Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

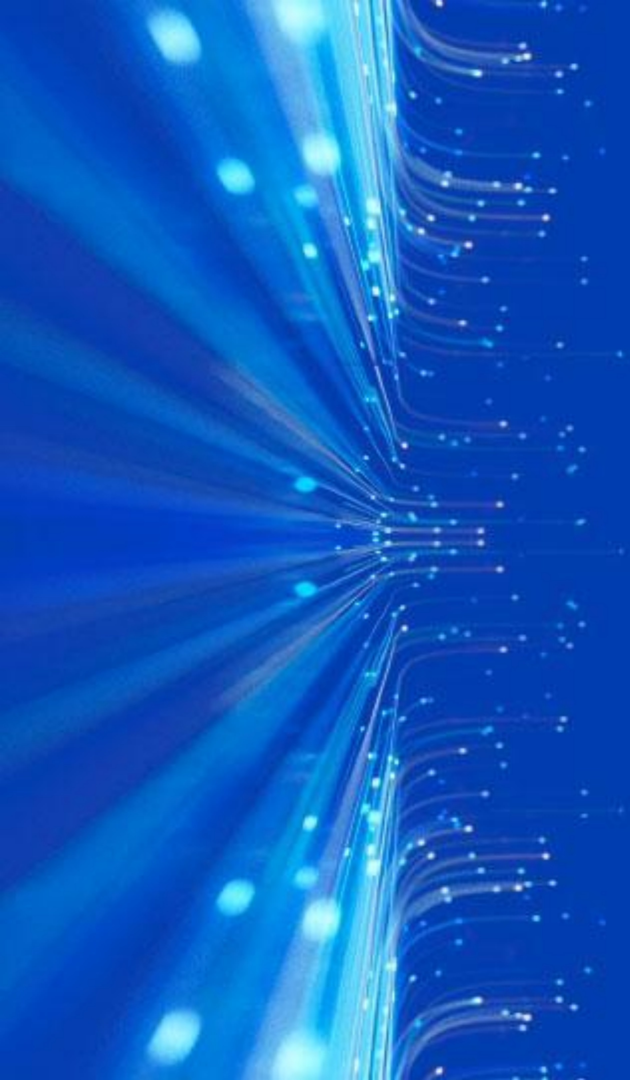
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing

Towards AI-driven User Support for CNAF Tier-1

M. Barbetti, A. Trashaj, E. Ronchieri, C. Pellegrino, and the contribution of many others

CNAF seminar | 21 November 2024

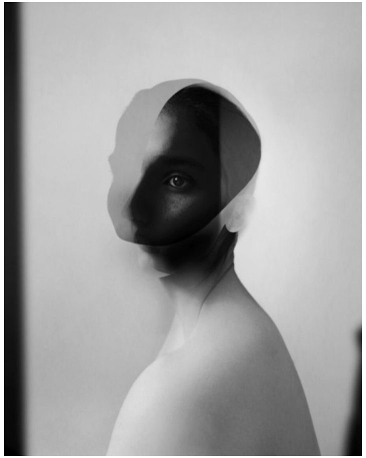


1.0

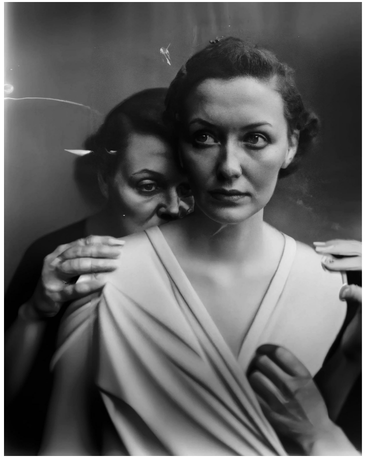
THE EVOLUTION OF ARTIFICIAL INTELLIGENCE

A sensationalistic overview of the history of Artificial Intelligence

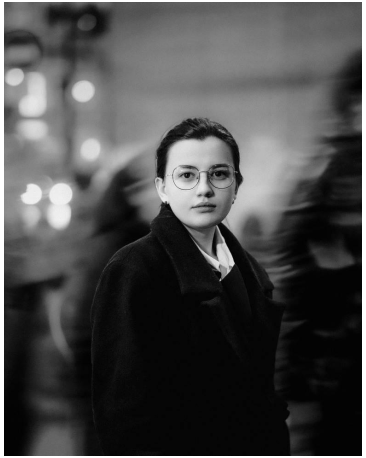
Can you find the (AI)ntruder?



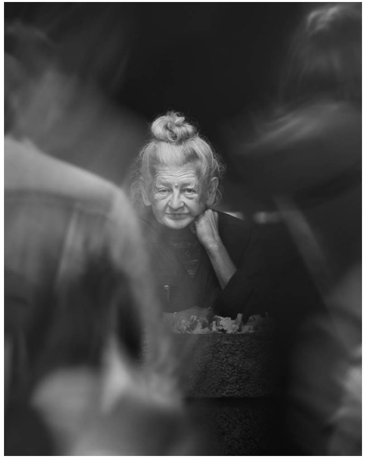
Credits: Jorg Karg



Credits: Boris Eldagsen



Credits: Aleksei Andreev

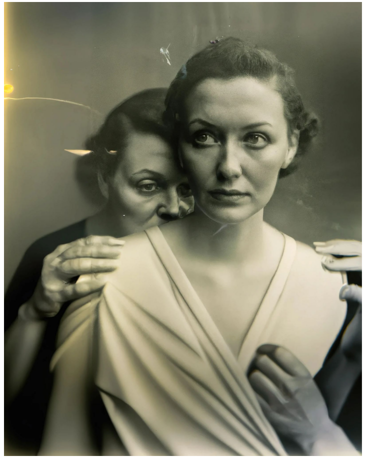


Credits: Evgeny Lutsko

Can you find the (AI)ntruder?

Credits: Jorg Karg

PSEUDOMNESIA: The Electrician, 2023



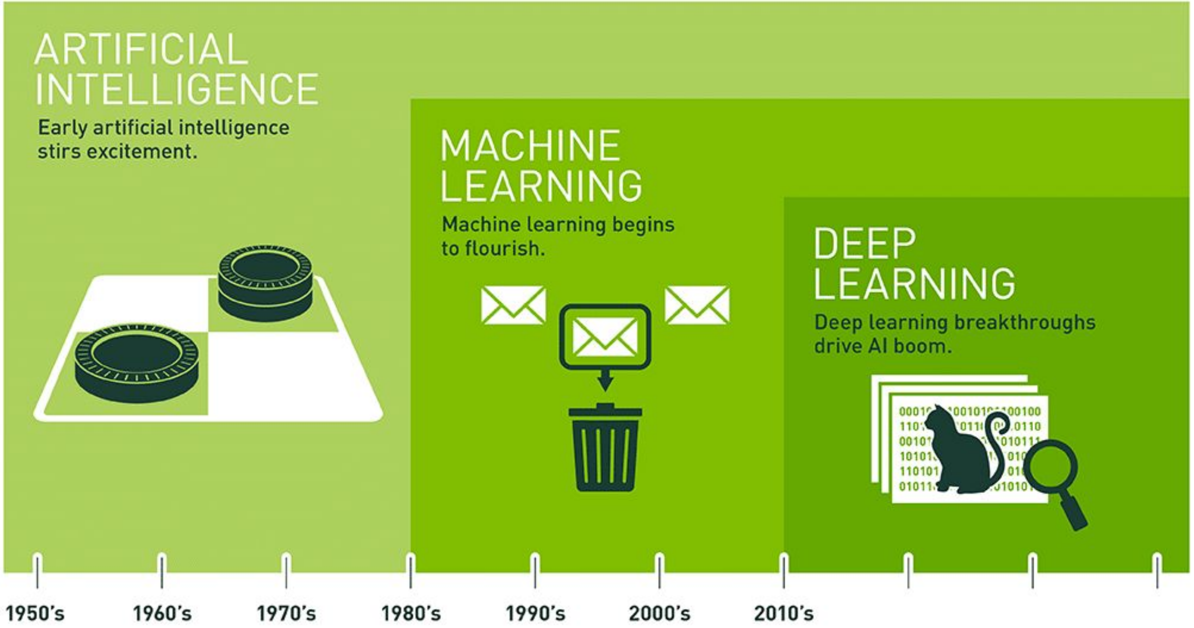
Credits: Boris Eldagsen
(by using DALL-E 2)

WINNER
of the creative open category at the
Sony World Photography Awards

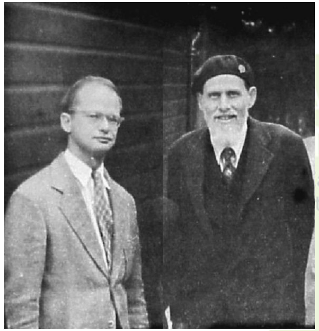
Credits: Aleksei Andreev

Credits: Evgeny Lutsko

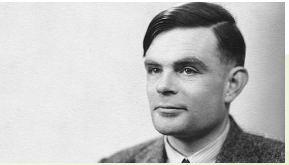
History of Artificial Intelligence


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

History of Artificial Intelligence

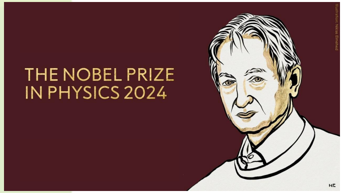


First model of an
Artificial Neural Network
by W. McCulloch and W. Pitts

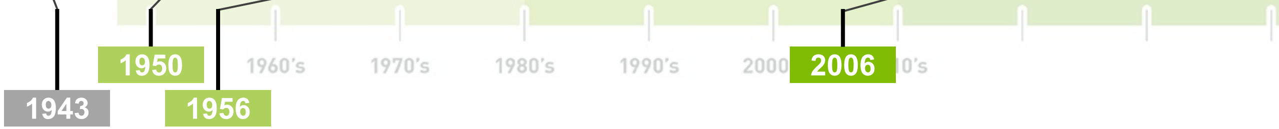


Alan Turing introduced
a test to answer the question
"Can machines think?"

The term ***Artificial Intelligence***
was coined for a thematic
workshop hosted at Dartmouth



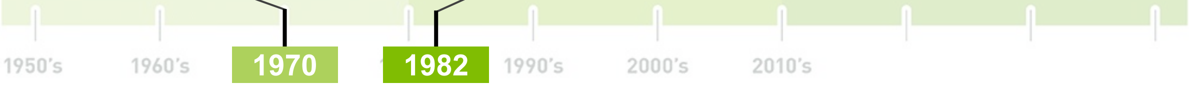
Geoffrey Hinton coined the term
Deep Learning to refer to
deep ANN-powered algorithms



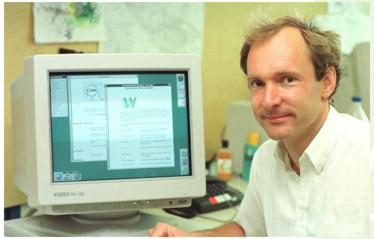
History of Artificial Intelligence

The modern technique for training MLPs, known as ***backpropagation***, was first published by S. Linnainmaa

P. Werbos and D.E. Rumelhart independently developed and applied backpropagation to MLPs in the way that has become **standard**



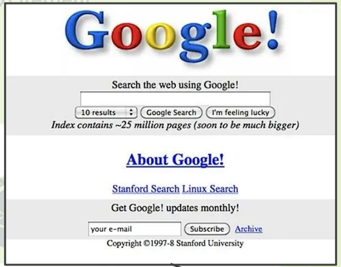
History of Artificial Intelligence



The birth of the
World Wide Web (WWW)
by Tim Berners-Lee

1989

The launch of **Google**
by Larry Page and Sergey Brin



Google!

