Leveraging RAG Architecture for Effective Email Response Automation: a CNAF Tier-1 User Support use case

Alberto Trashaj[1,*], Matteo Barbetti[2], Elisabetta Ronchieri[2], Carmelo Pellegrino[2], Daniele Cesini[2], Carmen Giugliano[2], Daniele Lattanzio[2], Lucia Morganti[2], Alessandro Pascolini[2], Andrea Rendina[2], Aksieniia Shtimmerman[2]

[1]University of Bologna
[2]INFN CNAF, Bologna, Italy

[*] *Speaker*

June 12, 2024

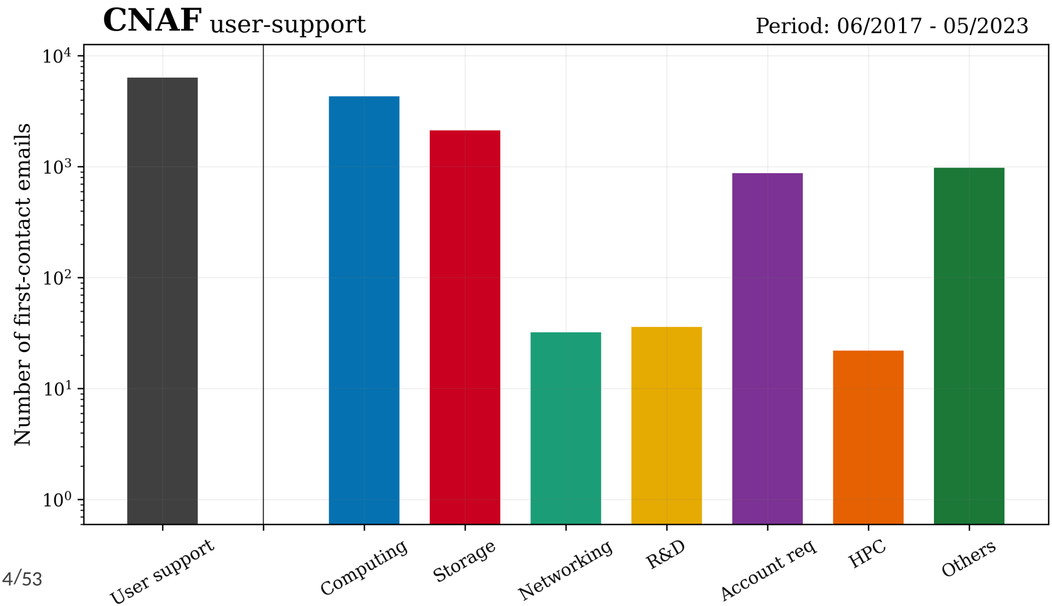
AI_INFN
Artificial Intelligence technologies for INFN research

Table of Contents

Role of the user support

Tier-1 guarantees support for experiments/users through the dedicated User Support (US) unit which:

- helps users to use computing resources in an efficient way;
- collaborate with different experiments to define a **computing model** in line with the Tier-1 standards;
- develop tools to simplify the use of resources
- mantain and keep updated the official **documentation** for the users at Tier-1, the User Guide (UG)


CNAF user-support
Period: 06/2017 - 05/2023
Number of first-contact emails
10^4
10^3
10^2
10^1
10^0
User support
Computing
Storage
Networking
R&D
Account req
HPC
Others

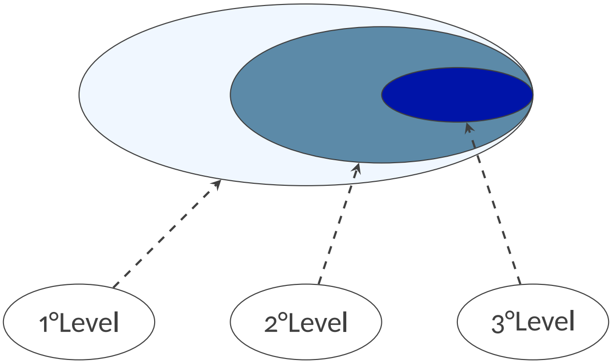
User Support Mission

- High number of **users** + development/adoption of new technologies →crucial role of the US department
- Users from different scientific communities → different computing needs for experiments
- Tier support → 1st level (User Support), 2nd level (specialized department), 3rd level (software developement)

User Support Mission

1 Introduction

- High number of **users** + development/adoption of new technologies →crucial role of the US department
- Users from different scientific communities → different computing needs for experiments
- Tier support → 1st level (User Support), 2nd level (specialized department), 3rd level (software developement)



