

# An Artificial Intelligence-based service to automatize the INFN CNAF User Support

<sup>1,\*</sup>Elisabetta Ronchieri, <sup>1</sup>Matteo Barbetti, <sup>1</sup>Alberto Trashaj, <sup>1</sup>Carmelo Pellegrino, <sup>1</sup>Daniele Cesini, <sup>1</sup>Federico Fornari, <sup>1</sup>Daniele Lattanzio, <sup>1</sup>Lucia Morganti, <sup>1</sup>Alessandro Pascolini, <sup>1</sup>Andrea Rendina

<sup>1</sup>INFN CNAF, Bologna, Italy

*\* Speaker*

March 27, 2024

International Symposium on  
Grids and Clouds 2024  
24 - 29 March 2024  
Academia Sinica, Taipei, Taiwan

▶ Background

▶ Methodology

▶ Task1

▶ Task2

▶ Future Plan

▶ Conclusions

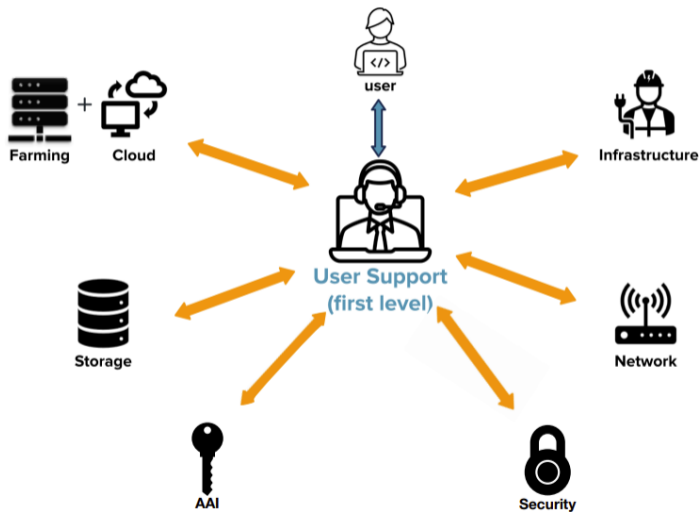
It focuses on technological development and knowledge transfer to heterogeneous experiments, national and European projects, and industry (whenever possible).



*Image provided by Pier Paolo Ricci, INFN CNAF.*

### Facts

- over 1700 active users
- over 60 experiments (not only LHC and not only from the Physics field)
- over 65,000 CPU cores
- 70 PB of disk
- 130 PB of tapes for long-term storage



- It solves most of the basic problems
- also writes **documentation** to improve the usage of **solutions** and **standard tools** the Data Centre provides
  - e.g. HTCondor, is the batch system for HTC, and SLURM for HPC
  - e.g. gfal2-util, is the tool for data transfer/management via Grid

- **On-boarding of new scientific communities** (projects, experiments, others)
- **Supporting the use of specific software**
  - personalising support on certain, specific, use cases
  - e.g. through user scripts, and environment
- NOTE** different scientific communities need different software
- **Managing the user registration procedure** (recognition, authorisation, account creation)
- **Writing documentation for users**
  - e.g. **INFN-T1 user guide** <https://l.infn.it/t1guide>
  - e.g. Automatically updated useful pages <https://www.cnaf.infn.it/~usersupport/>
- **Communicating with users**
  - e.g. **Direct user communication** (personal emails, chat)
  - e.g. Announces (mailing list, gocdb)
  - e.g. Periodic presentations (comitato di gestione (CdG), special events)
  - e.g. Dedicated meetings with experiments' people (on-boarding, special requests)

**RQ1** : Can Artificial Intelligence-based technologies successfully be integrated into the daily user-support activity?

