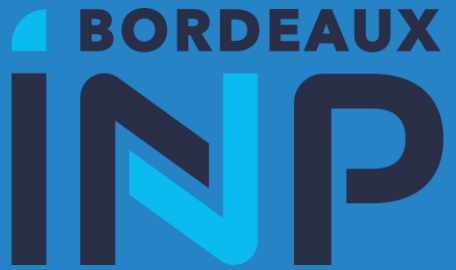
Knowledge Graph Generation

Project S9 Term 2023 - 2024

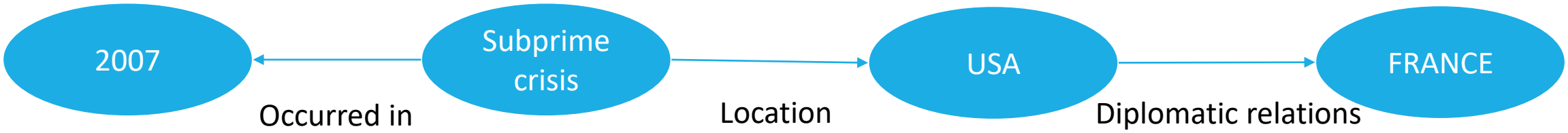


PLAN :

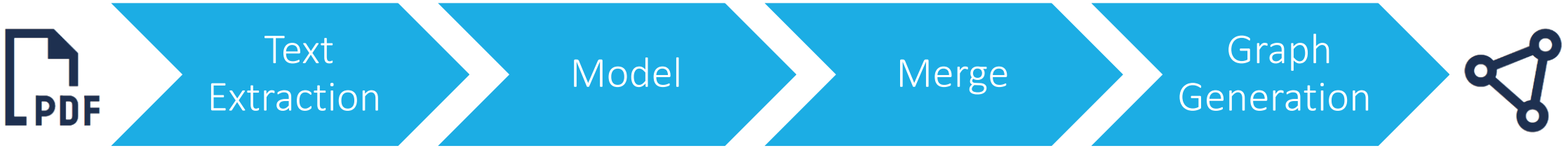
- INTRODUCTION – MOTIVATION
- PIPELINE
- M_REBEL
- R&D MERGE
- ALL_MINI
- FINETUNING + METRICS
- IMPLEMENTATION
- STORING KB
- USER/ADMIN INTERFACE
- PROJECT MANAGEMENT
- CODE CARBON
- CONCLUSION

Introduction – Motivation

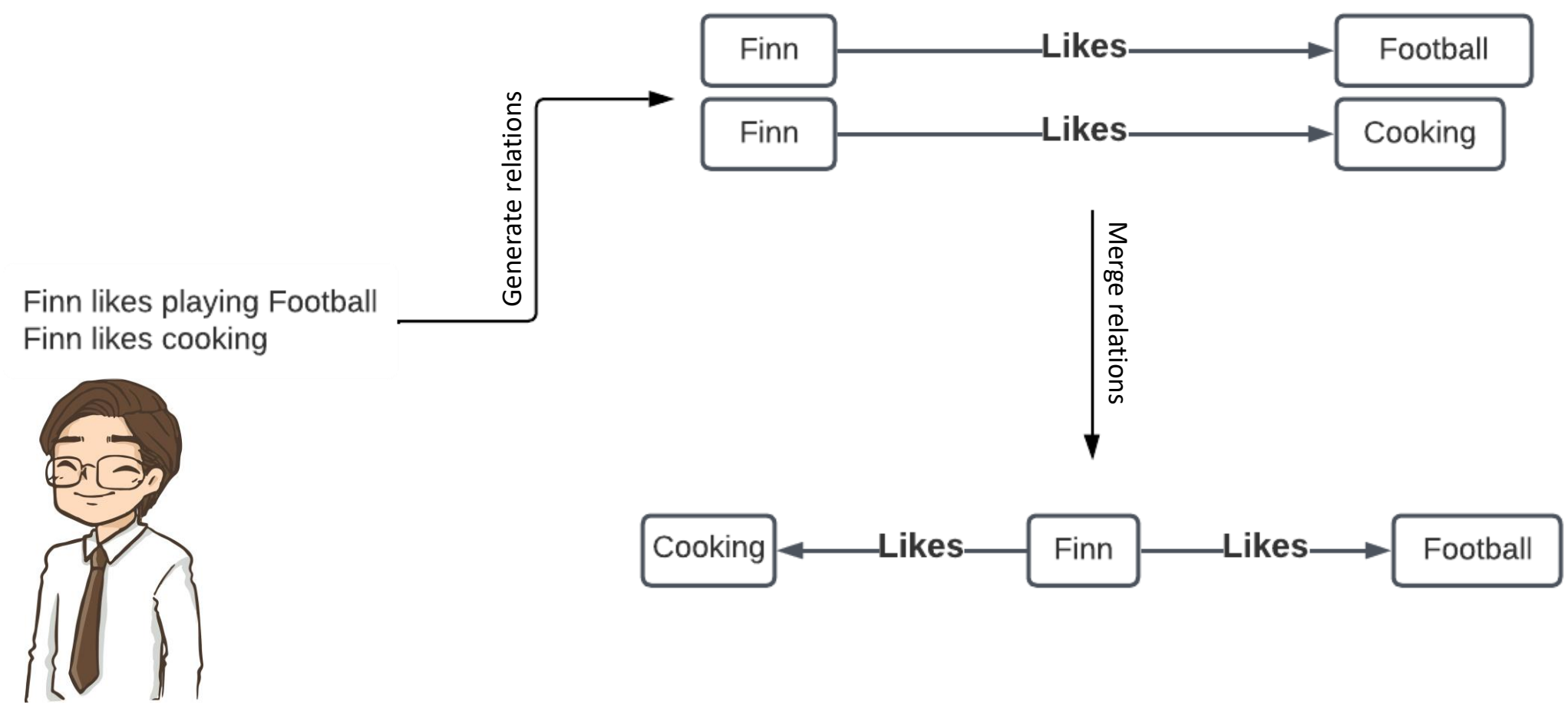
What is a Knowledge Graph?