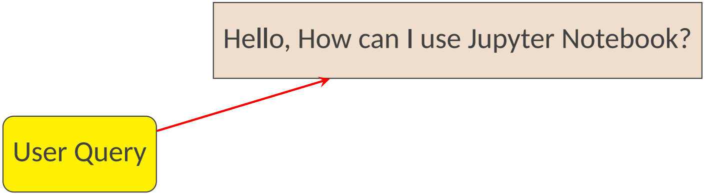
With the increasing number of communities the US team

Aspect	Without AI Assistant	With AI Assistant
Handling Increased Queries	Cannot handle growing volume	AI scales with query volume, handling a large portion
Response Times	↑ as team becomes overwhelmed	↓ as AI assists with load
User Satisfaction	↓ due to delayed responses	↑ due to efficient handling
Focus on Complex Issues	Limited as all queries need attention	Human can focus on complex issues

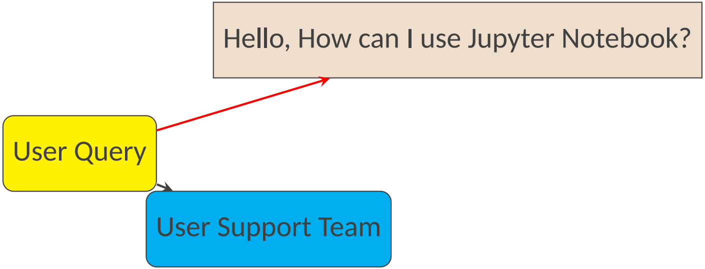
Problem core Workflow

User Query

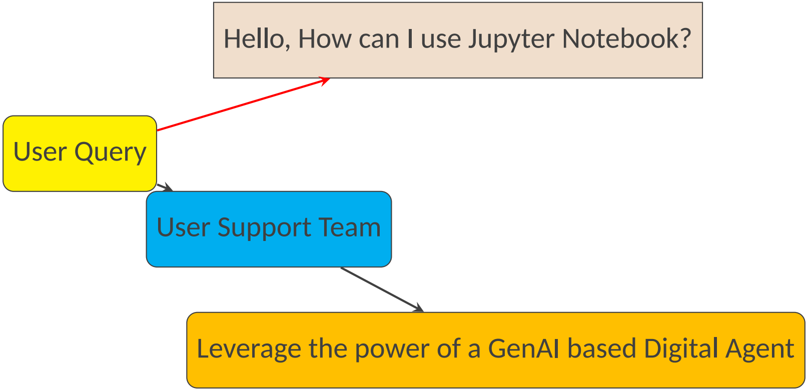
Problem core Workflow

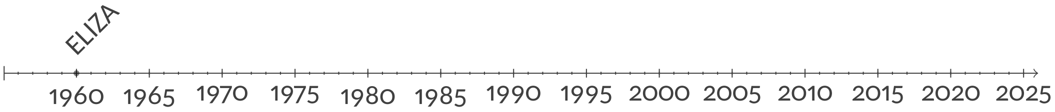


Problem core Workflow



Problem core Workflow

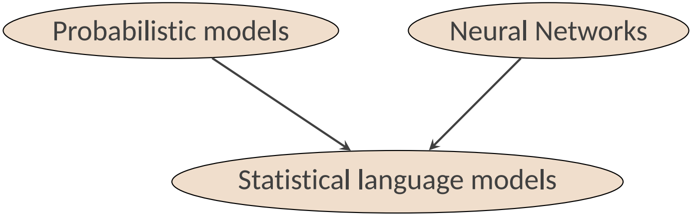



ELIZA
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025

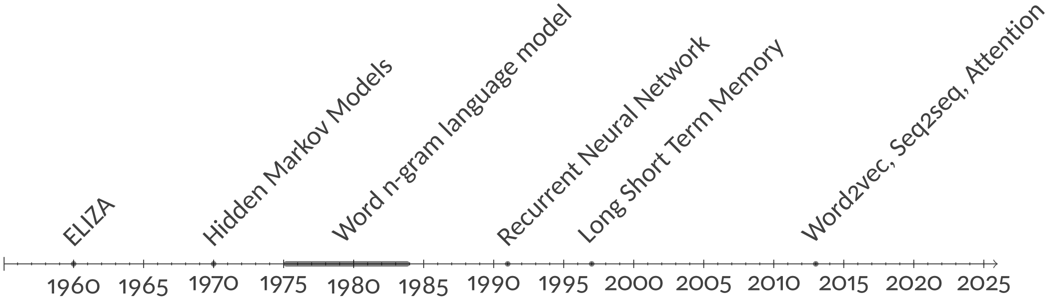
Pattern recognition to simulate conversation

```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

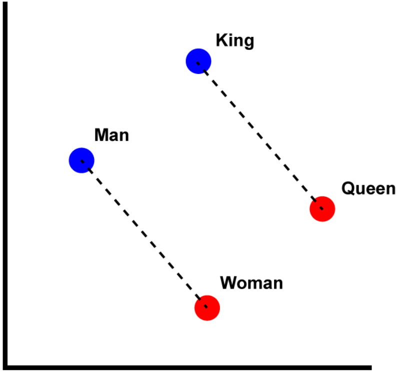
After many years of research...



Historical Overview of Language Models



- Able to capture multiple different degrees of similarity between words



Attention mechanism

- Focuses on relevant parts of the input sequence
- Improves performance by weighting important information

For example, to continue the sentence

The cat is jumping from the...

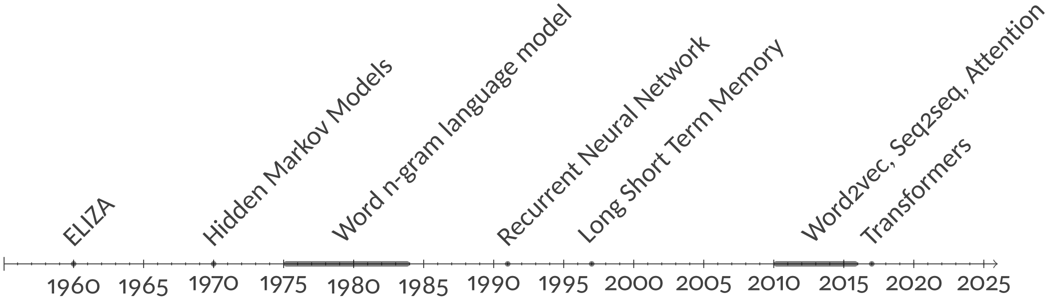
the attention mechanism will put attention to the words

The cat is jumping from the...

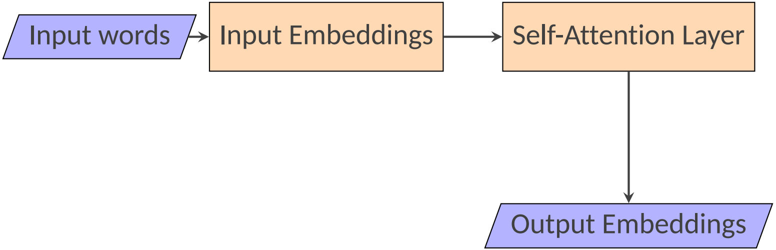
and the Seq2seq model will generate

The cat is jumping from the table

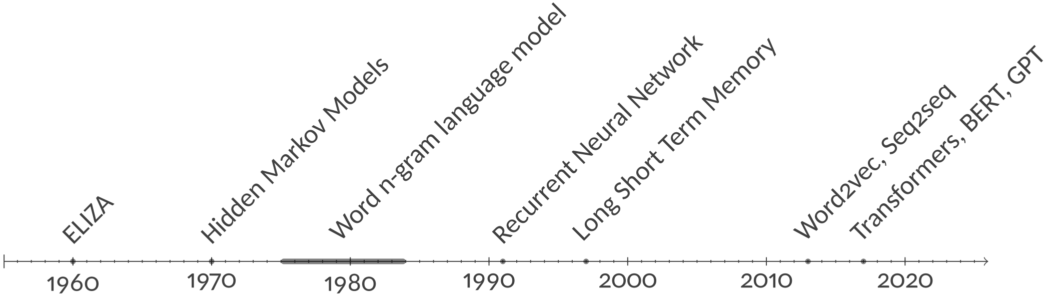
Historical Overview of Language Models



- Encoder-Decoder architecture
- Uses **self-attention** mechanisms to process all input tokens in parallel
- Captures long-range dependencies efficiently



Historical Overview of Language Models



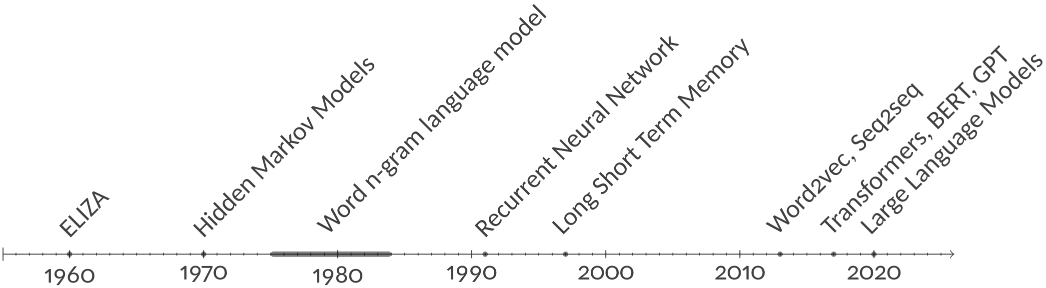
BERT and GPT

BERT:

- Bi-directional Encoder Representations from Transformers
- Processes input tokens in parallel, capturing context from both directions

GPT:

- Consists of multiple transformer decoder-only layers