**RQ2** : Can Artificial Intelligence-based technologies efficiently support INFN-Tier1 users?

▶ Background

▶ **Methodology**

▶ Task1

▶ Task2

▶ Future Plan

▶ Conclusions



### Task 1

- Automate the e-mail classification
- Guarantee the inclusion of specific Tier1 units to deal with specific problems

### Task 2

- Manage a high-latency reply to users' e-mails based on the collected e-mails and the knowledge base
- Define an e-mail-based framework

▶ Background

▶ Methodology

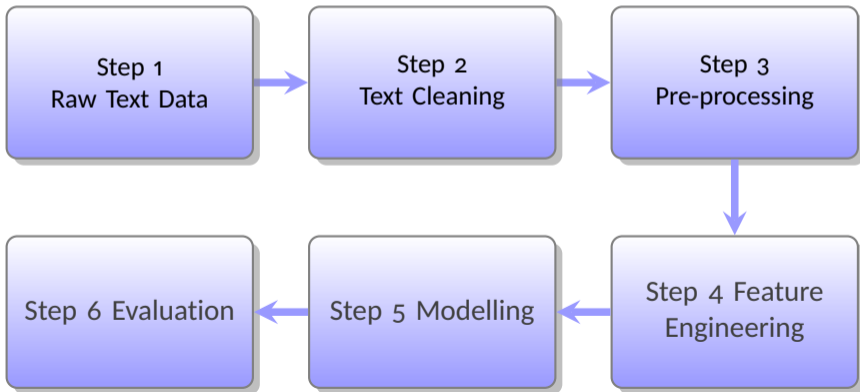
▶ **Task1**

▶ Task2

▶ Future Plan

▶ Conclusions

It is based on **Natural Language Processing (NLP)** steps.



- The considered e-mails cover the period 06/2017 - 05/2023.
- The original e-mails have been transformed into JSON files.
- Data e-mails are 28,500 JSON files for a total of 260 MB.

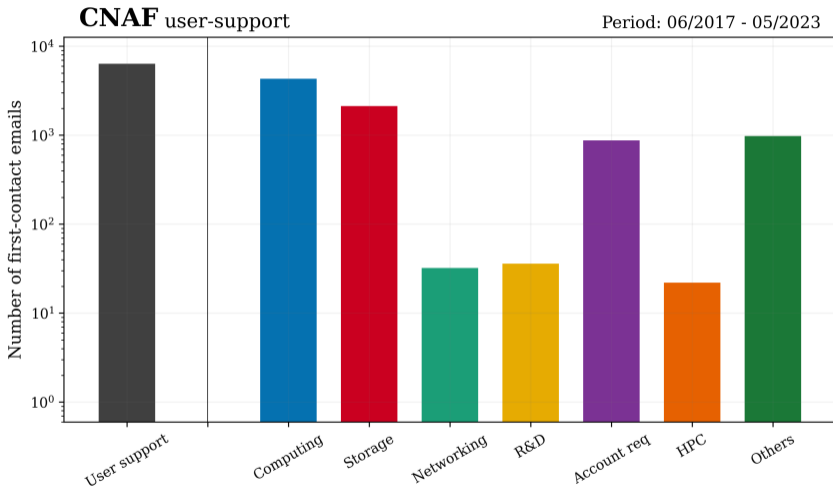
- E-mail example is:

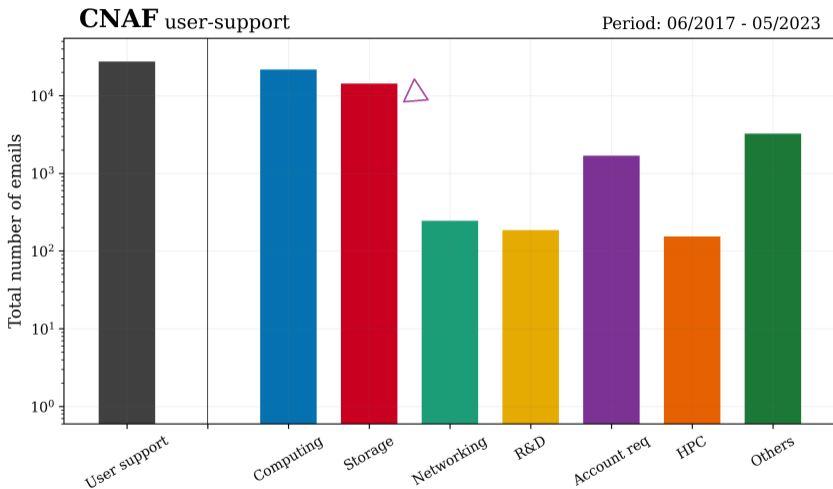
```

"from": "7c4351dc-63a9-4d7a-bde5-c94b3183e36d",
"to": "c77ab8f7-ab8d-4a1e-8d98-206957010e49",
"subject": "Fwd: [exp-sup-cnaf-gr2] [user-support] INFN-T1 Down on ...",
"content": "Just a reminder for the intervention of today....",
"type": "text/plain",
"date": "Thu, 1 Oct 2020 08:03:41 +0000",
"cc": [ "" ],
"id": "<E3F638AC-FC4A-4473-922F-93D5C29BB87D@cern.ch>",
"parent": null