Pipeline



Example

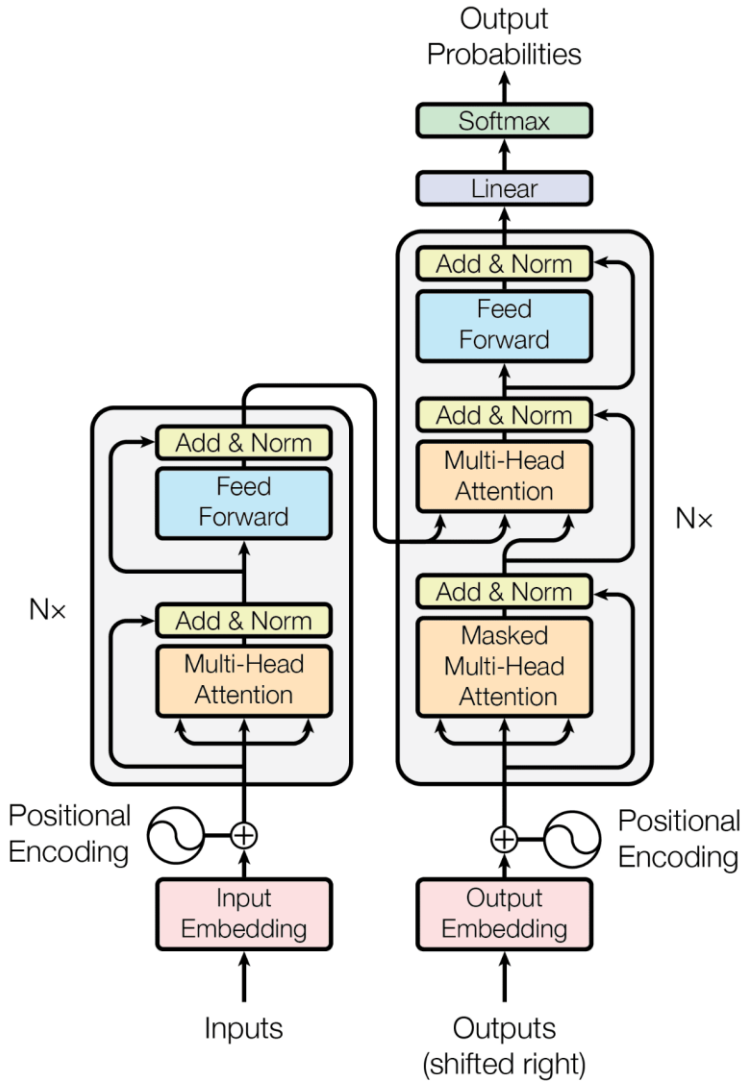

Finn likes playing Football
Finn likes cooking
Generate relations
Finn
Likes
Football
Finn
Likes
Cooking
Merge relations
Cooking
Likes
Finn
Likes
Football

MRebel

MREBEL'S Foundation:
Built on BART: A hybrid model combining bidirectional context and autoregressive generation.

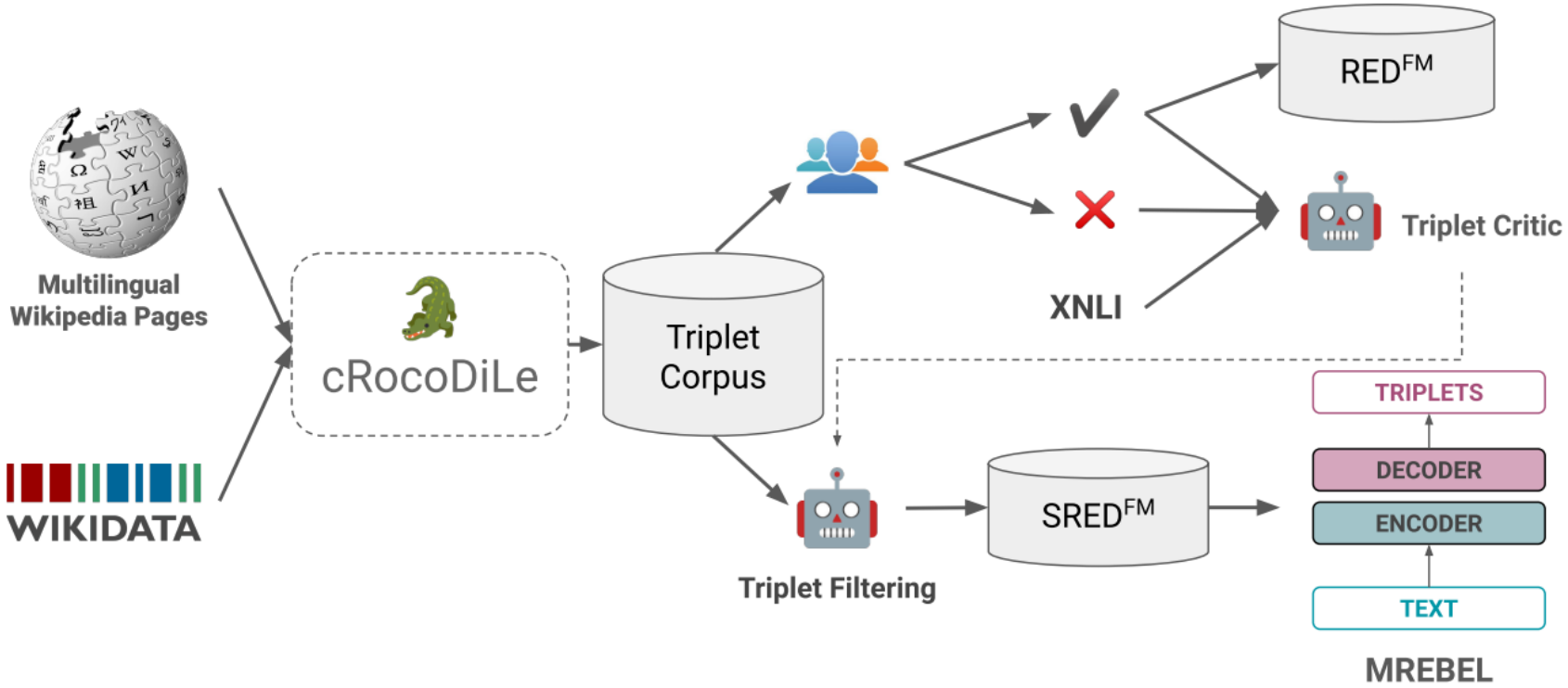
Relation Extraction as seq2seq:
Reframes Relation Extraction: Converts the task into a sequence-to-sequence language generation problem