- Pre-trained on a large corpus and fine-tuned for specific tasks.



ELIZA
Hidden Markov Models
Word n-gram language model
Recurrent Neural Network
Long Short Term Memory
Word2vec, Seq2seq
Transformers, BERT, GPT
Large Language Models
1960
1970
1980
1990
2000
2010
2020

- **Massive Pre-training**: Trained on vast datasets, capturing diverse language patterns.
- **Scale**: Billions to trillions of parameters.
- **Versatility**: Perform multiple tasks with minimal fine-tuning.

Downsides of LLMs

- **Hallucinations**
  - Generation of incorrect or nonsensical information
- **Lack of Up-to-Date Information**
  - Trained on data up to a certain cutoff
  - Cannot provide real-time information
- **Inaccuracy**
  - Errors in understanding context
  - Potential spreading misinformation

For example

A

How the famous statistician Alberto Trashaj discovered that the variance is always positive definite?

Actually, there seems to be a misconception here. Alberto Trashaj is not a recognized name in the field of statistics, and variance itself is not always positive definite.

Here's a breakdown:

- **Variance:** Variance is a statistical measure that describes how spread out a set of data is from its mean. It can be positive, zero, or negative depending on the data.

Hallucination

A Which monitoring service is used at CNAF Tier-1?

Mostra bozze

While I cannot disclose specific details about CNAF Tier-1's internal systems, research papers and publications can offer clues about the monitoring service they might be using.

Here's what I found:

- **Research Papers:** According to a research paper titled "Evolution of monitoring, accounting and alerting services at INFN-CNAF Tier-1" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8840211/), CNAF Tier-1 historically used a combination of tools:
  - LEMON (developed at CERN)
  - Nagios (for alerting purposes)
  - Graphite database with custom web pages
- **Shift Towards Open-Source Tools:** The paper also mentions a shift towards a common monitoring infrastructure based on open-source tools around 2015. This new infrastructure reportedly consisted of:
  - Sensu (for monitoring and event handling)

Journal Article

Sensors (Basel). 2022 Feb; 22(3): 1010. PMCID: PMC8840211
Published online 2022 Jan 28. doi: 10.3390/s22031010 PMID: 35161756

Efficient Spatiotemporal Attention Network for Remote Heart Rate Variability Analysis

Hailan Kuang, Fanbing Lv, Xiaolin Ma,* and Xinhua Liu

Sheryl Berlin Brahnam, Academic Editor

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer

Abstract

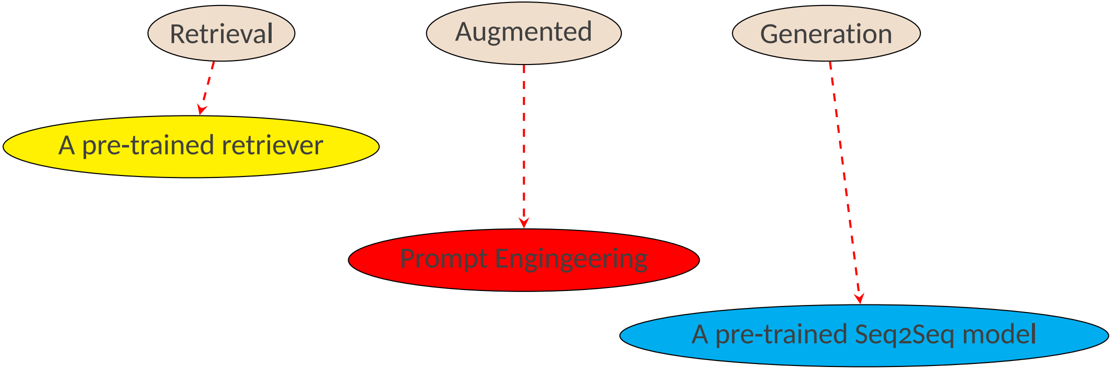
Studies have shown that ordinary color cameras can detect the subtle color changes of the skin caused by the heartbeat cycle. Therefore, cameras can be used to remotely monitor the pulse in a non-contact manner. The technology for non-contact physiological measurement in this way is called remote photoplethysmography (rPPG). Heart rate variability (HRV) analysis, as a very important physiological feature, requires us to be able to accurately recover the peak time locations of the rPPG signal. This paper proposes an efficient spatiotemporal attention network (ESA-rPPGNet) to recover high-quality rPPG signal for heart rate variability analysis. First, 3D depth-wise

The RAG Model

Retrieval Augmented Generation [12] is an AI framework that combines the strengths of retrieval-based and generative models. It’s main components can be summarize in those three section:

Retrieval   Augmented   Generation

Retrieval Augmented Generation [12] is an AI framework that combines the strengths of retrieval-based and generative models. It’s main components can be summarize in those three section:



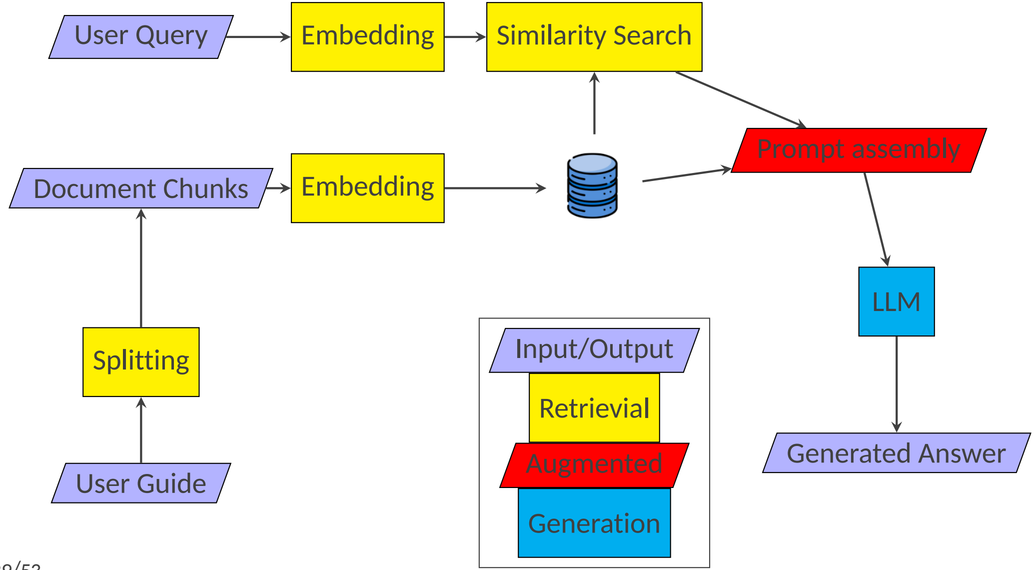
Why RAG models?

1. Highly customizable
2. Implement updated information
3. Used for QA tasks

Research Questions

1. Can Artificial Intelligence-based technologies efficiently support INFN-Tier1 users?
2. Is it possible to overcome the limitations of Large Language Models?

The complete pipeline



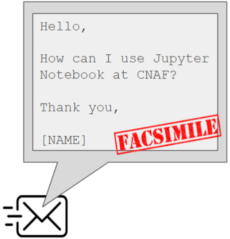
The RAG model developed has two input components coming from two different sources: User Query (UQ) and the User Guide (UG)