Search the web using Google!

10 results | Google Search | I'm feeling lucky

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

Get Google! updates monthly!

your e-mail | Subscribe | Archive

Copyright ©1997-8 Stanford University

1998

The launch of
Amazon Web Service (AWS)

2002

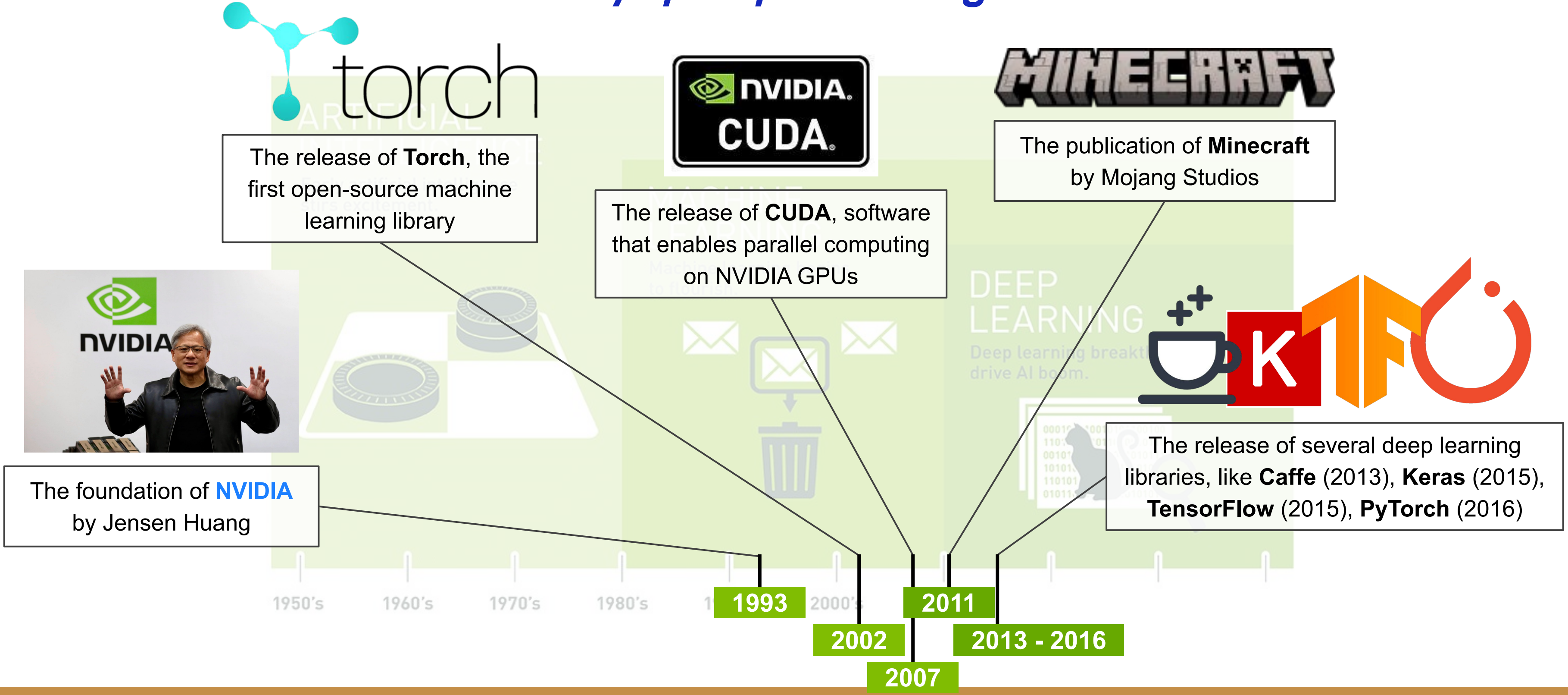
The launch of **Facebook**
by Mark Zuckerberg

2004

The launch of
Google Cloud Platform (GCP)

2008

History of Artificial Intelligence



1993 – The foundation of **NVIDIA** by Jensen Huang

2002 – The release of **Torch**, the first open-source machine learning library

2007 – The release of **CUDA**, software that enables parallel computing on NVIDIA GPUs

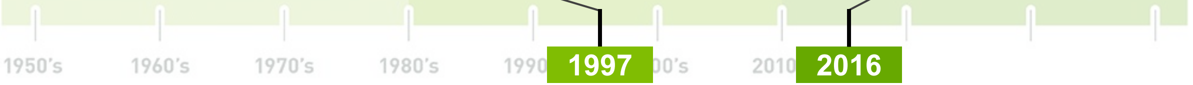
2011 – The publication of **Minecraft** by Mojang Studios

2013 - 2016 – The release of several deep learning libraries, like **Caffe** (2013), **Keras** (2015), **TensorFlow** (2015), **PyTorch** (2016)

History of Artificial Intelligence

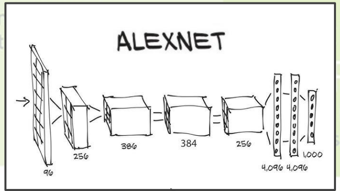
The defeat of Garry Kasparov by **IBM's Deep Blue**, the first computer to win against a world chess champion

The defeat of Lee Sedol by **DeepMind's AlphaGo**, the first computer to win against a 9-dan professional Go player

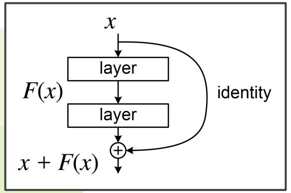


History of Artificial Intelligence

The victory of the *ImageNet 2012 Challenge* by a CNN-based model called **AlexNet**



The invention of ***Generative Adversarial Nets*** by Ian Goodfellow and colleagues



The victory of the *ImageNet 2015 Challenge* by a novel architecture introduced with **ResNet**

The first **ImageNet Large Scale Visual Recognition Challenge** (ILSVRC), an annual contest on image classification and detection via software programs

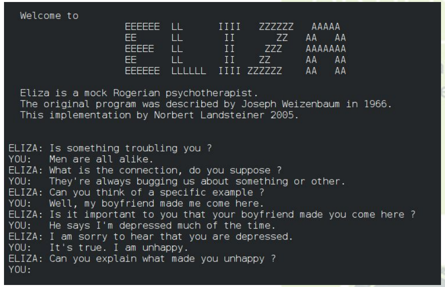
The launch of **DALL-E** by OpenAI

DALL·E

1950's 1960's 1970's 1980's 1990's 2000's

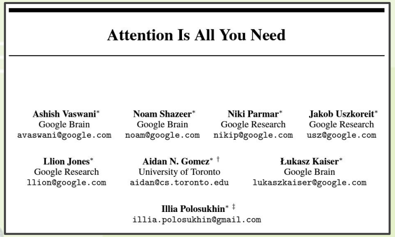
2010 **2012** **2014** **2015** **2021**

History of Artificial Intelligence



```
  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  LLLLLLL 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:
```

The release of **ELIZA** by Joseph Weizenbaum, the computer program to engage conversations with humans



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com	**Noam Shazeer*** Google Brain noam@google.com	**Niki Parmar*** Google Research nikip@google.com	**Jakob Uszkoreit*** Google Research usz@google.com
Llion Jones* Google Research llion@google.com	**Aidan N. Gomez*** † University of Toronto aidan@cs.toronto.edu	**Łukasz Kaiser*** Google Brain lukaszkaiser@google.com	

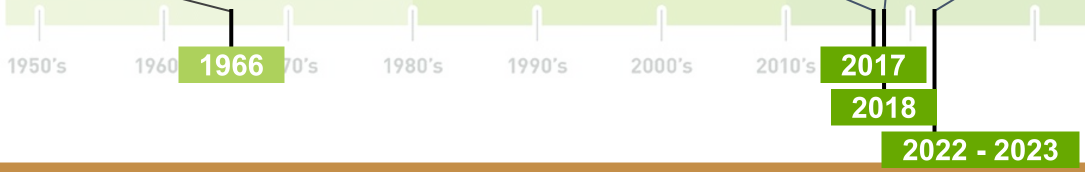
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

The introduction of the **Transformer** architecture with the paper "Attention is all you need"

The introduction of **BERT** by Google and **GPT-1** by OpenAI, models for NLP problems

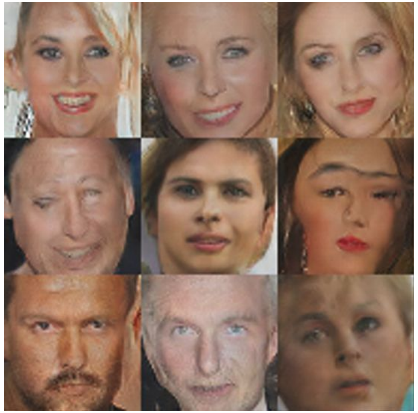
The launch of several chatbot platforms, like **ChatGPT** by OpenAI (2022), **Bard/Gemini** by Google (2023), **Claude** by Anthropic (2023)

ChatGPT



The evolution of image generation

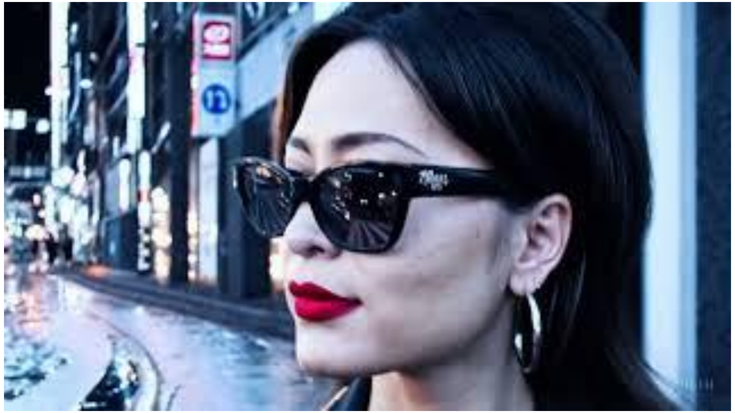
DCGAN (arXiv:1511.06434), 2015



Credits:

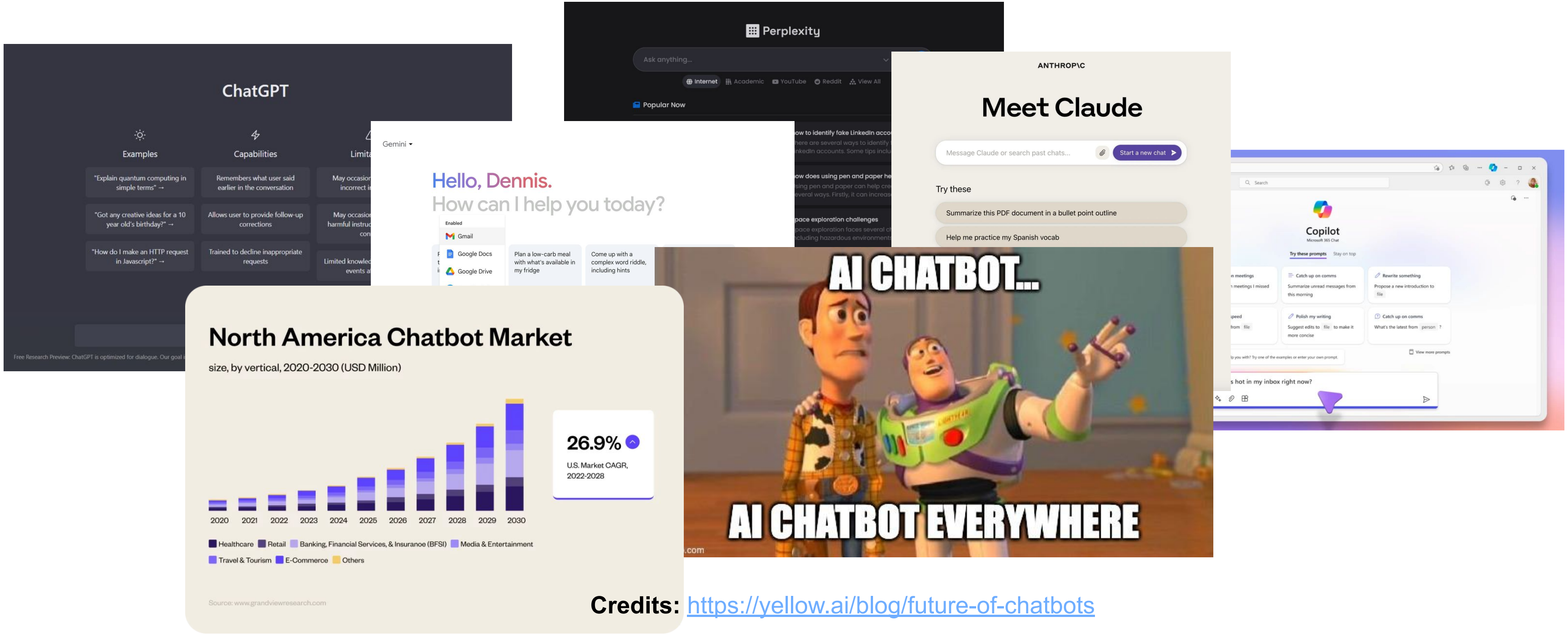
- Alec Radford
- Luke Metz
- Soumith Chintala

Sora (OpenAI text-to-video model), 2024



`<PROMPT>` A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about. `</PROMPT>`

Chatbot, chatbot everywhere!



ChatGPT

Examples
"Explain quantum computing in simple terms" →
"Got any creative ideas for a 10 year old's birthday?" →
"How do I make an HTTP request in Javascript?" →

Capabilities
Remembers what user said earlier in the conversation
Allows user to provide follow-up corrections
Trained to decline inappropriate requests

Perplexity
Ask anything...
Popular Now

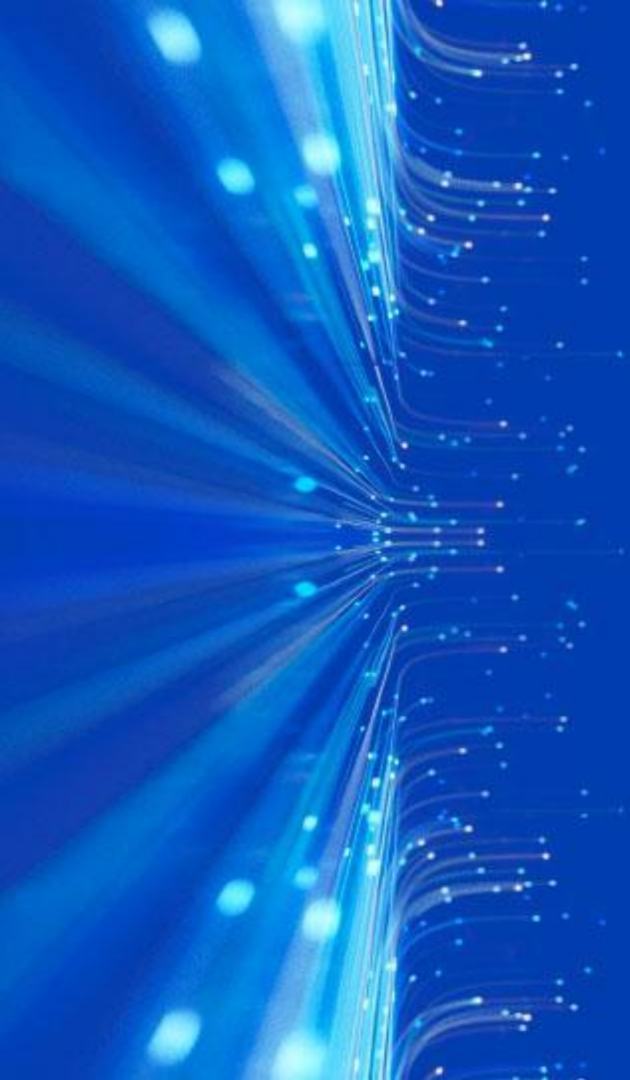
Gemini
Hello, Dennis.
How can I help you today?
Plan a low-carb meal with what's available in my fridge
Come up with a complex word riddle, including hints

ANTHROP\C
Meet Claude
Message Claude or search past chats...
Start a new chat
Try these
Summarize this PDF document in a bullet point outline
Help me practice my Spanish vocab

Copilot

North America Chatbot Market
size, by vertical, 2020-2030 (USD Million)
26.9% U.S. Market CAGR, 2022-2028
Source: www.grandviewresearch.com

Credits: https://yellow.ai/blog/future-of-chatbots



2.0

EVOLVING THE CNAF USER SUPPORT THROUGH AI-DRIVEN SOLUTIONS

How AI technologies can help in supporting Tier-1 users

Scientific computing at CNAF

As INFN national center dedicated to R&D on **Information and Communication Technologies**, CNAF host the Italian Tier-1 data center since 2003, providing computing resources to a **wide scientific community**:

- collaboration with **60+ experiments** (not only the four major experiments of LHC)
- support to **1500+ active users** accessing the Tier-1 resources (including communities different from the High Energy and Nuclear Physics one)

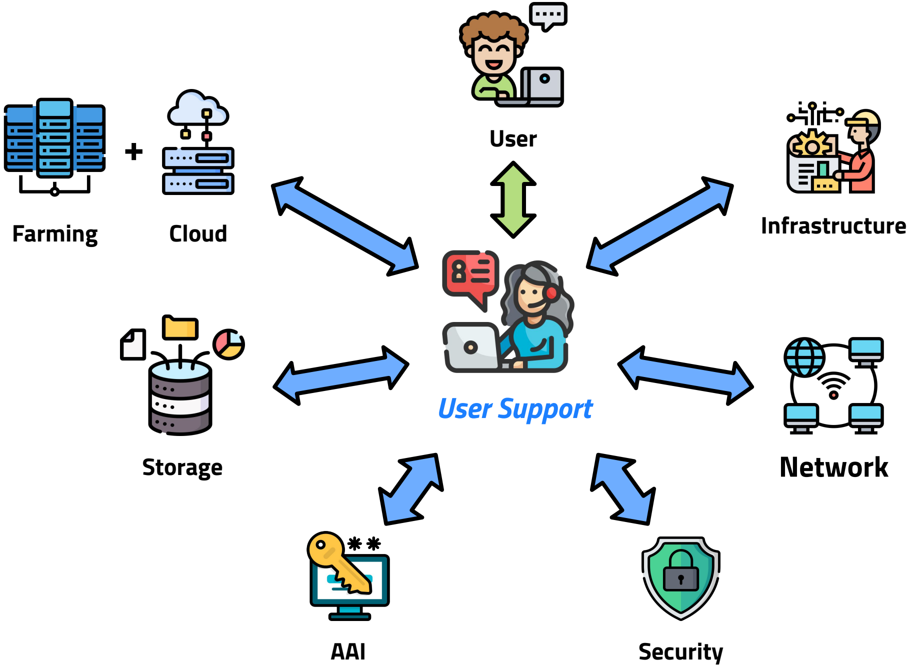
LHC

ATLAS EXPERIMENT
CMS
ALICE
LHCb

NON LHC

PIERRE AUGER OBSERVATORY
AMS-02
DARKSIDE
KM3NeT
Argo-YBJ
VIRGO
HERD
JUNO
BOREXINO GRAN SASSO
Fermi Gamma-ray Space Telescope
CUPID
ICARUS
CDF II
Belle II
cta cherenkov telescope array
NA62
GERDA
LHAASO
XENON Dark Matter Project

Support for experiments at Tier-1



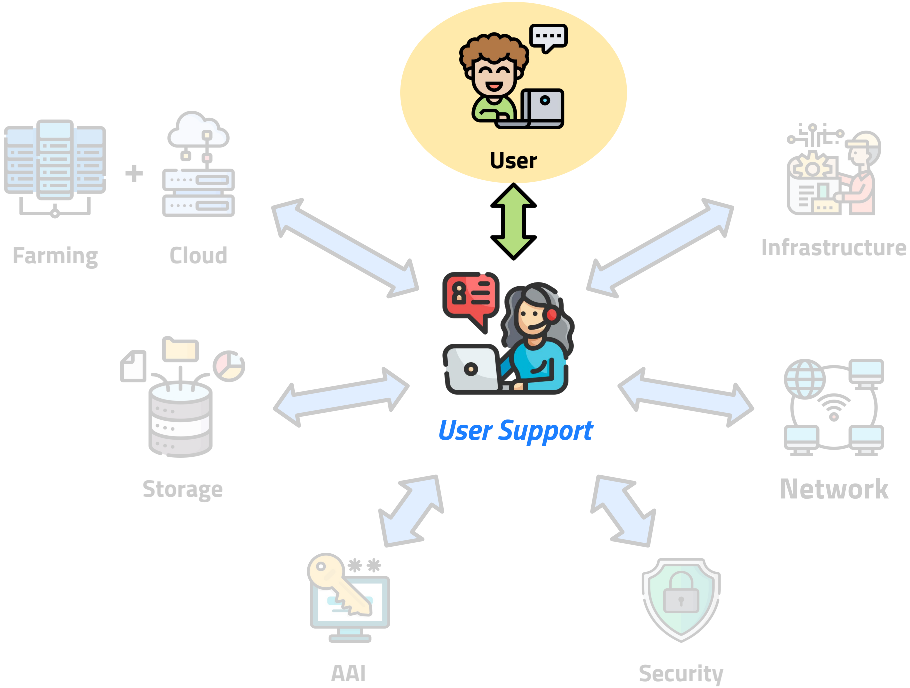
The Italian Tier-1 ensures support for experiments and users through the dedicated ***User Support*** (US) unit:

- **principal contact point** between users and other specialized Tier-1 units
- development of tools/procedures to **simplify** the access to computing resources
- preparation and updates of the **Tier-1 User Guide** (https://l.infn.it/t1guide)

Three levels of support:

1. User Support unit (entry point)
2. involvement of other specialized units
3. involvement of software developers

Support for experiments at Tier-1



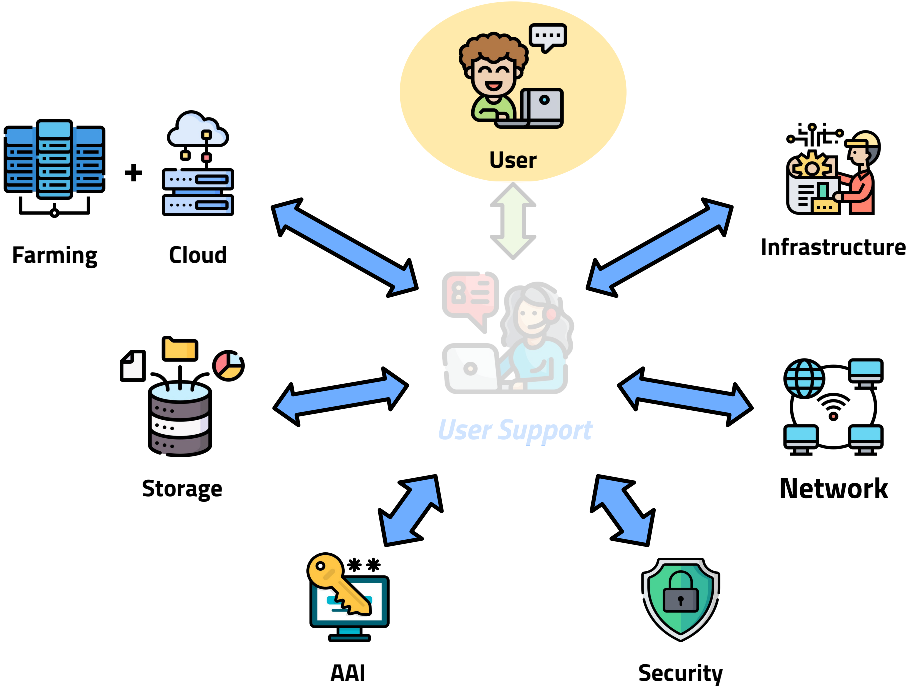
The Italian Tier-1 ensures support for experiments and users through the dedicated ***User Support*** (US) unit:

- **principal contact point** between users and other specialized Tier-1 units
- development of tools/procedures to **simplify** the access to computing resources
- preparation and updates of the **Tier-1 User Guide** (https://l.infn.it/t1guide)

Three levels of support:

1. User Support unit (entry point)
2. involvement of other specialized units
3. involvement of software developers

Support for experiments at Tier-1



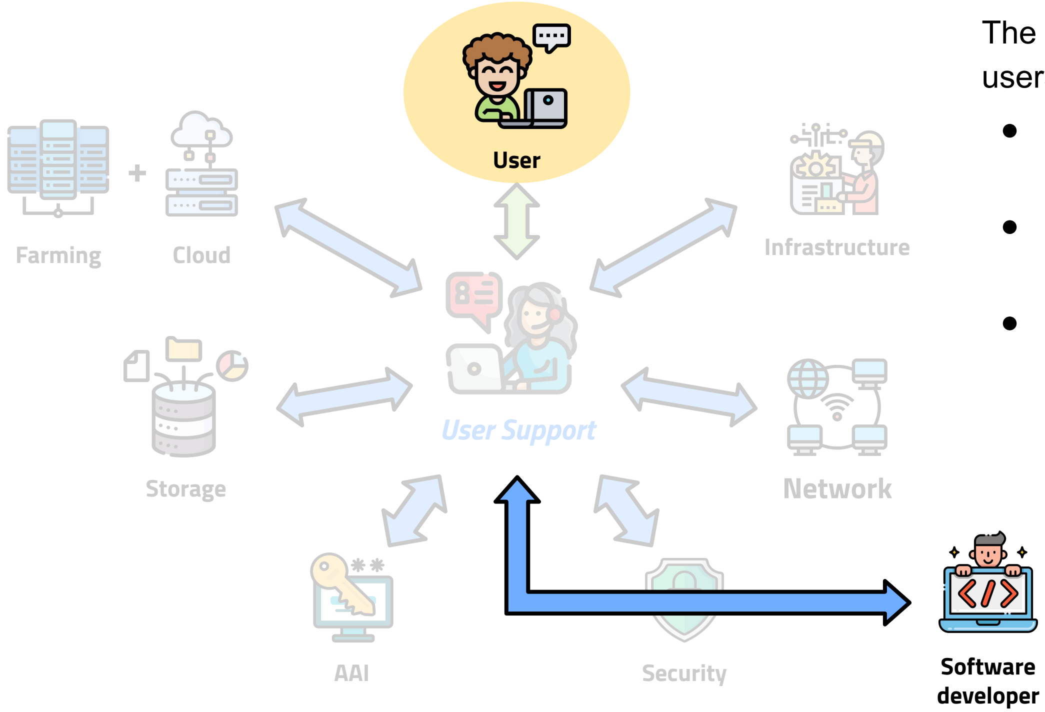
The Italian Tier-1 ensures support for experiments and users through the dedicated ***User Support*** (US) unit:

- **principal contact point** between users and other specialized Tier-1 units
- development of tools/procedures to **simplify** the access to computing resources
- preparation and updates of the **Tier-1 User Guide** (https://l.infn.it/t1guide)

Three levels of support:

1. User Support unit (entry point)
2. involvement of other specialized units
3. involvement of software developers

Support for experiments at Tier-1



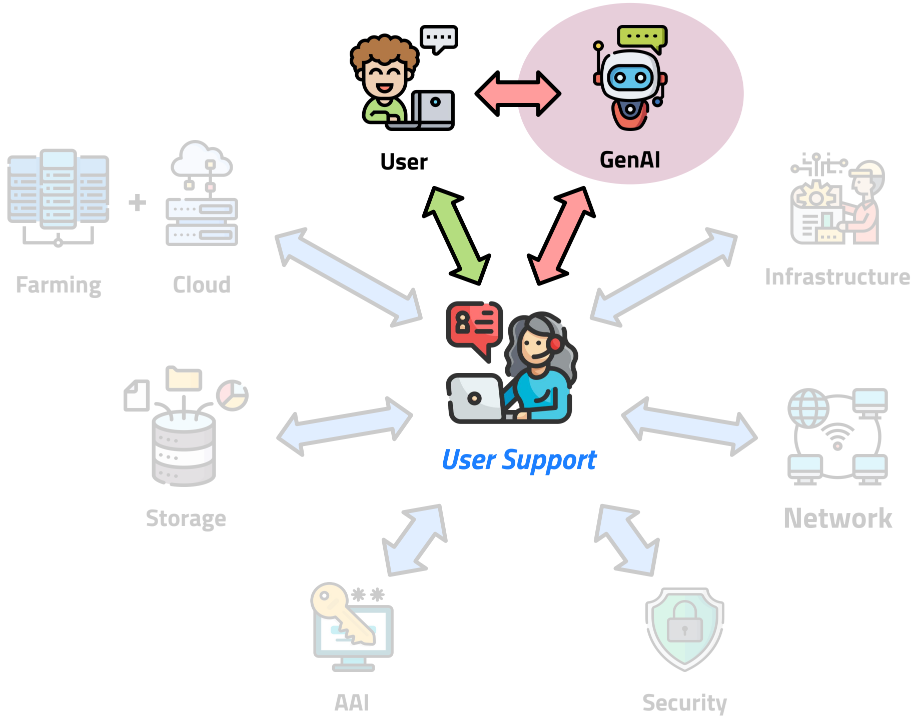
The Italian Tier-1 ensures support for experiments and users through the dedicated ***User Support*** (US) unit:

- **principal contact point** between users and other specialized Tier-1 units
- development of tools/procedures to **simplify** the access to computing resources
- preparation and updates of the **Tier-1 User Guide** (https://l.infn.it/t1guide)

Three levels of support:

1. User Support unit (entry point)
2. involvement of other specialized units
3. involvement of software developers

Evolving the User Support using Artificial Intelligence



In the context of NRRP and thanks to the action of the ICSC and TeRABIT projects, the **Italian Tier-1 will evolve** significantly in the next few years:

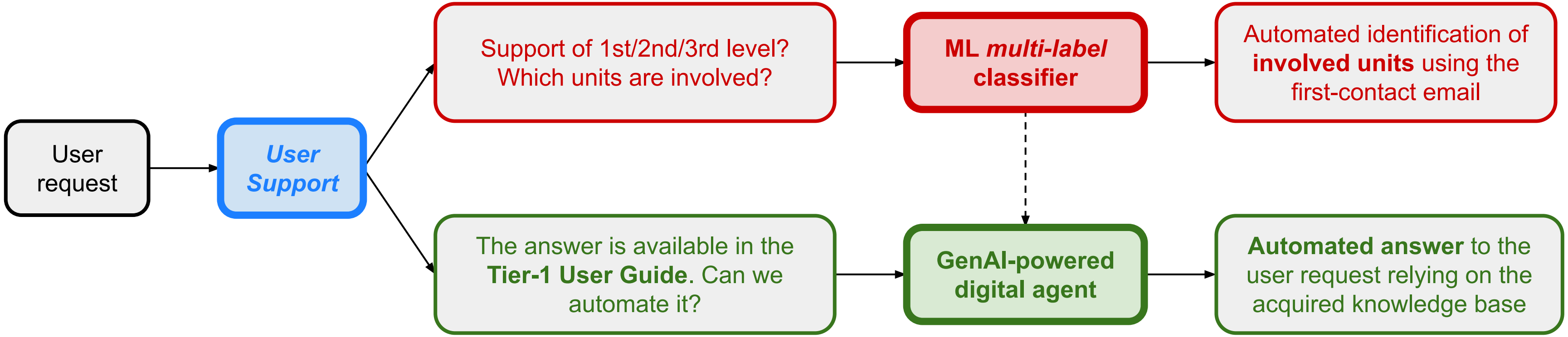
- acquisition of **more and more users and experiments**
- wide scientific community with **various requirements** in terms of infrastructure, hardware, and software
- need for an **evolution** of the CNAF data center and its units

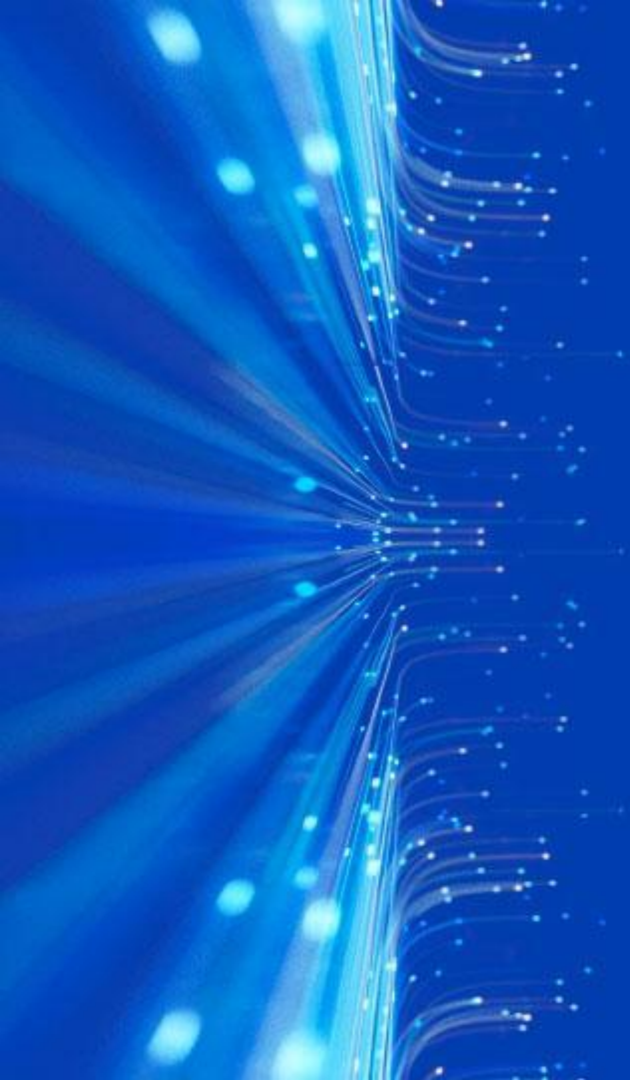
IDEA

Employing AI models to build a **Digital User Supporter** able to help in supporting a growing number of new users and in adopting the latest software technologies

Prototyping a Digital User Supporter

The process of developing an **AI-powered digital agent** for managing User Support duties has been organized into two parallel and complementary sub-projects:


User request
User Support
Support of 1st/2nd/3rd level? Which units are involved?
ML *multi-label* classifier
Automated identification of **involved units** using the first-contact email
The answer is available in the **Tier-1 User Guide**. Can we automate it?
GenAI-powered digital agent
Automated answer to the user request relying on the acquired knowledge base

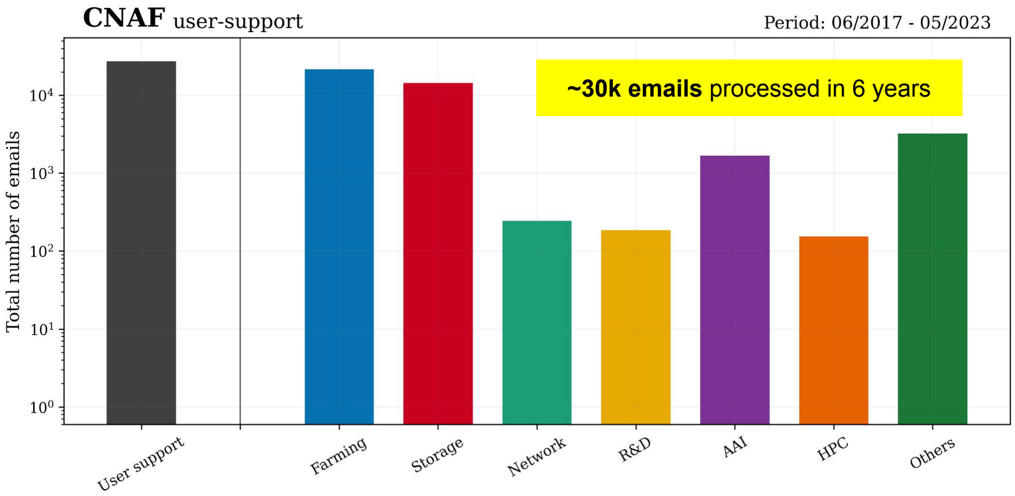


2.1 Multi-label classification

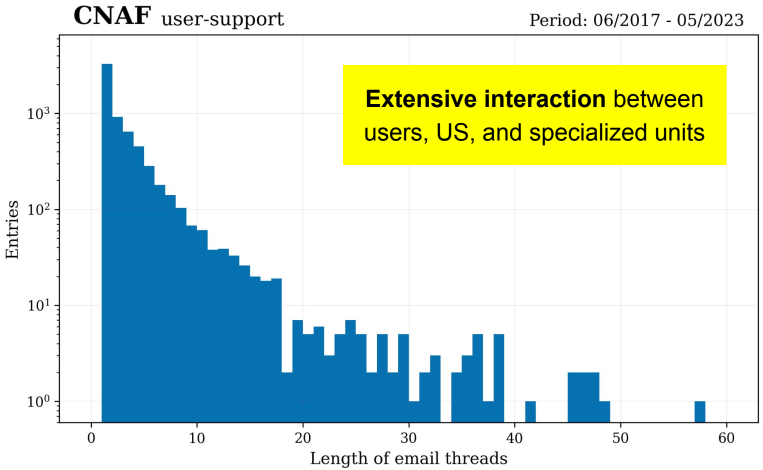
Steps for building an email classifier

1. Preparation of the **dataset of emails** for training
2. Mapping of the textual features into a **numerical representation** (*embedding*)
3. **Classifier training** using differente ML algorithms
4. Benchmark of the **classification performance** and **results combination** (*ensemble*)

A quick look at the data sample



~30k emails processed in 6 years



Extensive interaction between users, US, and specialized units

Preparation of the training dataset

Data collection

The sample contains **~30k emails** received/sent in the period 06/2017 – 05/2023

Each has been saved as an individual JSON file (**~260 MB**)

No attachments, and (almost) no HTML content

Each file includes:

- `from` – email address
- `to` – email address
- `date` – datetime
- `subject` – text
- `content` – text
- `parent` – email address

Anonymization

All the **email addresses** have been **anonymized** by replacing them with UUIDs

Any **references to real individuals** (e.g., names or usernames) found in the email subject/body have been **anonymized** and replaced with the placeholder `[NAME]`

The anonymization process was carried out using **automated scripts** or by **editing manually** the email metadata

Labelling

Each email is assigned a **label** based on the text contained in its subject/body

The **email thread** between a user and the various Tier-1 units involved is reconstructed using the information in the `parent` field

All labels associated to the same thread are propagated to the first contact email as a part of the **labelling** process

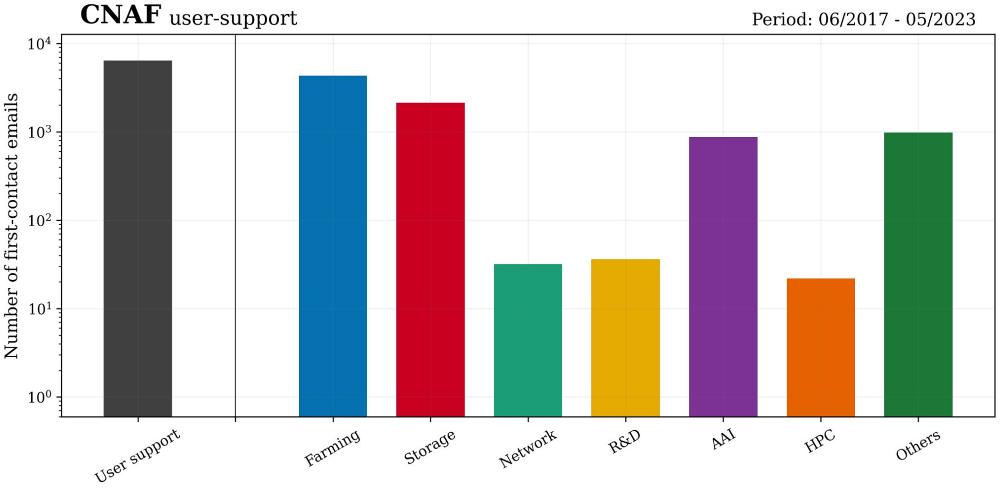
Data cleaning

The email body may contain **escape sequences** (e.g., `\n`, `\t`) o **HTML tags**

The emails are **multilingual** (English/Italian) and may include **special characters** (e.g., accented letters, emojis)

The **data cleaning** process transforms the text to retain only words, placeholder, and punctuation

The prepared training set



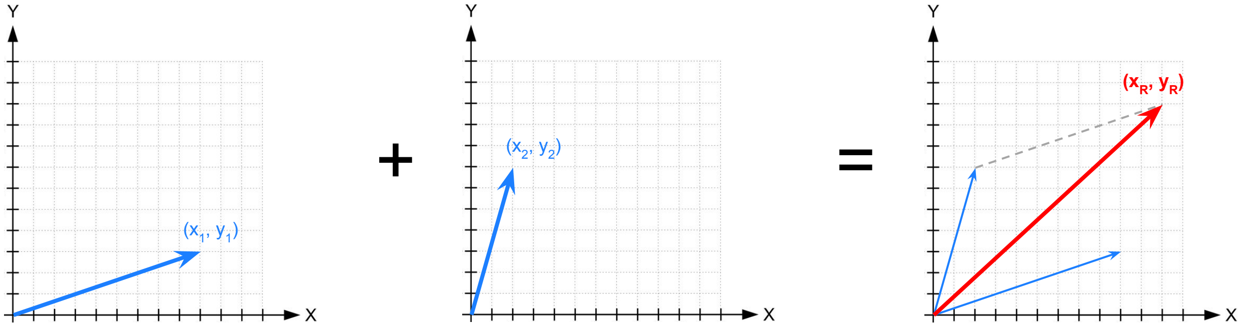
`[user-support]` `[farming]`

```
Hello,

How can I use Jupyter
Notebooks at CNAF?

Thank you,

[NAME]
```

FACSIMILE

How to represent the textual features

- The meaning of a text is encoded in word combinations → ML models operate on vectors (**features**)
  - Word (or token) combinations can be mapped into a **"representation space"** (a high-dimensional vector space) through an ***embedding model***
- The embedding model can be designed to preserve **semantics**, even across multiple languages
  - Sentence-Transformers is a Python framework that provides state-of-the-art models for text embedding

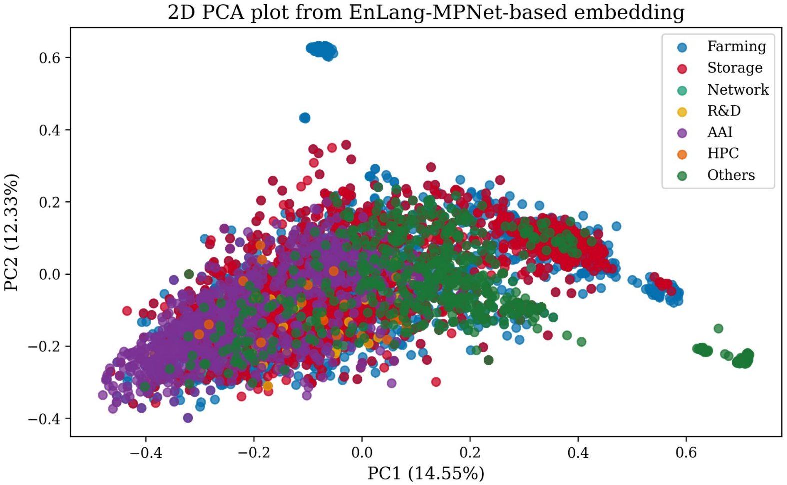


Representation for "👩🏻"

Representation for "👑"

Representation for "👸🏻"

How emails look like in the representation space



all-mpnet-base-v2

- **base model:** MPNet
- **embedding dim:** 768
- **max seq length:** 384
- **inference speed:** 1
- **language:** English only

paraphrase-multilingual-mpnet-base-v2

- **base model:** XLM-RoBERTa
- **embedding dim:** 768
- **max seq length:** 128
- **inference speed:** ~1
- **language:** multiple

all-MiniLM-L6-v2

- **base model:** MiniLM
- **embedding dim:** 384
- **max seq length:** 256
- **inference speed:** x5
- **language:** English only

paraphrase-multilingual-MiniLM-L12-v2

- **base model:** MiniLM
- **embedding dim:** 384
- **max seq length:** 126
- **inference speed:** x2.5
- **language:** multiple

Performance of the trained classifiers

Different ML models were trained to perform **multi-label classification**

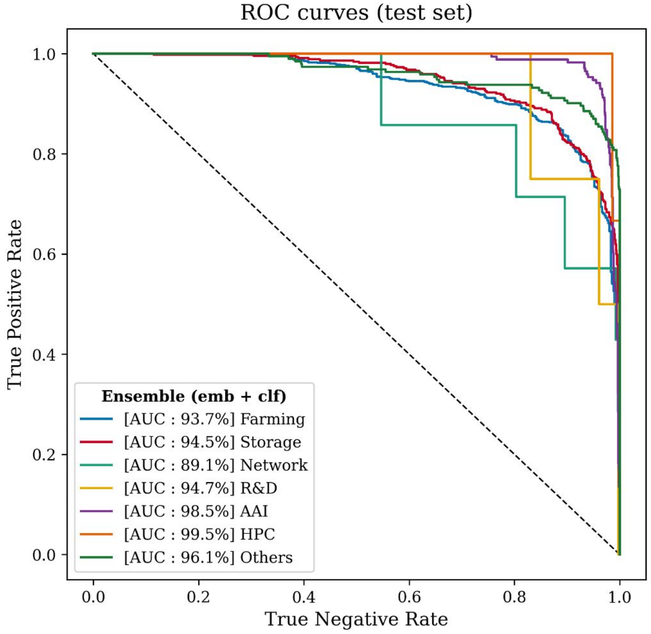
- ***k-Nearest Neighbors*** (kNN)
- ***Random Forest*** (RF)
- ***Extreme Gradient Boosting*** (XGBoost)
- ***Feed-forward Neural Network*** (FNN)