The Transformer Architecture



The encoder-decoder structure of the Transformer architecture
Taken from "Attention Is All You Need"

MRebel



Complete pipeline for the creation of REDFM, SREDFM and mREBEL.
Retrieved from *https://doi.org/10.48550/arXiv.2306.09802*

Wikipedia-Based Entity Merging Approach

Approach Evaluation:

- Time consuming page fetching.
- Repetitive network requests to external servers.
- Scalability issues

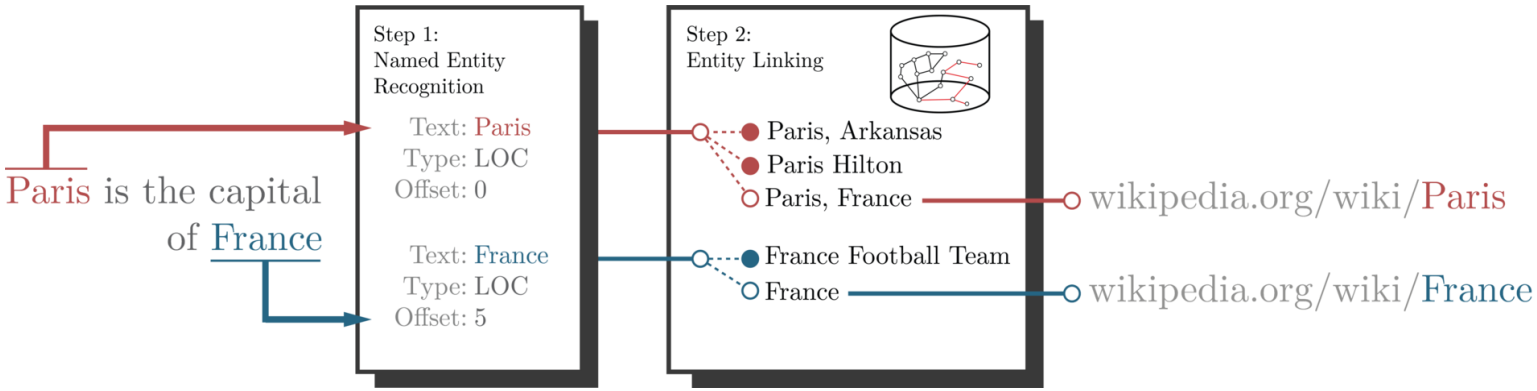
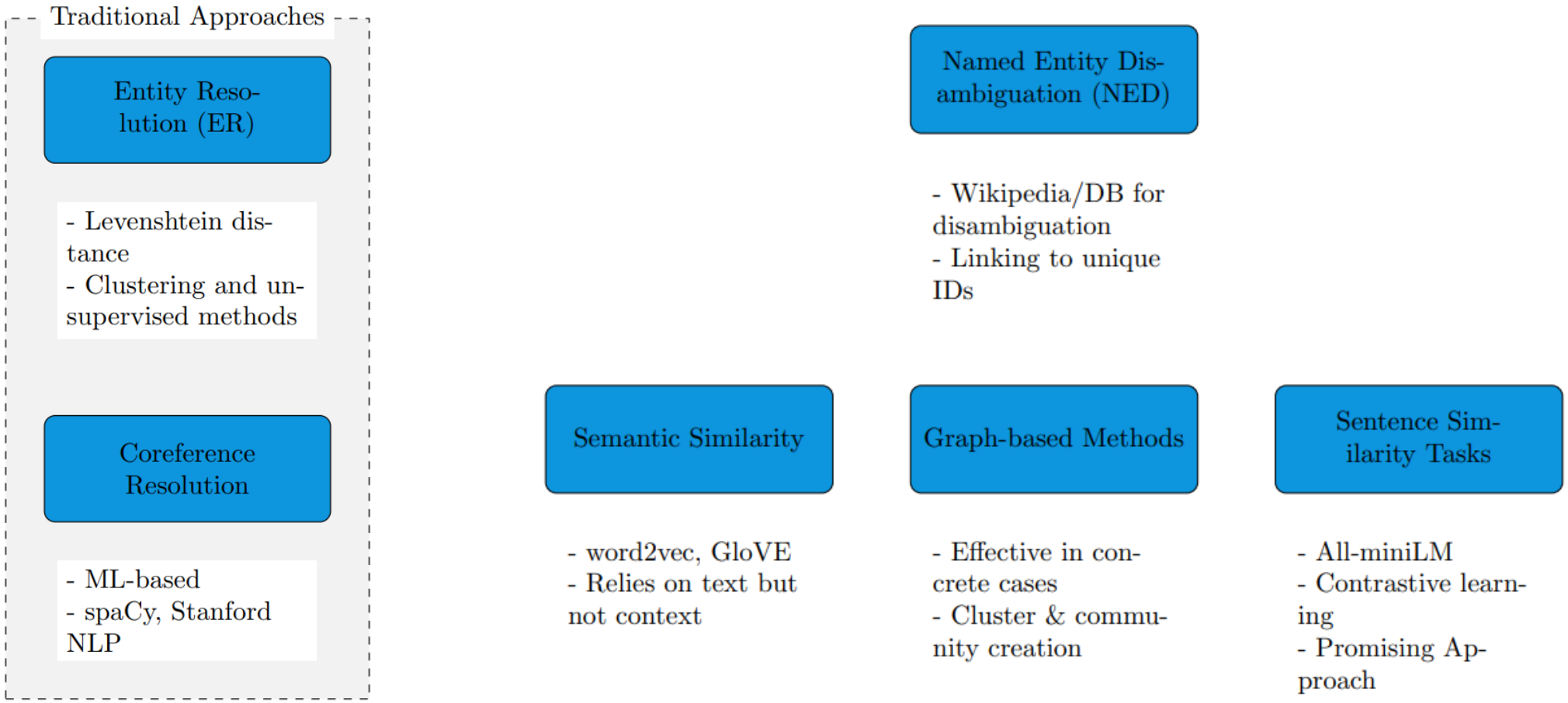