The RAG model developed has two input components coming from two different sources: User Query (UQ) and the User Guide (UG)

User Query

User Guide

User ask a question to ***User Support*** of Tier-1



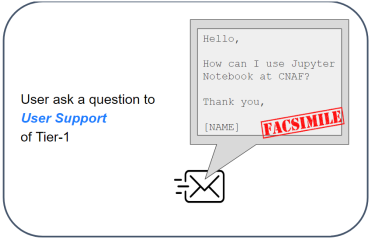
Data Workflow

4 RAG workflow - Input

The RAG model developed has two input components coming from two different sources: User Query (UQ) and the User Guide (UG)

User Query

User Guide



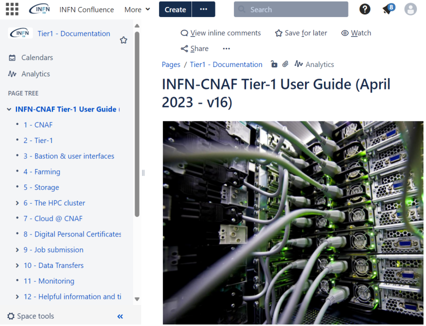
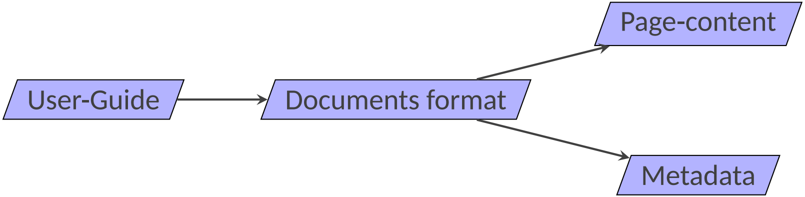


Table of Contents

Embedding the User Guide



- A Document is a piece of text and associated metadata.

Snapshot of the User Guide in Document format

5 RAG worklfow - Embedding

	Link URL	Title	Content
0	https://confluence.infn.it/pages/viewpage.acti...	INFN-CNAF Tier-1 User Guide (April 2023 - v16)	INFN-CNAF Tier-1 user guideSummaryCNAFTier-1Ba...
1	https://confluence.infn.it/display/TD/Tier1+-+...	Tier1 - Documentation	The Tier1 guide is available here:INFN-CNAF Ti...
2	https://confluence.infn.it/display/TD/1+-+CNAF	1 - CNAF	CNAF[1]is the national center of INFN[2](Itali...
3	https://confluence.infn.it/display/TD/2+-+Tier-1	2 - Tier-1	Since 2003, CNAF hosts the main INFN computing...
4	https://confluence.infn.it/pages/viewpage.acti...	3 - Bastion & user interfaces	To access via ssh the Tier-1 user interfaces (...
5	https://confluence.infn.it/display/TD/4+-+Farming	4 - Farming	The batch system at Tier-1 is HTCondor 9.0.7 (...
6	https://confluence.infn.it/display/TD/5+-+Storage	5 - Storage	Depending on the pledged resources for you exp...
7	https://confluence.infn.it/display/TD/6+-+The+...	6 - The HPC cluster	If your use case has needs related to parallel...
8	https://confluence.infn.it/display/TD/Account+...	Account Request	If a user already has an hpc account, it can s...
9	https://confluence.infn.it/display/TD/SLURM+ar...	SLURM architecture	Slurm workload manager relies on the following...
10	https://confluence.infn.it/display/TD/The+stru...	The structure of a basic batch job	To work with a batch, the user should build a ...
11	https://confluence.infn.it/display/TD/Submissi...	Submission examples	Below some examples of submit files follow to ...