```

- If the **parent** is NOT *null*, the e-mail message is a reply to another JSON file.
- E-mails have been opportunely anonymized whenever possible with a unique identifier.





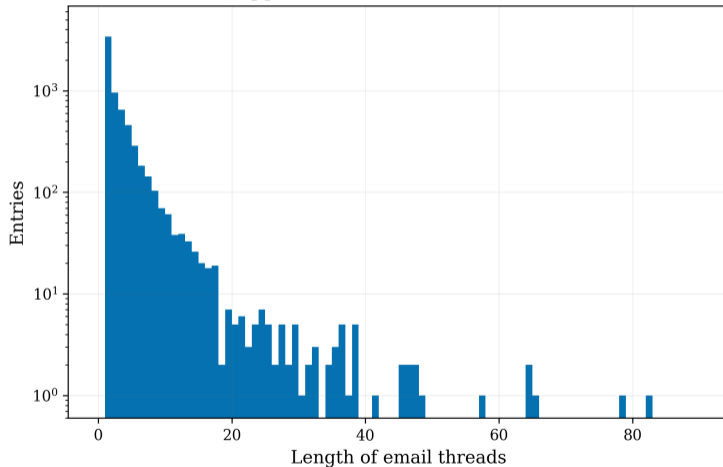
**NOTE:** E-mail category can change over time in the same e-mail thread.

It involves the following aspects:

- restructuring **email body**
  - E-mail body can be written in plain text or contain HTML tags.
- considering handling **e-mail text** with both a sequence of sentences and pieces of code
  - **Text language** can be both in English and Italian or just in one of the two languages.
  - **Code** is usually written in English or uses English terms.
- removing any reference to **users and user-support team**
  - Making e-mail anonymous is essential.
- considering **different lengths for the e-mail thread**

**CNAF** user-support

Period: 06/2017 - 05/2023

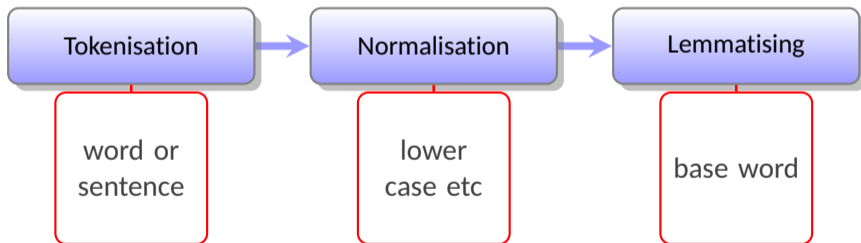


**NOTE:** Distribution of e-mail threads (length of 96 is the maximum).



Pre-processing transforms unstructured text into a structured one to prepare it for analysis by NLP standard procedure:

- removing noise and other inconsistencies in the text
- making the text easier to understand
- performing



**NOTE:** After pre-processing text, we have variable-length sequences of tokens.

- What Artificial Intelligence algorithm requires are fixed-length numeric vectors.
- Therefore, the pre-processed text is translated into numeric vectors through **vectorization**.

### Bag of Words (BoW)

- describes e-mails by word occurrences
- throws out word order

### *n*-Grams

- describes e-mails by *n*-contiguous words occurrences
- keeps word order and some context info

### TF-IDF

- highlights uncommon words in e-mail, and more common words across e-mails

### Conti. BoW

- predicts a word, leveraging all words in its neighborhood context

### Skip-GRAM

- predicts the context words based on a word

**NOTE:** After the vectorization phase, we can identify features relevant to our study.

It considers **sentence-transformers models** available on **HuggingFace**.