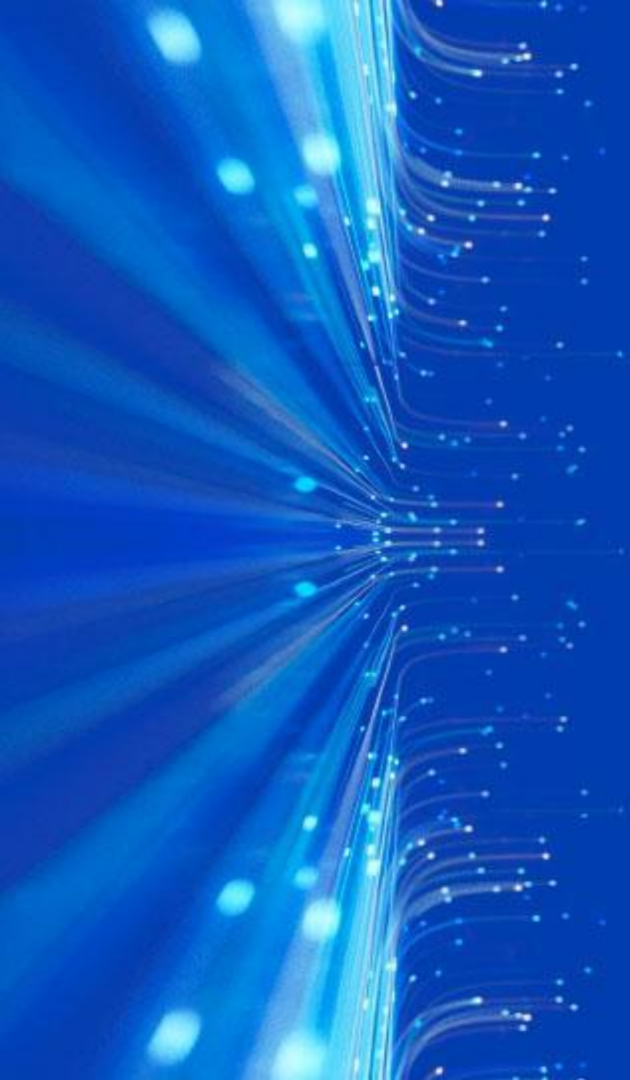
Various embedding models were used in combination with these classification models

- `all-mpnet-base-v2`
- `all-MiniLM-L6-v2`
- `paraphrase-multilingual-mpnet-base-v2`
- `paraphrase-multilingual-MiniLM-L12-v2`

Different preprocessing strategies → **4 (emb) x 4 (prep) x 4 (clf) = 64 models in total**

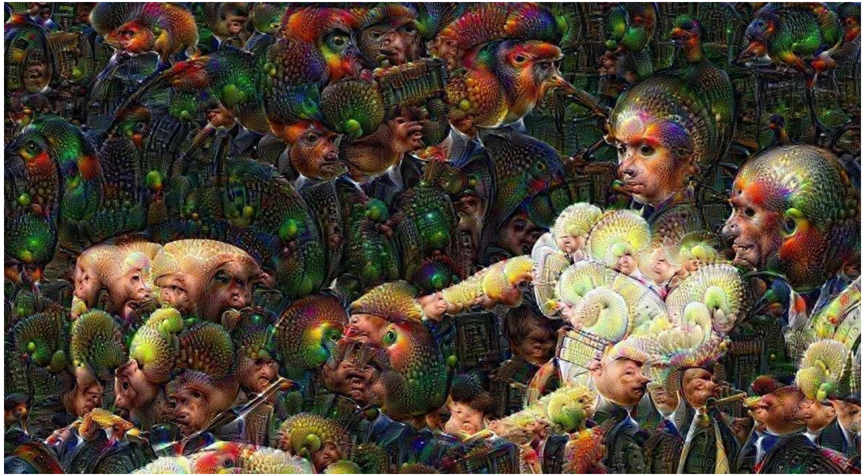
By combining the outputs of the most promising models using an **ensemble approach**, the best performance was achieved: **~95% of ROC AUC score** on the test set





2.2 Digital User Supporter

The problem of artificial hallucinations



Credits: https://fisicisenzapalestra.com/artificiale-come-lintelligenza-e-gli-allucinogeni.html

One of the major challenges in working with AI chatbots and Computer Vision tools is the issue of ***artificial hallucinations***

- responses that contain **false or misleading information** presented as **facts**

A potential solution to limit the "creativity" of Large Language Models (LLMs) is to integrate them into an architecture designed to **retrieve relevant information** and compose **answers based on that data**

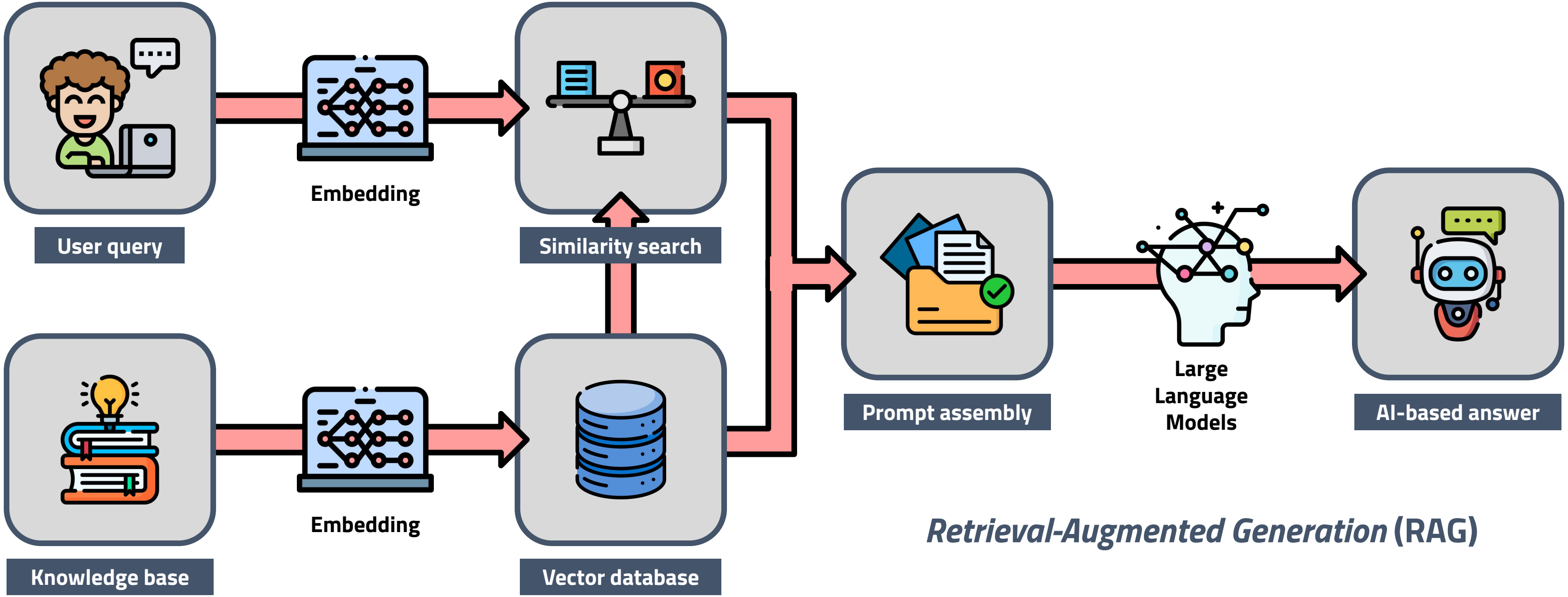
The ***Retrieval-Augmented Generation*** (RAG) architecture provides a **semantic-aware pipeline** that retrieves information from a knowledge base and generates text grounded in the retrieved data

PROTOTYPE

The first prototype of the **AI-powered Digital User Supporter** has been developed upon a RAG model, implemented with LangChain and provisioning the LLM relying on ollama

Credits: Alberto Trashaj

How to build a Digital User Supporter

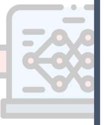


Retrieval-Augmented Generation (RAG)

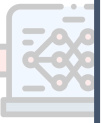
Credits: Alberto Trashaj

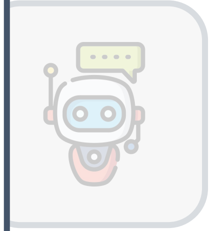
How to build a Digital User Supporter

User query

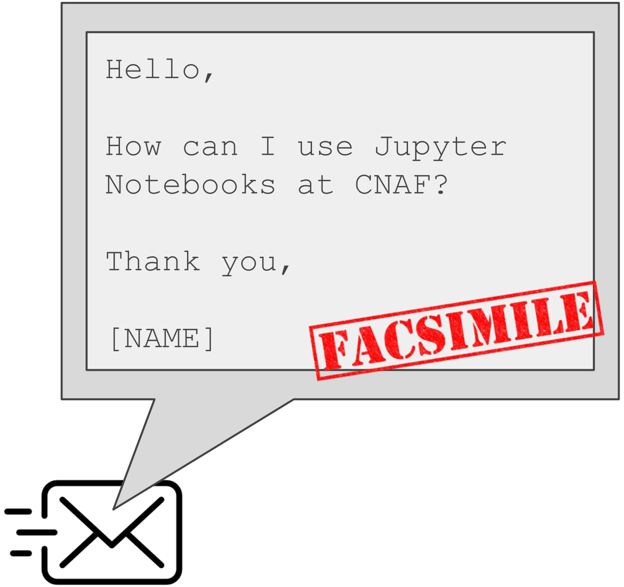


Knowledge base





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

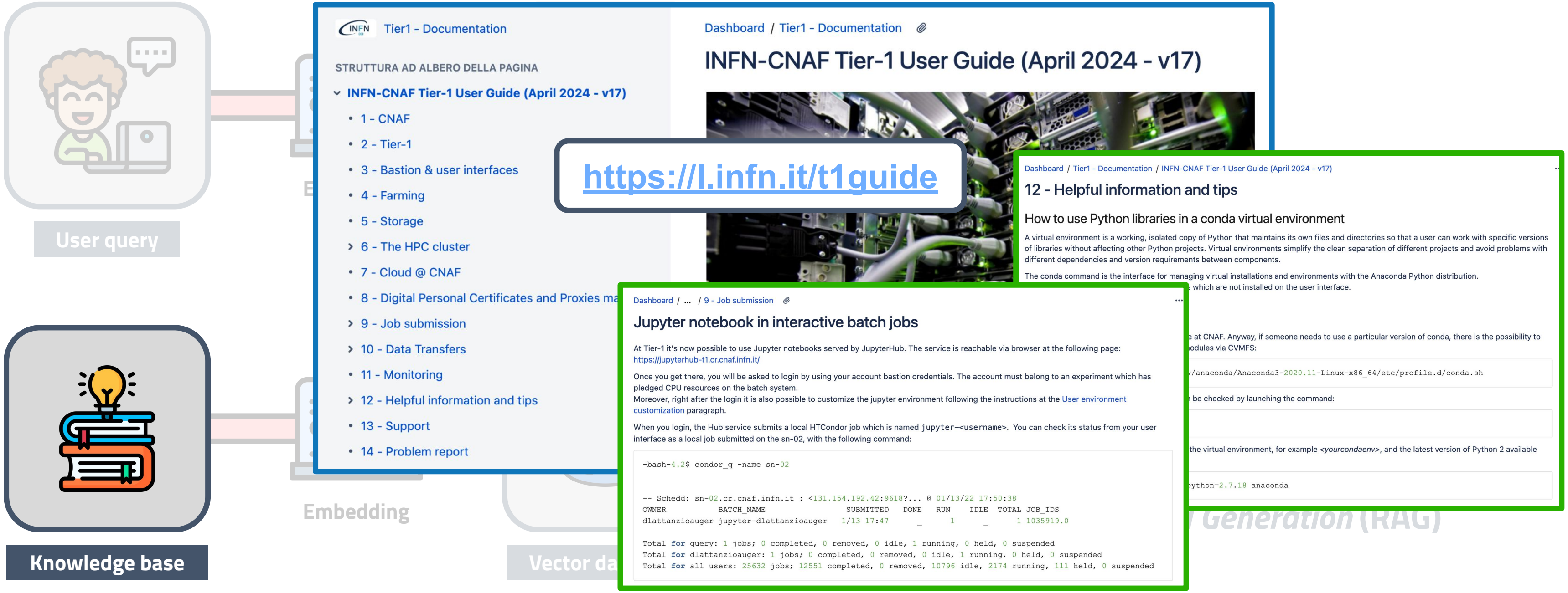


```
Hello,

How can I use Jupyter
Notebooks at CNAF?

Thank you,

[NAME]
```