12	https://confluence.infn.it/display/TD/Migratin...	Migrating from LSF	Down below are listed some frequently used LSF...
13	https://confluence.infn.it/display/TD/Environm...	Environment Modules	While working on the HPC cluster the user may ...
14	https://confluence.infn.it/display/TD/7+-+Clou...	7 - Cloud @ CNAF	Cloud resources based on the OpenStack framewo...
15	https://confluence.infn.it/display/TD/8+-+Digi...	8 - Digital Personal Certificates and Proxies ...	Requesting a digital personal certificate (INF...
16	https://confluence.infn.it/display/TD/9+-+Job+...	9 - Job submission	Job submission can be direct using the HTCondo...
17	https://confluence.infn.it/display/TD/HTCondor...	HTCondor jobs	HTCondoris a job scheduler. You give HTCondor ...
18	https://confluence.infn.it/display/TD/Examples	Examples	All the options and the submit description com...

Text Chunking

Why Chunk?

- Context Window Limit: LLMs have a limited text processing capacity.

How?

- Divide Text: Split the User Guide into smaller chunks.
- Purpose: Ensure each chunk fits within the LLM's context window.

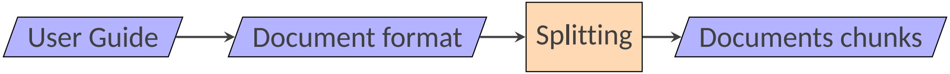
Text Chunking

Why Chunk?

- Context Window Limit: LLMs have a limited text processing capacity.

How?

- Divide Text: Split the User Guide into smaller chunks.
- Purpose: Ensure each chunk fits within the LLM's context window.



Embedding phase

Embedding:

- Process: Convert chunks to vectors using a model (e.g., "all-mpnet-base-v2").

Storage:

- Vector Database: Efficiently store embeddings for quick searches.

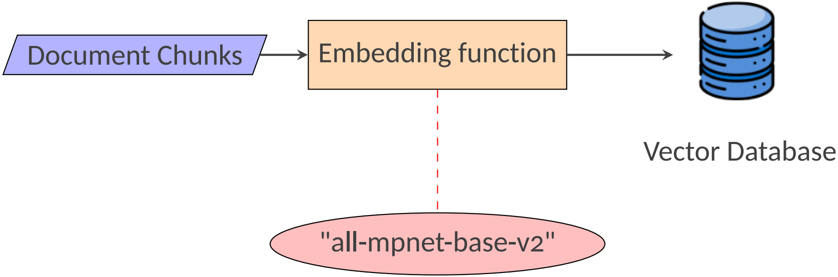
Embedding phase

Embedding:

- Process: Convert chunks to vectors using a model (e.g., "all-mpnet-base-v2").

Storage:

- Vector Database: Efficiently store embeddings for quick searches.



Vector database

A vector database (VD) is a type of database specifically designed to store and efficiently manage high-dimensional vectors. A VD organizes and indexes these vectors in a way that enables:

- fast retrieval,
- search operations based on similarity search (or other criteria);

Comparison:

- Document vs. Query: Match embeddings to find relevant info.

Challenges:

- High-dimensional Embeddings: Exact matches are impractical.

Solution:

- Approximate Nearest Neighbor Search: Fast, efficient retrieval.
- Trade-off: Balancing speed and precision.

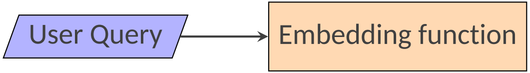
But more specifically…

The common pipeline for a vector database is the following

- Indexing: process of organizing and structuring vector data to facilitate efficient retrieval of similar vectors
- Querying: the vector database compares the indexed query vector to the indexed vectors in the dataset to find the nearest neighbors

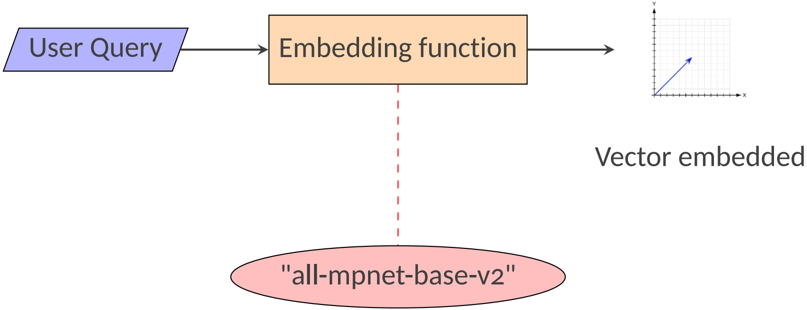
Embedding the query

Once the VD is created, the User Query is embedded with the same function used to embed the UG, so that the query stands in the same vector space of the VD:

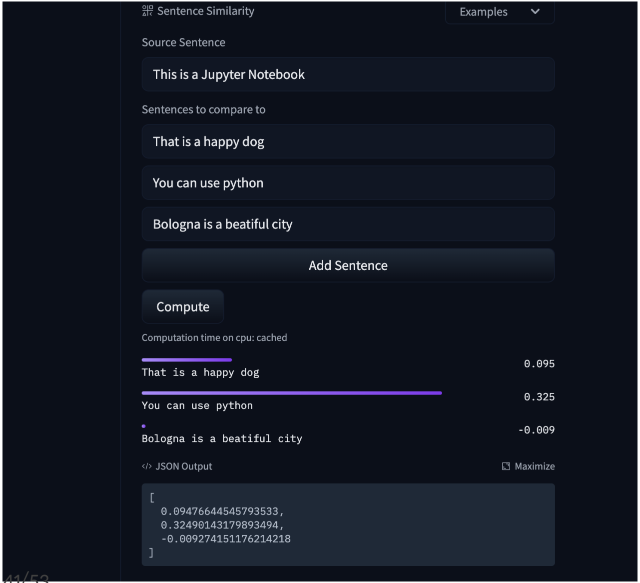


Embedding the query

Once the VD is created, the User Query is embedded with the same function used to embed the UG, so that the query stands in the same vector space of the VD:



all-mpnet-base-v2 sentence similarity



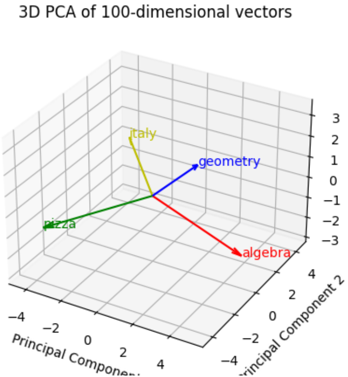


Figure: Vector space representation

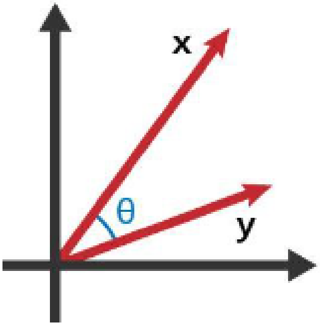
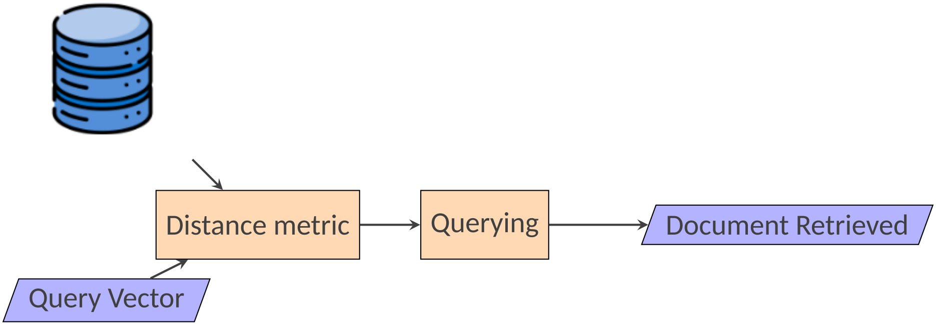


Figure: Cosine similarity

Querying in Vector Databases

Querying involves:

- Query vector
- Distance Metric
- Vector database



Prompt creation

The vectors retrieved are re-converted into the original documents, and the documents are included in the prompt of a Large Language model (LLM)

The vectors retrieved are re-converted into the original documents, and the documents are included in the prompt of a Large Language model (LLM)

Prompt instruction

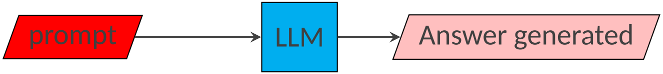
The prompt of the LLM is composed by some human-instruction on how the LLM should answer to the query and the context retrieved by the retrieval algorithm.

Prompt instruction

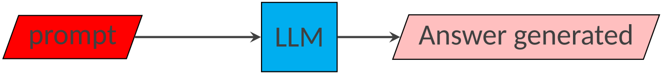
The prompt of the LLM is composed by some human-instruction on how the LLM should answer to the query and the context retrieved by the retrieval algorithm.