[https://www.sbert.net/docs/pretrained\\_models.html](https://www.sbert.net/docs/pretrained_models.html)

- They result from BERT-based *fine-tuned* models.
- The models have been designed for semantic textual similarity, semantic search, or paraphrase mining.
- The multilingual models support over 50 languages (Italian included).

<b>Model Name</b>	all-mpnet- base-v2	all-MiniLM-L6- v2	paraphrase- multilingual-mpnet- base-v2	paraphrase- multilingual-MiniLM- L12-v2
<b>Model Base</b>	MPNet	MiniLM	XLM-RoBERTa	MiniLM
<b>Quality</b>	high	high	good	good
<b>Speed</b>	1	×5	~ 1	×2.5
<b>Dimension</b>	768	384	768	384
<b>Max num tokens</b>	384	256	128	128

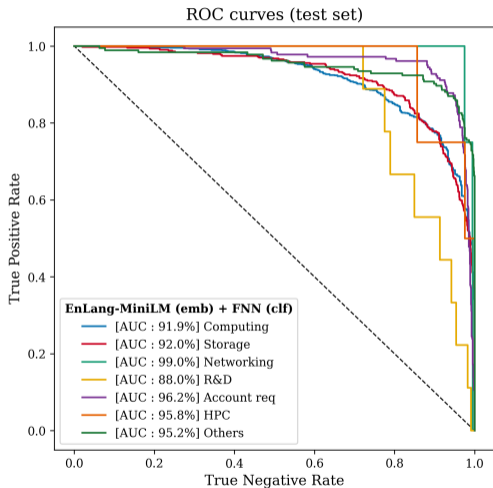
The embeddings resulting from Step 4 are employed as *input features* for the following multivariate classifiers:

k-Nearest (kNN)	Neighbors	Random Forest (RF)	Extreme Gradient Boosting (XGBoost)	Feed-forward Neural Network (FNN)
--------------------	-----------	--------------------	--	--------------------------------------

- The performances achieved by the previous models are further improved by relying on an **ensemble (EN)** model.
- The latter results by combining the output of all the trained models:
  - **balanced**: plain average of the output probabilities;
  - **randomized**: random combination of the output probabilities;
  - **weighted**: the FNN and XGB models dominates the classification decision except for the low represented classes (*i.e.*, HPC, R&D and Networking) for which RF is preferred.

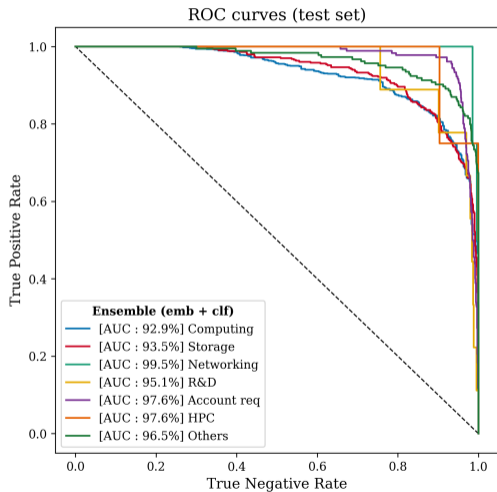
- **Receiver Operating Characteristics (ROC)** measures the performance of a classification model at all classification thresholds.
  - It plots *True Positive Rate* (sensitivity) vs. *False Positive Rate* (specificity) at different classification thresholds.
- **Area Under the Curve (AUC)** represents the area under the ROC curve, assigning a score to the overall performance of the binary classification model.
  - AUC provides an aggregate measure of performance across all possible classification thresholds.

	$kNN_{AUC}$	$RF_{AUC}$	$XGB_{AUC}$	$FNN_{AUC}$
<b>BoW</b>	0.843	0.916	<b>0.938</b>	0.909
n-Grams	0.860	0.887	0.886	0.880
TF-IDF	0.892	0.897	0.919	0.904
CBOw	0.807	0.837	0.856	0.761
SGRAM	0.816	0.843	0.855	0.787
MPNet (en-only)	0.914	0.895	0.924	0.914
<b>MiniLM (en-only)</b>	0.917	0.895	0.926	<b>0.927</b>
XLM-RoBERTa (multi)	0.908	0.890	0.922	0.916
MiniLM (multi)	0.902	0.888	0.915	0.915



	$EN_{balanced}$	$EN_{random}$	$EN_{weighted}$
AUC	0.939	0.939	<b>0.940</b>

The  $EN_{weighted}$  model performances are depicted in the plot.



