompose

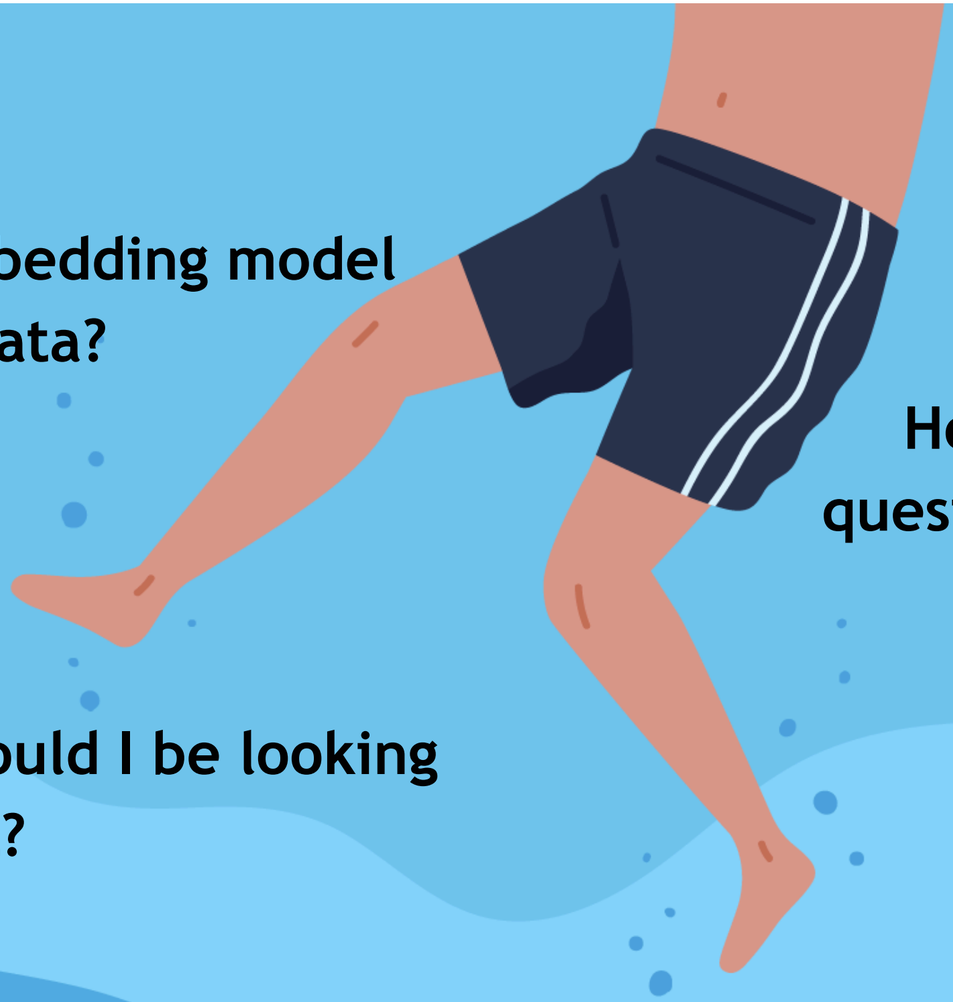
Question 1

Question 2

Question 3

Don’t drawn in evals

AstraBert/diRAGnosis



What’s the best embedding model for my data?

How can I come up with all the question-answers I need to evaluate my RAG?

What metrics should I be looking at?

What’s the best LLM for RAG?

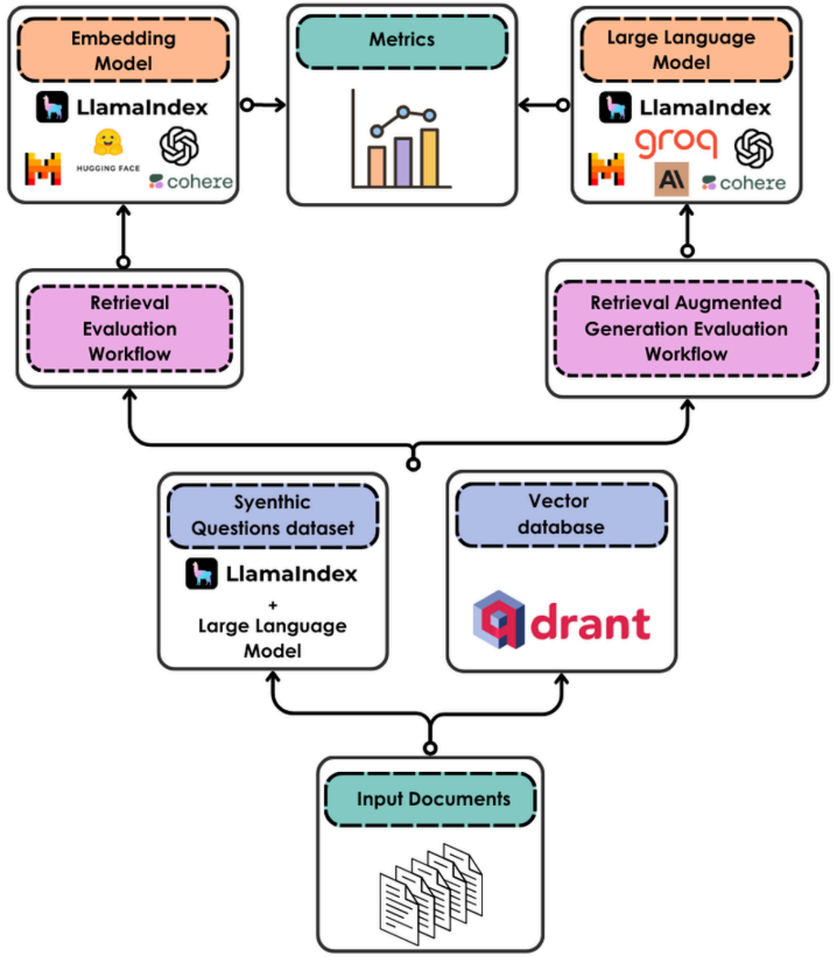
Don’t drawn in evals

AstraBert/diRAGnosis

Intuitive Metrics

- Faithfulness (LLM)
- Relevancy (LLM)
- Hit rate (embed)
- MRR (embed)

Synthetic test data generation (automatic)



Trial-and-error is key!

Choose *your own* LLM and embeddings model

Thank you for your attention!

If you don't wanna miss out
on my journey in the AI industry space
feel free to follow me on my social media accounts!