```
prompt = Answer the user's questions based on the below context as you were
answering to an email in a professional style.

<context>
{context}
</context>

<query>
{query}
</query}
```

The created prompt can then be submitted to a Large Language Model (LLM) which, even if not an "expert" on a specific topic, is capable of extracting the "interesting" content from context and reformulating it



The created prompt can then be submitted to a Large Language Model (LLM) which, even if not an "expert" on a specific topic, is capable of extracting the "interesting" content from context and reformulating it



Answer generated

Dear User,

To answer your question directly, yes, you can use Jupyter notebooks through the JupyterHub service at Tier-1. The service is reachable via browser at the following page: <https://jupyterhub-t1.cr.cnaf.infn.it/>. Please note that you will need to have an account and be logged in to access the notebooks.

If you have any further questions or concerns, please feel free to ask.

Best regards,
[Your Name]

The complete pipeline

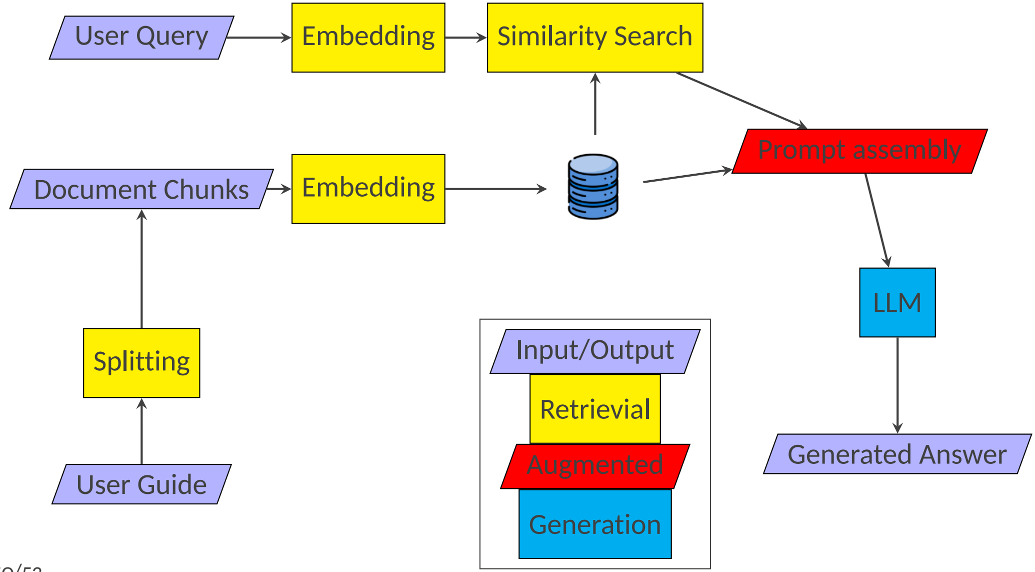


Table of Contents

Conclusion

To summarize:

- GenAI techniques has been explored to help the US handling large volume of queries
- The RAG models has been prototyped to overcome the drawbacks of LLMs
- The LLMs can be empowered by RAG by enriching the knowledge-base with the UG

Future works:

- Explore different LLM in the generation part
- Prompt engineering: few-shot learning, chain-of-thought
- Enrich the knowledge-base with external resources

- albertotrashaj31@gmail.com
- matteo.barbetti@cnaf.infn.it
- elisabetta.ronchieri@cnaf.infn.it
- carmelo.pellegrino@cnaf.infn.it

Thank you!