▶ Background

▶ Methodology

▶ Task1

▶ **Task2**

▶ Future Plan

▶ Conclusions



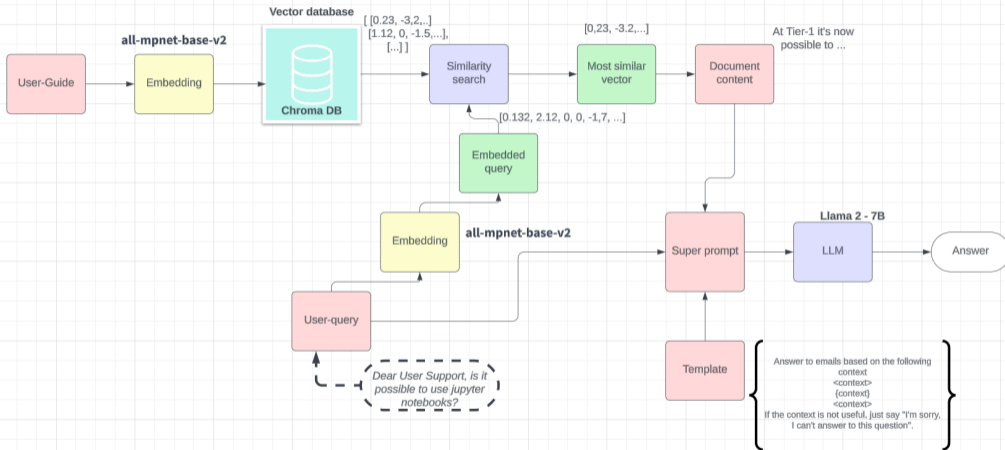
- RAG model combines **information retrieval** and **natural language generation**.
- It first **retrieves relevant information** based on **user queries** and then **employs a language model to generate contextually appropriate responses**.
- It is essentially composed by a parametric-memory generation model with a non-parametric memory retriever.

## Retrieval component

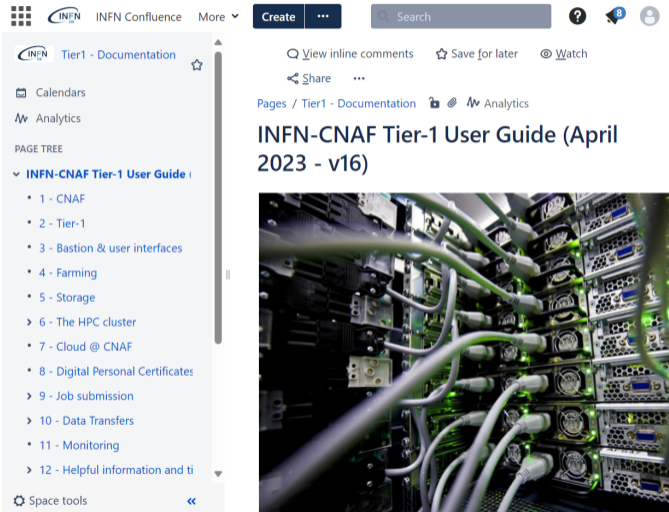
- It is responsible for retrieving relevant information from a predefined knowledge base.
- It employs techniques such as semantic search to find the most relevant pieces of information related to the input query.

## Generation component

- It is responsible for generating text based on the retrieved information and the input query.
- It employs techniques from AI generative models such as transformer architecture.



It has been properly scraped to be stored in a file.