Illustration of entity alignment using WikiPedia pages

Exploring Solutions For Entity Merging Challenges

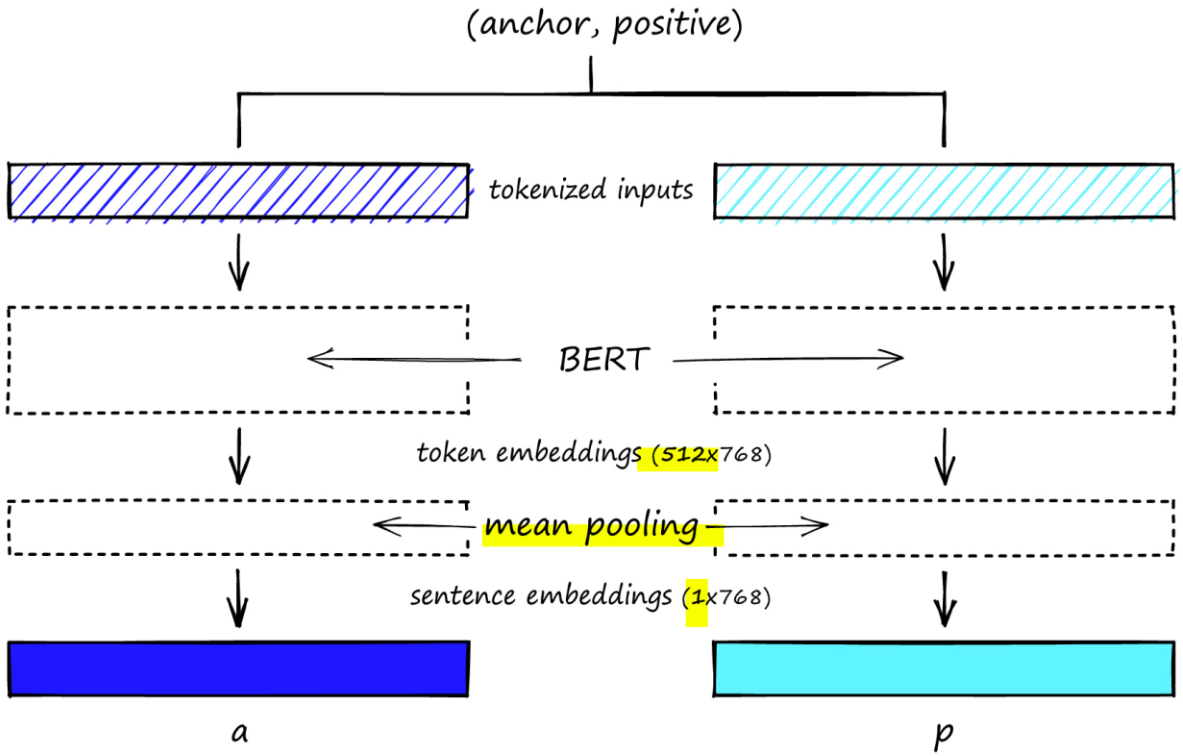


Introduction to the All-MiniLM Model

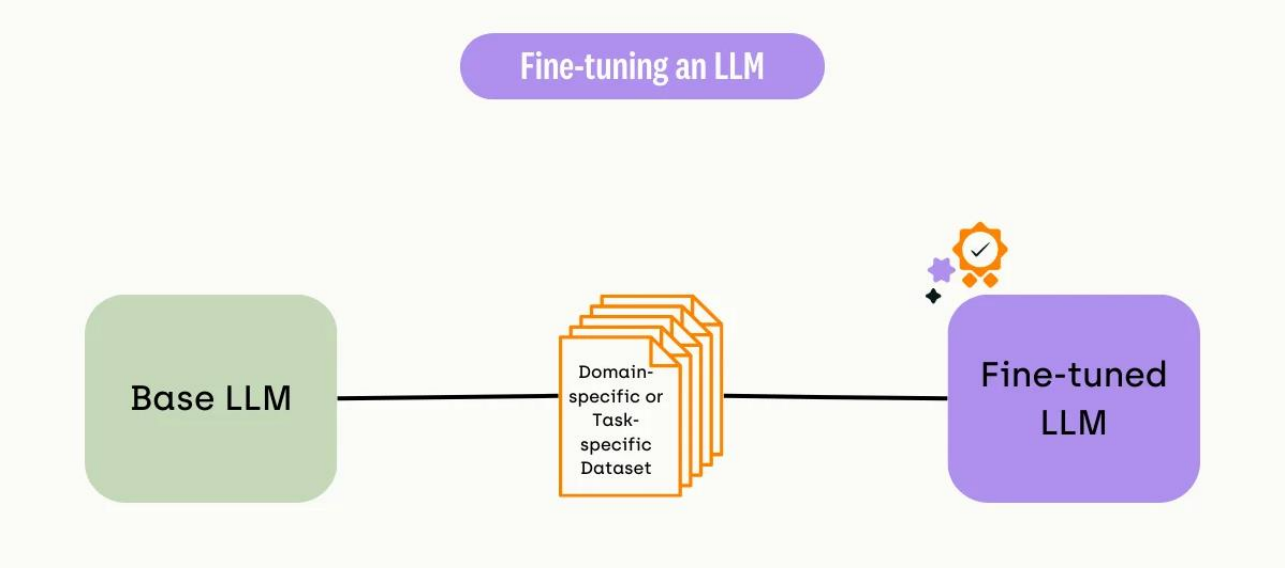
Sentence Transformers family
Sentences --> 384-dimensional vector space.

Semantically Encode meaningful sentence embeddings
Semantic similarity comparison...

Advantages:
Fast inference, small size
Great balance between performance and efficiency



Introduction to the All-MiniLM Model



Fine-Tuning Scheme

Fine-tuned on a 1 billion sentence pair dataset during an HF event

The fine-tuning process involved a **contrastive learning** objective.