FACSIMILE

Credits: Alberto Trashaj

How to build a Digital User Supporter



INFN-CNAF Tier-1 User Guide (April 2024 - v17)

https://l.infn.it/t1guide

Dashboard / Tier1 - Documentation / INFN-CNAF Tier-1 User Guide (April 2024 - v17)

12 - Helpful information and tips

How to use Python libraries in a conda virtual environment

A virtual environment is a working, isolated copy of Python that maintains its own files and directories so that a user can work with specific versions of libraries without affecting other Python projects. Virtual environments simplify the clean separation of different projects and avoid problems with different dependencies and version requirements between components.

The conda command is the interface for managing virtual installations and environments with the Anaconda Python distribution.
s which are not installed on the user interface.

e at CNAF. Anyway, if someone needs to use a particular version of conda, there is the possibility to
odules via CVMFS:

```
/anaconda/Anaconda3-2020.11-Linux-x86_64/etc/profile.d/conda.sh
```

be checked by launching the command:

the virtual environment, for example *<yourcondaenv>*, and the latest version of Python 2 available

```
ython=2.7.18 anaconda
```

Dashboard / ... / 9 - Job submission

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/

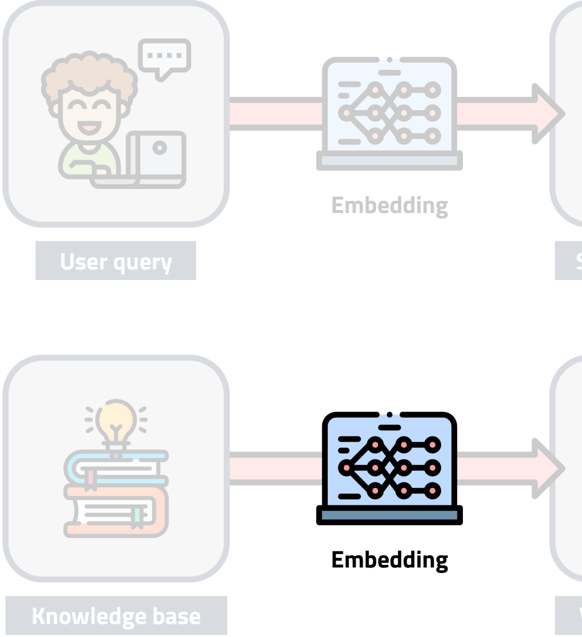
Once you get there, you will be asked to login by using your account bastion credentials. The account must belong to an experiment which has pledged CPU resources on the batch system.
Moreover, right after the login it is also possible to customize the jupyter environment following the instructions at the User environment customization paragraph.

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

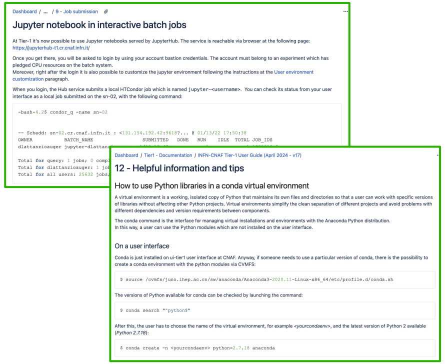
```
-bash-4.2$ condor_q -name sn-02


-- Schedd: sn-02.cr.cnaf.infn.it : <131.154.192.42:9618?... @ 01/13/22 17:50:38
OWNER           BATCH_NAME                    SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
dlattanzioauger jupyter-dlattanzioauger   1/13 17:47      _      1      _      1 1035919.0

Total for query: 1 jobs; 0 completed, 0 removed, 0 idle, 1 running, 0 held, 0 suspended
Total for dlattanzioauger: 1 jobs; 0 completed, 0 removed, 0 idle, 1 running, 0 held, 0 suspended
Total for all users: 25632 jobs; 12551 completed, 0 removed, 10796 idle, 2174 running, 111 held, 0 suspended
```

Credits: Alberto Trashaj

How to build a Digital User Supporter



The knowledge base (**Tier-1 User Guide**) is mapped to a vector space through an **embedding model**, such as `all-mpnet-base-v2`



Dashboard / ... / 9 - Job submission

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-<username>`. You can check its status from your user interface as a local job submitted on the sn-02, with the following command:

```
-bash-4.2$ condor_q -name sn-02


-- Schedd: sn-02.cr.cnaf.infn.it : <131.154.192.42:9618?... @ 01/13/22 17:50:38
OWNER           BATCH_NAME                  SUBMITTED   DONE   RUN    IDLE  TOTAL JOB_IDS
dlattanzioauger jupyter-dlattan

Total for query: 1 jobs; 0 compl
Total for dlattanzioauger: 1 job
Total for all users: 25632 jobs;
```

Dashboard / Tier1 - Documentation / INFN-CNAF Tier-1 User Guide (April 2024 - v17)

12 - Helpful information and tips

How to use Python libraries in a conda virtual environment

A virtual environment is a working, isolated copy of Python that maintains its own files and directories so that a user can work with specific versions of libraries without affecting other Python projects. Virtual environments simplify the clean separation of different projects and avoid problems with different dependencies and version requirements between components.

The conda command is the interface for managing virtual installations and environments with the Anaconda Python distribution.
In this way, a user can use the Python modules which are not installed on the user interface.

On a user interface

Conda is just installed on ui-tier1 user interface at CNAF. Anyway, if someone needs to use a particular version of conda, there is the possibility to create a conda environment with the python modules via CVMFS:

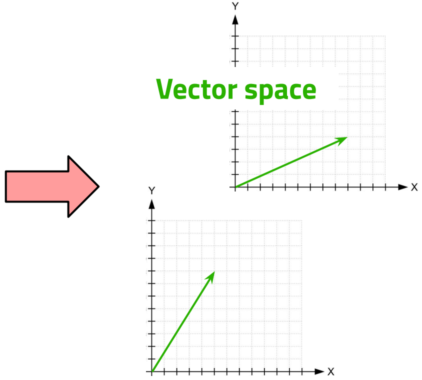
```
$ source /cvmfs/juno.ihep.ac.cn/sw/anaconda/Anaconda3-2020.11-Linux-x86_64/etc/profile.d/conda.sh
```

The versions of Python available for conda can be checked by launching the command:

```
$ conda search "^python$"
```

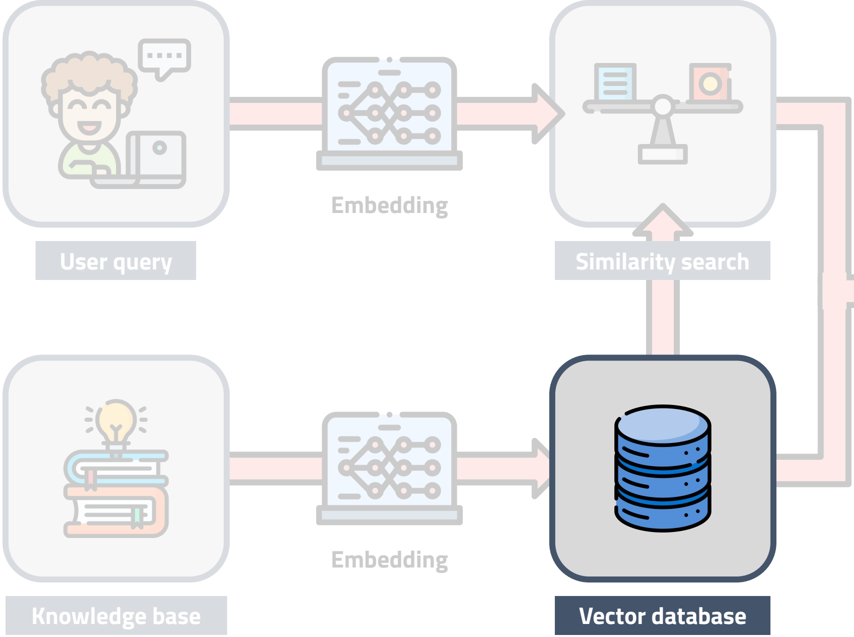
After this, the user has to choose the name of the virtual environment, for example *<yourcondaenv>*, and the latest version of Python 2 available (*Python 2.7.18*):

```
$ conda create -n <yourcondaenv> python=2.7.18 anaconda
```



Credits: Alberto Trashaj

How to build a Digital User Supporter

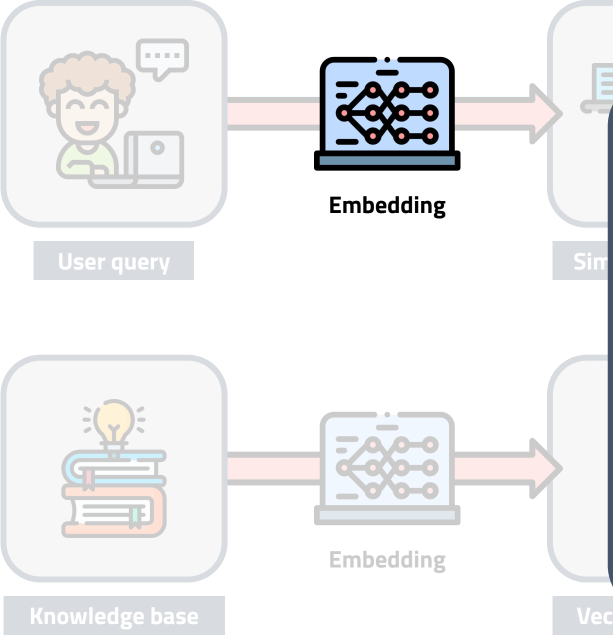


To store the embedding results, a **database** optimized for **handling large vectors** and their **metadata** (e.g., source text) is required. It should also support **efficient searches** within the vector space (features like these are at the core of Chroma)

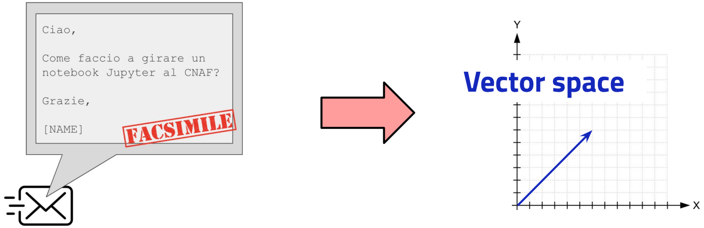
Chroma

Credits: Alberto Trashaj

How to build a Digital User Supporter



As well as the knowledge base, also the user request is mapped to a vector space relying on the same **embedding model** ([all-mpnet-base-v2]())


Ciao,

Come faccio a girare un notebook Jupyter al CNAF?