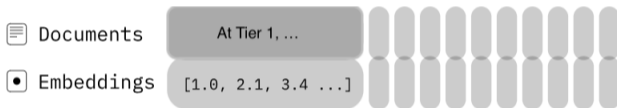


The screenshot displays the Confluence interface for the 'INFN-CNAF Tier-1 User Guide (April 2023 - v16)'. The left sidebar contains a 'PAGE TREE' with the following items:

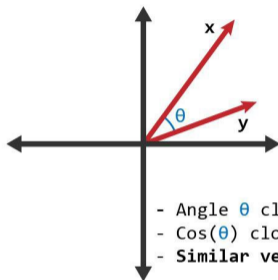
- INFN Tier1 - Documentation
- Calendars
- Analytics
- PAGE TREE
  - INFN-CNAF Tier-1 User Guide
    - 1 - CNAF
    - 2 - Tier-1
    - 3 - Bastion & user interfaces
    - 4 - Farming
    - 5 - Storage
    - > 6 - The HPC cluster
    - 7 - Cloud @ CNAF
    - 8 - Digital Personal Certificates
    - > 9 - Job submission
    - > 10 - Data Transfers
    - 11 - Monitoring
    - > 12 - Helpful information and ti
- Space tools

The main content area includes a search bar, navigation options like 'View inline comments', 'Save for later', and 'Watch', and a breadcrumb trail: 'Pages / Tier1 - Documentation / Analytics'. Below the breadcrumb is the title 'INFN-CNAF Tier-1 User Guide (April 2023 - v16)' and a photograph of a server rack with numerous cables and glowing green lights.

- The user guide is transformed in a set of vectors.
- Used the **all-mpbed-base-v2** model from huggingface.  
<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>



- User question: *Dear User Support, is it possible to use Jupyter Notebook?*
- This request is embedded by **mpbed-base** model.
- Once the request has transformed, a **query** to the vector database is performed to **retrieve the most similar vector** to the question.
- From the similar vector it is **retrieved the user guide page** containing the reply to the user question.



- Angle  $\theta$  close to  $0$
- $\text{Cos}(\theta)$  close to 1
- **Similar vectors**

- 5 - Storage
- 6 - The HPC cluster
- 7 - Cloud @ CNAF
- 8 - Digital Personal Certificates
- ▾ 9 - Job submission
  - HTCondor jobs
  - Examples
  - Singularity in batch jobs
  - **Jupyter notebook in interact**

## Jupyter notebook in interactive batch jobs

At Tier-1 it's now possible to use Jupyter notebooks served by JupyterHub. The service is reachable via browser at the following page: <https://jupyterhub-t1.cr.cnaf.infn.it/>

Once you get there, you will be asked to login by using your account bastion credentials. The account must belong to an experiment which has pledged CPU resources on the batch system.

Moreover, right after the login it is also possible to customize the jupyter environment following the instructions at the [User environment customization](#) paragraph.

When you login, the Hub service submits a local HTCondor job which is named `jupyter- $\langle$ username $\rangle$` . You can check its status from your user interface as a local job submitted on the sn-02, with the following command:

- LLM **Llama 2 - 7B** takes in input the prompt with the e-mail structure
- including the user guide page related to the Jupyter Notebook
- **producing** the following reply

*'Dear User, you for reaching out to us regarding your query about using Jupyter notebooks. Based on the context provided, it appears that at Tier-1, Jupyter notebooks are now available through JupyterHub. To access these notebooks, you will need to log in with your account bastion credentials and select the experiment that has pledged CPU resources on the batch system. Once logged in, you can customize the Jupyter environment as per the instructions provided. 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. you have any further questions or concerns, please feel free to ask. regards, [Your Name]'*

▶ Background

▶ Methodology

▶ Task1

▶ Task2

▶ **Future Plan**

▶ Conclusions



### Task 1 - Classification

- Fine-tune a modern Large Language Model, e.g. Gemma (state-of-the-art open-models by Google DeepMind)

### Task 2 - Question-reply framework

- Extend the knowledge base including also the 28k emails collected by the User-support unit in the latest 5 years

▶ Background

▶ Methodology

▶ Task1

▶ Task2

▶ Future Plan

▶ **Conclusions**

**RQ1** : Can Artificial Intelligence-based technologies successfully be integrated into the daily user-support activity?

- Yes, our results show that information stored in the emails and knowledge base can be retrieved by AI techniques to reply users' questions and support the user-support team in their daily activities.
- We got 0.94 AUC with Ensemble<sub>weighted</sub>.
- The defined question-reply framework is promising for the user-support activity.
- Test activity by user-support team is on the way.

**RQ2** : Can Artificial Intelligence-based technologies efficiently support INFN-Tier1 users?

- Yes, our preliminary results show that the defined question-reply framework can retrieve a correct reply to user's question with a score value.