Custom Fine-tuning of all-MiniLM

LLM used

Macroeconomy

Market Finance

Categories

Monetary Policy

Tax Policy

Banking and Credit

Capital Markets

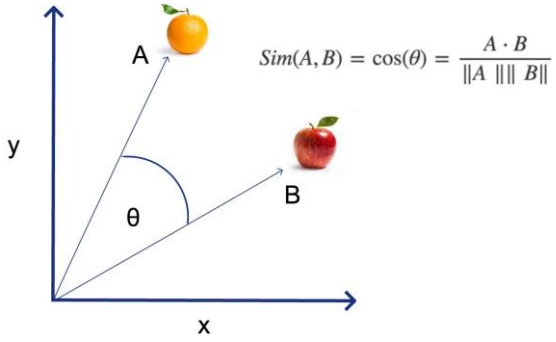
Sub-Categories

Bank loan interest rates have decreased this quarter.

Interest rates on bank loans have seen a reduction recently.

Pair of Sentences

Cosine Similarity



Cosine similarity on the Embeddings

Custom Fine-tuning of all-MiniLM

Sample:
The Federal Reserve cut interest rates
Interest rates were reduced by the central bank

Logits outputs

Base Model : **0,765**
FT Model : **0,964**

Assessing the results:
350 Test Sentences (10% of the train set's size)
Balanced Class Distribution
Merge Threshold: **0,8** *(if above then merge)*

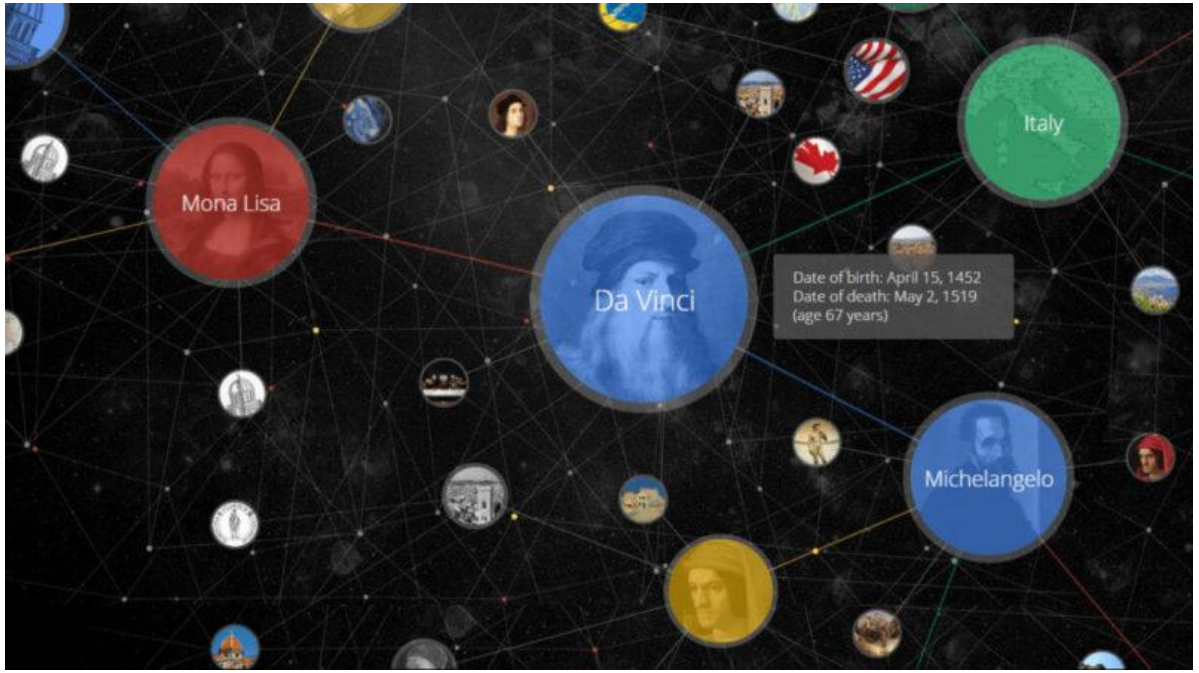
Model accuracy

Base Model : **76,5%**
FT Model : **96,4%**

Evaluating our Pipeline: Finding efficient Metrics

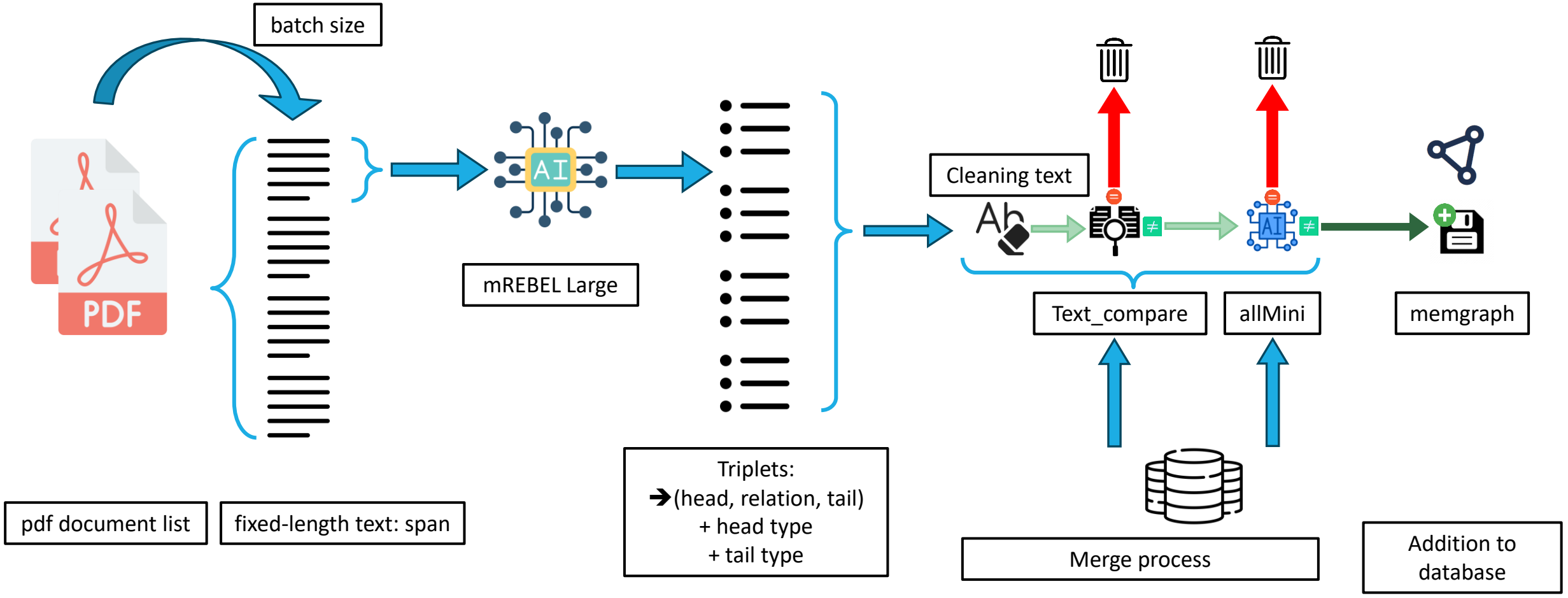
Unsupervised Metrics:

- **Mean Rank (MR)**: Reflects the average ranking of correct semantic triples in a list sorted by their likelihood. *A lower number indicates better performance.*
- **Hits@K**: Measures how often the correct semantic triples are ranked within the top K positions in the list. *Higher values mean more accurate predictions.*
- **Mean Reciprocal Rank (MRR)**: Focuses on the top results and is less dependant to misleading information.



Google Knowledge Graph

Detailed pipeline implementation

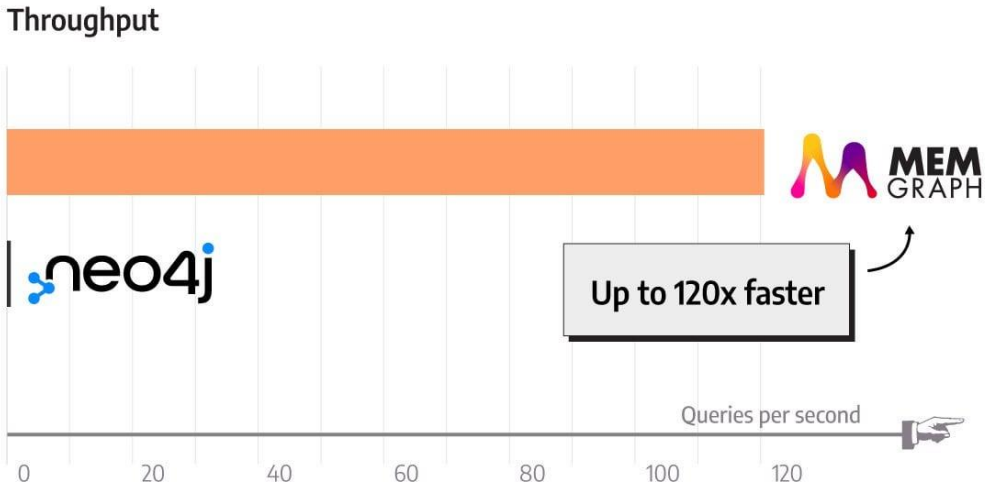


Storing the Knowledge Base