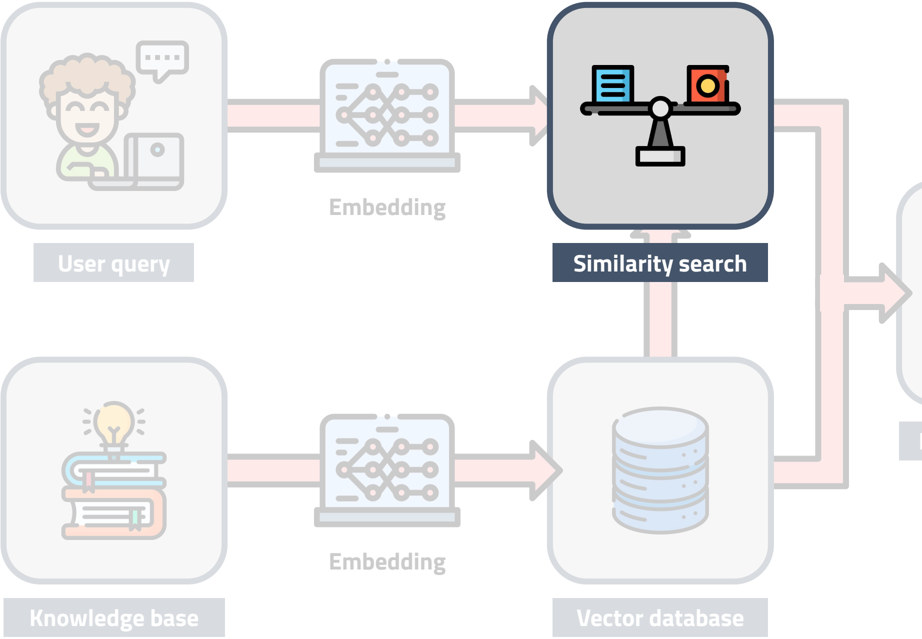
Grazie,

[NAME]

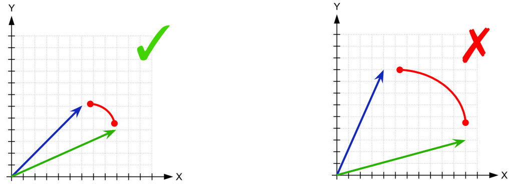
FACSIMILE

Credits: Alberto Trashaj

How to build a Digital User Supporter



To retrieve relevant information from the **guide** to respond to the **user request**, we can compare the embeddings from both cases based on ***similarity*** (e.g., dot product)



Once the guide embeddings that best match the request embedding are identified, the **vector database** can be queried to **retrieve the corresponding metadata**

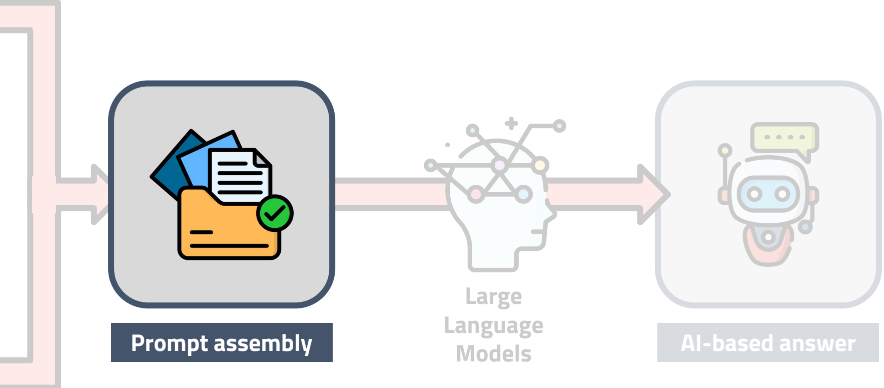
Credits: Alberto Trashaj

How to build a Digital User Supporter

The retrieved text segments are used to tailor the *template* for the textual **prompt** submitted to the **Language Model**

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

<context>
{context}
</context>
```

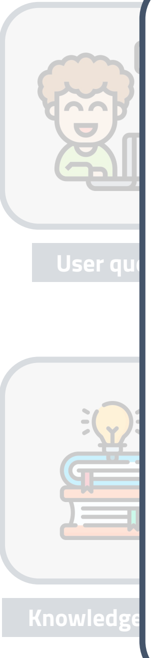
FACSIMILE



Retrieval-Augmented Generation (RAG)

Credits: Alberto Trashaj

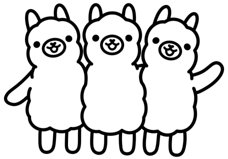
How to build a Digital User Supporter



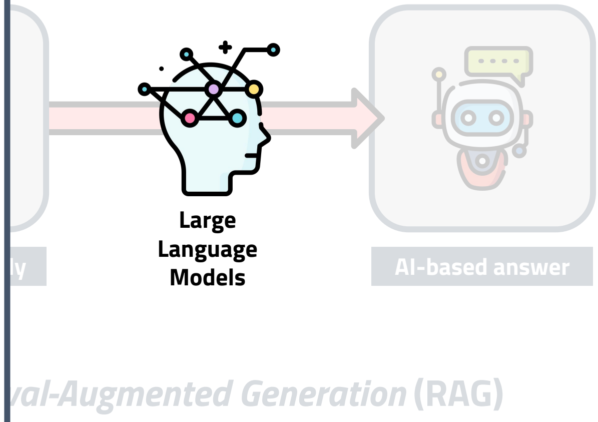
The generated prompt can then be submitted to a ***Large Language Model*** (LLM), which, even if not an "expert" in a specific subject, is capable of extracting the **"relevant" content** from `{context}` and rephrasing it

Since fine-tuning is not required, LLMs are often described as **general-purpose** technologies and are thus also referred to as **Foundation Models**

Meta



```
ollama run llama3
```



Credits: Alberto Trashaj

How to build a Digital User Supporter

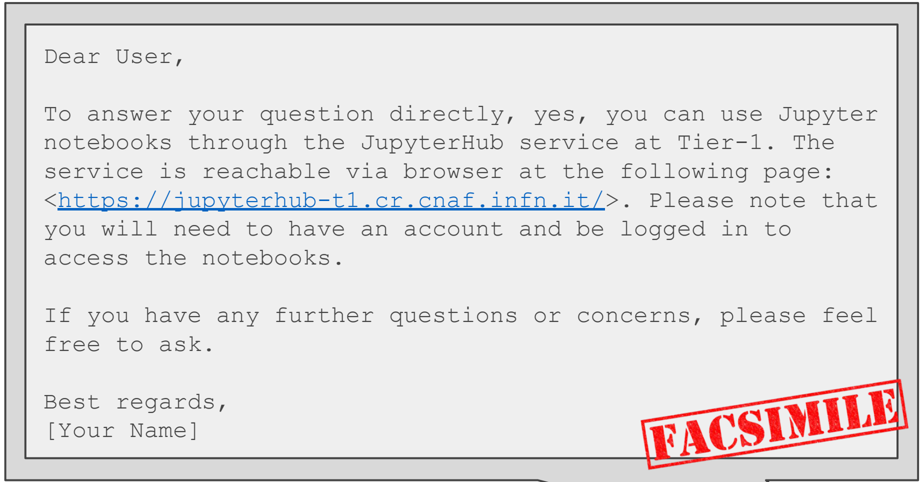
User qu

Knowledg

arge
guage
odels

AI-based answer

ed Generation (RAG)



Dear User,

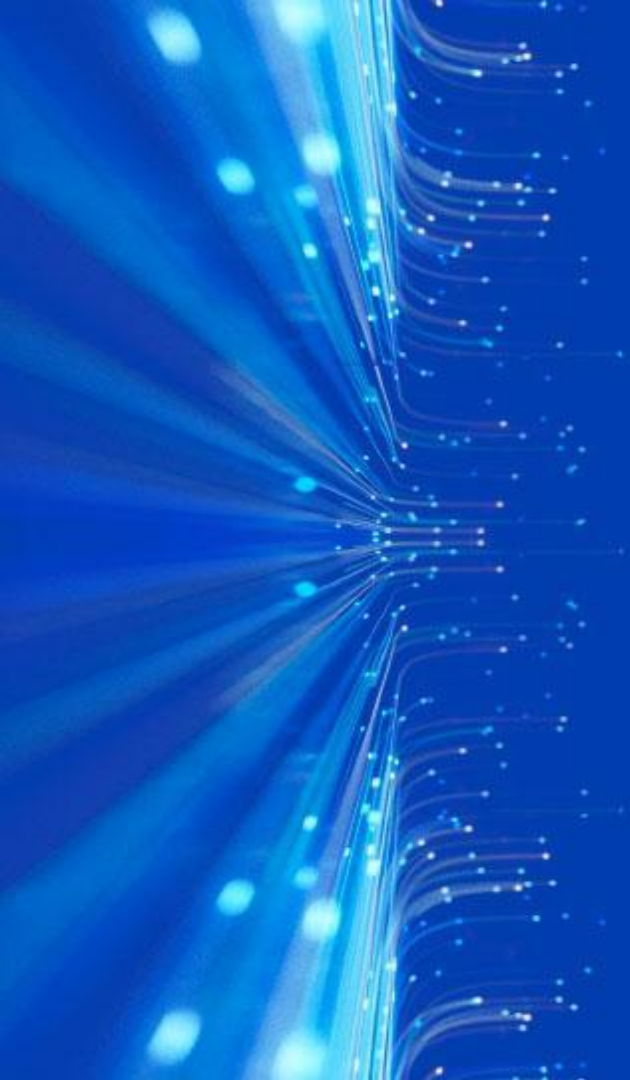
To answer your question directly, yes, you can use Jupyter notebooks through the JupyterHub service at Tier-1. The service is reachable via browser at the following page: <https://jupyterhub-t1.cr.cnaf.infn.it/>. Please note that you will need to have an account and be logged in to access the notebooks.

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

Best regards,
[Your Name]

FACSIMILE

Note: automated answer generated with ***Llama 2***



3.0

MOVING TO THE CONCLUSIONS

Summary, conclusions, and final remarks

Here are my conclusions!

WE HAVE OLLAMA SERVING LLAMA 3 ON GPUS
WE CAN STILL USE GPUS FOR DOING SCIENCE, RIGHT?
RIGHT?
imgflip.com

- AI is reshaping the computing landscape, driving technological evolution, influencing market trends, and dominating software development worldwide
- **CNAF should ride the AI wave**, exploring and implementing new technologies to drive its evolution within an AI-dominated world
  - *the development of a Digital User Supporter is a step in this direction, but also **other units can benefit from using AI***
- The results obtained with the first prototype of the Digital User Supporter are promising but there is still room for improvements
  - *e.g., multi-expert model, enlarged knowledge base (also including emails)*
- As AI companies develop increasingly large models **requiring massive GPU resources**, it is essential to implement tailored solutions
  - *this is crucial to avoid that AI chatbots **monopolize GPU resources** needed for scientific research*

Centro Nazionale di Ricerca in HPC
Big Data and Quantum Computing

Thanks!

Any questions or comments?

Matteo Barbetti (INFN CNAF)
email: matteo.barbetti@cnaf.infn.it