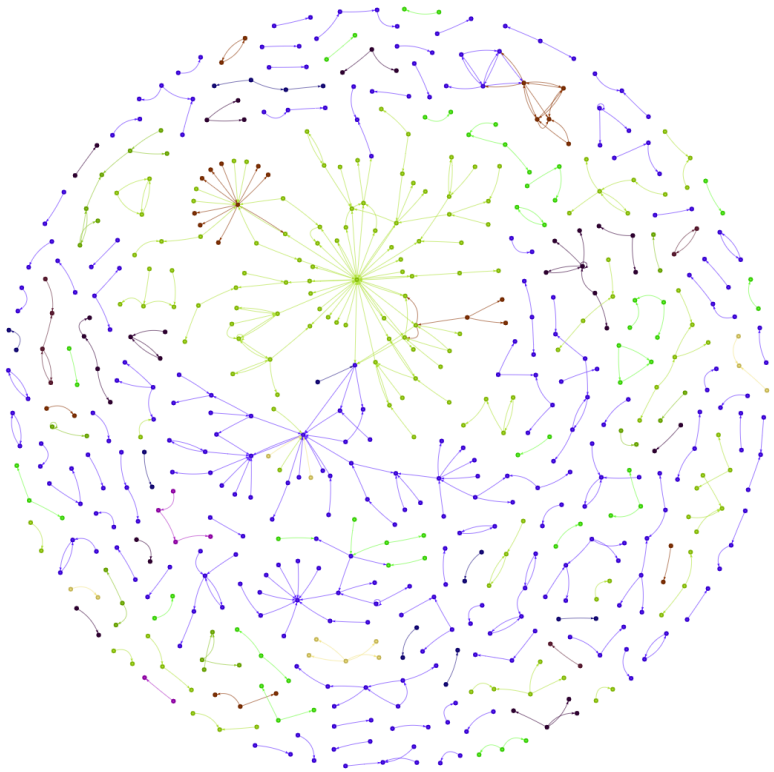
Multiple ways

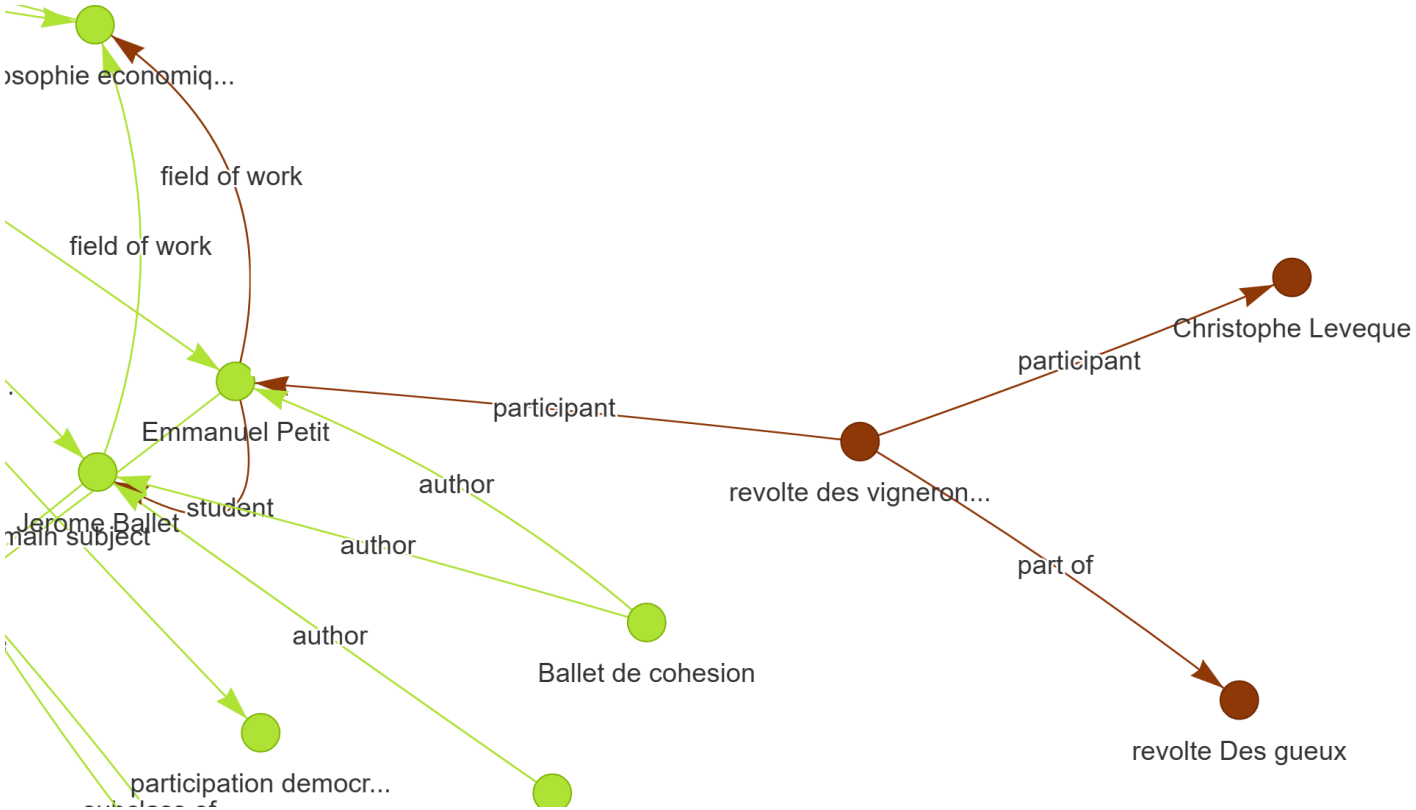
PROS :
- Fast write to file
- Accessibility for file edit

CONS :
- Slow read of data
- Security Issues
- Stability Issues
- Computationally expensive



memgraph-vs-neo4j-performance-benchmark-comparison

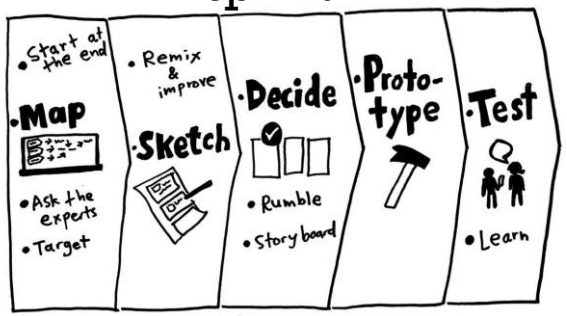



field of work
field of work
Emmanuel Petit
participant
author
student
author
author
Jerome Ballet
Ballet de cohesion
participation democr...
participant
Christophe Leveque
revolte des vigneron...
part of
revolte Des gueux

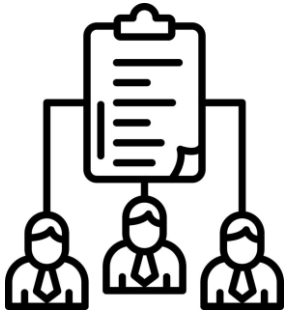
Project Management

Kanban

sprint

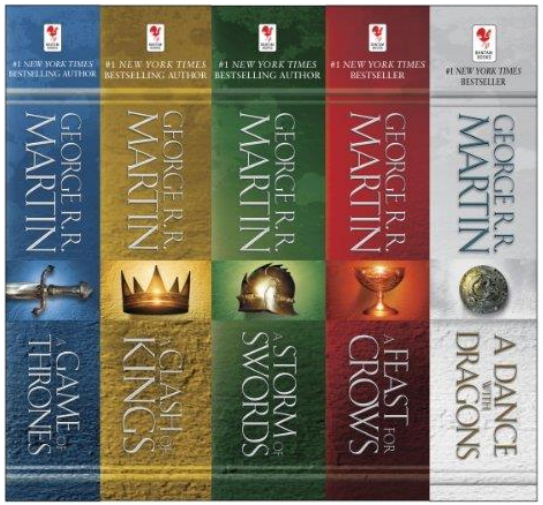


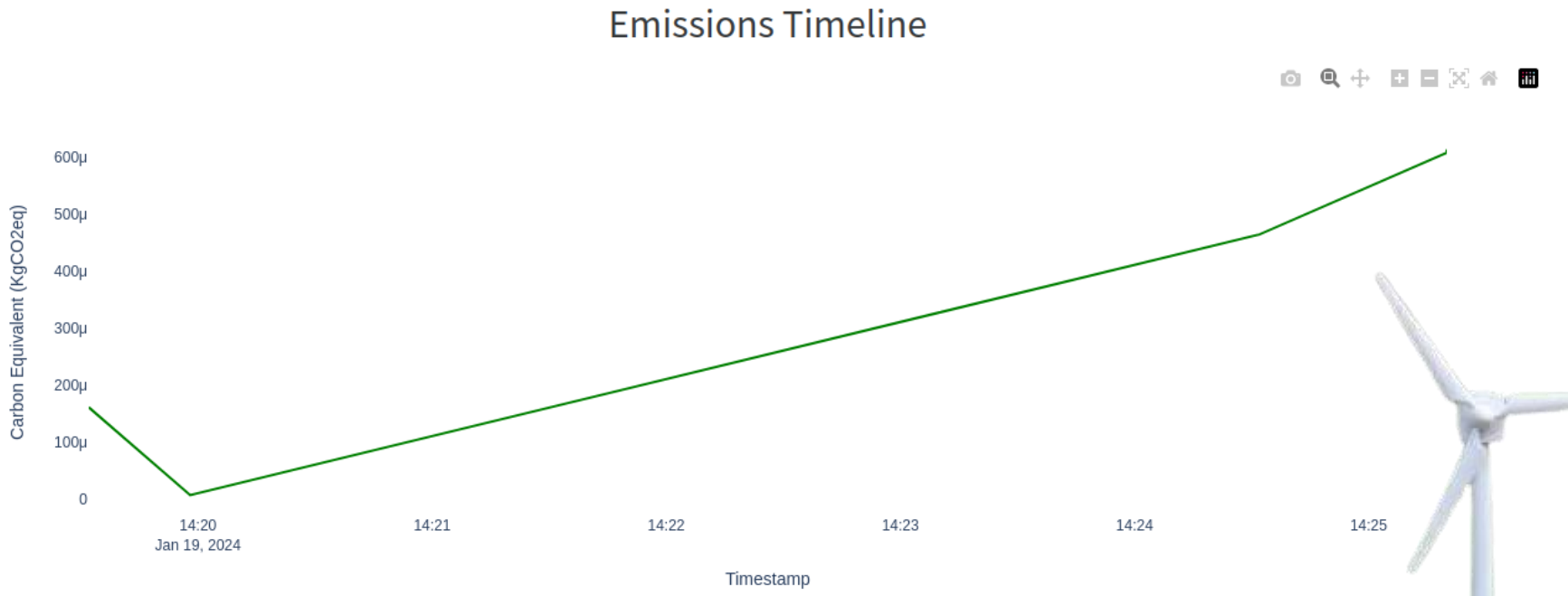
Weekly meeting - Friday



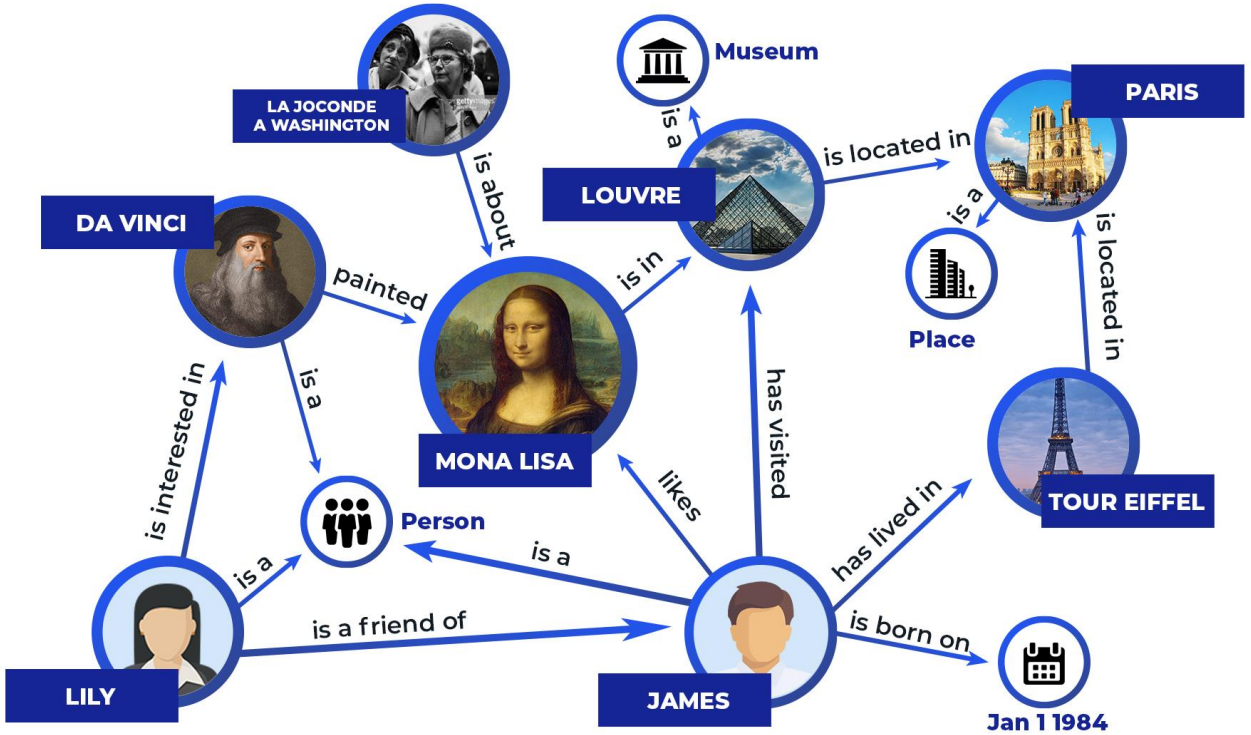
- **Optimize and match our work with the client's wishes.**
- **Quality over quantity concept and Divide and conquer model**

Environmental study





Conclusion


LA JOCONDE A WASHINGTON
is about
Museum
is a
LOUVRE
is located in
PARIS
is a
DA VINCI
painted
is in
Place
is located in
is interested in
is a
MONA LISA
likes
has visited
Person
is a
has lived in
TOUR EIFFEL
is a friend of
is born on
LILY
JAMES
Jan 1 1984

Team

Jad El Karchi
AI & LLM
student-engineer
@ Enseirb-Matmeca

Félicien Fichet
AI & OPTIMIZATION
student-engineer
@ Enseirb-Matmeca

Olha Nahorna
Research Engineer &
Project Leader
@ BSE

Mohamed Seddiq Elalaoui
AI & Data Science
Student-engineer
@ Enseirb-Matmeca

Joseph Beasse
AI & NLP
Student-engineer
@ Enseirb-Matmeca

Q & A

Thank you for your attention !